The Annals of Statistics
2013, Vol. 41, No. 5, 2639–2667
DOI: 10.1214/13-AOS1165

HIGH-DIMENSIONAL INFLUENCE MEASURE

BY JUNLONG ZHAO[1], CHENLEI LENG[2], LEXIN LI[3] AND HANSHENG WANG[4]

Beihang University, University of Warwick and National University of Singapore, North Carolina State University and Peking University

Influence diagnosis is important since presence of influential observations could lead to distorted analysis and misleading interpretations. For high-dimensional data, it is particularly so, as the increased dimensionality and complexity may amplify both the chance of an observation being influential, and its potential impact on the analysis. In this article, we propose a novel high-dimensional influence measure for regressions with the number of predictors far exceeding the sample size. Our proposal can be viewed as a high-dimensional counterpart to the classical Cook's distance. However, whereas the Cook's distance quantifies the individual observation's influence on the least squares regression coefficient estimate, our new diagnosis measure captures the influence on the marginal correlations, which in turn exerts serious influence on downstream analysis including coefficient estimation, variable selection and screening. Moreover, we establish the asymptotic distribution of the proposed influence measure by letting the predictor dimension go to infinity. Availability of this asymptotic distribution leads to a principled rule to determine the critical value for influential observation detection. Both simulations and real data analysis demonstrate usefulness of the new influence diagnosis measure.

1. Introduction. An observation is flagged influential if some important features of the analysis are substantially altered after this observation is removed [13]. Presence of influential observations would possibly lead to distorted analysis and misleading results [18], and therefore it is important to be alert to influential observations and take them into consideration when interpreting the results. In the classical normal linear model setup, regression coefficient estimate was chosen, naturally, as the feature whose substantial change defines influential observations. Toward that end, [12] proposed a difference measure between the OLS estimate

Received July 2013.

[1]Supported by National Natural Science Foundation of China (11101022) and Ministry of Education Humanities and Social Science Foundation Youth project (10YJC910013).

[2]Supported by National University of Singapore research grants.

[3]Supported by NSF Grant DMS-11-06668.

[4]Supported by National Natural Science Foundation of China (11131002, 11271032), Fox Ying Tong Education Foundation, the Business Intelligence Research Center of Peking University and the Center for Statistical Science of Peking University.

MSC2010 subject classifications. Primary 62J20; secondary 62E20.

Key words and phrases. Cook's distance, high-dimensional diagnosis, influential observation, LASSO, marginal correlations, variable screening.

on the full data and that on the subset of data without the observation in question. This measure, which is later on referred in the statistical literature as the *Cook's distance*, quantifies the contribution, or influence, of individual data observation on the regression coefficient estimate. Consequently an observation with a large Cook's distance is deemed as influential. Since its introduction, the Cook's distance has been routinely employed in regression analysis, due to its clear interpretation from the case deletion point of view, and its easy computation without having to re-estimate the model for each removed observation. The topic is covered in most standard regression textbooks, and it is implemented in popular statistical software such as R and SAS.

The problem of influence diagnosis has since attracted considerable attention and been systematically investigated for various models and analyses. Examples include linear regression models [9, 12, 14], categorical data analyses [1], generalized linear models [16, 33, 38], generalized estimation equations [30], linear mixed models [2, 3, 11], generalized linear mixed models [39], semiparametric mixed models [25], growth curve models [29], incomplete data analysis [44], perturbation theory [15, 42, 43], among others. For an excellent review on the latest developments in the field of influence diagnosis, we refer to [42].

Thanks to the aforementioned works, substantial insights have been gained on influence diagnosis. However, it is important to note that, all existing diagnosis approaches have been developed under the assumption that the number of predictors in regression is fixed. As such, none is immediately applicable to high-dimensional regression analysis, where the number of predictors p far exceeds the sample size n. On the other hand, nowadays prevailing in both science and business are data with unprecedented size and dimensionality, calling for the development of high-dimensional influence diagnosis. Detection of influential observations in high-dimensional data analysis, in our opinion, is equally, or to some extent, even more important than in a classical setup. This is partly because the increased dimensionality and complexity of the data may amplify both the chance of an observation being influential as well as its potential impact on the analysis. Moreover, the peculiar data observations themselves may be of practical importance in addition to data modeling. The diagnosis task, nevertheless, is more challenging in high-dimensional data analysis, and is far from a direct extension of existing diagnosis approaches. To the best of our knowledge, influence diagnosis in a high-dimensional setting has received little attention despite its evident importance.

The first challenge is the definition of influential observation. In other words, which feature of the analysis should one choose such that its substantial alternation defines an influential observation? In the classical setup, an observation is deemed influential if it incurs serious change in regression coefficient estimate. In high-dimensional regression where $p > n$, the ordinary least squares estimator is highly unstable as the gram matrix is not invertible. On the other hand, we recognize that variable selection and variable screening are of particular importance in high-dimensional regression analysis. There has been a vast literature on variable selection in recent years, including the LASSO [34], the adaptive LASSO [36, 40, 45], the SCAD [21], the bridge estimator [24, 26], the LARS algorithm [19], the Dantzig selector [8], the sure independence screening rule [22], SIS, the forward regression [35], FR, among many others. Underlying all those selection methods, one statistic plays a critical role and, that is, the *marginal* covariance, or equivalently, *marginal* correlation between the response and the individual covariates. To clarify, we note that, SIS is directly defined based on this statistic, whereas the first step of the forward regression hinges on the estimated marginal covariance too. In addition, the sample marginal covariance, in addition to the Gram matrix, is an important input for the well celebrated LARS algorithm, as well as the LASSO, the adaptive LASSO and the Dantzig selector.

Motivated by this vital observation, we choose the marginal correlation as the feature that defines influential observation. We propose a new influence diagnosis measure, which continues to utilize the leave-one-out idea of the classical Cook's distance, but is based on the combined marginal correlations between the response and all predictors. The new measure is applicable to high-dimensional setting where $p > n$, and is very fast and easy to compute. Unlike the classical Cook's distance that quantifies the individual observation's influence on the least squares coefficient estimate, the new measure captures the influence on the marginal correlation, which *in turn* exerts serious impact on variable selection and other downstream analysis. The choice of the marginal correlation as the defining feature of our influence diagnosis does not imply that the marginal correlation is our ultimate goal of interest. Instead, it reflects influence on important analysis features including parameter estimation, variable selection and screening. This definition of influential observation in a high-dimensional setting can be viewed as our first contribution.

Our second contribution is that the explicit asymptotic distribution for the proposed influence measure is derived. Availability of this asymptotic theory offers a principled guidance to determine the critical value for the influence measure. Subsequently, we propose a false discovery rate based procedure for that purpose [5, 6]. We remark that, in the classical setup where p is fixed, a standard Taylor's expansion type analysis [12] revealed that the classical Cook's distance's major variability is due to the observation under investigation and its sample size is only one. This rules out the possibility of establishing a standard asymptotic theory for the classical Cook's distance. To determine an appropriate threshold value for the classical Cook's distance, its distribution can be obtained by bootstrap if the true model is a parametric linear model. However, such a bootstrap procedure requires a parametric model assumption and can be computationally expensive especially for high-dimensional data. By contrast, the asymptotic distribution of the proposed influence measure is attainable in our setup, since the predictor dimension goes to infinity along with the sample size, and the threshold is easy to obtain.

When facing high-dimensional data diagnosis, an intuitive solution is to continue using the classical Cook's distance but to replace the OLS coefficient estimate with a regularized estimate, for instance, a LASSO estimate. This modified Cook's distance approach could be particularly useful when data perturbation concentrates on the nonzero coefficients, as it avoids unnecessary variability caused by irrelevant covariates. However, it also has several limitations. First, this solution interweaves influence diagnosis with variable selection, which can be flawed if the influence is reflected on variable selection itself. For instance, an influential observation may substantially alter the chosen tuning parameter of the LASSO, resulting in a totally different regularized coefficient estimate, which in turn affects the modified Cook's distance. Second, the tuning parameter of the LASSO, in principle, should be updated for every reduced data set, and this re-estimation requirement can be very expensive computationally, especially when the regression dimension p is large. Third, the asymptotic properties of the modified Cook's distance seem intractable analytically, which makes the thresholding of influential data difficult, whereas a bootstrap alternative to choose the thresholding value is again computationally expensive. Moreover, while there exist many competing variable selection methods, it is unclear which selection method is the best choice in the context of influence diagnosis. By contrast, our influence measure is not constrained by any particular variable selection method, and this flexibility could benefit downstream analysis. In Section 3, we carry out an intensive numerical study to compare this modified Cook's distance with our proposal, and this detailed comparison can be viewed as the third contribution of this article.

Before we proceed, we quickly show a simulated example to illustrate two points: first, how various aspects of a high-dimensional regression analysis, including regression coefficient estimation, variable selection and variable screening, can be seriously affected by influential observations, and second, how our proposed measure can help limit such influence. The data was generated from a linear model with $p = 1000$ predictors, $n = 100$ observations, among which 10 observations were influential. The magnitude of the influence was dictated by a scalar κ with a larger value indicating a larger influence. More details can be found in the setup of model 1 in Section 3. Evaluations include error in coefficient estimation, error in variable selection after applying the LASSO [34], and error in variable screening after applying the SIS [22]. The results are averaged over 200 simulation replicates, and are reported in Figure 1. It is clearly seen from the plot that, influential observations could have drastic effects on various features for high-dimensional data analysis. Meanwhile, our marginal correlation based diagnosis could greatly help control the adverse effects after detecting and removing those influential data points.

The rest of the artlicle is organized as follows. Section 2 begins with a review of the classical Cook's distance, then presents our new high-dimensional influence measure, along with a comparison with the Cook's distance, the asymptotic properties and a power study. Section 3 includes an intensive simulation study and a

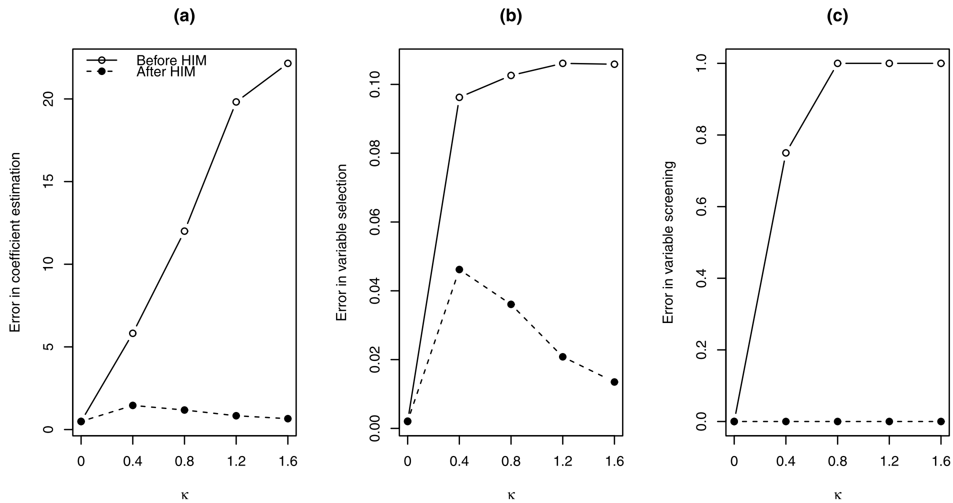

FIG. 1. *Effect of influential points on parameter estimation* (a), *variable selection* (b) *and variable screening* (c), *as the perturbation parameter* κ *varies. "Before HIM" denotes the analysis on the full data*, *and "After HIM" denotes the analysis on the reduced data after removing the influential observations flagged by our proposed high-dimensional measure* (*HIM*).

microarray data analysis. Section 4 presents a generalization of our proposal from the normal linear model to the generalized linear model. Section 5 concludes the paper with a discussion. All technical proofs are given in the Appendix and the supplementary material [41].

