Submitted to the Annals of Statistics
arXiv: math.PR/0000000

DEGREE OF PERTURBATION AND SCALED COOK'S DISTANCE

By Hongtu Zhu , Joseph G Ibrahim and Hyunsoon Cho

Department of Biostatistics, University of North Carolina at Chapel Hill

Cook's (Cook, 1977) distance is one of the most important diagnostic tools for detecting influential individual or subsets of observations in linear regression for cross-sectional data. However, for many complex data structures (e.g., longitudinal data), no rigorous approach has been developed to address a fundamental size issue: deleting subsets with different numbers of observations introduces different degrees of perturbation to the current model fitted to the data and the size of Cook's distance is associated with the degree of the perturbation. The aim of this paper is to address this "size issue" in general parametric models with complex data structures. We propose a new quantity for measuring the degree of the perturbation introduced by deleting a subset. We use stochastic ordering to quantify the stochastic relationship between the degree of the perturbation and the size of Cook's distance. We develop several scaled Cook's distances to resolve the size issue of Cook's distance. Theoretical and numerical examples are examined to highlight the broad spectrum of applications of these scaled Cook's distances in a formal influence analysis.

1. Introduction. Influence analysis assesses whether a modification of a statistical analysis, called a perturbation, seriously affects specific key inferences, such as parameter estimates. Such perturbation schemes include the deletion of individual or a subset of observations, case weight perturbation, and covariate perturbation among many others [9, 10, 30]. If a small perturbation has a small effect on the analysis, our analysis is relatively stable, while if a large perturbation has a small effect on the analysis, we learn that our analysis is robust [12, 18]. If a small perturbation seriously influences key results of the analysis, we want to know the cause [10, 12]. For instance, in influence analysis, a set of observations is flagged as 'influential' if its removal from the dataset produces a significant difference in the parameter estimates or equivalently a large value of Cook's distance for the current statistical model [9, 6]. Sometimes, these influential observations are also outliers, which are defined as a discordant individual or a set

AMS 2000 subject classifications: Primary 62J20
Keywords and phrases: Cook's distance; Perturbation; Relative influential; Conditionally scaled Cook's distance; Scaled Cook's distance; Size issue.

of discordant observations that is not a realization from the current model [6].

Since the seminal work of Cook [9] on Cook's distance in linear regression for cross-sectional data, considerable research has been devoted to developing Cook's distance for detecting influential observations (or clusters) in more complex data structures under various statistical models [9, 11, 7, 2, 13, 24, 16, 29, 15]. For example, for longitudinal data, Preisser and Qaqish [20] developed Cook's distance for generalized estimating equations, while Christensen, Pearson and Johnson [8], Banerjee and Frees [5], and Banerjee [4] considered case deletion and subject deletion diagnostics for linear mixed models. Furthermore, in the presence of missing data, Zhu et al. [29] developed deletion diagnostics for a large class of statistical models with missing data. Cook's distance has been widely used in statistical practice and can be calculated in popular statistical software, such as SAS and R.

A critical size issue with Cook's distance has been largely neglected in the existing literature on developing Cook's distance for general statistical models with more complex data structures. The size issue is that the magnitude of Cook's distance is positively associated with the amount of perturbation to the current model introduced by deleting a subset of observations. Specifically, a large value of Cook's distance can be caused by deleting a subset with a larger number of observations and/or other causes such as the presence of outlier(s) in the deleted subset. To delineate the cause of a large Cook's distance for a specific subset, it is more useful to compute Cook's distance relative to the degree of the perturbation introduced by deleting the subset [12, 30].

The aim of this paper is to address the size issue of Cook's distance for complex data structures in general parametric models. The main contributions of this paper are summarized as follows.

(a.1) We propose a quantity to measure the degree of perturbation introduced by deleting a subset in general parametric models. This quantity satisfies several attractive properties including uniqueness, non-negativity, monotonicity, and additivity.

(a.2) We use stochastic ordering to quantify the relationship between the degree of the perturbation and the size of Cook's distance. Particularly, in linear regression for cross-sectional data, we first show the stochastic relationship between the Cook's distances for any two subsets with possibly different numbers of observations.

(a.3) We develop several scaled Cook's distances and their first-order approximations to address the size issue while fixing some covariates of interest.

(a.4) We illustrate our development with various parametric models.

The rest of the paper is organized as follows. In Section 2, we quantify the degree of the perturbation for set deletion and delineate the stochastic relationship between Cook's distance and the degree of perturbation. We develop several scaled Cook's distances and derive their first-order approximations to address the size issue. In Section 3, we analyze simulated data and a real dataset using the proposed scaled Cook's distances. We give some final remarks in Section 4.

2. Scaled Cook's Distance.

2.1. *Cook's distance.* Consider the probability function of a random vector $\mathbf{Y}^T = (Y_1^T, \ldots, Y_n^T)$, denoted by $p(\mathbf{Y}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^T$ is a $q \times 1$ vector in an open subset Θ of R^q and $Y_i = (y_{i,1}, \ldots, y_{i,m_i})$, in which the dimension of Y_i, denoted by m_i, may vary across all i. For instance, in longitudinal studies, if our interest focuses on detecting influential clusters, then Y_i includes all responses and covariates of interest in the ith cluster. Thus, the number of observations in the ith cluster may vary significantly across clusters.

Cook's distance and many other deletion diagnostics measure the distance between the maximum likelihood estimators of $\boldsymbol{\theta}$ with and without Y_i [11, 9]. A subscript '[I]' denotes the relevant quantity with all observations in I deleted. Let $\mathbf{Y}_{[I]}$ be a subsample of $\mathbf{Y}$ with $\mathbf{Y}_I = \{Y_{(i,j)} : (i,j) \in I\}$ deleted and $p(\mathbf{Y}_{[I]}|\boldsymbol{\theta})$ be its probability function. We define the maximum likelihood estimators of $\boldsymbol{\theta}$ for the full sample $\mathbf{Y}$ and a subsample $\mathbf{Y}_{[I]}$ as

$$\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} \log p(\mathbf{Y}|\boldsymbol{\theta}) \quad \text{and} \quad \hat{\boldsymbol{\theta}}_{[I]} = \operatorname{argmax}_{\boldsymbol{\theta}} \log p(\mathbf{Y}_{[I]}|\boldsymbol{\theta}), \tag{2.1}$$

respectively. Cook's distance for I, denoted by $\mathrm{CD}(I)$, can be defined as follows:

$$\mathrm{CD}(I) = (\hat{\boldsymbol{\theta}}_{[I]} - \hat{\boldsymbol{\theta}})^T G_{n\theta} (\hat{\boldsymbol{\theta}}_{[I]} - \hat{\boldsymbol{\theta}}), \tag{2.2}$$

where $G_{n\theta}$ is chosen to be a positive definite matrix. Throughout the paper, $G_{n\theta}$ is set as $-\partial_\theta^2 \log p(\mathbf{Y}|\hat{\boldsymbol{\theta}})$ or its expectation, where ∂_θ^2 represents the second-order derivative with respect to $\boldsymbol{\theta}$. More generally, suppose that one is interested in a subset of $\boldsymbol{\theta}$ or q_1 linearly independent combinations of $\boldsymbol{\theta}$, say $\mathbf{L}^T\boldsymbol{\theta}$, where $\mathbf{L}$ is a $q \times q_1$ matrix with $\mathrm{rank}(\mathbf{L}) = q_1$ [5, 11]. The partial influence of the subset I on $\mathbf{L}^T\hat{\boldsymbol{\theta}}$, denoted by $\mathrm{CD}(I|\mathbf{L})$, can be defined as

$$\mathrm{CD}(I|\mathbf{L}) = (\hat{\boldsymbol{\theta}}_{[I]} - \hat{\boldsymbol{\theta}})^T \mathbf{L} \{\mathbf{L}^T G_{n\theta}^{-1} \mathbf{L}\}^{-1} \mathbf{L}^T (\hat{\boldsymbol{\theta}}_{[I]} - \hat{\boldsymbol{\theta}}). \tag{2.3}$$

For notational simplicity, even though we may focus on a subset of θ, we do not distinguish $\mathrm{CD}(I|\mathbf{L})$ from $\mathrm{CD}(I)$ throughout the paper.

Based on (2.2), we know that Cook's distance $\mathrm{CD}(I)$ is explicitly determined by three components including the current model fitted to the data, denoted by $\mathcal{M}$, the dataset $\mathbf{Y}$, and the subset I itself. Cook's distance is also implicitly determined by $\mathcal{M}$, $\mathbf{Y}$ and the degree of the perturbation to $\mathcal{M}$ introduced by deleting the subset I, denoted by $\mathcal{P}(I|\mathcal{M})$. Thus, we may represent $\mathrm{CD}(I)$ as follows:

$$\mathrm{CD}(I) = F_1(I, \mathcal{M}, \mathbf{Y}) = F_2(\mathcal{P}(I|\mathcal{M}), \mathcal{M}, \mathbf{Y}), \tag{2.4}$$

where $F_1(\cdot,\cdot,\cdot)$ and $F_2(\cdot,\cdot,\cdot)$ represent nonlinear functions.

We may use the values of $\mathrm{CD}(I)$ to assess the influential level of the subset I. We may regard a subset I as influential if either the value of $\mathrm{CD}(I)$ is relatively large compared with other Cook's distances or the magnitude of $\mathrm{CD}(I)$ is greater than the critical points of the χ^2 distribution [11]. However, for complex data structures, we will show that it is useful to compare Cook's distance relative to its associated degree of perturbation.

2.2. *Degree of perturbation.* Consider the subset I and the current model $\mathcal{M}$. We are interested in answering the following questions below.

(c.1) How do we measure the degree of the perturbation to $\mathcal{M}$ introduced by deleting the subset I?

Abstractly, $\mathcal{P}(I|\mathcal{M})$ should be defined as a mapping from a subset I and $\mathcal{M}$ to a nonnegative number. However, according to the best of our knowledge, no quantities have ever been developed to define a workable $\mathcal{P}(I|\mathcal{M})$ for an arbitrary subset I in general parametric models due to many conceptual difficulties [12]. Although [12] placed the Euclidean geometry on perturbation space for one-sample problems, such geometrical structure cannot be easily generalizable to general parametric models, since it does not account for the inherent data structure (e.g., correlation among observations) and $\mathcal{M}$ itself [1].

Our choice of $\mathcal{P}(I|\mathcal{M})$ is motivated by four principles as follows.

- (P.a) (non-negativity) For any subset I, $\mathcal{P}(I|\mathcal{M})$ is always non-negative.
- (P.b) (uniqueness) $\mathcal{P}(I|\mathcal{M}) = 0$ if and only if I is an empty set.
- (P.c) (monotonicity) If $I_2 \subset I_1$, then $\mathcal{P}(I_2|\mathcal{M}) \leq \mathcal{P}(I_1|\mathcal{M})$.
- (P.d) (additivity) If $I_2 \subset I_1$, $I_{1\cdot 2} = I_1 - I_2$, and $p(\mathbf{Y}_{I_{1\cdot 2}}|\mathbf{Y}_{[I_1]}, \boldsymbol{\theta}) = p(\mathbf{Y}_{I_{1\cdot 2}}|\mathbf{Y}_{[I_{1\cdot 2}]}, \boldsymbol{\theta})$ for all $\boldsymbol{\theta}$, then we have $\mathcal{P}(I_1|\mathcal{M}) = \mathcal{P}(I_2|\mathcal{M}) + \mathcal{P}(I_{1\cdot 2}|\mathcal{M})$.

Principles (P.a) and (P.b) indicates that deleting any nonempty subset always introduces a positive degree of perturbation. Principle (P.c) indicates that deleting a larger subset always introduces a larger degree of perturbation. Principle (P.d) presents the condition for ensuring the additivity

property of the perturbation. Since $\mathbf{Y}_{[I_{1\cdot 2}]}$ is the union of $\mathbf{Y}_{[I_1]}$ and $\mathbf{Y}_{I_2}$, $p(\mathbf{Y}_{I_{1\cdot 2}}|\mathbf{Y}_{[I_1]},\boldsymbol{\theta}) = p(\mathbf{Y}_{I_{1\cdot 2}}|\mathbf{Y}_{[I_{1\cdot 2}]},\boldsymbol{\theta})$ is equivalent to that of $\mathbf{Y}_{I_{1\cdot 2}}$ being independent of $\mathbf{Y}_{I_2}$ given $\mathbf{Y}_{[I_1]}$. The additivity property has important implications in cross-sectional, longitudinal, and family data. For instance, in longitudinal data, the degree of perturbation introduced by simultaneously deleting two independent clusters equals the sum of their degrees of individual cluster perturbations. Besides these principles, $\mathcal{P}(I|\mathcal{M})$ should naturally arise from $\mathcal{M}$. We propose $\mathcal{P}(I|\mathcal{M})$ based on the well-known Kullback-Leibler divergence below.

We consider a model for characterizing the deletion of $\mathbf{Y}_I$ given by

$$p(\mathbf{Y}|\boldsymbol{\theta}, I) = p(\mathbf{Y}_{[I]}|\boldsymbol{\theta})p_0(\mathbf{Y}_I|\mathbf{Y}_{[I]}), \tag{2.5}$$

where $p_0(\mathbf{Y}_I|\mathbf{Y}_{[I]})$ is a fixed conditional density of $\mathbf{Y}_I$ given $\mathbf{Y}_{[I]}$ independent of $\boldsymbol{\theta}$. In (2.5), $\mathbf{Y}_I$ does not provide any information on $\boldsymbol{\theta}$, and thus $\hat{\boldsymbol{\theta}}_{[I]}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ under $p(\mathbf{Y}|\boldsymbol{\theta}, I)$. To avoid any arbitrary specification of $p_0(\mathbf{Y}_I|\mathbf{Y}_{[I]})$, we suggest setting $p_0(\mathbf{Y}_I|\mathbf{Y}_{[I]}) = p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \boldsymbol{\theta}_*)$, where $\boldsymbol{\theta}_*$ is the true value of $\boldsymbol{\theta}$ under $\mathcal{M}$. Moreover, if $\mathcal{M}$ is correctly specified, then $p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \boldsymbol{\theta}_*)$ is the true data generator for $\mathbf{Y}_I$ given $\mathbf{Y}_{[I]}$.

Note that $p(\mathbf{Y}|\boldsymbol{\theta}) = p(\mathbf{Y}_{[I]}|\boldsymbol{\theta})p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \boldsymbol{\theta})$, where $p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \boldsymbol{\theta})$ is the conditional density of $\mathbf{Y}_I$ given $\mathbf{Y}_{[I]}$. We then consider the Kullback-Leibler distance between $p(\mathbf{Y}|\boldsymbol{\theta})$ and $p(\mathbf{Y}|\boldsymbol{\theta}, I)$, denoted by $\mathrm{KL}(\mathbf{Y}, \boldsymbol{\theta}|I)$, as follows:

$$\int p(\mathbf{Y}|\boldsymbol{\theta}) \log\left(\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y}|\boldsymbol{\theta}, I)}\right) d\mathbf{Y} = \int p(\mathbf{Y}|\boldsymbol{\theta}) \log\left(\frac{p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \boldsymbol{\theta})}{p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \boldsymbol{\theta}_*)}\right) d\mathbf{Y}. \tag{2.6}$$

We use $\mathrm{KL}(\mathbf{Y}, \boldsymbol{\theta}|I)$ to measure the effect of deleting $\mathbf{Y}_I$ on $\boldsymbol{\theta}$. If $\mathbf{Y}_I$ is independent of $\mathbf{Y}_{[I]}$, then $\mathrm{KL}(\mathbf{Y}, \boldsymbol{\theta}|I)$ reduces to $\int p(\mathbf{Y}_I|\boldsymbol{\theta}) \log\left(p(\mathbf{Y}_I|\boldsymbol{\theta})/p(\mathbf{Y}_I|\boldsymbol{\theta}_*)\right) d\mathbf{Y}_I$, which is independent of $\mathbf{Y}_{[I]}$. Since $\boldsymbol{\theta}$ is unknown and $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*)$ is asymptotically normal, we focus on these $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_*$ by assuming a Gaussian prior for $\boldsymbol{\theta}$ with mean $\boldsymbol{\theta}_*$ and positive definite covariance matrix Σ_* (e.g., the Fisher information matrix), denoted by $p(\boldsymbol{\theta}|\boldsymbol{\theta}_*, \Sigma_*)$. Finally, we define $\mathcal{P}(I|\mathcal{M})$ as the weighted Kullback-Leibler distance between $p(\mathbf{Y}|\boldsymbol{\theta})$ and $p(\mathbf{Y}|\boldsymbol{\theta}, I)$ as follows:

$$\mathcal{P}(I|\mathcal{M}) = \int \mathrm{KL}(\mathbf{Y}, \boldsymbol{\theta}|I) p(\boldsymbol{\theta}|\boldsymbol{\theta}_*, \Sigma_*) d\boldsymbol{\theta}. \tag{2.7}$$

Furthermore, if we are interested in a particular set of components of $\boldsymbol{\theta}$ and treat others as nuisance parameters, we may fix these nuisance parameters in their true value.