2. High-dimensional influence measure.

2.1. *Linear models and classical Cook's distance.* In this article, we focus on influence diagnosis in the context of the classical linear regression model. Meanwhile, we note that the proposed idea can be readily extended to a much broader class of regression models, and we will discuss one such extension in Section 4. Consider the following model:

$$Y_i = \beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta}_1 + \varepsilon_i, \tag{2.1}$$

where the pair $(Y_i, \mathbf{X}_i)$, $1 \le i \le n$, denote the observation of the ith subject, $Y_i \in \mathbb{R}$ is the response variable, $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^\top \in \mathbb{R}^p$ is the associated p-dimensional predictor vector, and $\varepsilon_i \in \mathbb{R}$ is a mean zero normally distributed random noise. Let $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^\top)^\top$ denote the coefficient vector. Under the classical setup of $n > p$, the OLS estimate of $\boldsymbol{\beta}$ is obtained by minimizing the objective function $\sum_{i=1}^n (Y_i - \beta_0 - \mathbf{X}_i^\top \boldsymbol{\beta}_1)^2$, and the solution is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$, where $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ denotes the $n \times 1$ response vector, and $\mathbf{X}$ denotes the $n \times (p+1)$ design matrix with the ith row being $p+1$ dimensional vector $(1, \mathbf{X}_i^\top)$, $i = 1, \dots, n$.

To quantify the influence of the kth observation on regression, $1 \le k \le n$, [12] employed the leave-one-out idea by studying the OLS estimate of β while the kth observation is excluded from estimation. That is, one minimizes the modified objective function $\sum_{i=1,i\neq k}^{n}(Y_i - \beta_0 - \mathbf{X}_i^\top \boldsymbol{\beta}_1)^2$. The new estimate is of the form $\hat{\boldsymbol{\beta}}^{(k)} = (\mathbf{X}_{(k)}^\top \mathbf{X}_{(k)})^{-1}\mathbf{X}_{(k)}^\top \mathbf{Y}_{(k)}$, where $\mathbf{Y}_{(k)}$ is the $(n-1)\times 1$ response vector with Y_k removed, and $\mathbf{X}_{(k)}$ is the $(n-1)\times(p+1)$ design matrix with the kth row $\mathbf{X}_k$ removed. Cook [12] naturally chose the estimate of $\boldsymbol{\beta}$ to define influence, and intuitively, if an observation is influential, the difference between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{(k)}$ is expected to be large. This leads to the following discrepancy measure, that is, the Cook's distance:

$$D_k = \frac{\{\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}\}^\top \mathbf{X}^\top \mathbf{X}\{\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}\}}{(p+1)\hat{\sigma}^2}, \tag{2.2}$$

where $\hat{\sigma}^2 = (n-p-1)^{-1}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \mathbf{X}_i^\top \hat{\boldsymbol{\beta}})^2$.

In the high-dimensional regression setting, the classical Cook's distance (2.2) encounters some difficulties. When p is close to n, the OLS estimate is known to be unstable, which would in turn cause D_k to be unstable. When $p > n$, the classical Cook's distance is not directly computable, because the OLS estimator $\hat{\beta}$ becomes unstable. For those reasons, the regression coefficient estimate may no longer be the best choice to define influence in high-dimensional analysis. This motivates us to consider an alternative influence measure for high-dimensional data.

2.2. *High-dimensional influence measure.* In high-dimensional regression analysis where $p \approx n$ or $p > n$, variable selection (screening) plays a central role, whereas marginal covariance or correlation is crucial to the majority of variable selection approaches. Motivated by this observation, for high-dimensional data influence diagnosis, we choose marginal correlation, instead of regression coefficient, as the feature that defines influence. Individual observation's influence on marginal correlation is to transmit to various features of downstream analysis, such as variable selection and coefficient estimation.

More specifically, we first define the marginal correlation as $\rho_j = E\{(X_j - \mu_{xj})(Y - \mu_y)\}/(\sigma_{xj}\sigma_y)$, where $\mu_{xj} = E(X_j)$, $\mu_y = E(Y)$, $\sigma_{xj}^2 = \operatorname{var}(X_j)$ and $\sigma_y^2 = \operatorname{var}(Y)$. We then obtain the sample estimate, $\hat{\rho}_j = \{\sum_{i=1}^{n}(X_{ij} - \hat{\mu}_{xj})(Y_i - \hat{\mu}_y)\}/\{n\hat{\sigma}_{xj}\hat{\sigma}_y\}$, for $j = 1,\dots,p$, where $\hat{\mu}_{xj}, \hat{\mu}_y, \hat{\sigma}_{xj}$ and $\hat{\sigma}_y$ are the sample estimates of $\mu_{xj}, \mu_y, \sigma_{xj}$ and σ_y, respectively. Next, we continue to use the leave-one-out principle as in the classical Cook's distance case, and compute the marginal correlation with the kth observation removed as

$$\hat{\rho}_j^{(k)} = \frac{\sum_{i=1,i\neq k}^{n}(X_{ij} - \hat{\mu}_{xj}^{(k)})(Y_i - \hat{\mu}_y^{(k)})}{(n-1)\hat{\sigma}_{xj}^{(k)}\hat{\sigma}_y^{(k)}}, \qquad j = 1,\dots,p, k = 1,\dots,n,$$

where $\hat{\mu}_{xj}^{(k)}, \hat{\mu}_{y}^{(k)}, \hat{\sigma}_{xj}^{(k)}$ and $\hat{\sigma}_{y}^{(k)}$ are the corresponding sample estimates with the kth observation removed. Finally, we define the influence measure based on the marginal correlation as

$$\mathcal{D}_k = \frac{1}{p}\sum_{j=1}^{p}(\hat{\rho}_j - \hat{\rho}_j^{(k)})^2. \tag{2.3}$$

We refer to $\mathcal{D}_k$ as the *high-dimensional influence measure*, or HIM for brevity. We make a few remarks. First, we note that the marginal correlation can be easily computed regardless of the predictor dimension, and such computational advantage is practically very useful for high-dimensional data analysis. Second, the proposed influence measure is built upon the marginal correlation coefficient, and is effectively scale invariant. However, it does *not* imply that marginal correlation is the ultimate feature of interest in our influence diagnosis. Instead, a substantial change on the marginal correlation caused by a data point is to exert influence on important features such as variable selection and parameter estimation, as we have seen in Figure 1. As such, for an estimation method to be robust to unexpected *perturbation* [15, 42, 43], the sample marginal correlation should be sufficiently robust. This is an important and necessary condition, although not necessarily sufficient. Finally, use of the *marginal* correlation to define the influence measure does *not* imply that we assume a *marginal model*. Instead, we still assume the joint model (2.1). As it may seem unclear how a marginal measure can capture the influence for a joint model, we will demonstrate through a simple joint model later in Section 2.5 that, the newly defined $\mathcal{D}_k$ can indeed identify the influential observation with probability one. This use of marginal correlation is also similar in spirit to the sure independence screening procedure for a joint normal model [22], but is in a different context. Fan and Lv [22] use marginal correlation for the variable screening purpose, while we use it for influence diagnosis.

The proposed high-dimensional influence measure also shares some similarity as the classical Cook's distance. Note that the Cook's distance can be reformulated as

$$D_k = \frac{\hat{\epsilon}_k^2}{p\hat{\sigma}^2}\frac{h_{kk}}{(1-h_{kk})^2}, \qquad k = 1, \dots, n, \tag{2.4}$$

where $\hat{\epsilon}_k = \hat{Y}_k - Y_k$ is the residual and $h_{kk} = \mathbf{X}_k^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_k$, $k = 1, \dots, n$ is the (k)th diagonal element of the hat matrix $\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. Clearly, D_k is an increasing function of both $|\hat{\epsilon}_k|$ and h_{kk}. As such, an observation has a large value in Cook's distance, if it has a large residual or it is a high leverage point in terms of h_{kk}. Our proposed information measure shares a similar spirit. In Section 2.3, we will derive a decomposition of our influence measure $\mathcal{D}_k$ under some conditions, and will show that $\mathcal{D}_k$ is mainly dominated by a term called B_2, which is of the form

$$B_2 = \frac{(n-2)}{pn(n-1)^2} \sum_{j=1}^{p} Y_k^2 X_{kj}^2 = \frac{(n-2)}{pn(n-1)^2} Y_k^2 \|\mathbf{X}_k\|^2.$$

Consequently the kth data point $(\mathbf{X}_k, Y_k)$ is more likely to be marked influential, if it has a large response and a large value of $\|\mathbf{X}_k\|^2$. Here $\|\mathbf{X}_k\|^2$ plays a similar role as h_{kk} in the classical Cook's distance, for detecting influential points induced mainly by covariates, whereas Y_k plays a similar role as the residual in the Cook's distance, for detecting the influential point induced by abnormal responses.

2.3. *Theoretical properties.* We next establish the asymptotic distribution of the proposed high-dimensional influence measure $\mathcal{D}_k$ as both the sample size n and the dimensionality p go to infinity. Toward that end, we impose the following conditions.

(C.1) For any fixed $j = 1, \dots, p$, ρ_j is constant and does not change as p increases.

(C.2) For the covariance matrix $\mathbf{\Sigma} = \operatorname{cov}(\mathbf{X})$, with the eigen-decomposition $\mathbf{\Sigma} = \sum_{j=1}^{p} \lambda_j \mathbf{u}_j \mathbf{u}_j^\top$, it is assumed that $l_p = \sum_{j=1}^{p} \lambda_j^2 = O(p^r)$ for some $0 \le r < 2$.

(C.3) The predictor X_i follows a multivariate normal distribution and the random noise ε_i follows a normal distribution.

Condition (C.1) is very general, since it only requires that for any fixed j, ρ_j is a constant independent of p. A sufficient condition for condition (C.2) to hold is that all eigenvalues of $\mathbf{\Sigma}$ are finite. This condition also permits eigenvalues of $\mathbf{\Sigma}$ to diverge to infinity but at a slower rate compared to the dimensionality. The normality assumption on $\mathbf{X}$ is mainly for convenience, and can be relaxed, for instance, to distributions with sub-Gaussian tails, at the expense of more lengthy proofs. In addition, since the error term is assumed normal, Y is normally distributed.

Next, we derive a decomposition of $\mathcal{D}_k$, that is, to serve as a basis for its asymptotic distribution. The result is presented in a way such that μ_y, μ_{xj} are assumed to be 0 and σ_{xj}, σ_y are 1 for $1 \le j \le p$. This leads to simplified estimates $\hat{\rho}_j = n^{-1} \sum_{1 \le i \le n} X_{ij} Y_i$ and $\hat{\rho}_j^{(k)} = n^{-1} \sum_{i \ne k} X_{ij} Y_i$. On the other hand, we note that this standardization is only for the purpose of simplifying the presentation and it loses no generality. As we will show later in Proposition 2, replacing the unknown quantities μ_{xj}, μ_y, σ_{xj} and σ_y with their consistent sample estimates would not alter $\mathcal{D}_k$'s asymptotic distribution. For $t, s = 1, \dots, n$, let $K_{p,ts} = \sum_j X_{tj} X_{sj} / p$ and $c_p = \max_{1 \le j \le p} \lambda_j$. After some algebraic computation, we obtain that

$$
\begin{aligned}
\mathcal{D}_k &= \frac{1}{p}\sum_{j=1}^{p}\left\{\frac{1}{n(n-1)}\sum_{1\le t\le n}^{t\neq k} Y_t X_{tj} - \frac{1}{n} Y_k X_{kj}\right\}^2 \\
&= \frac{1}{\{n(n-1)\}^2}\sum_{t=1}^{n} Y_t^2 K_{p,tt} + \frac{(n-2)}{n(n-1)^2} Y_k^2 K_{p,kk} \\
&\quad + \frac{1}{[n(n-1)]^2}\sum_{t\neq s} Y_t Y_s K_{p,ts} - \frac{2}{n(n-1)^2}\sum_{t=1,t\neq k}^{n} Y_k Y_t K_{p,tk} \\
&:= B_1 + B_2 + B_3 - 2B_4.
\end{aligned}
\tag{2.5}
$$

Then we have the following result on the expectation of $\mathcal{D}_k$ along with the variance of its decomposition in terms of B's.

PROPOSITION 1. *Suppose that* $(\mathbf{X}_i, Y_i)$ *are i.i.d. observations and that* (C.1) *and* (C.3) *hold. Then it holds that*

$$
E(\mathcal{D}_k) = \left[n(n-1)\right]^{-1} E\left(Y_k^2\right) E(K_{p,kk}) + O\left(n^{-2} p^{-1} l_p^{1/2}\right).
$$

In addition, $\operatorname{var}(B_1) = O(n^{-7})$, $\operatorname{var}(B_2) = O(n^{-4})$, $\operatorname{var}(B_3) = O(c_p^2 n^{-5} p^{-2}) + O(p^{-2} n^{-6})$ *and* $\operatorname{var}(B_4) = O(l_p p^{-2} n^{-5}) + O(c_p^2 p^{-2} n^{-4})$.

Now we return to the asymptotic distribution of $\mathcal{D}_k$. Proposition 1 helps to derive the asymptotic distribution of $\mathcal{D}_k$. We first present the result assuming μ_{xj}, μ_y, σ_{xj} and σ_y are all known. Then we obtain the asymptotic distribution when μ_{xj}, μ_y, σ_{xj} and σ_y are replaced by their sample estimates.

THEOREM 1. *Suppose that* (C.1)–(C.3) *hold. When there is no influential point and* $\min\{n, p\} \longrightarrow \infty$, *we have*

$$
n^2 \mathcal{D}_k \longrightarrow \chi^2(1),
$$

where $\chi^2(1)$ *is the chi-square distribution with one degrees of freedom.*

Next, we consider the asymptotic distribution of $\mathcal{D}_k$ when μ_{xj}, μ_y, σ_j and σ_y are unknown. A natural choice is to replace them by their corresponding sample moment estimates as $\hat{\mu}_y = \sum_i Y_i/n$, $\hat{\mu}_{xj} = \sum_i X_{ij}/n$, $\hat{\sigma}_{xj}^2 = \sum_i (X_{ij} - \hat{\mu}_{xj})^2/(n-1)$ and $\hat{\sigma}_y^2 = \sum_i (Y_i - \hat{\mu}_y)^2/(n-1)$. Another choice is to employ robust estimators, for example, the median in place of the mean, and the median absolute deviation in place of the standard deviation. The following proposition shows that the conclusion of Theorem 1 continues to hold as long as u_{xj}, u_y, σ_{xj} and σ_y are replaced by $\sqrt{n}$-consistent estimates under certain moment assumptions. Let $\dot{Y}_t = (Y_t - \mu_y)/\sigma_y$, $\dot{X}_{tj} = (X_{tj} - u_{tj})/\sigma_{tj}$, $t = 1, \dots, n$, $j = 1, \dots, p$ and $(Q_{xj}, R_{xj}) =$

$((\hat{\mu}_{xj} - \mu_{xj})/\sigma_{xj}, \sigma_{xj}/\hat{\sigma}_{xj})$ and (Q_y, R_y) are defined similarly. Furthermore, let $S_{Qx} = \limsup_{n\to\infty} E(n^{1/2}Q_{x1})^8$, $S_{Rx} = \limsup_{n\to\infty} E[n^{1/2}(R_{x1} - 1)]^8$, $S_{Qy} = \limsup_{n\to\infty} E(n^{1/2}Q_y)^8$ and $S_{Ry} = \limsup_{n\to\infty} E[n^{1/2}(R_y - 1)]^8$. We make the following additional assumption.

(C.4) For all $1 \le j \le p$, (Q_{xj}, R_{xj}) are the same symmetric function of $\{\dot{X}_{tj}$, for $t = 1, \ldots, n\}$; and (Q_y, R_y) are also the same symmetric function of $\dot{Y}_t$ for $t = 1, \ldots, n$. We assume that S_{Qx}, S_{Rx}, S_{Qy} and S_{Ry} are finite.

Condition (C.4) indicates that, for all $1 \le j \le p$, $((\hat{\mu}_{xj} - \mu_{xj})/\sigma_{xj}, \hat{\sigma}_{xj}/\sigma_{xj}) = f(\dot{X}_{1j}, \ldots, \dot{X}_{nj})$, where $f(x_1, \ldots, x_p) = (f_1(x_1, \ldots, x_p), f_2(x_1, \ldots, x_p))$ and f_1 and f_2 are symmetric functions. Condition (C.4) is a mild condition. Recall that $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$ are i.i.d. normal in Theorem 1. When $\hat{\mu}_{xj}, \hat{\sigma}_{xj}$ are the moment estimates, we have $Q_{xj} = n^{-1}\sum_{1\le t\le n} \dot{X}_{tj} \sim N(0, 1/n)$ and consequently S_{Qx} is finite. Moreover, we have $R_{xj} = 1/S_{nj}$ where S_{nj}^2 is the sample variance of $\{\dot{X}_{tj}, t = 1, \ldots, n\}$. Since $S_{n1}^2 \sim \chi^2_{n-1}/(n - 1)$, it is easy to verify that S_{Rx} is also finite. Similarly, S_{Qy} and S_{Ry} are also finite with moment estimates $\hat{\mu}_y$ and $\hat{\sigma}_y$. Under the normality of $(\mathbf{X}, Y)$, (C.4) also holds for some robust estimates.

PROPOSITION 2. *Assume that $\hat{\mu}_{xj}, \hat{\sigma}_{xj}, \hat{\mu}_y, \hat{\sigma}_y$ are $\sqrt{n}$-consistent and satisfy* (C.4). *Substituting $\mu_{xj}, \mu_y, \sigma_j, \sigma_y$ with their corresponding estimates in $\mathcal{D}_k$, Theorem* 1 *continues to hold under the same conditions.*

We remark that the asymptotic distribution of the high-dimensional influence measure $\mathcal{D}_k$ is obtained as the number of predictor p goes to infinity. This is different from the case of classical Cook's distance D_k where p is fixed, for which a standard asymptotic distribution is not attainable. We view this as a *blessing* of dimensionality in contrast to the usually conceived *curse* of dimensionality. For more examples of blessing of dimensionality, see [17] and [28].

2.4. *Influence diagnosis.* An important implication of Theorem 1 is that we can now obtain a p-value for influence diagnosis. Specifically, for the hypothesis that the kth observation is not influential versus its alternative, the p-value is $P(\chi^2(1) > n^2\mathcal{D}_k)$. Given that the number of predictors p is usually large and multiple hypotheses are tested simultaneously, we employ the false discovery rate based multiple testing procedure of [5] to determine which hypothesis should be rejected while controlling the family-wise error. Denote n_{infl} as the number of influential observations among the n observations, n_{tp} and n_{fp} as the number of the observations that are correctly rejected and incorrectly rejected, respectively, and r as the total number of rejections in the n hypotheses testing. Then the power and the false discovery rate are denoted as Power $= n_{\text{tp}}/n_{\text{infl}}$ and FDR $= n_{\text{fp}}/r$, respectively. We will set FDR level being small, such as 0.05, and report the power and other quantities in the numerical study section. We also remark that more sophisticated alternative multiple testing procedure, for example, in [6], [20] and [32], can be used in conjunction with our approach, but, that is, not the focus of this article.

2.5. *A power comparison of two influence measures.* We next study the power property of both the new diagnosis measure and the Cook's distance via a simple model. This study serves two purposes. First, we can gain insight about difference between the two diagnosis measures. Second, it offers evidence that the marginal correlation based measure is capable of detecting influential observation in a joint model with a large probability.

More specifically, we consider the model (2.1), but drop the intercept for simplicity. The predictors $\mathbf{X}_i, i = 1, \ldots, n$, are i.i.d. observations from a multivariate normal distribution $N_p(0, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is a $p \times p$ covariance matrix with all its diagonal elements $\sigma_{jj} = 1$. The error term ϵ_i is of the structure $\epsilon_i = e_i + c_i$, where e_i follows a standard normal distribution and c_i is constant, $c_2 = \cdots = c_n = 0$. Under this setup, the first observation is an influential point as long as $c_1 \neq 0$, and we aim to establish the power of both the classical and our proposed high-dimensional influence measure in identifying this influential observation. Let D_i be the Cook's distance defined in (2.2) for the ith observation, $\mathcal{D}_i^{(c)}$ be the proposed high-dimensional measure in (2.3), and $\mathcal{T}_i^{(c)} = n^2 \mathcal{D}_i^{(c)}$ be the statistic defined in Theorem 1. Moreover, consider the following condition:

(C.5) All eigenvalues of $\mathbf{\Sigma}$ are positive and bounded.

Then the next theorem states that, both the classical and the high-dimensional Cook's distance have the power of detecting the influential observation approaching one under appropriate yet different conditions.