It is easy to compute $\mathcal{P}(I|\mathcal{M})$ in real data analysis. Specifically, we only need to specify $\mathcal{M}$ and $(\boldsymbol{\theta}_*, \Sigma_*)$ and then use some numerical integration

methods to compute $\mathcal{P}(I|\mathcal{M})$. Although $(\boldsymbol{\theta}_*, \Sigma_*)$ is unknown, we suggest substituting $\boldsymbol{\theta}_*$ by an estimator of $\boldsymbol{\theta}$, denoted by $\tilde{\boldsymbol{\theta}}$, and Σ_* by the covariance matrix of $\tilde{\boldsymbol{\theta}}$. For instance, we may set $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$, since $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_*$ even though $\mathcal{M}$ may be mis-specified [25, 26].

We can obtain the following theorem.

Theorem 1. *Suppose that* $L(\{\mathbf{Y} : p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \boldsymbol{\theta}) = p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \boldsymbol{\theta}_*)\}) > 0$ *for any* $\boldsymbol{\theta} \neq \boldsymbol{\theta}_*$, *where* $L(A)$ *is the Lebesgue measure of a set* A. *Then,* $\mathcal{P}(I|\mathcal{M})$ *defined in (2.7) satisfies the four principles (P.a)-(P.d).*

Proof of Theorem 1. (P.a) directly follows from the Jensen inequality, (2.6) and (2.7).

For (P.b), if I is an empty set, then $\mathrm{KL}(\mathbf{Y}, \boldsymbol{\theta}|I) \equiv 0$ and thus $\mathcal{P}(I|\mathcal{M}) = 0$. On the other hand, if $\mathcal{P}(I|\mathcal{M}) = 0$, then $\mathrm{KL}(\mathbf{Y}, \boldsymbol{\theta}|I) \equiv 0$ for almost every $\boldsymbol{\theta}$. Thus, by using the Jensen inequality, we have $p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \boldsymbol{\theta}) \equiv p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \boldsymbol{\theta}_*)$ for all $\boldsymbol{\theta} \in \Theta$. Based on the identifiability condition, we know that I must be an empty set.

Let $I_{1\cdot 2} = I_1 - I_2$. It is easy to show that

$$\begin{aligned} p(\mathbf{Y}_{I_1}|\mathbf{Y}_{[I_1]}, \boldsymbol{\theta}) &= p(\mathbf{Y}_{I_2}, \mathbf{Y}_{I_{1\cdot 2}}|\mathbf{Y}_{[I_1]}, \boldsymbol{\theta}) = \frac{p(\mathbf{Y}_{I_2}, \mathbf{Y}_{I_{1\cdot 2}}, \mathbf{Y}_{[I_1]}|\boldsymbol{\theta})}{p(\mathbf{Y}_{[I_1]}|\boldsymbol{\theta})} \\ &= p(\mathbf{Y}_{I_2}|\mathbf{Y}_{[I_2]}, \boldsymbol{\theta}) p(\mathbf{Y}_{[I_2]}|\mathbf{Y}_{[I_1]}, \boldsymbol{\theta}). \end{aligned}$$

Thus, by substituting the above equation into (2.6), we have

(2.8)
$$\mathcal{P}(I_1|\mathcal{M}) = \mathcal{P}(I_2|\mathcal{M}) + \int p(\boldsymbol{\theta}|\boldsymbol{\theta}_*, \Sigma_{n*}) p(\mathbf{Y}|\boldsymbol{\theta}) \log \left(\frac{p(\mathbf{Y}_{[I_2]}|\mathbf{Y}_{[I_1]}, \boldsymbol{\theta})}{p(\mathbf{Y}_{[I_2]}|\mathbf{Y}_{[I_1]}, \boldsymbol{\theta}_*)} \right) d\boldsymbol{\theta} d\mathbf{Y},$$

in which the second term on the right hand side can be written as

$$\int p(\boldsymbol{\theta}|\boldsymbol{\theta}_*, \Sigma_{n*}) p(\mathbf{Y}_{I_2}|\mathbf{Y}_{[I_2]}, \boldsymbol{\theta}) \left\{ \int p(\mathbf{Y}_{[I_2]}|\boldsymbol{\theta}) \log \left(\frac{p(\mathbf{Y}_{[I_2]}|\mathbf{Y}_{[I_1]}, \boldsymbol{\theta})}{p(\mathbf{Y}_{[I_2]}|\mathbf{Y}_{[I_1]}, \boldsymbol{\theta}_*)} \right) d\mathbf{Y}_{[I_2]} \right\} d\boldsymbol{\theta} d\mathbf{Y}_{I_2} \geq 0,$$

which yield (P.c).

Based on the assumption of (P.d), we know that

$$p(\mathbf{Y}_{[I_2]}|\mathbf{Y}_{[I_1]}, \boldsymbol{\theta}) = p(\mathbf{Y}_{I_{1\cdot 2}}|\mathbf{Y}_{[I_1]}, \boldsymbol{\theta}) = p(\mathbf{Y}_{I_{1\cdot 2}}|\mathbf{Y}_{[I_{1\cdot 2}]}, \boldsymbol{\theta})$$

for all $\boldsymbol{\theta}$. Thus, the second term on the right hand side of (2.8) reduces to $\mathcal{P}(I_{1\cdot 2}|\mathcal{M})$, which finishes the proof of (P.d).

As an illustration, we show how to calculate $\mathcal{P}(I|\mathcal{M})$ under the standard linear regression model for cross-sectional data as follows.

Example 1. Consider the linear regression model $y_i = \mathbf{x}_i^T \boldsymbol{\beta}_* + \epsilon_i$, where $\mathbf{x}_i$ is a $p \times 1$ vector and the ϵ_i are independently and identically distributed (i.i.d)

as $N(0,\sigma_*^2)$. Let $\mathbf{y}=(y_1,\ldots,y_n)^T$ and $\mathbf{X}$ be an $n\times p$ matrix of rank p with i-th row $\mathbf{x}_i^T$. In this case, $\boldsymbol{\theta}=(\boldsymbol{\beta}^T,\sigma^2)^T$. Recall that $\hat{\boldsymbol{\beta}}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, $\hat{\sigma}^2=\mathbf{y}^T(\mathbf{I}_n-H_x)\mathbf{y}/n$, $\mathrm{Cov}(\hat{\boldsymbol{\beta}})=\sigma_*^2(\mathbf{X}^T\mathbf{X})^{-1}$, and $\mathrm{var}(\hat{\sigma}^2)=2\sigma_*^4/n$, where $\mathbf{I}_n$ is an $n\times n$ identity matrix and $H_x=(h_{ij})=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. We first compute the degree of the perturbation for deleting each $(y_i,\mathbf{x}_i)$. We consider two scenarios: fixed and random covariates. For the case of fixed covariates, $\mathcal{M}$ assumes $y_i\sim N(\mathbf{x}_i^T\boldsymbol{\beta},\sigma^2)$. After some algebraic calculations, it can be shown that $\mathcal{P}(\{i\}|\mathcal{M})$ equals

$$0.5E_\theta[\log(\sigma_*^2/\sigma^2)]+0.5\frac{\mathbf{x}_i^TE_\theta[(\boldsymbol{\beta}-\boldsymbol{\beta}_*)(\boldsymbol{\beta}-\boldsymbol{\beta}_*)^T]\mathbf{x}_i}{\sigma_*^2}\approx\frac{1}{2}h_{ii}+\frac{1}{2n}, \tag{2.9}$$

where E_θ is taken with respect to $p(\boldsymbol{\theta}|\boldsymbol{\theta}_*,G_{n\theta}^{-1})$. If we are only interested in $\boldsymbol{\beta}$ and treat σ^2 as a nuisance parameter, the term $0.5E_\theta[\log(\sigma_*^2/\sigma^2)]$ and $1/(2n)$ can be dropped from $\mathcal{P}(\{i\}|\mathcal{M})$ in (2.9).

Furthermore, for the case of random covariates, we assume that the $\mathbf{x}_i$'s are identically distributed with mean μ_x and covariance matrix Σ_x. It can be shown that $\mathcal{P}(\{i\}|\mathcal{M})$ equals

$$0.5E_\theta[\log(\sigma_*^2/\sigma^2)]+0.5\sigma_*^{-2}\mathrm{tr}\{\Sigma_xE_\theta[(\boldsymbol{\beta}-\boldsymbol{\beta}_*)(\boldsymbol{\beta}-\boldsymbol{\beta}_*)^T]\}\approx\frac{p+1}{2n}. \tag{2.10}$$

If σ^2 is a nuisance parameter, then $\mathcal{P}(\{i\}|\mathcal{M})$ reduces to $p/(2n)$. Furthermore, consider deleting a subset of observations $\{(y_{i_k},\mathbf{x}_{i_k}):k=1,\cdots,n(I)\}$ and $I=\{i_1,\ldots,i_{n(I)}\}$. It follows from Theorem 1 that $\mathcal{P}(\{i_1,\cdots,i_{n(I)}\}|\mathcal{M})=\sum_{k=1}^{n(I)}\mathcal{P}(\{i_k\}|\mathcal{M})$. Furthermore, for the case of random covariates, we have $\mathcal{P}(I|\mathcal{M})=n(I)\mathcal{P}(\{1\}|\mathcal{M})$ for any subset I with $n(I)$ observations. Thus, in this case, deleting any two subsets I_1 and I_2 with the same number of observations, that is $n(I_1)=n(I_2)$, has the same degree of perturbation. An important implication of these calculations in real data analysis is that we can directly compare $\mathrm{CD}(I_1)$ and $\mathrm{CD}(I_2)$ when $n(I_1)=n(I_2)$.

2.3. *Size issue.* Given $\mathcal{P}(I|\mathcal{M})$ and $\mathrm{CD}(I)$, we are interested in solving the second question below.

(c.2) Is there any relationship between $\mathcal{P}(I|\mathcal{M})$ and $\mathrm{CD}(I)$? If any, how do we quantify such a relationship?

To have a better understanding of Cook's distance, we consider the standard linear regression model for cross-sectional data as follows.

Example 1 (continued). We are interested in $\boldsymbol{\beta}$ and treat σ^2 as a nuisance parameter. We first consider deleting individual observations in linear regression. Cook's distance [9] for case i, $(y_i,\mathbf{x}_i)$, is given by

$$\mathrm{CD}(\{i\})=\frac{(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{[i]})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{[i]})}{\hat{\sigma}^2}=\frac{\sigma^2}{\hat{\sigma}^2}t_i^2\frac{h_{ii}}{1-h_{ii}}, \tag{2.11}$$

where $\hat{\boldsymbol{\beta}}$ is the least squares estimate of $\boldsymbol{\beta}$, $\hat{\sigma}^2$ is a consistent estimator of σ^2, $t_i = \hat{e}_i/(\sigma\sqrt{1-h_{ii}})$ and $\hat{\boldsymbol{\beta}}_{[i]} = \hat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\hat{e}_i/(1-h_{ii})$, in which $\hat{e}_i = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$. It should be noted that except for a constant p, CD($\{i\}$) is almost the same as the original Cook's distance (Cook, 1977). As shown in (2.9) and (2.10), regardless of the exact value of $(y_i, \mathbf{x}_i)$, deleting any $(y_i, \mathbf{x}_i)$ has the approximately same degree of perturbation to $\mathcal{M}$. Moreover, the CD($\{i\}$) are comparable regardless of i. Specifically, if $\epsilon_i \sim N(0, \sigma^2)$, then t_i^2 follows the $\chi^2(1)$ distribution for all i. For the case of random covariates, if $\mathbf{x}_i$ are identically distributed, then all CD($\{i\}$) are truly comparable, since they follow the same distribution.

We consider deleting multiple observations in the linear model. Cook's distance for deleting the subset I with $n(I)$ is given by

(2.12)
$$\frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[I]})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[I]})}{\hat{\sigma}^2} = \frac{1}{\hat{\sigma}^2}\hat{\mathbf{e}}_I^T(\mathbf{I}_{n(I)} - H_I)^{-1}H_I(\mathbf{I}_{n(I)} - H_I)^{-1}\hat{\mathbf{e}}_I,$$

where $\hat{\mathbf{e}}_I$ is an $n(I) \times 1$ vector containing all $\hat{e}_i$ for $i \in I$ and $H_I = \mathbf{X}_I(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_I^T$, in which $\mathbf{X}_I$ is an $n(I) \times p$ matrix whose rows are $\mathbf{x}_i^T$ for all $i \in I$. Similar to the deletion of a single case, deleting any subset with the same number of observations introduces approximately the same degree of perturbation to $\mathcal{M}$, and the CD(I) are comparable among all subsets with the same $n(I)$. We will make this statement precise in Theorem 2 given below.

Generally, we want to compare CD(I_1) and CD(I_2) for any two subsets with $n(I_1) \neq n(I_2)$. As shown in Example 1, when $n(I_1) > n(I_2)$, deleting I_1 introduces larger degree of perturbation to model $\mathcal{M}$ compared to deleting I_2. To compare Cook's distances among arbitrary subsets, we need to understand the relationship between $\mathcal{P}(I|\mathcal{M})$ and CD(I) for any subset I. Surprisingly, in linear regression for cross-sectional data, we can show the stochastic relationship between $\mathcal{P}(I|\mathcal{M})$ and CD(I) as follows.

Theorem 2. *For the standard linear model, where* $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$ *and* $\epsilon \sim N(\mathbf{0}, \sigma^2 I_n)$, *we have the following results:*

(a) for any $I_2 \subset I_1$, *CD*(I_1) *is stochastically larger than CD*(I_2) *for any* $\mathbf{X}$, *that is,* $P(CD(I_1) > t|\mathcal{M}) \geq P(CD(I_2) > t|\mathcal{M})$ *holds for any* $t \geq 0$.

(b) Suppose that the components of $\mathbf{X}_I$ *and* $\mathbf{X}_{I'}$ *are identically distributed for any two subsets* I *and* I' *with* $n(I) = n(I')$. *Thus, CD*(I) *and CD*(I') *follow the same distribution when* $n(I) = n(I')$ *and CD*(I_1) *is stochastically larger than CD*(I_2) *for any two subsets* I_2 *and* I_1 *with* $n(I_1) > n(I_2)$.

Proof of Theorem 2. (a) Let $I_3 = I_1 \setminus I_2$, I_1 is a union of two disjoint sets I_3

and I_2. Without loss of generality, H_{I_1} can be decomposed as

$$H_{I_1} = \mathbf{X}_{I_1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{I_1}^T = \begin{pmatrix} \mathbf{X}_{I_2}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{I_2}^T & \mathbf{X}_{I_2}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{I_3}^T \\ \mathbf{X}_{I_3}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{I_2}^T & \mathbf{X}_{I_3}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{I_3}^T \end{pmatrix}.$$

Let $\lambda_{1,1} \geq \ldots \geq \lambda_{1,n(I_1)} \geq 0$ and $\lambda_{2,1} \geq \ldots \geq \lambda_{2,n(I_2)} \geq 0$ be ordered eigenvalues of H_{I_1} and H_{I_2}, respectively, where $n(I_k)$ denotes the number of observations in I_k for $k = 1, 2$. It follows from Wielandt's eigenvalue inequality [14] that $\lambda_{1,l} \geq \lambda_{2,l}$ for all $l = 1, \ldots, n(I_2)$. For $k = 1, 2$, we define $\Gamma_k\Lambda_k\Gamma_k^T$ as the spectral decomposition of H_{I_k} and $\mathbf{h}_k = (\mathbf{I}_{n(I_k)} - \Lambda_k)^{-1/2}\Gamma_k^T\hat{\mathbf{e}}_{I_k} = (h_{k,1}, \ldots, h_{k,n(I_k)})^T$, where Γ_k is an orthnormal matrix and $\Lambda_k = \text{diag}(\lambda_{k,1}, \ldots, \lambda_{k,n(I_k)})$. It can be shown that for $k = 1, 2$,

$$\mathbf{h}_k \sim N(\mathbf{0}, \sigma^2\mathbf{I}_{n(I_k)}) \text{ and } \text{CD}(I_k) = \frac{1}{\hat{\sigma}^2}\sum_{j=1}^{n(I_k)} \frac{\lambda_{k,j}}{1-\lambda_{k,j}} h_{k,j}^2.$$

Since $f(x) = x/(1-x)$ is an increasing function of $x \in (0, 1)$, this completes the proof of Theorem 2 (a).

Note that $\text{CD}(I) = (\hat{\sigma}^2)^{-1}\sum_{j=1}^{n(I)} \lambda_j(1-\lambda_j)^{-1}h_j^2$, where λ_j are the eigenvalues of H_I and $\mathbf{h} = (h_1, \ldots, h_{n(I)})^T \sim N(\mathbf{0}, \sigma^2\mathbf{I}_{n(I)})$. Moreover, the distribution of λ is uniquely determined by H_I. Combining $\mathbf{h} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_{n(I)})$ with the assumptions of Theorem 2 (b) yields that $\text{CD}(I)$ and $\text{CD}(I')$ follow the same distribution when $n(I) = n(I')$. Furthermore, we can always choose an I_2' such that $n(I_2') = n(I_2)$ and $I_1 \subset I_2'$. Following arguments in Theorem 2 (a), we can then complete the proof of Theorem 2 (b).

Theorem 2 (a) shows that the Cook's distances for two nested subsets satisfy the stochastic ordering property. Theorem 2 (b) indicates that for random covariates, the Cook's distances for any two subsets also satisfy the stochastic ordering property under some mild conditions.

According to Theorem 2, for more complex data structures and models, it may be natural to use the stochastic order to stochastically quantify the positive association between the degree of the perturbation and the size of Cook's distance. Specifically, we consider two possibly overlapping subsets I_1 and I_2 with $\mathcal{P}(I_1|\mathcal{M}) > \mathcal{P}(I_2|\mathcal{M})$. Although $\text{CD}(I_1)$ may not be greater than $\text{CD}(I_2)$ for a fixed dataset $\mathcal{D}$, $\text{CD}(I_1)$, as a random variable, should be *stochastically larger* than $\text{CD}(I_2)$ if $\mathcal{M}$ is the true model. We make the following assumption:

Assumption A1. For any two subsets I_1 and I_2 with $\mathcal{P}(I_1|\mathcal{M}) > \mathcal{P}(I_2|\mathcal{M})$,

$$\text{P}(\text{CD}(I_1) > t|\mathcal{M}) \;\geq\; \text{P}(\text{CD}(I_2) > t|\mathcal{M}) \tag{2.13}$$

holds for any $t > 0$, where the probability is taken with respect to $\mathcal{M}$.

Assumption A1 is essentially saying that if $\mathcal{M}$ is the true data generator, then $\mathrm{CD}(I_1)$ stochastically dominates $\mathrm{CD}(I_2)$ whenever $\mathcal{P}(I_1|\mathcal{M}) > \mathcal{P}(I_2|\mathcal{M})$. According to the definition of stochastic ordering [21], we can now obtain the following proposition.

Proposition 1. *Under Assumption A1, for any two subsets I_1 and I_2 with $\mathcal{P}(I_1|\mathcal{M}) > \mathcal{P}(I_2|\mathcal{M})$, Cook's distance satisfies*

$$E[h(CD(I_1))|\mathcal{M}] \geq E[h(CD(I_2))|\mathcal{M}] \tag{2.14}$$

holds for all increasing functions $h(\cdot)$. In particular, we have $E[CD(I_1)|\mathcal{M}] \geq E[CD(I_2)|\mathcal{M}]$ and $Q_{CD(I_1)}(\alpha|\mathcal{M})$ is greater than the α-quantile of $Q_{CD(I_2)}(\alpha|\mathcal{M})$ for any $\alpha \in [0, 1]$, where $Q_{CD(I)}(\alpha|\mathcal{M})$ denotes the $\alpha-$quantile of the distribution of $CD(I)$ for any subset I.

Proposition 1 formally characterizes the size issue for Cook's distance. Specifically, for any two subsets I_1 and I_2 with $\mathcal{P}(I_1|\mathcal{M}) > \mathcal{P}(I_2|\mathcal{M})$, $\mathrm{CD}(I_1)$ has a high probability of being greater than $\mathrm{CD}(I_2)$. Thus, Cook's distance for subsets with different degrees of perturbation are not directly comparable. More importantly, it indicates that $\mathrm{CD}(I)$ cannot be simply expressed as a linear function of $\mathcal{P}(I|\mathcal{M})$ even for linear regression for cross-sectional data. Thus, the standard solution, which standardizes $\mathrm{CD}(I)$ by calculating the ratio of $\mathrm{CD}(I)$ over $\mathcal{P}(I|\mathcal{M})$, is not desirable.

2.4. *Scaled Cook's distances.* After characterizing the stochastic relationship between $\mathcal{P}(I|\mathcal{M})$ and $\mathrm{CD}(I)$, we are interested in answering a third question as follows.

(c.3) How do we compare Cook's distance relative to $\mathcal{P}(I|\mathcal{M})$ for different subsets I?

We focus on developing several scaled Cook's distance for a subset I, denoted by $\mathrm{SCD}(I)$, to detect relatively influential subsets. From here on, we call a subset I as *relatively influential*, if its $\mathrm{CD}(I)$ is large relative to $\mathcal{P}(I|\mathcal{M})$. We consider matching several features (e.g., mean, median, or quantiles) of $\mathrm{SCD}(I)$, when the current model $\mathcal{M}$ is the true data generator. Specifically, we consider two pairs of features including (mean, Std) and (median, Mstd), where Std and Mstd, respectively, denote the standard deviation and the median standard deviation. By matching any of the two pairs, we can at least ensure that the centers and scales of the scaled Cook's distances for different subsets are the same. Therefore, for any two subsets I_1 and I_2, the probability of observing the event $\mathrm{SCD}(I_1) > \mathrm{SCD}(I_2)$ and that of the event $\mathrm{SCD}(I_1) < \mathrm{SCD}(I_2)$ should be reasonably close to each other. Thus, the $\mathrm{SCD}(I)$ are roughly comparable.

We introduce two scaled Cook's distance measures, called scaled Cook's distances, as follows.

Definition 1. The *scaled Cook's distances* for matching (mean, Std) and (median, Mstd) are, respectively, defined as

(2.15)
$$\mathrm{SCD}_1(I) = \frac{\mathrm{CD}(I) - E[\mathrm{CD}(I)|\mathcal{M}]}{\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M}]} \text{ and } \mathrm{SCD}_2(I) = \frac{\mathrm{CD}(I) - Q_{CD(I)}(0.5|\mathcal{M})}{\mathrm{Mstd}[\mathrm{CD}(I)|\mathcal{M}]},$$

where both the expectation and the quantile are taken with respect to $\mathcal{M}$.

We can use $\mathrm{SCD}_1(I)$ and $\mathrm{SCD}_2(I)$ to evaluate the relatively influential level for different subsets I. A large value of $\mathrm{SCD}_1(I)$ (or $\mathrm{SCD}_2(I)$) indicates that the subset I is relatively influential. Note that the scaled Cook's distances do not provide a "per unit" effect of removing one observation within the set I, whereas they measure the standardized influential level of the set I when $\mathcal{M}$ is true.

The next task is how to compute $E[\mathrm{CD}(I)|\mathcal{M}]$, $\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M}]$, $\mathrm{Mstd}[\mathrm{CD}(I)|\mathcal{M}]$, and $Q_{CD(I)}(0.5|\mathcal{M})$ for each subset I under the assumption that $\mathcal{M}$ is the true data generator. Computationally, based on $p(\mathbf{Y}|\hat{\boldsymbol{\theta}})$, we suggest using the parametric bootstrap to approximate the four quantities of $\mathrm{CD}(I)$ as follows.

Step 1. We use $\hat{\mathcal{M}} = \{p(\mathbf{Y}|\hat{\boldsymbol{\theta}})\}$ to approximate the model $\mathcal{M} = \{p(\mathbf{Y}|\boldsymbol{\theta}_*)\}$, generate a random sample $\mathbf{Y}^s$ from $p(\mathbf{Y}|\hat{\boldsymbol{\theta}})$ and then calculate $\mathrm{CD}(I)^{(s)} = F_1(I, \hat{\mathcal{M}}, \mathbf{Y}^s)$ for each s and each subset I.

Step 2. By repeating this process S times, we can obtain a sample $\{\mathrm{CD}(I)^{(s)} : s = 1, \cdots, S\}$ and then we use its empirical mean $\overline{\mathrm{CD}(I)} = \sum_{s=1}^{S} \mathrm{CD}(I)^{(s)}/S$ to approximate $E[\mathrm{CD}(I)|\mathcal{M}]$.

Step 3. We approximate $\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M}]$, $Q_{CD(I)}(0.5|\mathcal{M})$, and $\mathrm{Mstd}[\mathrm{CD}(I)|\mathcal{M}]$ by using their corresponding empirical quantities of $\{\mathrm{CD}(I)^{(s)} : s = 1, \cdots, S\}$.

In this process, we have used $\hat{\mathcal{M}}$ to approximate $\mathcal{M}$ [25] and simulated data $\mathbf{Y}^s$ from $\hat{\mathcal{M}}$ in the standard parametric bootstrap method. If $\mathbf{Y}$ truly comes from $\mathcal{M}$, then the simulated data $\mathbf{Y}^s$ should resemble $\mathbf{Y}$. Since $\hat{\boldsymbol{\theta}}$ is a consistent estimate of $\boldsymbol{\theta}_*$, $E[F_1(I, \hat{\mathcal{M}}, \mathbf{Y})|\hat{\mathcal{M}}] \approx E[F_1(I, \mathcal{M}, \mathbf{Y})|\mathcal{M}]$ and thus $\overline{\mathrm{CD}(I)}$ is a consistent estimate of $E[F_1(I, \mathcal{M}, \mathbf{Y})|\mathcal{M}]$. Similar arguments hold for the other three quantities of $\mathrm{CD}(I)$. In Steps 2 and 3, we can use relatively large S, say $S = 1,000$, in order to accurately approximate all four quantities of $\mathrm{CD}(I)$. According to our experience, such approximation is very accurate even for moderate S and small n. See simulation studies in Section 3.1 for details. However, for most statistical models with complex data structures, it can be computationally intensive to compute $\hat{\boldsymbol{\theta}}^s$ for each $\mathbf{Y}^s$. We will address this issue in Section 2.6.

In the following, we will derive the scaled Cook's distances for generalized linear models.

Example 2. We consider Cook's distance in generalized linear models [19] as follows. Suppose that the components of $\mathbf{y} = (y_1, \ldots, y_n)^T$ given $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$ are mutually independent, and the conditional density of y_i given $\mathbf{x}_i$ is given by

$$p(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \tau) = \exp\left\{a(\tau)^{-1}[y_i\eta_i - b(\eta_i)] + c(y_i, \tau)\right\}, \tag{2.16}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$ are known functions, $\eta_i = \eta(\mu_i)$ and $\mu_i = \mu_i(\boldsymbol{\beta}) = g(\mathbf{x}_i^T\boldsymbol{\beta})$, in which $g(\cdot)$ is a known monotonic function and twice continuously differentiable and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. Throughout this example, the parameter of interest is $\boldsymbol{\beta}$ and τ is a nuisance parameter and is fixed at $\hat{\tau}$. Let $V(\boldsymbol{\beta}) = \text{diag}(\ddot{b}(\eta(\mu_1(\boldsymbol{\beta}))), \ldots, \ddot{b}(\eta(\mu_n(\boldsymbol{\beta}))))$ and $D(\boldsymbol{\beta})^T = (\partial_\beta\mu_1(\boldsymbol{\beta}), \ldots, \partial_\beta\mu_n(\boldsymbol{\beta}))$, where ∂_β denotes differentiation with respect to $\boldsymbol{\beta}$ and $\ddot{b}(\eta)$ denotes the second derivative of $b(\eta)$ with respect to η. Using a first-order approximation, we can show that Cook's distance for deleting subset I with $\text{size}(I) = n(I)$ can be approximated by

$$\widetilde{\text{CD}}(I) = \frac{1}{a(\hat{\tau})}\hat{\mathbf{e}}^T\hat{V}^{-1/2}U_I(\mathbf{I}_{n(I)} - \hat{H}_I)^{-1}\hat{H}_I(\mathbf{I}_{n(I)} - \hat{H}_I)^{-1}U_I^T\hat{V}^{-1/2}\hat{\mathbf{e}}, \tag{2.17}$$

where $\hat{D} = D(\hat{\boldsymbol{\beta}})$, $\hat{V} = V(\hat{\boldsymbol{\beta}})$, $\hat{\mathbf{e}}$ is an $n \times 1$ vector containing all $\hat{e}_i = y_i - \mu_i(\hat{\boldsymbol{\beta}})$, and $\hat{H}_I = \tilde{\mathbf{X}}_I(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}_I^T$. In addition, $\tilde{\mathbf{X}} = \hat{V}^{-1/2}\hat{D}$ and $\tilde{\mathbf{X}}_I$ is an $n(I) \times p$ matrix containing the $i-$th row of $\tilde{\mathbf{X}}$ for all $i \in I$, and $U_I = (\mathbf{u}_{i_1}, \ldots, \mathbf{u}_{i_{n(I)}})$, in which $i_k \in I$ and $\mathbf{u}_{i_k}$ is an $n \times 1$ vector with i_k-th element equal to 1 and zero otherwise.

For generalized linear models, we can calculate the scaled Cook's distance and thus obtain the following theorem.

Theorem 3. *Suppose that Assumptions A2-A5 in the appendix hold for the generalized linear model (2.16). We have the following results:*

(a) $\widetilde{CD}(I) = CD_*(I)[1 + o_p(1)]$, *and* $CD_*(I) = \mathbf{e}_*^T\mathbf{W}_*\mathbf{e}_*/[a(\tau_*)]$, *where* $\mathbf{W}_* = (w_{ij*})$ *is an* $n \times n$ *matrix and given by*

$$\mathbf{W}_* = V_*^{-1/2}(\mathbf{I}_n - H_*)U_I(\mathbf{I}_{n(I)} - H_{*,I})^{-1}H_{*,I}(\mathbf{I}_{n(I)} - H_{*,I})^{-1}U_I^T(\mathbf{I}_n - H_*)V_*^{-1/2}, \tag{2.18}$$

in which $\mathbf{e}_* = (e_{1*}, \ldots, e_{n*})^T$ *and* $e_{i*} = y_i - \mu_i(\boldsymbol{\beta}_*)$, $D_* = D(\boldsymbol{\beta}_*)$, $V_* = V(\boldsymbol{\beta}_*)$, $H_* = \mathbf{X}_*(\mathbf{X}_*^T\mathbf{X}_*)^{-1}\mathbf{X}_*^T$, $\mathbf{X}_* = V_*^{-1/2}D_*$, $H_{*,I} = U_I^T H_* U_I$ *and* $\boldsymbol{\beta}_*$ *is the true value of* $\boldsymbol{\beta}$.