THEOREM 2. *Consider the model stated above.*

1. *Suppose that* (C.1) *and* (C.5) *hold. If* $\max\{n^{-1}p^6, |c_1|^{-1}n^{2/3}\} \to 0$, *then we have that for the Cook's distance* D_i, $P(nD_1 - \max_{2 \le i \le n} nD_i > M) \to 1$ *for any* $M > 0$, *when* $n \to \infty$.
2. *Suppose that* (C.1) *and* (C.2) *hold. If* $\max\{|c_1|^{-1}(\log n)^{1/2}, l_p p^{-2} c_1^{-4} n\} \to 0$, *then we have that for the proposed high-dimensional influence measure* $\mathcal{D}_i^{(c)}$, $P(\mathcal{T}_1^{(c)} - \max_{2 \le i \le n} \mathcal{T}_i^{(c)} > M) \to 1$ *for any* $M > 0$, *when* $\min(n, p) \to \infty$.

The proof is given in the supplementary material [41]. Here we compare the two sets of conditions to gain some insight about the difference of the two diagnosis measures. First, we examine the condition $\max\{n^{-1}p^6, |c_1|^{-1}n^{2/3}\} \to 0$, that is, required by the Cook's distance. The condition $|c_1|^{-1}n^{2/3} \to 0$ here is to ensure that the influence of the first observation does not vanish as n goes to infinity. Moreover, in terms of the predictor dimension p, the classical Cook's distance is defined when $p < n$. Consequently, the condition $n^{-1}p^6 \to 0$, or equivalently, $p = o(n^{1/6})$, constrains the growing rate of p with n at a much slower rate. We note that although this rate may not be the optimal one, the condition $p = o(n)$ is clearly necessary for the classical Cook's distance to be feasible. Next, we examine the condition $\max\{|c_1|^{-1}(\log n)^{1/2}, l_p p^{-2} c_1^{-4} n\} \to 0$, that is, required by our new influence measure. For illustration, we consider a simple case with all the eigenvalues of Σ bounded and $p > n$. We know immediately that both l_p/p and n/p are bounded. Accordingly, we should have $l_p p^{-2} c_1^{-4} n \to 0$ as long as $c_1 \to \infty$. As $\log n \to \infty$ when $n \to \infty$, then a sufficient condition for $\max\{|c_1|^{-1}(\log n)^{1/2}, l_p p^{-2} c_1^{-4} n\} \to 0$ is that $(\log n)^{1/2}/|c_1| \to 0$. This suggests that the influence point can be consistently detected, as long as c_1 diverges to infinity at a speed faster than $(\log n)^{1/2}$. This is clearly a rate much slower than $n^{2/3}$. Finally, the bounded eigenvalue condition (C.5) is commonly used in the literature for estimating covariance matrices [7]. Here it is assumed for the Cook's distance case. For the new diagnosis measure, (C.2) is required instead, which is weaker than (C.5).

3. Numerical studies. We have carried out an intensive simulation study, along with a microarray data analysis, to examine the empirical performance of our proposed high-dimensional influence measure. Since the classical Cook's distance depends on both leverage points and outliers, in our simulation study, we consider three different scenarios where there exist outliers only (model 1), leverage points only (model 2), or mixed leverage points and outliers (model 3). For the scenarios with leverage points (models 2 and 3), we further consider sub-scenarios where important covariates contribute to leverage observations, or noisy covariates contribute to leverage observations. Below we present the summary of the analysis.

3.1. *Simulation models.* For all simulations, we set the sample size $n = 100$, and the number of predictors $p = 1000$. We set 10% of total observations as influential, so that $\tilde{n} = 10$. We consider the model

$$Y_i = \mathbf{X}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \dots, n,$$

where $\mathbf{X}_i$ is multivariate normal with $\operatorname{cov}(X_{ij}, X_{ij'}) = 0.5^{|i-j|}$, ε_i follows the standard normal distribution, and $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, \dots, 0)^\top$. We simulated $n = 100$ i.i.d. observations from this model. Next, we reset the first $\tilde{n} = 10$ data observations as coming from another model,

$$\tilde{Y}_i = \tilde{\mathbf{X}}_i^\top \tilde{\boldsymbol{\beta}} + \varepsilon_i, \qquad i = 1, \dots, \tilde{n},$$

where perturbations are to be introduced on the regression coefficient, the covariates and their combination. In particular, we have considered three perturbation models of generating influential points.

Model 1. The perturbation was introduced on the response. That is, for $i = 1, \dots, \tilde{n}$, $\tilde{\mathbf{X}}_i = \mathbf{X}_i$, and $\tilde{\boldsymbol{\beta}} = (3, 1.5, \kappa, \kappa, 2, \kappa, \dots, \kappa)^\top$. In other words, the influential observations are generated according to $\tilde{Y}_i = \mathbf{X}_i^\top \boldsymbol{\beta} + \kappa Z_i + \varepsilon_i$, where $Z_i = \mathbf{X}_i^\top \gamma$ and $\gamma = (0, 0, 1, 1, 0, 1, 1, \dots, 1)^\top$. In this case, the responses of the influential observations are contaminated by a random perturbation κZ_i. Consequently, the corresponding responses admit a different pattern, whereas the predictors of influential observations follow the same distribution as the rest.

Model 2. The perturbation was introduced on the predictors and keep the response uncontaminated. That is, for $i = 1, \dots, \tilde{n}$, $\tilde{Y}_i = Y_i$ and $\tilde{X}_{ij} = X_{ij} +$

$30\kappa I_{\{j\in S\}}$, $j = 1, \ldots, p$. In other words, a set S of predictors admit a different pattern, and its magnitude is controlled by the scalar κ. We examined three choices of S: $S_1 = \{1, \ldots, 100\}$, and in this case, the influenced predictors overlap with those truly relevant ones $\{1, 2, 5\}$ in $\boldsymbol{\beta}$; $S_2 = \{p - 100, \ldots, p\}$, and as such there is no overlap; and $S_3 = \{1, \ldots, p\}$, and in this case, all predictors are subjected to potential influence.

Model 3. The perturbation was introduced on both the response and the predictors. That is, $\tilde{\boldsymbol{\beta}} = (3, 1.5, \kappa, \kappa, 2, \kappa, \ldots, \kappa)^\top$ and $\tilde{X}_{ij} = X_{ij} + 30\kappa I_{\{j\in S\}}$, $j = 1, \ldots, p$. Again, we considered three sets of S as described earlier.

It is clear that κ is the parameter that dictates the magnitude of the influential points. When $\kappa = 0$, there is no influential point. We used $\kappa = 0$, 0.4, 0.8, 1.2 and 1.6 in our experiment.

3.2. *Performance evaluation*. We evaluate and compare our proposed influence measure in several ways. First, we study the potential impact of influential data and how the proposed diagnosis measure could help limit such impact. Toward that end, we first applied the LASSO or SIS to the full data. Then we computed the proposed high-dimensional influence measure, evaluated the corresponding p-value, and applied the multiple testing procedure of [5], with the false discovery rate fixed at $\alpha = 5\%$. We then obtained a reduced data set by removing those flagged influential points and applied the LASSO or SIS to the reduced data set. We evaluated the impact of influential data in terms of coefficient estimation, variable selection, and variable screening. For coefficient estimation, we report the error between the estimated and true $\boldsymbol{\beta}$, $\text{ERR} = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\text{true}}\|_2$; for variable selection, we report the false positive rate, FPR = #False Positive/#True Negative; and for variable screening, we report the coverage probability CP. In addition, we also report the empirical power of our influence identification procedure.

Second, we compare our method to two potentially competing solutions in high-dimensional influence diagnosis. One is a modified Cook's distance based on the LASSO. That is, we continue to employ the classical Cook's distance, but estimate the regression coefficient $\boldsymbol{\beta}$ under a LASSO penalty and as such avoid the difficulty of the OLS estimate when $p > n$. This seems a very natural solution. We compare it with our proposal in terms of estimation accuracy, selection accuracy and power. On the other hand, we note the lack of asymptotic theory for this modified Cook's distance. To determine the threshold for influential data, one may use bootstrap. However, in our comparison, we simply label the observations with the largest $\tilde{n}$ modified Cook's distance as influential. This is not feasible in practice, but provides a useful benchmark for comparison. The other competing solution is the penalized least absolute deviation via the LASSO penalty (LAD + LASSO) [4, 37]. Due to the use of the least absolute deviation as the loss function, this method is designed to handle heavy tailed errors in linear regression, and as such a potentially useful way to limit impact of the influence observations.

TABLE 1
Simulation results for perturbation model 1. *HIM denotes our proposed high-dimensional diagnosis measure, and CD denotes the classical Cook's distance*

Method	**Criterion**	κ **0**	**0.4**	**0.8**	**1.2**	**1.6**
SIS	CP	1	0.25	0	0	0
SIS + HIM	CP	1	1	1	1	1
LASSO	ERR	0.510	4.917	9.553	14.636	18.478
	FPR	0.002	0.094	0.103	0.107	0.106
LASSO + HIM	ERR	0.519	1.296	1.020	0.872	0.769
	FPR	0.002	0.045	0.029	0.015	0.012
	Power	–	0.6	0.765	0.865	0.865
LASSO + CD	ERR	0.535	1.136	2.176	2.565	4.182
	FPR	0.003	0.034	0.066	0.072	0.076
	Power	–	0.630	0.670	0.700	0.660
LAD + LASSO	ERR	0.642	1.920	2.073	2.406	1.769

3.3. *The results*. The averages of a total of 200 random replications are reported in Tables 1–3. We make the following observations.

(1) First, the presence of influential points significantly affects variable selection and screening accuracy. This can be seen by comparing the results between SIS and SIS + HIM in terms of CP. Consider, for example, Table 1. As κ increases, the coverage probability of the SIS method deteriorates quickly from 1 with $\kappa = 0$ to 0 with $\kappa = 1.6$. This confirms that influential observations do affect variable screening consistency. Meanwhile, the performance of SIS + HIM is quite encouraging as its CP values maintains at 1 for every κ value considered. This suggests that the proposed HIM method helps SIS in removing the influential observations.

(2) Second, the presence of influential observations does affect estimation accuracy seriously. This can be seen clearly by comparing the results of LASSO and LASSO + HIM in terms of ERR values. For instance, the ERR values in Table 3 for LASSO with S_1 increases quickly from 0.446 with $\kappa = 0$ to 14.498 with $\kappa = 1.6$. This confirms that influential observations do affect the accuracy of the LASSO estimate in a negative way. However, we find that the ERR values of LASSO + HIM are always well controlled with ERR < 1.5. In fact, as κ increases, the power for HIM to detect influential observation increases. Thus, those influential observations are more likely to be detected and eliminated from the data analysis. This makes the ERR values of LASSO + HIM eventually converges to a level around ERR ≈ 0.5, as κ increases. This confirms the usefulness of the HIM method for LASSO estimation, even though its definition only involves marginal correlation coefficients.