(b) Let $\lambda_{I,1} \geq \ldots \lambda_{I,n(I)} \geq 0$ be the ordered eigenvalues of $H_{*,I}$. We have

$$E[CD_*(I)|\mathcal{M}] = E\{tr[(\mathbf{I}_{n(I)} - H_{*,I})^{-1}]|\mathcal{M}\} - n(I) = \sum_{j=1}^{n(I)} E[(1-\lambda_{I,j})^{-1}|\mathcal{M}] - n(I),$$

$$Var[CD_*(I)|\mathcal{M}] = a(\tau_*)\sum_{i=1}^{n} w_{ii*}b^{(4)}(\eta_{i*}) + Var\{tr[(\mathbf{I}_{n(I)} - H_{*I})^{-1}]|\mathcal{M}\}$$
$$+ \quad 2E\{tr[(\mathbf{I}_{n(I)} - H_{*,I})^{-2}]|\mathcal{M}\} - 4E\{tr[(\mathbf{I}_{n(I)} - H_{*,I})^{-1}]|\mathcal{M}\} + 2n(I),$$

where $\eta_{i*} = \eta(\mu_i(\boldsymbol{\beta}_*))$ and $b^{(4)}(\eta_{i*})$ denotes the fourth derivative of $b(\eta)$ with respect to η. If $n(I) \geq p$, then $\sum_{j=1}^{n(I)} E[(1-\lambda_{I,j})^{-1}|\mathcal{M}] - n(I) = \sum_{j=1}^{p} E[(1-\lambda_{I,j})^{-1}|\mathcal{M}] - p$.

(c) If the $\mathbf{x}_i$ are independently and identically distributed and $0 < E[||\ddot{b}(\eta(g(\mathbf{x}^T\boldsymbol{\beta})))^{-1/2}\partial_\beta g(\mathbf{x}^T\boldsymbol{\beta})||_2^{1+s}] < \infty$ for an arbitrary $s > 0$, then $\lambda_{I,j} - n(I)/n = o(1)$ for $j \leq p$ as $n(I) \to \infty$ and $n(I)/n \to \gamma \in [0,1)$.

Proof of Theorem 3. (a). Let $\mu(\boldsymbol{\beta}) = (\mu_1(\boldsymbol{\beta}), \ldots, \mu_n(\boldsymbol{\beta}))^T$. If the model $\mathcal{M}$ is true, then $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*) = (D_*^T V_*^{-1} D_*)^{-1} D_*^T V_*^{-1}\mathbf{e}_* + o_p(n^{-1/2})$. Thus, under Assumptions A2-A5, we have

$$U_I^T V_*^{-1/2}\hat{\mathbf{e}} = U_I^T V_*^{-1/2}[\mathbf{y} - \mu(\boldsymbol{\beta}_*) + \mu(\boldsymbol{\beta}_*) - \mu(\hat{\boldsymbol{\beta}})]$$
$$= U_I^T V_*^{-1/2}[\mathbf{e}_* - D_*(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)] = U_I^T(\mathbf{I}_n - H_*)V_*^{-1/2}\mathbf{e}_*[1 + o_p(1)],$$

where $\mathbf{e}_* = \mathbf{y} - \mu(\boldsymbol{\beta}_*)$. This yields Theorem 3 (a).

(b). We consider two scenarios including both random and fixed covariates. For the case of random covariate, the current model $\mathcal{M}$ includes the specifications of the distribution on $\mathbf{X}$ and the conditional distribution of $\mathbf{y}$ given $\mathbf{X}$, which are, respectively, represented as $\mathcal{M}_{\mathbf{X}}$ and $\mathcal{M}_{\mathbf{y}|\mathbf{X}}$. Since $E[\mathbf{e}_*^{\otimes 2}|\mathcal{M}] = a(\tau_*)E[V_*|\mathcal{M}_{\mathbf{X}}]$, it can be shown that $E[\mathrm{CD}_*(I)|\mathcal{M}]$ equals

$$p^{-1}E\{\mathrm{tr}[H_{*,I}(\mathbf{I}_{n(I)} - H_{*,I})^{-1}]|\mathcal{M}\} = p^{-1}E\{\mathrm{tr}[(\mathbf{I}_{n(I)} - H_{*,I})^{-1}]|\mathcal{M}\} - p^{-1}n(I),$$

where $E[\cdot|\mathcal{M}_{\mathbf{X}}]$ denotes the expectation taken with respect to the distribution of $\mathbf{X}$. Recall that $E[e_{i*}|\mathcal{M}] = 0$, $E[e_{i*}^2|\mathcal{M}] = a(\tau_*)E[\ddot{b}(\eta_{i*})|\mathcal{M}_{\mathbf{X}}]$, and

$$E[e_{i*}^4|\mathcal{M}] = 3a(\tau_*)^2 E[\ddot{b}(\eta_{i*})^2|\mathcal{M}_{\mathbf{X}}] + a(\tau_*)^3 E[b^{(4)}(\eta_{i*})|\mathcal{M}_{\mathbf{X}}].$$

With some algebraic calculation, it can be shown that

$$
\begin{aligned}
& E[\sum_{i,j=1}^{n} w_{ij*}e_{i*}e_{j*}|\mathcal{M}] = a(\tau_*)\sum_{i=1}^{n} E[w_{ii*}\ddot{b}(\eta_{i*})|\mathcal{M}_{\mathbf{X}}], \\
& E\{[\sum_{i,j=1}^{n} w_{ij*}e_{i*}e_{j*}]^2|\mathcal{M}\} = a(\tau_*)^3\sum_{i=1}^{n} E[w_{ii*}^2 b^{(4)}(\eta_{i*})|\mathcal{M}_{\mathbf{X}}] \\
+ \quad & a(\tau_*)^2 E(\{[\sum_{i=1}^{n} w_{ii*}\ddot{b}(\eta_{i*})]^2 + 2\sum_{i,j=1}^{n} w_{ij*}^2\ddot{b}(\eta_{i*})\ddot{b}(\eta_{j*})\}|\mathcal{M}_{\mathbf{X}}), \\
& \mathrm{Var}\{[\sum_{i,j=1}^{n} w_{ij*}e_{i*}e_{j*}]^2|\mathcal{M}\} = a(\tau_*)^3\sum_{i=1}^{n} E[w_{ii*}^2 b^{(4)}(\eta_{i*})|\mathcal{M}_{\mathbf{X}}] \\
+ \quad & a(\tau_*)^2\mathrm{Var}[\sum_{i=1}^{n} w_{ii*}\ddot{b}(\eta_{i*})|\mathcal{M}_{\mathbf{X}}] + 2a(\tau_*)^2 E[\sum_{i,j=1}^{n} w_{ij*}^2\ddot{b}(\eta_{i*})\ddot{b}(\eta_{j*})|\mathcal{M}_{\mathbf{X}}].
\end{aligned}
$$

Furthermore, we have

$$
\begin{aligned}
& \sum_{i,j=1}^{n} w_{ij*}^2\ddot{b}(\eta_{i*})\ddot{b}(\eta_{j*}) = \mathrm{tr}[(\mathbf{I}_{n(I)} - H_{*I})^{-1}H_{*I}(\mathbf{I}_{n(I)} - H_{*I})^{-1}H_{*I}] \\
= \quad & \mathrm{tr}[(\mathbf{I}_{n(I)} - H_{*I})^{-2}] - 2\mathrm{tr}[(\mathbf{I}_{n(I)} - H_{*I})^{-1}] + n(I), \\
& \sum_{i=1}^{n} w_{ii*}\ddot{b}(\eta_{i*}) = \mathrm{tr}[(\mathbf{I}_{n(I)} - H_{*I})^{-1}H_{*I}] = \mathrm{tr}[(\mathbf{I}_{n(I)} - H_{*I})^{-1}] - n(I).
\end{aligned}
$$

In addition, since H_* only has p non-zero eigenvalues and $H_{*,I}$ is a submatrix of H_*, it follows from Wielandt's eigenvalue inequality that $\lambda_{I,1} \geq \ldots \geq \lambda_{I,p} \geq 0 = \lambda_{I,p+1} = \ldots = \lambda_{I,n(I)}$ for $n(I) \geq p$. This yields Theorem 3 (b).

(c). Note that the matrices $H_{*,I}$ and $(\mathbf{X}_*^T\mathbf{X}_*)^{-1}\mathbf{X}_{*,I}^T\mathbf{X}_{*,I}$ have the same set of nonzero eigenvalues. Since $n^{-1}\mathbf{X}_*^T\mathbf{X}_*$ and $n(I)^{-1}\mathbf{X}_{*,I}^T\mathbf{X}_{*,I}$ converge to the same matrix almost surely, $n(I)n^{-1}[(n^{-1}\mathbf{X}_*^T\mathbf{X}_*)^{-1}n(I)^{-1}\mathbf{X}_{*,I}^T\mathbf{X}_{*,I} - \mathbf{I}_p]$ converges to $\mathbf{0}$ almost surely as $n, n(I) \to \infty$. This completes the proof of Theorem 3 (c).

Theorem 3 (a) characterizes the stochastic behavior of $\widetilde{\mathrm{CD}}(I)$, which depends on both the responses and the covariates in the set I. To ensure that $E[\mathrm{CD}(I)|\mathcal{M}]$ and $Q_{CD(I)}(0.5|\mathcal{M})$ depend only on the size of the perturbation, not the set I itself, we need to bootstrap the randomness in both the responses and the covariates. Specifically, we can generate a new set of responses from the fitted model and draw an I_s at random from the original covariate data without (or with) replacement, where size(I_s)=size(I). Then, we calculate the CD(I_s) based on the bootstrapped data for $s = 1, \ldots, S$

and use their sample median to approximate $Q_{CD(I)}(0.5|\mathcal{M})$. Theorem 3 (b) gives an approximation of $E[\widetilde{\mathrm{CD}}(I)|\mathcal{M}]$ and $\mathrm{Var}[\widetilde{\mathrm{CD}}(I)|\mathcal{M}]$. We can draw a sample of sets $\{I_s : s = 1, \ldots, S\}$ of size(I) at random from the original covariate data without (or with) replacement and approximate them. Moreover, it should be noted that $\sum_{j=1}^{n(I)} E[(1-\lambda_{I,j})^{-1}|\mathcal{M}] - n(I)$ increases with the size of I even for $n(I) \geq p$. Theorem 3 (c) shows the asymptotic consistency of $\lambda_{I,j}$ for $j \leq p$. As $n(I)/n \to \gamma \in [0,1)$, $\sum_{j=1}^{p} E[(1-\lambda_{I,j})^{-1}|\mathcal{M}] - p$ converges to $p\gamma/(1-\gamma)$.

We consider the general linear model with correlated errors (LMCE).

Example 3. Consider the LMCE given by $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}+\epsilon$, where $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{R})$. By choosing various $\mathbf{R}$'s, LMCE includes the linear model with independent data, the multivariate linear model, time series models, geostatistical models, and mixed effects models as special cases [16, 17]. Similar to Haslett [16], we fix $\mathbf{R}$ at an appropriate estimate $\hat{\mathbf{R}}$ throughout the example. We can calculate the generalized least squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{Y} = \mathbf{B}\mathbf{Y}$, $\mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X})^{-1}$, and $\hat{\sigma}^2 = \mathbf{Y}^T\mathbf{Q}\mathbf{Y}/(n-p) = \hat{\mathbf{e}}^T\mathbf{R}^{-1}\hat{\mathbf{e}}/(n-p)$, where $\mathbf{Q} = \mathbf{R}^{-1} - \mathbf{H}$, $\hat{\mathbf{e}} = \mathbf{R}\mathbf{Q}\mathbf{Y}$, and $\mathbf{H} = \mathbf{R}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{R}^{-1}$. It has been shown in Haslett [16] that Cook's distance for deleting the subset I is given by

$$\mathrm{CD}(I) = \frac{1}{\hat{\sigma}^2}\epsilon^T\mathbf{Q}U_I\mathbf{Q}_{II}^{-1}(\mathbf{R}^{II} - \mathbf{Q}_{II})\mathbf{Q}_{II}^{-1}U_I^T\mathbf{Q}\epsilon, \tag{2.19}$$

where $\mathbf{Q}_{II}$ is the (I,I) subset of $\mathbf{Q}$ and $\mathbf{R}^{II}$ is the (I,I) subset of $\mathbf{R}^{-1}$. After some algebraic calculations, it can be shown that

$$\mathrm{E}[\mathrm{CD}(I)|\mathcal{M}] \approx \mathrm{E}[\mathrm{tr}(\mathbf{Q}_{II}^{-1}\mathbf{R}^{II})|\mathcal{M}] - n(I) = \sum_{j=1}^{n(I)} \mathrm{E}[(1-\lambda_{I,j})^{-1}|\mathcal{M}] - n(I),$$

$$\mathrm{Var}[\mathrm{CD}(I)|\mathcal{M}] \approx 2\mathrm{E}(\mathrm{tr}\{[\mathbf{Q}_{II}^{-1}\mathbf{R}^{II} - \mathbf{I}_{n(I)}]^2\}|\mathcal{M}) + \mathrm{Var}[\mathrm{tr}(\mathbf{Q}_{II}^{-1}\mathbf{R}^{II})|\mathcal{M}_X],$$

where $\mathcal{M}_X$ represents the distribution of $\mathbf{X}$ and $\lambda_{I,1} \geq \ldots \geq \lambda_{I,n(I)}$ are the ordered eigenvalues of $(\mathbf{R}^{II})^{-1/2}\mathbf{H}_{II}(\mathbf{R}^{II})^{-1/2}$, in which $\mathbf{H}_{II}$ is the (I,I) subset of $\mathbf{H}$. Similar to Theorem 3 (b), when $n(I) \geq p$, $\mathrm{E}[\mathrm{CD}(I)|\mathcal{M}]$ reduces to $\sum_{j=1}^{p} E[(1-\lambda_{I,j})^{-1}|\mathcal{M}] - p$. In many scenarios such as the multivariate linear model, we can follow the strategies in Example 3 to approximate $\mathrm{E}[\mathrm{CD}(I)|\mathcal{M}]$ and $\mathrm{Var}[\mathrm{CD}(I)|\mathcal{M}]$. However, for time series data, since the elements in $\mathbf{X}$ are responses in an autoregressive model, such as the AR(1) model, we can use the parametric bootstrap to generate random samples from the fitted model and then approximate $\mathrm{E}[\mathrm{CD}(I)|\mathcal{M}]$ and $\mathrm{Var}[\mathrm{CD}(I)|\mathcal{M}]$.

2.5. *Conditionally scaled Cook's distances.* In certain research settings (e.g., regression), it may be better to perform influence analysis while fixing some covariates of interest, such as measurement time. For instance, in longitudinal data, if different subjects can have different numbers of measurements and measurement times, which are not covariates of interest in an influence analysis, it may be better to eliminate their effect in calculating Cook's distance. We are interested in answering a fourth question as follows.

(c.4) How do we compare Cook's distance relative to $\mathcal{P}(I|\mathcal{M})$ while fixing some covariates?

To eliminate the effect of some fixed covariates, we introduce two other scaled Cook's distances, called conditionally scaled Cook's distance, as follows.

Definition 2. The *conditionally scaled Cook's distances* (CSCD) for matching (mean, Std) and (median, Mstd) while controlling for $\mathbf{Z}$ are, respectively, defined as

(2.20)
$$\begin{aligned}\mathrm{CSCD}_1(I,\mathbf{Z}) &= \{\mathrm{CD}(I) - E[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]\}/\{\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]\}\\ \mathrm{CSCD}_2(I,\mathbf{Z}) &= \{\mathrm{CD}(I) - \mathrm{Q}_{CD(I)}(0.5|\mathcal{M},\mathbf{Z})\}/\{\mathrm{Mstd}[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]\},\end{aligned}$$

where $\mathbf{Z}$ is the set of some fixed covariates in $\mathbf{Y}$ and the expectation and quantiles are taken with respect to $\mathcal{M}$ given $\mathbf{Z}$.

According to Definition 2, these conditionally scaled Cook's distances can be used to evaluate the relative influential level of different subsets I given $\mathbf{Z}$. Similar to $\mathrm{SCD}_1(I)$ and $\mathrm{SCD}_2(I)$, a large value of $\mathrm{CSCD}_1(I,\mathbf{Z})$ (or $\mathrm{CSCD}_2(I,\mathbf{Z})$) indicates a large influence of the subset I after controlling for $\mathbf{Z}$. It should be noted that because $\mathbf{Z}$ is fixed, the $\mathrm{CSCD}_k(I,\mathbf{Z})$ do not reflect the influential level of $\mathbf{Z}$ and the $\mathrm{CSCD}_k(I,\mathbf{Z})$ may vary across different $\mathbf{Z}$. The conditionally scaled Cook's distances measure the difference of the observed influence level of the set I given $\mathbf{Z}$ to the expected influence level of a set with the same size when the current model $\mathcal{M}$ is true and $\mathbf{Z}$ is fixed.