TABLE 2
Simulation results for perturbation model 2. HIM denotes our proposed high-dimensional diagnosis measure, and CD denotes the classical Cook's distance

			κ				
Subset	**Method**	**Criterion**	**0**	**0.4**	**0.8**	**1.2**	**1.6**
S_1	SIS	CP	1	0.05	0	0	0
	SIS + HIM	CP	1	0.05	0.1	0.3	0.25
	LASSO	ERR	0.439	4.917	4.972	4.971	4.954
		FPR	0.002	0.086	0.090	0.089	0.089
	LASSO + HIM	ERR	0.455	4.803	4.591	3.055	3.136
		FPR	0.002	0.080	0.060	0.055	0.044
		Power	–	0.620	0.775	0.892	0.930
	LASSO + CD	ERR	0.513	4.566	4.568	4.603	4.533
		FPR	0.004	0.073	0.073	0.070	0.070
		Power	–	0.095	0.085	0.105	0.115
	LAD + LASSO	ERR	0.642	1.339	1.303	1.320	1.330
S_2	SIS	CP	1	1	1	1	1
	SIS + HIM	CP	1	1	1	1	1
	LASSO	ERR	0.509	0.456	0.439	0.450	0.469
		FPR	0.001	0.001	0.001	0.002	0.002
	LASSO + HIM	ERR	0.521	0.494	0.493	0.494	0.506
		FPR	0.001	0.001	0.001	0.002	0.002
		Power	–	0.695	0.8	0.85	0.895
	LASSO + CD	ERR	0.548	0.523	0.532	0.556	0.551
		FPR	0.001	0.002	0.002	0.002	0.002
		Power	–	0.065	0.085	0.135	0.115
	LAD + LASSO	ERR	0.642	0.650	0.645	0.647	0.634
S_3	SIS	CP	1	0.35	0.45	0.30	0.25
	SIS + HIM	CP	1	0.50	0.60	0.62	0.65
	LASSO	ERR	0.473	1.567	1.545	1.598	1.609
		FPR	0.003	0.051	0.053	0.051	0.055
	LASSO + HIM	ERR	0.490	1.517	1.456	1.221	1.115
		FPR	0.003	0.034	0.031	0.023	0.033
		Power	–	0.735	0.86	0.95	0.95
	LASSO + CD	ERR	0.560	1.751	1.700	1.743	1.871
		FPR	0.003	0.047	0.042	0.042	0.048
		Power	–	0.115	0.085	0.115	0.110
	LAD + LASSO	ERR	0.642	0.608	0.573	0.580	0.581

(3) Third, the performance of LASSO + CD is mixed. If the perturbation is due to the response only as in Table 1, it does yield much better performance than LASSO with much smaller ERR values. This suggests that LASSO + CD can perform well to limit the effect of influential points. However, even for this

TABLE 3

Simulation results for perturbation model 3. *HIM denotes our proposed high-dimensional diagnosis measure, and CD denotes the classical Cook's distance*

			κ				
Subset	**Method**	**Criterion**	**0**	**0.4**	**0.8**	**1.2**	**1.6**
S_1	SIS	CP	1	1	0.65	0.10	0.05
	SIS + HIM	CP	1	0.90	1	1	1
	LASSO	ERR	0.446	1.559	5.308	9.628	14.498
		FPR	0.002	0.062	0.093	0.099	0.098
	LASSO + HIM	ERR	0.447	1.278	0.771	0.499	0.542
		FPR	0.002	0.046	0.027	0.003	0.002
		Power	–	0.185	0.94	1	1
	LASSO + CD	ERR	0.559	0.686	2.149	5.623	10.926
		FPR	0.002	0.009	0.063	0.084	0.090
		Power	–	0.555	0.720	0.675	0.585
	LAD + LASSO	ERR	0.642	1.416	4.367	8.740	13.252
S_2	SIS	CP	1	1	0.05	0	0
	SIS + HIM	CP	1	1	1	1	1
	LASSO	ERR	0.479	2.090	6.619	11.997	17.279
		FPR	0.002	0.072	0.095	0.101	0.101
	LASSO + HIM	ERR	0.494	1.836	0.696	0.475	0.501
		FPR	0.002	0.062	0.009	0.002	0.002
		Power	–	0.145	0.955	1	1
	LASSO + CD	ERR	0.501	0.769	3.702	7.676	14.585
		FPR	0.003	0.016	0.078	0.087	0.091
		Power	–	0.605	0.680	0.685	0.520
	LAD + LASSO	ERR	0.642	1.859	5.855	10.829	16.157
S_3	SIS	CP	1	1	0.1	0	0
	SIS + HIM	CP	1	1	1	1	1
	LASSO	ERR	0.464	1.682	5.720	10.943	17.384
		FPR	0.002	0.065	0.098	0.103	0.105
	LASSO + HIM	ERR	0.484	1.479	1.262	0.557	0.515
		FPR	0.002	0.057	0.034	0.003	0.002
		Power	–	0.1	0.87	1	1
	LASSO + CD	ERR	0.586	0.726	1.874	4.504	7.566
		FPR	0.002	0.013	0.055	0.074	0.087
		Power	–	0.465	0.765	0.810	0.855
	LAD + LASSO	ERR	0.642	1.635	5.264	10.662	17.023

example, it is still outperformed by LASSO + HIM. However, the story changes if the perturbation is due to the predictors as in Table 2. This is to be expected because, with contaminated predictors, LASSO is no longer a stable method for variable selection. If predictors are selected incorrectly, the subsequent modified

Cook's distance cannot be calculated appropriately. This makes the performance of LASSO + CD unsatisfactory.

(4) Fourth, as a robust regression method, we find that LAD + LASSO performs quite well. Its ERR values are smaller than those of the LASSO estimates in all the tables. However, in most cases, it is still outperformed by LASSO + HIM as seen from Tables 1 and 3.

(5) Lastly, we find that for most cases, the reported FPR values are well controlled. Furthermore, as κ increases, the corresponding empirical power increases toward 100%. These findings are consistent with the theoretical claims in Theorems 1 and 2.

To summarize, our simulation experiments confirm that the proposed HIM method is useful in controlling the effects of the influential observations in terms of parameter estimation and variable screening.

3.4. *A real data example*. We applied our proposed influence diagnosis approach to a microarray data of [31], and noted that the analysis results become substantially different when the detected influential observations are removed. For this dataset, F1 animals were intercrossed and then 120 twelve-week-old male offspring were selected for tissue harvesting from the eyes and for microarray analysis. The Affymetrix microarrays that were used to analyze the RNA from the eyes of those F2 animals contain over 31,042 different probe sets. Among them, one probe is for gene TRIM32, which was recently found to cause Bardet–Biedl syndrome [10], a genetically heterogeneous disease of multiple organ systems including the retina. One goal of interest of this data analysis is to find genes whose expressions are correlated with that of gene TRIM32. We first followed [27] to exclude probes that were not expressed in the eye or that lacked sufficient variation, which results in 18,975 probes as regressors. We then followed [22] to retain the top 1000 probes that are mostly correlated with the probe of TRIM32. The resulting analysis has $p = 1000$ predictors and a sample size $n = 120$. As a standard procedure [27], all the probes are standardized to have mean zero and standard deviation one.

We next applied our method with FDR rate $\alpha = 0.10$ to the data, and identified a total of 5 influential observations. Their corresponding p-values were 0, 0.0004, 0.0011, 0.0029 and 0.0033, respectively. We also show the logarithm of p-values versus the indices for all 120 observations in Figure 2. To assess the influence of the detected points, we again compared the LASSO estimate with and without those points. Since we used ten-fold cross-validation to select the tuning parameter and every run is random, we repeated this analysis 100 times and report the average results.

We summarize the difference of the estimates in the following aspects: the sparsity, the norm difference and the angle between the two estimates. First, by removing the identified influential observations, the resulting LASSO estimate is

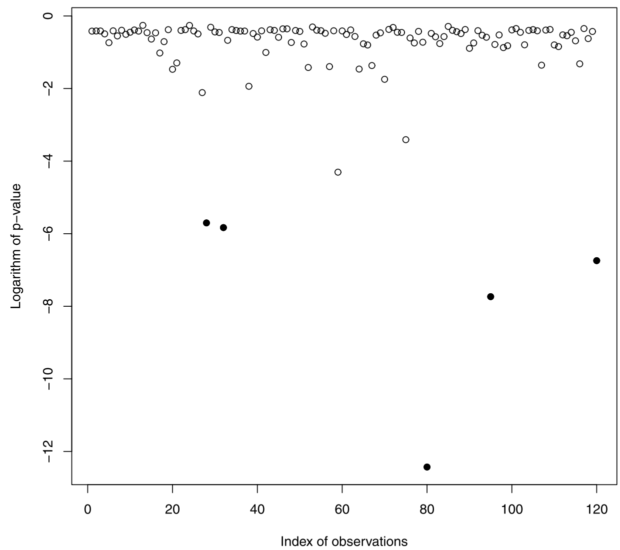

FIG. 2. *The logarithm of the* p*-value for each observation*: *the detected influential points are denoted by solid circles*.

considerably more sparse. The average model size with the full data is 63. By contrast, the average model size without the influential observations reduces to 27 on the average. The existence of the potential influential points clearly shows a significant effect on the model size. Besides that, the average number of the commonly selected predictors by fitting the full data and the reduced data, respectively, is only 8.67, which again shows clear discrepancy of the two estimates. Consequently, the influential points identified by our approach seem to have significant effect for subsequent analysis. Second, denote $d_0 = \|\hat{\boldsymbol{\beta}}_{\text{full}}\|_2$, $d_1 = \|\hat{\boldsymbol{\beta}}_{\text{redu}}\|_2$ and $d_2 = \|\hat{\boldsymbol{\beta}}_{\text{redu}} - \hat{\boldsymbol{\beta}}_{\text{full}}\|_2$, where $\hat{\boldsymbol{\beta}}_{\text{full}}$ is the LASSO estimate using all the observations and $\hat{\boldsymbol{\beta}}_{\text{redu}}$ is the estimate after removing the influential points identified by HIM. We observe that the average of $(d_0 - d_1)/d_0$ is 0.532 and that of d_2/d_0 is 0.972. Both show that the estimates without influential points are quite different in terms of the ℓ_2 norm. In addition, the angle between $\hat{\boldsymbol{\beta}}_{\text{full}}$ and $\hat{\boldsymbol{\beta}}_{\text{redu}}$, which is defined as $\hat{\boldsymbol{\beta}}_{\text{full}}^\top \hat{\boldsymbol{\beta}}_{\text{redu}}/d_0 d_1$, equals 0.262, averaged over 100 times. These numbers again indicate that the estimates change substantially after removing the influential observations. In summary, this analysis illustrates the importance of influence diagnosis, and the identified influential observations should be treated with extreme care.