The next problem is how to compute $E[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$, $\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$, $Q_{CD(I)}(0.5|\mathcal{M},\mathbf{Z})$, and $\mathrm{Mstd}[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$ for each subset I when $\mathcal{M}$ is the true data generator and $\mathbf{Z}$ is fixed. Similar to the computation of the scaled Cook's distances, we can essentially use almost the same approach to approximate the four quantities for $\mathrm{CSCD}_1(I,\mathbf{Z})$ and $\mathrm{CSCD}_2(I,\mathbf{Z})$. However, a slight difference occurs in the way that we simulate the data. Specifically, let $\mathbf{Y}_Z$ be the data $\mathbf{Y}$ with $\mathbf{Z}$ deleted. We need to simulate random samples $\mathbf{Y}_Z^s$ from $\hat{\mathcal{M}}_Z = \{p(\mathbf{Y}_Z|\mathbf{Z},\hat{\boldsymbol{\theta}})\}$ and then calculate $\mathrm{CD}(I)^{(s)} = F_1(I,\hat{\mathcal{M}}_Z,(\mathbf{Y}_Z^s,\mathbf{Z}))$ for each subset I.

As an illustration, we consider how to calculate the conditionally scaled Cook's distances in generalized linear models as follows.

Example 2 (continued). For generalized linear models, we fix all covariates, that is $\mathbf{Z} = \mathbf{X}$, and then calculate the CSCDs as follows. First, we can show that

$$E[\widetilde{\mathrm{CD}}(I)|\mathcal{M}, \mathbf{Z}] \approx \mathrm{tr}[(\mathbf{I}_{n(I)} - H_{*,I})^{-1}] - n(I),$$
$$\mathrm{Var}[\widetilde{\mathrm{CD}}(I)|\mathcal{M}, \mathbf{Z}] \approx a(\tau_*) \sum_{i=1}^{n} w_{ii*} b^{(4)}(\eta_{i*}) + \mathrm{tr}[(\mathbf{I}_{n(I)} - H_{*I})^{-1} H_{*I} (\mathbf{I}_{n(I)} - H_{*I})^{-1} H_{*I}].$$

Then, similar to the derivations of Theorem 3 (a) and (b), we can show that the conditionally scaled Cook's distance $\mathrm{CSCD}_1(I, \mathbf{X})$ can be approximated by

$$\frac{\hat{\mathbf{e}}^T \hat{V}^{-1/2} U_I (\mathbf{I}_{n(I)} - \hat{H}_I)^{-1} \hat{H}_I (\mathbf{I}_{n(I)} - \hat{H}_I)^{-1} U_I^T \hat{V}^{-1/2} \hat{\mathbf{e}} - [\sum_{j=1}^{n(I)} (1 - \lambda_{I,j})^{-1} - n(I)]}{\{a(\tau_*) \sum_{i=1}^{n} w_{ii*} b^{(4)}(\eta_{i*}) + \mathrm{tr}[(\mathbf{I}_{n(I)} - H_{*I})^{-1} H_{*I} (\mathbf{I}_{n(I)} - H_{*I})^{-1} H_{*I}]\}^{1/2}}.$$

To approximate $\mathrm{CSCD}_2(I, \mathbf{X})$, we can generate responses from the model fitted to the data and then substitute them into Theorem 3 (a) to obtain a sample of simulated $\mathrm{CD}(I)$'s given the covariates. Finally, we can use the empirical median and median standard deviation of the simulated $\mathrm{CD}(I)$'s to approximate $\mathrm{CSCD}_2(I, \mathbf{Z})$.

2.6. *First-order approximations.* We have focused on developing the scaled Cook's distances and their approximations for several parametric models, such as generalized linear models. More generally, we are interested in answering a fifth question as follows.

(c.5) How do we approximate the scaled Cook's distances for a large class of parametric models for both independent and dependent data?

We obtain the following theorem.

Theorem 4. *If Assumptions A2-A5 in the Appendix hold and $n(I)/n \to \gamma \in [0, 1)$, where $n(I)$ denotes the size of I, then we have the following results: (a) Let $\mathbf{F}_n(\boldsymbol{\theta}) = -\partial_\theta^2 \log p(\mathbf{Y}|\boldsymbol{\theta})$, $\mathbf{f}_I(\boldsymbol{\theta}) = \partial_\theta \log p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \hat{\boldsymbol{\theta}})$, and $\mathbf{s}_I(\boldsymbol{\theta}) = -\partial_\theta^2 \log p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \boldsymbol{\theta})$, $CD(I)$ can be approximated by*

$$\widetilde{CD}(I) = \mathbf{f}_I(\hat{\boldsymbol{\theta}})^T [\mathbf{F}_n(\hat{\boldsymbol{\theta}}) - \mathbf{s}_I(\hat{\boldsymbol{\theta}})]^{-1} \mathbf{F}_n(\hat{\boldsymbol{\theta}}) [\mathbf{F}_n(\hat{\boldsymbol{\theta}}) - \mathbf{s}_I(\hat{\boldsymbol{\theta}})]^{-1} \mathbf{f}_I(\hat{\boldsymbol{\theta}}); \tag{2.21}$$

(b) $E[\widetilde{CD}(I)|\mathcal{M}] \approx tr(\{E[\mathbf{F}_n(\hat{\boldsymbol{\theta}})|\mathcal{M}] - E[\mathbf{s}_I(\hat{\boldsymbol{\theta}})|\mathcal{M}]\}^{-1} E[\mathbf{s}_I(\hat{\boldsymbol{\theta}})|\mathcal{M}])$;
(c) $E[\widetilde{CD}(I)|\mathcal{M}, \mathbf{Z}] \approx tr(\{E[\mathbf{F}_n(\hat{\boldsymbol{\theta}})|\mathcal{M}, \mathbf{Z}] - E[\mathbf{s}_I(\hat{\boldsymbol{\theta}})|\mathcal{M}, \mathbf{Z}]\}^{-1} E[\mathbf{s}_I(\hat{\boldsymbol{\theta}})|\mathcal{M}, \mathbf{Z}])$.

Proof of Theorem 4. (a) It follows from a Taylor's series expansion and assumption A3 that

$$\partial_\theta \log p(\mathbf{Y}_{[I]}|\hat{\boldsymbol{\theta}}_{[I]}) = \mathbf{0} = \partial_\theta \log p(\mathbf{Y}_{[I]}|\hat{\boldsymbol{\theta}}) + \partial_\theta^2 \log p(\mathbf{Y}_{[I]}|\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{[I]} - \hat{\boldsymbol{\theta}}),$$

where $\tilde{\boldsymbol{\theta}} = t\hat{\boldsymbol{\theta}}_{[I]} + (1-t)\hat{\boldsymbol{\theta}}$ for $t \in [0,1]$. Combining this with Assumption A4 and the fact that $\partial_\theta \log p(\mathbf{Y}|\hat{\boldsymbol{\theta}}) = \partial_\theta \log p(\mathbf{Y}_{[I]}|\hat{\boldsymbol{\theta}}) + \partial_\theta \log p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \hat{\boldsymbol{\theta}}) = \mathbf{0}$, we get

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{[I]} - \hat{\boldsymbol{\theta}} &= [-\partial_\theta^2 \log p(\mathbf{Y}_{[I]}|\hat{\boldsymbol{\theta}})]^{-1} \partial_\theta \log p(\mathbf{Y}_{[I]}|\hat{\boldsymbol{\theta}})[1 + o_p(1)] \\ &= -[-\partial_\theta^2 \log p(\mathbf{Y}_{[I]}|\hat{\boldsymbol{\theta}})]^{-1} \partial_\theta \log p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \hat{\boldsymbol{\theta}})[1 + o_p(1)]. \end{aligned} \tag{2.22}$$

Substituting (2.22) into $\mathrm{CD}(I) = (\hat{\boldsymbol{\theta}}_{[I]} - \hat{\boldsymbol{\theta}})^T \mathbf{F}_n(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{[I]} - \hat{\boldsymbol{\theta}})$ completes the proof of Theorem 4 (a).

(b) It follows from Assumptions A2-A4 that

$$\begin{aligned} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_* &= \mathbf{F}_n(\boldsymbol{\theta}_*)^{-1} \partial_\theta \log p(\mathbf{Y}|\boldsymbol{\theta}_*)[1 + o_p(1)] \\ &= \mathbf{F}_n(\boldsymbol{\theta}_*)^{-1} [\partial_\theta \log p(\mathbf{Y}_{[I]}|\boldsymbol{\theta}_*) + \partial_\theta \log p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \boldsymbol{\theta}_*)][1 + o_p(1)]. \end{aligned}$$

Let $J_I(\boldsymbol{\theta}) = \partial_\theta \log p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \boldsymbol{\theta})$. Using a Taylor's series expansion along with Assumptions A4 and A5, we get

$$\begin{aligned} J_I(\hat{\boldsymbol{\theta}}) &= J_I(\boldsymbol{\theta}_*) - \mathbf{s}_I(\boldsymbol{\theta}_*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*)[1 + o_p(1)] \\ &= J_I(\boldsymbol{\theta}_*) - E[\mathbf{s}_I(\boldsymbol{\theta}_*)|\mathcal{M}](\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*)[1 + o_p(1)] \\ &= (\{\mathbf{I}_p - E[\mathbf{s}_I(\boldsymbol{\theta})|\mathcal{M}]\mathbf{F}_n(\boldsymbol{\theta}_*)^{-1}\} J_I(\boldsymbol{\theta}_*) - \\ &\qquad E[\mathbf{s}_I(\boldsymbol{\theta})|\mathcal{M}]\mathbf{F}_n(\boldsymbol{\theta}_*)^{-1} \partial_\theta \log p(\mathbf{Y}_{[I]}|\boldsymbol{\theta}_*))[1 + o_p(1)]. \end{aligned} \tag{2.23}$$

Since $E[J_I(\boldsymbol{\theta}_*)\partial_\theta \log p(\mathbf{Y}_{[I]}|\boldsymbol{\theta}_*)|\mathcal{M}] = \mathbf{0}$,

$$E[J_I(\hat{\boldsymbol{\theta}}) J_I(\hat{\boldsymbol{\theta}})^T|\mathcal{M}] = E[\mathbf{s}_I(\boldsymbol{\theta}_*)|\mathcal{M}]\mathbf{F}_n(\boldsymbol{\theta}_*)^{-1}\{\mathbf{F}_n(\boldsymbol{\theta}_*) - E[\mathbf{s}_I(\boldsymbol{\theta}_*)|\mathcal{M}]\}[1 + o_p(1)].$$

It follows from Assumption A4 that for $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_*$, $\mathbf{F}_n(\boldsymbol{\theta})$ and $\mathbf{F}_n(\boldsymbol{\theta}_*) - \mathbf{f}_I(\boldsymbol{\theta})$ can be replaced by $E[\mathbf{F}_n(\boldsymbol{\theta})|\mathcal{M}]$ and $E[\mathbf{F}_n(\boldsymbol{\theta}_*) - \mathbf{f}_I(\boldsymbol{\theta})|\mathcal{M}]$, respectively, which completes the proof of Theorem 4 (b).

(c) Similar to Theorem 4 (b), we can prove Theorem 4 (c).

Theorem 4 (a) establishes the first order approximation of Cook's distance for a large class of parametric models for both dependent and independent data. This leads to a substantial savings in computational time, since it is computationally easier to calculate $\mathbf{f}_I(\hat{\boldsymbol{\theta}})$, $\mathbf{F}_n(\hat{\boldsymbol{\theta}})$, and $\mathbf{s}_I(\hat{\boldsymbol{\theta}})$ compared to $\mathrm{CD}(I)$. Theorem 4 (b) and (c) give an approximation of $\mathrm{E}[\mathrm{CD}(I)|\mathcal{M}]$ and $\mathrm{E}[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}]$, respectively. Generally, it is difficult to give a simple approximation to $\mathrm{Var}[\mathrm{CD}(I)|\mathcal{M}]$ and $\mathrm{Var}[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}]$, since it involves the fourth moment of $\mathbf{f}_I(\hat{\boldsymbol{\theta}})$, which does not have a simple form.

Based on Theorem 4, we can approximate the scaled Cook's distance measures as follows.

Step 1. We generate a random sample $\mathbf{Y}^s$ from $p(\mathbf{Y}|\mathbf{Z}, \hat{\boldsymbol{\theta}})$ and calculate $\widetilde{\mathrm{CD}}(I)$ based on the simulated sample $\mathbf{Y}^s$ and fixed $\mathbf{Z}$, denoted by $\widetilde{\mathrm{CD}}(I)^s$. Explicitly, to calculate $\widetilde{\mathrm{CD}}(I)^s$, we replace $\mathbf{Y}$ in $\mathbf{f}_I(\hat{\boldsymbol{\theta}})$, $\mathbf{F}_n(\hat{\boldsymbol{\theta}})$, and $\mathbf{s}_I(\hat{\boldsymbol{\theta}})$ by $\mathbf{Y}^s$. The computational burden involved in computing $\widetilde{\mathrm{CD}}(I)^s$ is very minor.

Compared to the exact computation of the scaled Cook's distances, we have avoided computing the maximum likelihood estimate of $\boldsymbol{\theta}$ based on $\mathbf{Y}^s$, which leads to great computational savings in computing $\widetilde{\mathrm{CD}}(I)^s$ even for large S. Theoretically, since $\hat{\boldsymbol{\theta}}$ is a consistent estimate of $\boldsymbol{\theta}_*$, $E[\widetilde{\mathrm{CD}}(I)|\mathcal{M}]$ is a consistent estimate of $E[\mathrm{CD}(I)|\mathcal{M}]$. Compared with reestimating $\hat{\boldsymbol{\theta}}^s$ for each $\mathbf{Y}^s$, a drawback of using $\hat{\boldsymbol{\theta}}$ in calculating $\widetilde{\mathrm{CD}}(I)^s$ is that $\widetilde{\mathrm{CD}}(I)^s$ does not account for the variability in $\hat{\boldsymbol{\theta}}$. Similar arguments hold for the other three quantities of $\mathrm{CD}(I)$.

Step 2. By repeating Step 1 S times, we can use the empirical quantities of $\{\widetilde{\mathrm{CD}}(I)^s : s = 1, \ldots, S\}$ to approximate $E[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}]$, $\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}]$, $Q_{CD(I)}(0.5|\mathcal{M}, \mathbf{Z})$, and $\mathrm{Mstd}[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}]$. Subsequently, we can approximate $\mathrm{CSCD}_1(I, \mathbf{Z})$ and $\mathrm{CSCD}_2(I, \mathbf{Z})$ and determine their magnitude based on $\widetilde{\mathrm{CD}}(I)^s$.