4. Extension to generalized linear models. The main idea of the high-dimensional influence measure can be extended to a broad class of regression models. Here we briefly discuss one such extension to generalized linear models (GLM). Assume that the data $(\mathbf{X}_i, Y_i)$ follow an exponential distribution with the canonical probability density function, $f(y;\theta) = \exp\{y\theta - b(\theta) + c(y)\}$, and the conditional mean is of the form

$$E(Y_i \mid \mathbf{X}_i) = b'(\theta(\mathbf{X}_i)) = g^{-1}(\beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta}_1),$$

where g is a known link function. For the purpose of feature screening in ultra high-dimensional regressions, [23] introduced a marginal utility measure, the maximum marginal likelihood estimator, as

$$\hat{\boldsymbol{\beta}}_j = (\hat{\beta}_{j,0}, \hat{\beta}_j) = \arg\min E_n l(Y, \beta_{j0} + \beta_j X_j),$$

where $l(Y;\theta) = -Y\theta + b(\theta) + \log c(Y)]$ and $E_n f(X, Y) = n^{-1} \sum_{i=1}^n f(X_i, Y_i)$. That is, $\hat{\boldsymbol{\beta}}_j$ is the maximum likelihood estimator of fitting a GLM model of Y on the jth predictor X_j alone plus an intercept. As remarked by [23], this measure can be rapidly computed.

In the context of high-dimensional diagnosis, we define the high-dimensional influence measure for generalized linear models for the kth observation, $k = 1, \dots, n$, as

$$\mathcal{D}_k^{\mathrm{glm}} = \frac{1}{p} \sum_{j=1}^p \|\hat{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_j^{(k)}\|_2^2, \tag{4.1}$$

where $\hat{\boldsymbol{\beta}}_j^{(k)}$ denotes the maximum marginal likelihood estimator but with the kth observation removed. For GLM, the estimator $\hat{\boldsymbol{\beta}}_j$ and $\hat{\boldsymbol{\beta}}_j^{(k)}$ may not have a closed-form solution. Consequently, the exact distribution of the proposed statistic $\mathcal{D}_k^{\mathrm{glm}}$ is complicated and some approximation is necessary. The detailed derivation, however, is beyond the scope of this paper. In practice, one can always sort the values of $\{\mathcal{D}_k^{\mathrm{glm}}, k = 1, \dots, n\}$ and remove those observations associated with large values of $\mathcal{D}_k^{\mathrm{glm}}$.

We have conducted a small simulation study to examine the empirical performance of this measure for GLM. The simulation setup is similar to that of model 1 in Section 3.1, except that this time we adopt a binary response model, $P(Y_i = 1 \mid \mathbf{X}_i) = 1/[1 + \exp\{-(2 + \mathbf{X}_i^\top \boldsymbol{\beta})\}]$, where $\boldsymbol{\beta} = \boldsymbol{\beta}_{\mathrm{true}} = (5, 5, 0, \dots, 0)^\top$, and the outliers are generated from the model $\boldsymbol{\beta} = \boldsymbol{\beta}_{\mathrm{infl}} = (5, 5, 0, \dots, 0, -\kappa, \dots, -\kappa)^\top$ with $p/2$ many κ's. We set $n = 100$, with 10% influential observations, that is, $n_{\mathrm{infl}} = 10$, and we set $p = 50$ or 100. Since the asymptotic distribution of $\mathcal{D}_k^{\mathrm{glm}}$ is not available for the logistic regression, we flag the 10 observations with the largest p-values of $\mathcal{D}_k^{\mathrm{glm}}$ as influential. For a binary response, one is often interested in classification. As such we compare the misclassification error rate for the full data as E_{full} and for the reduced data as E_{redu} without the detected influential points. We also report the empirical power. The results out of 200 data replications are summarized in Table 4.

From Table 4, we note that the proposed method has some power for a logistic model, but it is lower than that in a linear model. This is probably due to the fact

TABLE 4
Simulation results for the logistic model

p	**Criterion**	κ **0**	**0.4**	**0.8**	**1.2**	**1.6**
50	Power	–	0.220	0.472	0.422	0.254
	E_{full}	0.037	0.062	0.088	0.083	0.064
	E_{redu}	0.018	0.022	0.031	0.049	0.033
100	Power	–	0.332	0.386	0.220	0.152
	E_{full}	0.047	0.069	0.065	0.045	0.029
	E_{redu}	0.020	0.042	0.028	0.018	0.019

that a binary response contains much less information, and thus detecting influential observations in a logistic model is much more challenging, especially in a high-dimensional setting. On the other hand, we also observe from Table 4 that removing those points with the largest values of $\mathcal{D}_k^{\text{glm}}$ improves the classification accuracy by a large margin. This again suggests that the usefulness of influence diagnosis. Meanwhile, we remark that further investigation into both theoretical and empirical properties of high-dimensional influence measure in GLM is warranted.

5. Conclusion. We perceive several future avenues to extend the proposed work in this article. First, we have employed the leave-one-out principle when quantifying influence of individual observations. We expect that our high-dimensional influence measure can also be generalized to the cases of leaving out pairs of observations, or triplets or more. Such a strategy can be useful when those observations conceal one another [18]. Second, we have focused on the classical linear model in our development, while extension to more sophisticated models, such as the generalized linear model, that is, briefly examined in Section 4, survival models, and semiparametric additive models, deserve further investigations. Finally, our proposal deals with the cross sectional data with i.i.d. observations. It is interesting to extend the proposed influence measure to complex correlated data such as longitudinal data where dependence among observations needs to be taken into consideration in influence diagnosis [42].

APPENDIX

We outline the main idea of the proof for the asymptotic distribution of $\mathcal{D}_k$ in Theorem 1. First, we decompose $\mathcal{D}_k$ as $\mathcal{D}_k = B_1 + B_2 + B_3 - 2B_4$ as given in Section 2.3. Then we compute the mean and variance of B_i, $i = 1, \dots, 4$ as presented in Proposition 1. This step builds on the assumption of normality of the predictors and benefits from the fact that the predictor dimension goes to infinity. Comparing the orders of the variance of B_i, we find that B_2 is the leading term. We then study the asymptotic distribution of B_2, which turns out to follow a $\chi^2(1)$ distribution. Recall in Section 2.3, we defined $K_{p,tl} = p^{-1}\sum_{j=1}^{p} X_{tj}X_{lj}$, for $t, l = 1, \dots, n$, $l_p = \sum_{j=1}^{p}\lambda_j^2 = O(p^r)$ and $c_p = \max_{1\le j\le p}\lambda_j$, where λ_j's are the eigenvalues of the covariance matrix $\mathbf{\Sigma}$. Furthermore, we define $a_p = \sum_{j=1}^{p}\lambda_j^4$, $C_1 = E(Y_tY_lK_{p,tl})^2$ and $C_2 = E[Y_t^2(\sum_{j=1}^{p}\rho_jX_{tj}/p)^2]$ for any $t \ne l$.

PROOF OF PROPOSITION 1. We break the proof into three parts: first, we obtain an expansion of $\mathcal{D}_k$; second, we derive $E(\mathcal{D}_k)$; and finally, we derive the asymptotic behavior of the components in the expansion of $\mathcal{D}_k$.

Step 1. First, we have the following expansion for $\mathcal{D}_k$, $k = 1, \dots, n$:

$$
\begin{aligned}
\mathcal{D}_k &= \frac{1}{p}\sum_{j=1}^{p}\left(\frac{1}{n-1}\sum_{t=1,t\ne k}^{n} Y_tX_{tj} - \frac{1}{n}\sum_{t=1}^{n} Y_tX_{tj}\right)^2 \\
&= \frac{1}{p}\sum_{j=1}^{p}\left\{\frac{1}{n(n-1)}\sum_{t=1,t\ne k}^{n} Y_tX_{tj} - \frac{1}{n}Y_kX_{kj}\right\}^2 \\
&= \frac{1}{p}\sum_{j=1}^{p}\left\{\frac{1}{n(n-1)}\sum_{t=1,t\ne k}^{n} Y_tX_{tj}\right\}^2 + \frac{1}{pn^2}Y_k^2\sum_{j=1}^{p}X_{kj}^2 \\
&\quad - \frac{2}{pn^2(n-1)}\sum_{t=1,t\ne k}^{n} Y_tY_k\left\{\sum_{j=1}^{p}X_{tj}X_{kj}\right\} \\
&= \frac{1}{p\{n(n-1)\}^2}\sum_{t=1,t\ne k}^{n} Y_t^2\left\{\sum_{j=1}^{p}X_{tj}^2\right\} + \frac{1}{pn^2}Y_k^2\left\{\sum_{j=1}^{p}X_{kj}^2\right\} \\
&\quad + \frac{1}{p\{n(n-1)\}^2}\sum_{t\ne s,t,s\ne k} Y_tY_s\left\{\sum_{j=1}^{p}X_{tj}X_{sj}\right\} \\
&\quad - \frac{2}{pn^2(n-1)}\sum_{t=1,t\ne k}^{n} Y_kY_t\left\{\sum_{j=1}^{p}X_{tj}X_{kj}\right\} \\
&= \frac{1}{\{n(n-1)\}^2}\sum_{t=1}^{n} Y_t^2K_{p,tt} + \left[\frac{1}{n^2} - \frac{1}{\{n(n-1)\}^2}\right]Y_k^2K_{p,kk} \\
&\quad + \frac{1}{\{n(n-1)\}^2}\sum_{t\ne s} Y_tY_sK_{p,ts} \\
&\quad - \left[\frac{2}{\{n(n-1)\}^2} + \frac{2}{n^2(n-1)}\right]\sum_{t=1,t\ne k}^{n} Y_kY_tK_{p,tk}
\end{aligned}
$$

$$= \frac{1}{\{n(n-1)\}^2} \sum_{t=1}^{n} Y_t^2 K_{p,tt} + \frac{(n-2)}{n(n-1)^2} Y_k^2 K_{p,kk}$$

$$+ \frac{1}{\{n(n-1)\}^2} \sum_{t \neq s} Y_t Y_s K_{p,ts} - \frac{2}{n(n-1)^2} \sum_{t=1, t \neq k}^{n} Y_k Y_t K_{p,tk}$$

$$:= B_1 + B_2 + B_3 - 2B_4.$$

Step 2. Next, we derive the expectation of $\mathcal{D}_k$. It is easy to see that

$$E(B_1) = \frac{1}{pn(n-1)^2} \sum_{j=1}^{p} E(Y_k^2 X_{kj}^2),$$

$$E(B_2) = \frac{n-2}{pn(n-1)^2} \sum_{j=1}^{p} E(Y_k^2 X_{kj}^2),$$

$$E(B_3) = \frac{1}{pn(n-1)} \sum_{j=1}^{p} \rho_j^2, \qquad E(B_4) = \frac{1}{pn(n-1)} \sum_{j=1}^{p} \rho_j^2.$$

Therefore, we have

$$E(\mathcal{D}_k) = E(B_1 + B_2 + B_3 - 2B_4) = \frac{1}{pn(n-1)} \sum_{j=1}^{p} \text{var}(Y_k X_{kj}).$$

By Lemmas 1 and 3, we have

$$E\{Y_k^2(K_{p,kk} - E(K_{p,kk}))\} \leq E^{1/2}(Y_k^4) E^{1/2}[\{K_{p,kk} - E(K_{p,kk})\}^2]$$
$$= O(p^{-1} l_p^{1/2})$$

and

$$p^{-1} \sum_{j=1}^{p} E^2(Y_k X_{kj}) = p^{-1} \sum_{j=1}^{p} \rho_j^2 = O(p^{-1} c_p).$$

In addition, noting that $c_p^2 \leq l_p$, we have

$$p^{-1} \sum_{j=1}^{p} \text{var}(Y_k X_{kj})$$

$$= p^{-1} \sum_{j=1}^{p} \{E(Y_k^2 X_{kj}^2) - E^2(Y_k X_{kj})\}$$

$$= E(Y_k^2) E(K_{p,kk}) + E\{Y_k^2(K_{p,kk} - E(K_{p,kk}))\} - p^{-1} \sum_{j=1}^{p} \rho_j^2$$

$$= E(Y_k^2) E(K_{p,kk}) + O(p^{-1} l_p^{1/2}).$$

Consequently, we have

$$E(\mathcal{D}_k) = \frac{1}{\{pn(n-1)\}} E(Y_k^2) \sum_{j=1}^{p} E(X_{kj}^2) + O(n^{-2} p^{-1} l_p^{1/2}).$$

Step 3. Next, we derive the asymptotic behavior of $B_i, i = 1, \dots, 4$.