For instance, let $\widehat{M}[\widetilde{\mathrm{CD}}(I)]$ and $\widehat{\mathrm{Std}}[\widetilde{\mathrm{CD}}(I)]$ be, respectively, the sample mean and standard deviation of $\{\widetilde{\mathrm{CD}}(I)^s : s = 1, \ldots, S\}$. We calculate

$$\widetilde{\mathrm{CSCD}}_1(I, \mathbf{Z}) = \frac{\{\widetilde{\mathrm{CD}}(I) - \widehat{M}[\widetilde{\mathrm{CD}}(I)]\}}{\widehat{\mathrm{Std}}[\widetilde{\mathrm{CD}}(I)]} \text{ and } \widetilde{\mathrm{CSCD}}_1(I, \mathbf{Z})^s = \frac{\{\widetilde{\mathrm{CD}}(I)^s - \widehat{M}[\widetilde{\mathrm{CD}}(I)]\}}{\widehat{\mathrm{Std}}[\widetilde{\mathrm{CD}}(I)]}.$$

We use $\widetilde{\mathrm{CSCD}}_1(I, \mathbf{Z})$ to approximate $\mathrm{CSCD}_1(I, \mathbf{Z})$ and then compare $\widetilde{\mathrm{CSCD}}_1(I, \mathbf{Z})$ across different I in order to determine whether a specific subset I is relatively influential or not. Moreover, since $\widetilde{\mathrm{CSCD}}_1(\tilde{I}, \mathbf{Z})^s$ can be regarded as the 'true' scaled Cook distance when $p(\mathbf{Y}|\mathbf{Z}, \hat{\boldsymbol{\theta}})$ is true, we can either compare $\widetilde{\mathrm{CSCD}}_1(I, \mathbf{Z})$ with $\widetilde{\mathrm{CSCD}}_1(\tilde{I}, \mathbf{Z})^s$ for all subsets $\tilde{I}$ and s or compare $\widetilde{\mathrm{CSCD}}_1(I, \mathbf{Z})$ with $\widetilde{\mathrm{CSCD}}_1(I, \mathbf{Z})^s$ for all s. Specifically, we calculate two probabilities as follows:

$$P_A(I, \mathbf{Z}) = \sum_{s=1}^{S} \mathbf{1}(\widetilde{\mathrm{CSCD}}_1(I, \mathbf{Z})^s \leq \widetilde{\mathrm{CSCD}}_1(I, \mathbf{Z}))/S, \tag{2.24}$$

$$P_B(I, \mathbf{Z}) = \sum_{\tilde{I}} \sum_{s=1}^{S} \frac{\mathbf{1}(\widetilde{\mathrm{CSCD}}_1(\tilde{I}, \mathbf{Z})^s \leq \widetilde{\mathrm{CSCD}}_1(I, \mathbf{Z}))}{S \times \#(\tilde{I})}, \tag{2.25}$$

where $\#(\tilde{I})$ is the total number of all possible sets and $\mathbf{1}(\cdot)$ is an indicator function of a set. We regard a subset I as influential if the value of $P_A(I, \mathbf{Z})$ (or $P_B(I, \mathbf{Z})$) is relatively large. Similarly, we can use the same strategy to quantify the size of $\mathrm{CSCD}_2(I, \mathbf{Z})$, $\mathrm{SCD}_1(I)$, and $\mathrm{SCD}_2(I)$.

Another issue is the accuracy of the first order approximation $\widetilde{\mathrm{CD}}(I)$ to the exact $\mathrm{CD}(I)$. For relatively influential subsets, even though the accuracy of the first-order approximation may be relatively low, $\widetilde{\mathrm{CD}}(I)$ can easily pick out these influential points. Thus, for diagnostic purposes, the first-order approximation may be more effective at identifying influential subsets compared to the true Cook's distance. We conduct simulation studies to investigate the performance of the first-order approximation $\widetilde{\mathrm{CD}}(I)$ relative to the exact $\mathrm{CD}(I)$. Numerical comparisons are given in Section 3.

We consider cluster deletion in generalized linear mixed models (GLMM).

Example 4. Consider a dataset that is composed of a response y_{ij}, covariate vectors $\mathbf{x}_{ij}(p \times 1)$ and $\mathbf{c}_{ij}(p_1 \times 1)$, for observations $j = 1, \ldots, m_i$ within clusters $i = 1, \ldots, n$. The GLMM assumes that conditional on a $p_1 \times 1$ random variable $\mathbf{b}_i$, y_{ij} follows an exponential family distribution of the form [19]

$$p(y_{ij}|\mathbf{b}_i) = \exp\{a(\tau)^{-1}[y_{ij}\eta_{ij} - b(\eta_{ij})] + c(y_{ij}, \tau)\}, \tag{2.26}$$

where $\eta_{ij} = k(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{c}_{ij}^T\mathbf{b}_i)$ in which $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)^T$ and $k(\cdot)$ is a known continuously differentiable function. The distribution of $\mathbf{b}_i$ is assumed to be $N(\mathbf{0}, \Sigma)$, where $\Sigma = \Sigma(\boldsymbol{\gamma})$ depends on a $p_2 \times 1$ vector $\boldsymbol{\gamma}$ of unknown variance components. In this case, we fix all covariates $\mathbf{x}_{ij}$ and $\mathbf{c}_{ij}$ and all m_i and include them in $\mathbf{Z}$. For simplicity, we fix $(\boldsymbol{\gamma}, \tau)$ at an appropriate estimate $(\hat{\boldsymbol{\gamma}}, \hat{\tau})$ throughout the example.

We focus here on cluster deletion in GLMMs. After some calculations, the first order approximation of Cook's distance for deleting the i-th cluster is given by

$$\widetilde{\mathrm{CD}}(I_i) = \partial_\beta \ell_i(\hat{\boldsymbol{\beta}})^T[\mathbf{F}_n(\hat{\boldsymbol{\beta}}) - \mathbf{f}_i(\hat{\boldsymbol{\beta}})]^{-1}\mathbf{F}_n(\hat{\boldsymbol{\beta}})[\mathbf{F}_n(\hat{\boldsymbol{\beta}}) - \mathbf{f}_i(\hat{\boldsymbol{\beta}})]^{-1}\partial_\beta \ell_i(\hat{\boldsymbol{\beta}}), \tag{2.27}$$

where $I_i = \{(i,1), \ldots, (i, m_i)\}$, $\ell_i(\boldsymbol{\beta})$ is the log-likelihood function for the $i-$th cluster, $\mathbf{f}_i(\boldsymbol{\beta}) = -\partial_\beta^2 \ell_i(\boldsymbol{\beta})$ and $\mathbf{F}_n(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{f}_i(\boldsymbol{\beta})$. Note that

$$\partial_\beta \ell_i(\hat{\boldsymbol{\beta}}) \approx \{\mathbf{I}_p - \mathbf{f}_i(\hat{\boldsymbol{\beta}})[\mathbf{F}_n(\boldsymbol{\beta}_*)]^{-1}\}\partial_\beta \ell_i(\boldsymbol{\beta}_*) + \mathbf{f}_i(\hat{\boldsymbol{\beta}})[\mathbf{F}_n(\boldsymbol{\beta}_*)]^{-1}\sum_{j \neq i} \partial_\beta \ell_j(\boldsymbol{\beta}_*).$$

Then, conditional on all the covariates and $\{m_1, \ldots, m_n\}$ in $\mathbf{Z}$, we can show that $\mathrm{E}[\widetilde{\mathrm{CD}}(I_i)|\mathcal{M}, \mathbf{Z}]$ can be approximated by $\mathrm{tr}\{(\mathrm{E}[\mathbf{F}_n(\hat{\boldsymbol{\beta}})|\mathcal{M}, \mathbf{Z}] - \mathrm{E}[\mathbf{f}_i(\hat{\boldsymbol{\beta}})|\mathcal{M}, \mathbf{Z}])^{-1}\mathrm{E}[\mathbf{f}_i(\hat{\boldsymbol{\beta}})|\mathcal{M}, \mathbf{Z}]\}$ when the fitted model $\mathcal{M}$ is true. Moreover, we may approximate $\mathrm{Var}[\widetilde{\mathrm{CD}}(I_i)|\mathcal{M}, \mathbf{Z}]$ by using the fourth moment of $\partial_\beta \ell_i(\boldsymbol{\beta}_*)$. It is not straightforward to approximate $Q_{CD(I_i)}(0.5|\mathcal{M}, \mathbf{Z})$ and $\mathrm{Mstd}[\mathrm{CD}(I_i)|\mathcal{M}, \mathbf{Z}]$. Computationally, we employ the parametric bootstrap method described above to approximate the conditionally scaled Cook's distances $\mathrm{CSCD}_1(I_i, \mathbf{Z})$ and $\mathrm{CSCD}_2(I_i, \mathbf{Z})$.

3. Simulation Studies and A Real Data Example. In this section, we illustrate our methodology with simulated data and a real data example.

3.1. *Simulated Studies.* The goals of our simulations were to evaluate the accuracy of the first-order approximations to Cook's distance and its associated quantities (e.g., mean) and to examine the finite sample performance of Cook's distance and the scaled Cook's distances for detecting influential clusters in longitudinal data. We generated 100 datasets from a linear mixed model. Specifically, each dataset contains n clusters. For each cluster, the random effect b_i was first independently generated from a $N(0, \sigma_b^2)$ distribution and then, given b_i, the observations y_{ij} $(j = 1, \cdots, m_i; i = 1, \ldots, n)$ were independently generated from a normal random generator such that $y_{ij} \sim N(\mathbf{x}_{ij}^T\boldsymbol{\beta} + b_i, \sigma_y^2)$ and the m_i were randomly drawn from $\{1, \ldots, 10\}$. The covariates $\mathbf{x}_{ij}$ were set as $(1, u_i, t_{ij})^T$, among which t_{ij} represents time and u_i denotes a baseline covariate. Moreover, $t_{ij} = \log(j)$ and the u_i's were independently generated from a $N(0, 1)$ distribution. For all 100 datasets, both the responses and covariates were repeatedly generated, while the true value of $(\boldsymbol{\beta}^T, \sigma_b, \sigma_y)$ was fixed at $(1, 1, 1, 1, 1)$. The sample size n was set at 30 to represent a relatively small sample size. We also explored other sample sizes and different degrees of correlation and obtained similar findings, and thus we did not report them here for the sake of space.

We carried out three experiments as follows. We treated (σ_b, σ_y) as nuisance parameters and $\boldsymbol{\beta}$ as the parameter vector of interest. The first experiment was to evaluate the accuracy of $\widetilde{\mathrm{CD}}(I)$ to $\mathrm{CD}(I)$. We considered two scenarios. In the first scenario, we directly used the simulated 100 datasets as the above linear mixed model. In the second scenario, for each simulated dataset, we deleted all the observations in clusters $n-1$ and n and then reset $(m_{n-1}, b_{n-1}) = (1, 4)$ and $(m_n, b_n) = (10, 3)$ to generate $y_{i,j}$ for $i = n-1, n$ and all j according to the above random effects model. Thus, the new $(n-1)$th and nth clusters can be regarded as influential clusters due to the extreme values of b_{n-1} and b_n. Moreover, the number of observations in these two clusters is extremely unbalanced.

For each dataset, we deleted each cluster one at a time and then calculated $\mathrm{CD}(I)$ and its first order approximation $\widetilde{\mathrm{CD}}(I)$ for each cluster. Moreover, we computed the average $\mathrm{CD}(I)$, and the biases and standard errors of the differences $\mathrm{CD}(I) - \widetilde{\mathrm{CD}}(I)$ for each I. Table 1 shows some selected results for each scenario. The average $\mathrm{CD}(I)$, is positively proportional to the cluster size $n(I)$. For the true 'good' clusters, the first-order approximation is very accurate and leads to small average biases and standard errors. Even for the influential clusters, $\widetilde{\mathrm{CD}}(I)$ is relatively close to $\mathrm{CD}(I)$.

In the second experiment, we considered the same two scenarios as the first experiment in order to examine the finite sample performance of $E[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$ and $\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$ and their first-order approximations. Specifically, for each dataset, we set $S = 100$ and simulated $S = 100$ random samples from the fitted linear mixed model. Then, we approximated $E[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$ and $\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$ by using their empirical ones, and calculated their first approximations $\widehat{M}[\widetilde{\mathrm{CD}}(I)]$ and $\widehat{\mathrm{Std}}[\widetilde{\mathrm{CD}}(I)]$.

Across all 100 data sets, for each cluster I, we computed the averages of $E[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$ and $\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$, and the biases and standard errors of the differences $E[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}] - \widehat{M}[\widetilde{\mathrm{CD}}(I)]$ and $\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}] - \widehat{\mathrm{Std}}[\widetilde{\mathrm{CD}}(I)]$. Table 1 shows some selected results for each scenario. The averages of $E[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$ and $\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$ are positively proportional to the cluster size $n(I)$. For the true 'good' clusters, the first-order approximations of $E[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$ and $\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$ are very accurate and lead to small average biases and standard errors, while for the influential clusters, their first-order approximations are relatively accurate.

The third experiment was to examine the finite sample performance of Cook's distance and the scaled Cook's distances for detecting influential cluster in longitudinal data. We considered two scenarios. In the first scenario, for each of the 100 simulated datasets, we deleted all the observations in cluster n and then reset $m_n = 1$ and varied b_n from 0.4 to 8.0 to generate $y_{n,1}$ according to the above random effects model. The second scenario is almost the same as the first scenario except that we reset $m_n = 10$.

For each dataset, we deleted each cluster one at a time and calculated $\mathrm{CD}(I)$. Then, we computed $P_C(I,\mathbf{Z}) = \sum_{I\neq\{n\}} \mathbf{1}(\mathrm{CD}(I) \leq \mathrm{CD}(\{n\}))/(n-1)$, which characterizes the probability that $\mathrm{CD}(\{n\})$ is greater than all the other $\mathrm{CD}(I)$. We set $S = 100$ and then we approximated $\mathrm{CSCD}_1(I,\mathbf{Z})$, $\mathrm{CSCD}_2(I,\mathbf{Z})$, $\widetilde{\mathrm{CSCD}}_1(I,\mathbf{Z})$, and $\widetilde{\mathrm{CSCD}}_2(I,\mathbf{Z})$. Subsequently, we calculated $P_A(I,\mathbf{Z})$ and $P_B(I,\mathbf{Z})$ in (2.24) based on $\widetilde{\mathrm{CSCD}}_1(I,\mathbf{Z})$ and $\widetilde{\mathrm{CSCD}}_2(I,\mathbf{Z})$.

Finally, across all 100 datasets, we calculated the averages and standard errors of all diagnostic measures for the nth cluster for each scenario. Figures 1 and 2 present some selected results. Comparing the two scenarios, we observed that deleting the n-th cluster with 10 observations causes larger effect than that with 1 observation (Fig 1 (a) and Fig 2 (a)). For the first scenario, $\mathrm{CD}(\{n\})$ is relatively smaller than the other $\mathrm{CD}(I)$ (Fig. 1 (d)), whereas for the second scenario, $\mathrm{CD}(\{n\})$ is relatively larger than other $\mathrm{CD}(I)$ (Fig. 2 (d)). This confirms the size issue discussed in Section 2.3. Furthermore, in the two scenarios, $P_A(\{n\},\mathbf{Z})$ and $P_B(\{n\},\mathbf{Z})$ for the scaled Cook's distances increase with b_n as expected, while they are quite close to each other across all values of b_n (Fig. 1 (d) and Fig. 2 (d)). It may indicate

TABLE 1

Selected results from simulation studies for $n = 30$ and the two scenarios: $n(I)$, M, SD, Mdif ($\times 10^{-2}$), and SDif ($\times 10^{-1}$) of the three quantities $CD(I)$, $E[CD(I)|\mathcal{M}, \mathbf{Z}]$, and $Std[CD(I)|\mathcal{M}, \mathbf{Z}]$. $n(I)$ denotes the cluster size of subset I; M denotes the mean; SD denotes the standard deviation; Mdif and SDdif, respectively, denote the mean and standard deviation of the differences between each quantity and its first-order approximation. In the first scenario, all observations were generated from the linear mixed model, while in the second scenario, clusters 29 and 30 were influential clusters. For each case, 100 simulated datasets were used.