Step 3.1. We start with the variance of B_1. Note that

$$\operatorname{var}(B_1) = \frac{n}{n^4(n-1)^4} \operatorname{var}(Y_t^2 K_{p,tt})$$

and that $E(Y_t^4 K_{p,tt}^2) \le E^{1/2}(Y_t^8) E^{1/2}(K_{p,tt}^4)$. Furthermore,

$$E(K_{p,tt})^4 = p^{-4} E(\mathbf{Z}_t^\top \mathbf{\Sigma} \mathbf{Z}_t)^4 = p^{-4} E\left[\sum_{j=1}^{p} \lambda_j (\mathbf{Z}_t^\top \mathbf{u}_j)^2\right]^4$$

$$\le p^{-4} E[(\mathbf{Z}_t^\top \mathbf{u}_j)^8] \left(\sum_{j=1}^{p} \lambda_j\right)^4 \le E(\mathbf{Z}_t^\top \mathbf{u}_j)^8.$$

The last equation holds because $\operatorname{tr}(\Sigma) = \sum_{j=1}^{p} \lambda_j = p$. As a result, we have

$$\operatorname{var}(B_1) = O(n^{-7}).$$

Step 3.2. Next, we consider the variance of B_4. By Lemma 3, we have

$$E(B_4) = \frac{1}{pn(n-1)} \sum_{j=1}^{p} \rho_j^2 = O(c_p \cdot p^{-1} n^{-2}).$$

In addition, it is easy to see

$$\begin{aligned} E(B_4^2) &= \frac{1}{n^2(n-1)^4} \left\{ \sum_{t=1, t\ne k}^{n} E(Y_k^2 Y_t^2 K_{p,tk}^2) + \sum_{t \ne s}^{t,s\ne k} E(Y_k^2 Y_t Y_s K_{p,tk} K_{p,sk}) \right\} \\ &= \frac{1}{n^2(n-1)^4} \Bigg[(n-1) E(Y_k^2 Y_t^2 K_{p,tk}^2) \\ &\qquad + (n-1)(n-2) E\left\{ Y_k^2 \left(\sum_{j=1}^{p} \rho_j X_{kj}/p \right)^2 \right\} \Bigg] \\ &= \frac{1}{n^2(n-1)^3} E(Y_k^2 Y_t^2 K_{p,tk}^2) + \frac{(n-2)}{n^2(n-1)^3} E\left\{ Y_k^2 \left(\sum_{j=1}^{p} \rho_j X_{kj}/p \right)^2 \right\} \\ &= \frac{1}{n^2(n-1)^3} C_1 + \frac{(n-2)}{n^2(n-1)^3} C_2, \end{aligned}$$

where C_1, C_2 are defined as in (1.1) and (1.2) in the supplementary material [41], respectively. From the proof of Lemma 2, we know $C_2 = O(c_p^2 p^{-2})$ and $C_1 = C_{11} + C_{12}$, with $C_{11} = O(p^{-2} a_p^{1/2})$ and $C_{12} = l_p p^{-2} E^2(Y_t^2)$. Therefore, we have

$$\begin{aligned} E(B_4^2) &= \frac{l_p}{p^2 n^2 (n-1)^3} E^2(Y_t^2) + O(n^{-5} p^{-2} a_p^{1/2}) + O(c_p^2 p^{-2} n^{-4}) \\ &= O(l_p p^{-2} n^{-5}) + O(c_p^2 p^{-2} n^{-4}). \end{aligned}$$

Consequently, we have

$$\operatorname{var}(B_4) = E(B_4^2) - E^2(B_4) = O(l_p p^{-2} n^{-5}) + O(c_p^2 p^{-2} n^{-4}).$$

Step 3.3. Next, we aim at $\operatorname{var}(B_3)$. First, we have

$$\begin{aligned} B_3 &= \frac{1}{p\{n(n-1)\}^2} \sum_{t \neq s} Y_t Y_s \left(\sum_{j=1}^p X_{sj} X_{tj} \right) \\ &= \frac{1}{pn(n-1)} \sum_{t \neq s} \phi(Y_t, Y_s, \mathbf{X}_s, \mathbf{X}_t) / \{(n-1)n\} := \frac{1}{pn(n-1)} \bar{B}_3, \end{aligned}$$

where $\phi(Y_t, Y_s, \mathbf{X}_s, \mathbf{X}_t) = \sum_{t \neq s} Y_t Y_s (\sum_{j=1}^p X_{sj} X_{tj})$. Let

$$\begin{aligned} \phi_1(Y_t, \mathbf{X}_t) &= E\{\phi(Y_t, Y_s, \mathbf{X}_s, \mathbf{X}_t) - E(\phi(Y_t, Y_s, \mathbf{X}_s, \mathbf{X}_t) \mid Y_t, \mathbf{X}_t\} \\ &= Y_t \sum_{j=1}^p X_{tj} \rho_j - \sum_{j=1}^p \rho_j^2. \end{aligned}$$

Noting that $\bar{B}_3$ is an U-statistic, and by the properties of the U-statistic, we have

$$\begin{aligned} \operatorname{var}(B_3) &= \frac{1}{\{pn(n-1)\}^2} \operatorname{var}(\bar{B}_3) \\ &= \frac{1}{\{pn(n-1)\}^2} \left[\frac{4}{n} \operatorname{var}\{\phi_1(Y_t, \mathbf{X}_t)\} + o(n^{-2}) \right] \qquad \text{(A.1)} \\ &= \frac{4}{p^2 n^3 (n-1)^2} \operatorname{var}\left\{ Y_t \sum_{j=1}^p X_{tj} \rho_j \right\} + o(p^{-2} n^{-4} (n-1)^{-2}). \end{aligned}$$

Furthermore, we have

$$\begin{aligned} & p^{-2} \operatorname{var}\left(Y_t \sum_{j=1}^p X_{tj} \rho_j \right) \\ &= E\left\{ Y_t^2 \left(\sum_{j=1}^p X_{tj} \rho_j / p \right)^2 \right\} - \left(\sum_{j=1}^p \rho_j^2 / p \right)^2 \qquad \text{(A.2)} \\ &\leq E^{1/2}(Y_t^4) E^{1/2}\left\{ \left(\sum_{j=1}^p X_{tj} \rho_j / p \right)^4 \right\} - \left(\sum_{j=1}^p \rho_j^2 / p \right)^2. \end{aligned}$$

In addition, we show in the proof of Lemma 2 that

$$E\left\{\left(\sum_{j=1}^{p} X_{tj}\rho_j/p\right)^4\right\} = O\left(c_p^4 p^{-4}\right) \tag{A.3}$$

and in Lemma 3 that

$$\sum_{j=1}^{p} \rho_j^2/p = O\left(c_p \cdot p^{-1}\right). \tag{A.4}$$

Combining (A.1)–(A.4), we have

$$\operatorname{var}(B_3) = O\left(c_p^2 n^{-5} p^{-2}\right) + O\left(p^{-2} n^{-6}\right).$$

Step 3.4. Finally we turn to B_2, which can be written as

$$\begin{aligned} B_2 &= \frac{(n-2)}{n(n-1)^2} Y_k^2 K_{p,kk} \\ &= \frac{(n-2)}{n(n-1)^2}\left[Y_k^2\{K_{p,kk} - E(K_{p,kk})\} + Y_k^2 E(K_{p,kk})\right] \\ &:= B_{21} + B_{22}. \end{aligned}$$

By Lemma 1, we have

$$\begin{aligned} E\left(Y_k^4\{K_{p,kk} - E(K_{p,kk})\}^2\right) &\le E^{1/2}\left(Y_k^8\right) E^{1/2}\left(\{K_{p,kk} - E(K_{p,kk})\}^4\right) \\ &= O\left(p^{-2} l_p\right), \\ E\left(Y_k^2\{K_{p,kk} - E(K_{p,kk})\}\right) &\le E^{1/2}\left(Y_k^4\right) E^{1/2}\left(\{K_{p,kk} - E(K_{p,kk})\}^2\right) \\ &= O\left(l_p^{1/2} p^{-1}\right). \end{aligned}$$

Therefore, we have

$$\begin{aligned} &\operatorname{var}(B_{21}) = \left\{\frac{(n-2)}{n(n-1)^2}\right\}^2 \operatorname{var}\left(Y_k^2[K_{p,kk} - E(K_{p,kk})]\right) = O\left(n^{-4} p^{-2} l_p\right), \\ &\operatorname{var}(B_{22}) = O\left(n^{-4}\right). \end{aligned}$$

This completes the proof. $\square$

PROOF OF THEOREM 1. Consider the behavior of $K_{p,kk}$, $k = 1, \dots, n$, for a sufficient large p

$$K_{p,kk} = \sum_{j=1}^{p} X_{kj}^2/p = \mathbf{X}_k^\top \mathbf{X}_k/p = \mathbf{Z}_k^\top \boldsymbol{\Sigma} \mathbf{Z}_k = \sum_{j=1}^{p} \lambda_j \left(\mathbf{Z}_k^\top \mathbf{u}_j\right)^2/p.$$

Its variance is $\mathrm{var}(K_{p,kk}) = 2\sum_{j=1}^{p}\lambda_j^2/p^2 = 2p^{-2}l_p$. Under the assumption $l_p = O(p^r)$ with $0 \le r < 2$, we have $K_{p,kk} = E(K_{p,kk}) + O_p(p^{r/2-1})$, and consequently,

$$Y_k^2 K_{p,kk} = Y_k^2\big[E(K_{p,kk}) + O_p\big(p^{r/2-1}\big)\big].$$

In addition, noting that $E[Y_k^2(K_{p,kk} - E(K_{p,kk}))] \le E^{1/2}(Y_k^4)(\mathrm{var}(K_{p,kk}))^{1/2} = O(p^{r/2-1})$, we have

$$\begin{aligned} E\big(Y_k^2 K_{p,kk}\big) &= E\big(Y_k^2\big)E(K_{p,kk}) + E\big[Y_k^2\big(K_{p,kk} - E(K_{p,kk})\big)\big] \\ &= E\big(Y_k^2\big)E(K_{p,kk}) + O\big(p^{r/2-1}\big). \end{aligned}$$