	CD(I)									
	Scenario I					Scenario II				
I	$n(I)$	M	SD	Mdif	SDdif	$n(I)$	M	SD	Mdif	SDdif
1	4	0.133	0.237	0.345	0.186	4	0.087	0.142	0.055	0.054
5	9	0.162	0.163	0.001	0.125	9	0.140	0.139	0.019	0.074
10	8	0.159	0.220	0.124	0.107	8	0.138	0.186	-0.0003	0.106
15	1	0.036	0.048	0.022	0.010	1	0.033	0.041	0.018	0.010
20	8	0.156	0.213	0.271	0.019	8	0.120	0.130	0.085	0.069
25	9	0.164	0.166	-0.027	0.102	9	0.143	0.149	-0.111	0.084
29	1	0.041	0.081	0.020	0.010	1	0.343	0.309	0.555	0.181
30	10	0.159	0.203	0.151	0.082	10	0.508	0.505	3.245	0.571
	E[CD(I)$\|\mathcal{M}, \mathbf{Z}$]									
	Scenario I					Scenario II				
I	$n(I)$	M	SD	Mdif	SDdif	$n(I)$	M	SD	Mdif	SDdif
1	4	0.083	0.057	0.016	0.010	4	0.070	0.048	0.030	0.008
5	9	0.165	0.066	0.211	0.031	9	0.159	0.068	0.170	0.022
10	8	0.137	0.056	0.106	0.018	8	0.140	0.078	0.113	0.019
15	1	0.050	0.059	-0.144	0.030	1	0.055	0.051	-0.116	0.026
20	8	0.141	0.056	0.118	0.022	8	0.130	0.062	0.089	0.015
25	9	0.174	0.086	0.194	0.027	9	0.177	0.081	0.170	0.025
29	3	0.067	0.055	0.003	0.010	1	0.056	0.045	-0.129	0.048
30	7	0.119	0.055	0.117	0.016	10	0.197	0.065	0.192	0.028
	Std[CD(I)$\|\mathcal{M}, \mathbf{Z}$]									
	Scenario I					Scenario II				
I	$n(I)$	M	SD	Mdif	SDdif	$n(I)$	M	SD	Mdif	SDdif
1	4	0.107	0.084	0.114	0.036	4	0.088	0.063	0.096	0.034
5	9	0.174	0.076	0.218	0.068	9	0.163	0.072	0.017	0.063
10	8	0.142	0.066	0.036	0.052	8	0.149	0.099	0.114	0.059
15	1	0.075	0.103	0.147	0.063	1	0.080	0.075	0.211	0.061
20	8	0.145	0.069	0.076	0.073	8	0.135	0.081	0.010	0.047
25	9	0.177	0.099	0.046	0.069	9	0.185	0.097	0.039	0.060
29	3	0.090	0.085	0.174	0.077	1	0.082	0.065	0.251	0.089
30	7	0.128	0.070	0.132	0.062	10	0.205	0.068	0.077	0.063

that all scaled Cook's distances are consistent with each other.

3.2. *Yale Infant Growth Data.* The Yale infant growth data were collected to study whether cocaine exposure during pregnancy may lead to the maltreatment of infants after birth, such as physical and sexual abuse. A total of 298 children were recruited from two subject groups (cocaine exposed group and unexposed group). The key feature of this dataset is that different children had different numbers and patterns of visits during the study period [23, 22]. The total number of data points is $\sum_{i=1}^{n} m_i = 3176$, whereas m_i varies from 2 to 30.

Following Zhang [27] and Zhu et al. [30], we consider a linear mixed model with a compound symmetry covariance structure as follows: $y_{i,j} = \mathbf{x}_{i,j}^T\boldsymbol{\beta} + \epsilon_{i,j}$, where $y_{i,j}$ is the weight (in kilograms) of the j-th visit from the i-th subject, $\mathbf{x}_{i,j} = (1, d, (d-120)^+, (d-200)^+, (g_a-28)^+, d(g_a-28)^+, (d-60)^+(g_a-28)^+, (d-490)^+(g_a-28)^+, sd, s(d-120)^+)^T$, in which d and g_a (days) are the age of visit and gestational age, respectively, and s is the indicator for gender, with one for a girl and zero for a boy. In addition, we assume $\epsilon_i \sim N_{m_i}(\mathbf{0}, \sigma^2 R_i)$ and consider a compound symmetry covariance structure for R_i.

By using PROC MIXED (SAS 9.1, Cary, NC), we calculated the restricted maximum likelihood estimates. We treated $\boldsymbol{\beta}$ as parameters of interest and all other parameters as nuisance parameters. We calculated $\mathrm{CD}(I)$ for each child, which relates more to the detection of influential clusters [5]. We computed the degree of the perturbation for deleting each subject and then we calculated the scaled Cook's distances and associated quantities. We then used 100 bootstrap samples to approximate $\mathrm{CSCD}_1(I, \mathbf{Z})$, $\mathrm{CSCD}_2(I, \mathbf{Z})$, $\widetilde{\mathrm{CSCD}}_1(I, \mathbf{Z})$, and $\widetilde{\mathrm{CSCD}}_2(I, \mathbf{Z})$. Subsequently, we calculated $P_A(I, \mathbf{Z})$ and $P_B(I, \mathbf{Z})$ in (2.24) based on $\widetilde{\mathrm{CSCD}}_1(I, \mathbf{Z})$ and $\widetilde{\mathrm{CSCD}}_2(I, \mathbf{Z})$.

We obtained a strong Pearson correlation of 0.363 between Cook's distance and the cluster size. This indicates that the bigger the cluster size, the larger the Cook's distance measure. Figure 3 (a) presents nine influential subjects 269, 217, 294, 289, 274, 90, 38, 285, and 280, whose $(\mathrm{CD}(i), m_i)$ are, respectively, given by (2.416, 21), (1.465, 19), (1.252, 13), (1.188, 18), (1.163, 22), (0.858, 17), (0.823, 24), (0.738, 8), and (0.695, 9) (Table 2 and Figure 3). There are several difficulties in using Cook's distance for this model [20, 8, 5, 4]. First, cluster sizes vary significantly across all clusters and deleting a larger cluster may have a higher probability of having a larger influence as discussed in Section 2.3. For instance, comparing subjects 274 and 285, we observe $(m_{285}, \mathrm{CD}(\{285\})) = (8, 0.738)$ and $(m_{274}, \mathrm{CD}(\{274\})) = (22, 1.163)$. A larger influence measure $\mathrm{CD}(\{274\})$ can

be caused by a larger perturbation $m_{274} = 22$ and/or a larger discrepancy between the deleted observations in subject 274 and the model fitted to the data. Since m_{274} is much larger than m_{285}, it is difficult to claim that subject 274 is more influential than subject 285. Secondly, there is no rule for determining whether a specific subject is influential relative to $\hat{\mathcal{M}}$. Although we have selected the first nine subjects as influential, it is unclear whether they are truly influential or not.

We computed the degree of the perturbation for deleting each cluster as follows. Since $\mathcal{M}$ assumes $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,m_i})^T \sim N(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2 R_i)$, where $\mathbf{x}_i$ is an $m_i \times 10$ matrix with the $j-$th row being $\mathbf{x}_{i,j}^T$. After some algebraic calculations, it can be shown that for the case of fixed covariates, we have

$$\mathcal{P}(\{i\}|\mathcal{M}) = 0.5\text{tr}\{\mathbf{x}_i^T\sigma^{-2}R_i^{-1}\mathbf{x}_i E_\beta[(\boldsymbol{\beta}-\boldsymbol{\beta}_*)(\boldsymbol{\beta}-\boldsymbol{\beta}_*)^T]\}, \tag{3.1}$$

where E_β is taken with respect to $p(\boldsymbol{\beta}|\boldsymbol{\beta}_*, G_{n\beta}^{-1})$. Let $G_{n\beta}^{-1} = [\sum_{i=1}^n \mathbf{x}_i^T\sigma^{-2}R_i^{-1}\mathbf{x}_i]^{-1}$ be the covariance matrix of $\hat{\boldsymbol{\beta}}$. We observed a strong positive correlation between $\mathcal{P}(\{i\}|\mathcal{M})$ and $\text{CD}(I)$ (Fig. 3 (b)). Particularly, Figure 3 (b) shows that subject 269 has the largest Cook's distance and the largest degree of perturbation. We also observed a strong positive correlation between $\mathcal{P}(\{i\}|\mathcal{M})$ and the cluster size (Fig. 3 (c)). That is, the bigger cluster size usually corresponds to the larger degree of perturbation. Figure 3 (c) presents nine subjects $269, 223, 58, 270, 165$, 103, 288, and 279 with large degree of perturbation, which may be caused by both large leverage value and moderate cluster size. Finally, we observed the positive correlation between Cook's distance and the conditionally scaled Cook's distance (Figure 3 (d)), but there are some discrepancies between them. For instance, the magnitude of $\text{CSCD}_1(\{269\}, \mathbf{Z})$ is only moderate, whereas $\text{CD}_1(\{269\}, \mathbf{Z})$ is the highest one.

Furthermore, all $\text{CSCD}_1(I, \mathbf{Z})^s$ calculated from the bootstrapped samples give a range of 'good' values of $\text{CSCD}_1(I, \mathbf{Z})$ across all subjects (Fig. 4 (a)). Thus, we can calculate $P_B(I, \mathbf{Z})$ for all subjects, which gives the magnitude of each subject I (Fig. 4 (b)). Specifically, subjects 269, 217, 294, 289, 274, 90, 38, 285, 280, 149, 109, and 224 are identified as the top 12 most influential observations by CD, whereas compared to $\text{CD}(I)$, $\text{CSCD}_1(I, \mathbf{Z})$ and $P_B(I, \mathbf{Z})$ identify a set of 31 influential observations with $P_B(I, \mathbf{Z}) = 1$. This indicates that if the fitted linear mixed model is true, it is almost impossible to observe such 31 subjects (Fig. 4). For instance, since $\text{CD}(\{246\}) = 0.253$, it is unclear whether subject 246 is influential or not according to CD (Table 2), whereas we have $\text{CSCD}_1(\{246\}, \mathbf{Z}) = 21.443$ and $P_B(\{246\}, \mathbf{Z}) = 1.0$. Thus, subject 246 is really influential after eliminating the effect of the cluster size (Table 2). Moreover, it is difficult to compare the influential levels

TABLE 2
Yale infant growth data. Top 12 influential subjects for single case deletion with the compound symmetry model.

ID	m_i	CD	ID	m_i	CSCD_1	$P_B(I, \mathbf{Z})$	ID	m_i	CSCD_2	$P_B(I, \mathbf{Z})$
269	12	2.416	274	22	43.593	1.000	217	19	62.639	1.000
217	19	1.465	217	19	27.359	1.000	274	22	60.809	1.000
294	13	1.252	90	17	27.273	1.000	90	17	51.969	1.000
289	18	1.188	109	12	25.520	1.000	109	12	48.173	1.000
274	22	1.163	289	18	24.610	1.000	294	13	45.117	1.000
90	17	0.858	294	13	23.950	1.000	149	17	43.843	1.000
38	24	0.823	149	17	22.217	1.000	38	24	40.753	1.000
285	**8**	**0.738**	**246**	**5**	**21.443**	**1.000**	289	18	36.529	1.000
280	9	0.695	38	24	16.508	1.000	**246**	**5**	**35.626**	**1.000**
149	17	0.668	62	13	16.455	1.000	269	12	33.447	1.000
109	12	0.625	269	12	16.172	1.000	280	9	25.034	1.000
224	22	0.591	280	9	15.098	1.000	62	13	24.483	1.000

Note that m_i represents cluster size and $P_B(I, \mathbf{Z})$ is computed by equation (2.24) .

of subjects 274 and 285 using CD. All of the scaled Cook's distances and associated quantities suggest that subject 274 is more influential than subject 285 after eliminating their size difference. Finally, given the large number of influential observations identified by $P_B(I, \mathbf{Z})$, it strongly indicates that further research may be needed to explore other statistical models and improve the model fitting for the Yale infant growth data.

4. Discussion. We have introduced a new quantity to quantify the degree of perturbation and examined its properties. We have used stochastic ordering to quantify the relationship between the degree of the perturbation and the size of Cook's distance. We have developed several scaled Cook's distances to address the size issue for deletion diagnostics in general parametric models. We have shown that the scaled Cook's distances provide important information about the relatively influential level of each subset. We have illustrated our development with linear regression, generalized linear models, general linear models with correlated errors, and generalized linear mixed models. We have analyzed simulated data and a real dataset using the scaled and conditionally scaled Cook's distance measure. Future work includes developing Bayesian analogs to the scaled Cook's distance measure and developing such a methodology for other types of models, such as survival models and models with missing covariate data.

Appendix. The following assumptions are needed to facilitate the technical details, although they are not the weakest possible conditions. Because we develop all results for general parametric models, we only assume several high-level assumptions as follows.

Assumption A2. $\hat{\boldsymbol{\theta}}_{[I]}$ for any I is a consistent estimate of $\boldsymbol{\theta}_*$, an interior point of Θ.

Assumption A3. All $p(\mathbf{Y}_{[I]}|\boldsymbol{\theta})$ are three times continuously differentiable on Θ and satisfy

$$\log p(\mathbf{Y}_{[I]}|\boldsymbol{\theta}) = \log p(\mathbf{Y}_{[I]}|\boldsymbol{\theta}_*)+\Delta(\boldsymbol{\theta})^T J_{n,[I]}(\boldsymbol{\theta}_*)-0.5\Delta(\boldsymbol{\theta})^T \mathbf{F}_{n,[I]}(\boldsymbol{\theta}_*)\Delta(\boldsymbol{\theta})+R_{[I]}(\boldsymbol{\theta}),$$

in which $|R_{[I]}(\boldsymbol{\theta})| = o_p(1)$ uniformly for all $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_*, \delta_0 n^{-1/2}) = \{\boldsymbol{\theta} : \sqrt{n}||\boldsymbol{\theta} - \boldsymbol{\theta}_*|| \le \delta_0\}$, where $\Delta(\boldsymbol{\theta}) = \boldsymbol{\theta} - \boldsymbol{\theta}_*$, $J_{n,[I]}(\boldsymbol{\theta}) = \partial_\theta \log p(\mathbf{Y}_{[I]}|\boldsymbol{\theta})$ and $\mathbf{F}_{n,[I]}(\boldsymbol{\theta}_*) = \partial^2_\theta \log p(\mathbf{Y}_{[I]}|\boldsymbol{\theta})$.

Assumption A4. For any set I and $\mathbf{Z}$, $\sup_{\theta \in B(\theta_*, n^{-1/2}\delta_0)} n^{-1/2} J_{n,[I]}(\boldsymbol{\theta}) = O_p(1)$,

$$\sup_{\theta \in B(\theta_*, n^{-1/2}\delta_0)} ||\mathbf{F}_{n,[I]}(\boldsymbol{\theta}) - E[\mathbf{F}_I(\boldsymbol{\theta})|\mathcal{M}, \mathbf{Z}]|| = O_p(\sqrt{n}),$$

$$\sup_{\theta, \theta' \in B(\theta_*, n^{-1/2}\delta_0)} n^{-1}||\mathbf{F}_{n,[I]}(\boldsymbol{\theta}) - \mathbf{F}_{n,[I]}(\boldsymbol{\theta}')|| = o_p(1),$$

and $0 < \inf_{\theta \in B(\theta_*, \delta_0 n^{-1/2})} \lambda_{\min}(n^{-1}\mathbf{F}_{n,[I]}(\boldsymbol{\theta})) \le \sup_{\theta \in B(\theta_*, \delta_0 n^{-1/2})} \lambda_{\max}(n^{-1}\mathbf{F}_{n,[I]}(\boldsymbol{\theta})) < \infty$.

Assumption A5. For any set I and $\mathbf{Z}$,

$$\sup_{\theta \in B(\theta_*, n^{-1/2}\delta_0)} J_I(\boldsymbol{\theta}) = O_p(\sqrt{n(I)}), \qquad \sup_{\theta \in B(\theta_*, n^{-1/2}\delta_0)} ||\mathbf{f}_I(\boldsymbol{\theta})|| = O_p(n(I)),$$

$$\sup_{\theta \in B(\theta_*, n^{-1/2}\delta_0)} ||\mathbf{f}_I(\boldsymbol{\theta}) - E[\mathbf{f}_I(\boldsymbol{\theta})|\mathcal{M}, \mathbf{Z}]|| = O_p(\sqrt{n(I)}).$$

Remarks: Assumptions A2-A5 are very general conditions and are generalizations of some higher level conditions for the extremum estimator, such as the maximum likelihood estimate, given in Andrews [3]. Assumption A2 assumes that the parameter estimators with and without deleting the observations in the subset I are consistent. Assumption A3 assumes that the log-likelihood functions for any I and $\mathbf{Y}_{[I]}$ admit a second-order Taylor's series expansion in a small neighborhood of $\boldsymbol{\theta}_*$. Assumptions A4 and A5 are standard assumptions to ensure that the first- and second-order derivatives of $p(\mathbf{Y}_{[I]}|\boldsymbol{\theta})$ and $p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \boldsymbol{\theta})$ have appropriate rates of n and n_I [3, 28]. Sufficient conditions of Assumptions A2-A5 have been extensively discussed in the literature [3, 28]. Moreover, for simplicity, we use the rates of n and $n(I)$ in Assumptions A3-A5, which can be modified to accommodate more intricate examples in Andrews [3] and Zhu and Zhang [28].

References.

[1] Amari, S. I. (1990). *Differential-Geometrical Methods in Statistics (2nd edition). Lecture Notes in Statistics 28*. Berlin: Springer-Verlag.