Therefore, we have

$$Y_k^2 K_{p,kk} - E\big(Y_k^2 K_{p,kk}\big) = \big[Y_k^2 - E\big(Y_k^2\big)\big]E(K_{p,kk}) + O_p\big(p^{r/2-1}\big).$$

As a result, it holds that

$$B_2 - E(B_2) = \frac{n-2}{n(n-1)^2}\big\{\big[Y_k^2 - E\big(Y_k^2\big)\big]E(K_{p,kk}) + O_p\big(p^{r/2-1}\big)\big\}. \tag{A.5}$$

Note that $c_p^2 \le l_p = O(p^r)$ under (C.2). Combined with Proposition 1, we have

$$\begin{aligned} B_1 - E(B_1) &= O_p\big(n^{-7/2}\big), \\ B_3 - E(B_3) &= O_p\big(n^{-5/2}p^{r/2-1}\big), \\ B_4 - E(B_4) &= O_p\big(p^{r/2-1}n^{-2}\big). \end{aligned}$$

Consequently, we have

$$\begin{aligned} &\frac{n(n-1)^2}{(n-2)}\Big\{\sum_{i=1,3}\big[B_i - E(B_i)\big] - 2\big(B_4 - E(B_4)\big)\Big\} \\ &\quad = O_p\big(n^{-3/2}\big) + O_p\big(p^{r/2-1}\big). \end{aligned} \tag{A.6}$$

Furthermore, by the results on $E(\mathcal{D}_k)$ in step 2 of the proof of Proposition 1, we have

$$E(\mathcal{D}_k) = \frac{1}{n(n-1)}E\big(Y_k^2\big)E(K_{p,kk}) + O\big(n^{-2}p^{-1}l_p^{1/2}\big).$$

Consequently, by $l_p = O(p^r)$, we have

$$\frac{n(n-1)^2}{(n-2)}E(\mathcal{D}_k) = \frac{n-1}{n-2}E\big(Y_k^2\big)E(K_{p,kk}) + O\big(p^{r/2-1}\big). \tag{A.7}$$

Combining (A.5)–(A.7), we have

$$\begin{aligned}
&\frac{n(n-1)^2}{(n-2)}\mathcal{D}_k \\
&\quad = \frac{n(n-1)^2}{(n-2)}E(\mathcal{D}_k) + \frac{n(n-1)^2}{(n-2)}[\mathcal{D}_k - E(\mathcal{D}_k)] \\
&\quad = \frac{n(n-1)^2}{(n-2)}E(\mathcal{D}_k) + \frac{n(n-1)^2}{(n-2)}\Bigl(\sum_{i=1,2,3}[B_i - E(B_i)] - 2\bigl(B_4 - E(B_4)\bigr)\Bigr) \\
&\quad = \frac{n-1}{n-2}E(Y_k^2)E(K_{p,kk}) + \{Y_k^2 - E(Y_k^2)\}E(K_{p,kk}) \\
&\qquad + O_p(p^{r/2-1}) + O_p(n^{-3/2}) \\
&\quad = Y_k^2 E(K_{p,kk}) + \frac{1}{n-2}E(Y_k^2)E(K_{p,kk}) + O_p(p^{r/2-1}) + O_p(n^{-3/2}) \\
&\quad = Y_k^2 + o_p(1),
\end{aligned}$$

where the last equation is from the fact that $E(X_{kj}^2) = 1, j = 1, \ldots, p$ and $E(Y_k^2) = 1$. Since $Y \sim N(0, 1)$, we have $\frac{n(n-1)^2}{(n-2)}\mathcal{D}_k \sim \chi^2(1)$; that is, $n^2\mathcal{D}_k \sim \chi^2(1)$. $\square$

Acknowledgements. We thank Professor Peter Bühlmann, Professor Peter Hall, an Associate Editor and three referees for their constructive comments.

SUPPLEMENTARY MATERIAL

Further proofs (DOI: 10.1214/13-AOS1165SUPP; .pdf). The supplementary file contains the proofs of four additional lemmas, Proposition 2 and Theorem 2.

REFERENCES

[1] ANDERSON, E. B. (1992). Diagnostics in categorical data analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **54** 781–791.

[2] BANERJEE, M. (1998). Cook's distance in linear longitudinal models. *Comm. Statist. Theory Methods* **27** 2973–2983. MR1659375

[3] BANERJEE, M. and FREES, E. W. (1997). Influence diagnostics for linear longitudinal models. *J. Amer. Statist. Assoc.* **92** 999–1005. MR1482130

[4] BELLONI, A. and CHERNOZHUKOV, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* **39** 82–130. MR2797841

[5] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57** 289–300. MR1325392

[6] BENJAMINI, Y. and HOCHBERG, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.* **25** 60–83. MR1325392

[7] BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

[8] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

[9] CHATTERJEE, S. and HADI, A. S. (1988). *Sensitivity Analysis in Linear Regression*. Wiley, New York. MR0939610

[10] CHIANG, A. P., BECK, J. S., YEN, H. J., TAYEH, M. K., SCHEETZ, T. E., SWIDERSKI, R., NISHIMURA, D., 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 a novel gene for Bardet–Biedl syndrome (BBS11). *Proc. Natl. Acad. Sci. USA* **103** 6287–6292.

[11] CHRISTENSEN, R., PEARSON, L. M. and JOHNSON, W. (1992). Case-deletion diagnostics for mixed models. *Technometrics* **34** 38–45. MR1157792

[12] COOK, R. D. (1977). Detection of influential observation in linear regression. *Technometrics* **19** 15–18. MR0436478

[13] COOK, R. D. (1979). Influential observations in linear regression. *J. Amer. Statist. Assoc.* **74** 169–174. MR0529533

[14] COOK, R. D. and WEISBERG, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall, London. MR0675263

[15] CRITCHLEY, F., ATKINSON, R. A., LU, G. and BIAZI, E. (2001). Influence analysis based on the case sensitivity function. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63** 307–323. MR1841417

[16] DAVISON, A. C. and TSAI, C. L. (1992). Regression model diagnostics. *Int. Stat. Rev.* **55** 337–353.

[17] DONOHO, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. Technical report, Stanford Univ.

[18] DRAPER, N. R. and SMITH, H. (1998). *Applied Regression Analysis*, 3rd ed. Wiley, New York. MR1614335

[19] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

[20] EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571

[21] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[22] FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 849–911. MR2530322

[23] FAN, J. and SONG, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38** 3567–3604. MR2766861

[24] FU, W. J. (1998). Penalized regressions: The bridge versus the Lasso. *J. Comput. Graph. Statist.* **7** 397–416. MR1646710

[25] FUNG, W.-K., ZHU, Z.-Y., WEI, B.-C. and HE, X. (2002). Influence diagnostics and outlier tests for semiparametric mixed models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 565–579. MR1924307

[26] HUANG, J., HOROWITZ, J. L. and MA, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36** 587–613. MR2396808

[27] HUANG, J., MA, S. and ZHANG, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica* **18** 1603–1618. MR2469326

[28] JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961

[29] PAN, J.-X. and FANG, K.-T. (2002). *Growth Curve Models and Statistical Diagnostics*. Springer, New York. MR1937691

[30] PREISSER, J. S. and QAQISH, B. F. (1996). Deletion diagnostics for generalised estimating equations. *Biometrika* **83** 551–562.

[31] SCHEETZ, T., KIM, K., SWIDERSKI, R., PHILP, A., BRAUN, T., KNUDTSON, K., DORRANCE, A., DIBONA, G., HUANG, J., CASAVANT, T. et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proc. Natl. Acad. Sci. USA* **103** 14429–14434.

[32] STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 479–498. MR1924302

[33] THOMAS, W. and COOK, R. D. (1989). Assessing influence on regression coefficients in generalized linear models. *Biometrika* **76** 741–749. MR1041419

[34] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288. MR1379242

[35] WANG, H. (2009). Forward regression for ultra-high dimensional variable screening. *J. Amer. Statist. Assoc.* **104** 1512–1524. MR2750576

[36] WANG, H. and LENG, C. (2007). Unified Lasso estimation via least squares approximation. *J. Amer. Statist. Assoc.* **101** 1418–1429.

[37] WANG, H., LI, G. and JIANG, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD–Lasso. *J. Bus. Econom. Statist.* **25** 347–355. MR2380753

[38] WILLIAMS, D. A. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **36** 181–191. MR0897457

[39] XIANG, L., TSE, S.-K. and LEE, A. H. (2002). Influence diagnostics for generalized linear mixed models: Applications to clustered data. *Comput. Statist. Data Anal.* **40** 759–774. MR1933008

[40] ZHANG, H. H. and LU, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94** 691–703. MR2410017

[41] ZHAO, J., LENG, C., LI, L. and WANG, H. (2013). Supplement to "High-dimensional influence measure." DOI:10.1214/13-AOS1165SUPP.

[42] ZHU, H., IBRAHIM, J. G. and CHO, H. (2012). Perturbation and scaled Cook's distance. *Ann. Statist.* **40** 785–811. MR2933666

[43] ZHU, H., IBRAHIM, J. G., LEE, S. and ZHANG, H. (2007). Perturbation selection and influence measures in local influence analysis. *Ann. Statist.* **35** 2565–2588. MR2382658

[44] ZHU, H., LEE, S.-Y., WEI, B.-C. and ZHOU, J. (2001). Case-deletion measures for models with incomplete data. *Biometrika* **88** 727–737. MR1859405

[45] ZOU, H. (2006). The adaptive LASSO and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

J. ZHAO
SCHOOL OF MATHEMATICS
AND SYSTEM SCIENCE
BEIHANG UNIVERSITY
BEIJING
CHINA
E-MAIL: zhaojunlong928@126.com

C. LENG
DEPARTMENT OF STATISTICS
UNIVERSITY OF WARWICK
COVENTRY
UNITED KINGDOM
AND
DEPARTMENT OF STATISTICS
AND APPLIED PROBABILITY
NATIONAL UNIVERSITY OF SINGAPORE
E-MAIL: C.Leng@warwick.ac.uk

L. LI
DEPARTMENT OF STATISTICS
NORTH CAROLINA STATE UNIVERSITY
RALEIGH, NORTH CAROLINA
USA E-MAIL: li@stat.ncsu.edu

H. WANG
GUANGHUA SCHOOL OF MANAGEMENT
PEKING UNIVERSITY
BEIJING
CHINA
E-MAIL: hansheng@gsm.pku.edu.cn