[2] Andersen, E. B. (1992). Diagnostics in Categorical Data Analysis. *Journal of the Royal Statistical Society, Series B: Methodological* **54** 781–791.

[3] Andrews, D. W. K. (1999). Estimation When a Parameter Is on a Boundary. *Econometrica* **67** 1341–1383.

[4] Banerjee, M. (1998). Cook's Distance in Linear Longitudinal Models. *Communications in Statistics: Theory and Methods* **27** 2973-2983.

[5] Banerjee, M. and Frees, E. W. (1997). Influence Diagnostics for Linear Longitudinal Models. *Journal of the American Statistical Association* **92** 999–1005.

[6] Beckman, R. J. and Cook, R. D. (1983). Outlier.....s. *Technometrics* **25** 119-149.

[7] Chatterjee, S. and Hadi, A. S. (1988). *Sensitivity Analysis in Linear Regression*. John Wiley & Sons.

[8] Christensen, R., Pearson, L. M. and Johnson, W. (1992). Case-deletion Diagnostics for Mixed Models. *Technometrics* **34** 38-45.

[9] Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics* **19** 15–18.

[10] Cook, R. D. (1986). Assessment of Local Influence (with Discussion). *Journal of the Royal Statistical Society, Series B: Methodological* **48** 133–169.

[11] Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall Ltd.

[12] Critchley, F., Atkinson, R. A., Lu, G. and Biazi, E. (2001). Influence Analysis Based on the Case Sensitivity Function. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **63** 307–323.

[13] Davison, A. C. and Tsai, C. L. (1992). Regression Model Diagnostics. *International Statistical Review* **60** 337–353.

[14] Eaton, M. L. and Tyler, D. E. (1991). On Wielandt's Inequality and Its Application to the Asymptotic Distribution of the Eigenvalues of a Random Symmetric Matrix. *The Annals of Statistics* **19** 260–271.

[15] Fung, W.-K., Zhu, Z.-Y., Wei, B.-C. and He, X. (2002). Influence Diagnostics and Outlier Tests for Semiparametric Mixed Models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **64** 565–579.

[16] Haslett, J. (1999). A Simple Derivation of Deletion Diagnostic Results for the General Linear Model with Correlated Errors. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **61** 603–609.

[17] Haslett, J. and Haslett, S. J. (2007). The Three Basic Types of Residuals for a Linear Model. *International Statistical Review* **75** 1-24.

[18] Huber, P. J. (1981). *Robust Statistics*. Wiley Series in Probability and Statistics.

[19] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall Ltd.

[20] Preisser, J. S. and Qaqish, B. F. (1996). Deletion Diagnostics for Generalised Estimating Equations. *Biometrika* **83** 551–562.

[21] Shaked, M. and Shanthikumar, G. J. (2006). *Stochastic Orders*. Springer.

[22] Stier, D. M., Leventhal, J. M., Berg, A. T., Johnson, L. and Mezger, J. (1993). Are Children Born to Young Mothers at Increased Risk of Maltreatment. *Pediatrics* **91** 642-648.

[23] Wasserman, D. R. and Leventhal, J. M. (1993). Maltreatment of Children Born

to Cocaine-Dependent Mothers. *American Journal of Diseases of Children* **147** 1324-1328.

[24] Wei, B.-C. (1998). *Exponential Family Nonlinear Models*. Springer: Singapore.

[25] White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica* **50** 1–26.

[26] White, H. (1994). *Estimation, Inference, and Specification Analysis*. Cambridge University Press.

[27] Zhang, H. (1999). Analysis of Infant Growth Curves Using Multivariate Adaptive Splines. *Biometrics* **55** 452–459.

[28] Zhu, H. and Zhang, H. (2006). Asymptotics for Estimation and Testing Procedures under Loss of Identifiability. *Journal of Multivariate Analysis* **97** 19–45.

[29] Zhu, H., Lee, S. Y., Wei, B. C. and Zhou, J. (2001). Case Deletion Measures for Models with Incomplete Data. *Biometrika* **88** 727–737.

[30] Zhu, H., Ibrahim, J. G., Lee, S.-Y. and Zhang, H. (2007). Perturbation Selection and Influence Measures in Local Influence Analysis. *The Annals of Statistics* **35** 2565–2588.

Department of Biostatistics,
University of North Carolina at Chapel Hill,
NC, USA, 27599-7420.
E-mail: hzhu@bios.unc.edu
ibrahim@bios.unc.edu

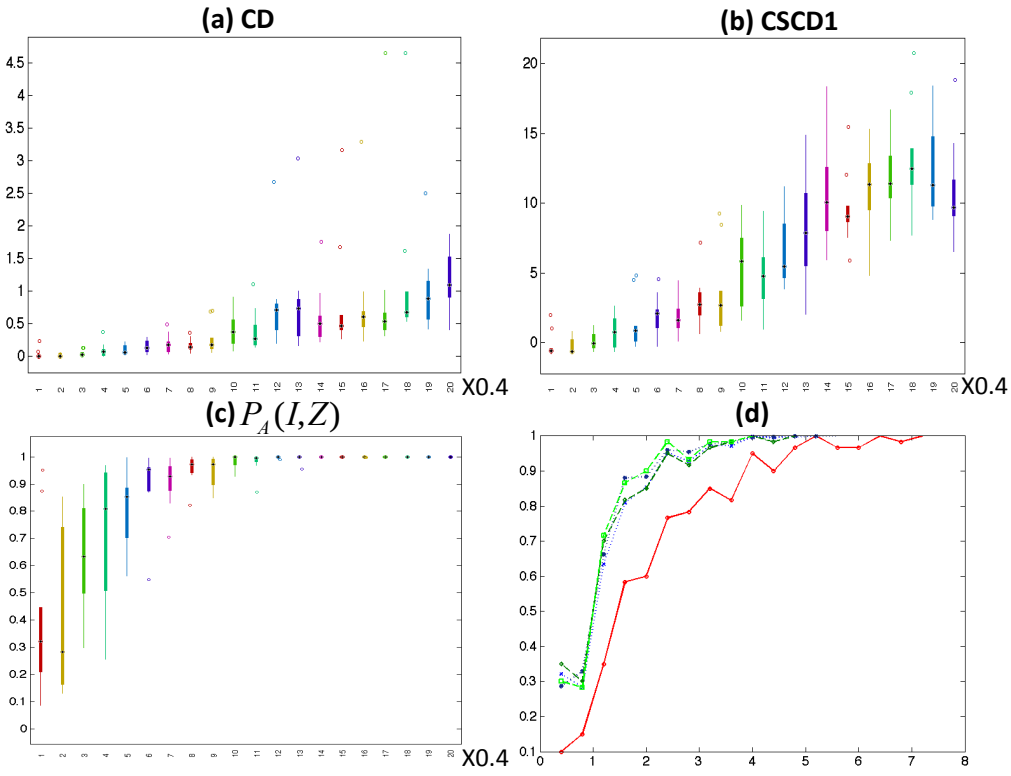

FIG 1. *Results from 100 datasets simulated from a linear mixed model, in which* $m_{30} = 1$ *and* b_{30} *varies from 0.4 to 8.0. Panel (a) shows the box plots of Cook's distances as a function of* b_{30}*; panel (b) shows the box plots of* $CSCD_1(I, \mathbf{Z})$ *as a function of* b_{30}*; panel (c) shows the box plots of* $P_A(I, \mathbf{Z})$ *as a function of* b_{30}*; panel (d) shows the mean curves of* $P_A(I, \mathbf{Z})$ *based on the four scaled Cook's distances, in which the green line is for* $CSCD_1(I, \mathbf{Z})$*, the dark green line is for* $CSCD_2(I, \mathbf{Z})$*, the blue line is for* $\widetilde{CSCD}_1(I, \mathbf{Z})$*, and the dark line is for* $\widetilde{CSCD}_1(I, \mathbf{Z})$*, and the mean curve of* $P_C(I, \mathbf{Z})$ *based on* $CD(I)$ *(red line) as functions of* b_{30}*.*

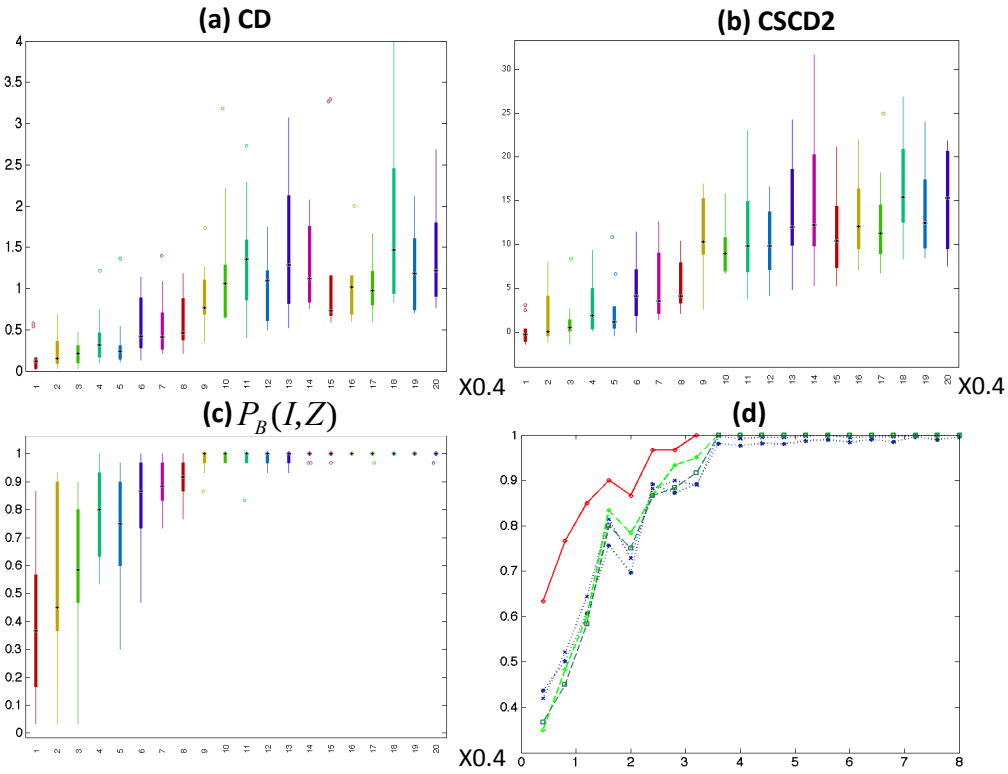

FIG 2. *Results from 100 datasets simulated from a linear mixed model, in which* $m_{30} = 1$ *and* b_{30} *varies from 0.4 to 8.0. Panel (a) shows the box plots of Cook's distances as a function of* b_{30}*; panel (b) shows the box plots of* $CSCD_1(I, \mathbf{Z})$ *as a function of* b_{30}*; panel (c) shows the box plots of* $P_B(I, \mathbf{Z})$ *as a function of* b_{30}*; panel (d) shows the mean curves of* $P_B(I, \mathbf{Z})$ *based on the four scaled Cook's distances, in which the green line is for* $CSCD_1(I, \mathbf{Z})$*, the dark green line is for* $CSCD_2(I, \mathbf{Z})$*, the blue line is for* $\widetilde{CSCD}_1(I, \mathbf{Z})$*, and the dark line is for* $\widetilde{CSCD}_1(I, \mathbf{Z})$*, and the mean curve of* $P_C(I, \mathbf{Z})$ *based on* $CD(I)$ *(red line) as functions of* b_{30}*.*

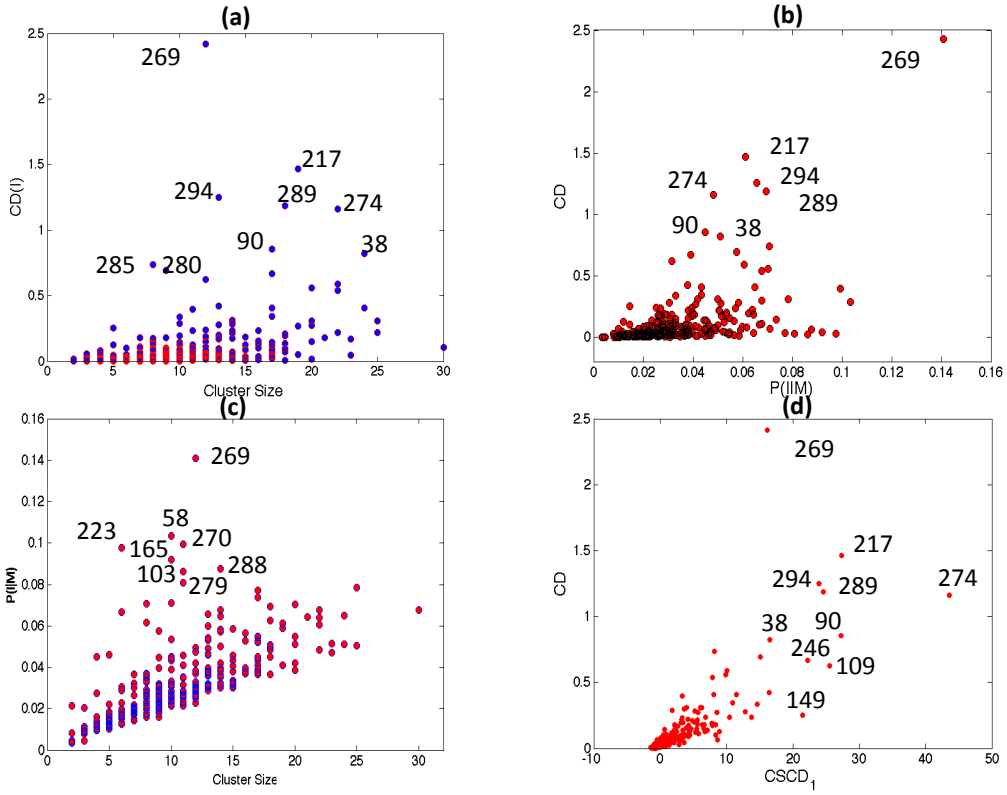

FIG 3. *Yale infant growth data. Panels (a) and (b) present cluster size versus $CD(I)$ and $P(I|\mathcal{M})$ versus $CD(I)$ for cluster deletion, respectively; panels (c) and (d), respectively, present cluster size versus $P(I|\mathcal{M})$ and $CD(I)$ versus $CSCD_1(I, \mathbf{Z})$ for cluster deletion.*

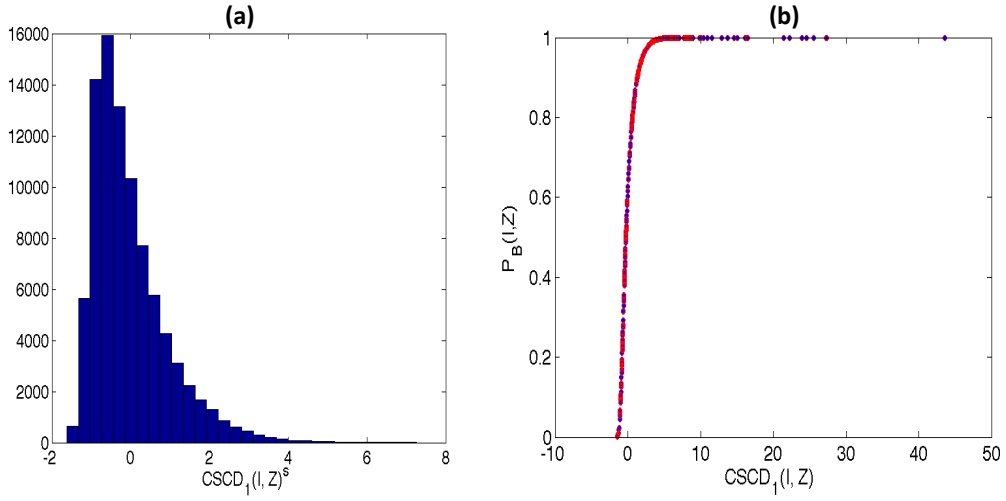

FIG 4. *Yale infant growth data. Panel (a) shows the histogram of* $CSCD_1(I, \mathbf{Z})^s$ *for all subjects and* s*; panel (b) shows* $CSCD_1(I, \mathbf{Z})$ *versus* $P_B(I, \mathbf{Z})$.