PERTURBATION AND SCALED COOK'S DISTANCE

BY HONGTU ZHU[1], JOSEPH G. IBRAHIM[1] AND HYUNSOON CHO

University of North Carolina at Chapel Hill

Cook's distance [*Technometrics* **19** (1977) 15–18] 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 issue: deleting subsets with different numbers of observations introduces different degrees of perturbation to the current model fitted to the data, and the magnitude of Cook's distance is associated with the degree of the perturbation. The aim of this paper is to address this 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 magnitude of Cook's distance. We develop several scaled Cook's distances to resolve the comparison of Cook's distance for different subset deletions. 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 (see Section 2.2), seriously affects specific key inferences, such as parameter estimates. Such perturbation schemes include the deletion of an individual or a subset of observations, case weight perturbation and covariate perturbation, among many others [8, 9, 28]. For example, for linear models, a perturbation measures the effect on the model of deleting a subset of the data matrix. In general, perturbation measures do not depend on the data directly, but rather on its structure via the model. 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 [11, 16]. If a small perturbation seriously influences key results of the analysis, we want to know the cause [9, 11]. 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 [5, 8].

Since the seminal work of Cook [8] 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 [1, 6, 8, 10, 12, 14, 15, 23, 29]. For example, for longitudinal data, Preisser and Qaqish [19] developed Cook's distance for generalized estimating equations, while Christensen, Pearson and Johnson [7], Banerjee and Frees [4] and Banerjee [3] 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 major research problem regarding Cook's distance that has been largely neglected in the existing literature is the development of Cook's distance for general statistical models with more complex data structures. The fundamental issue that arises here 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 influential observations 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 [11, 28].

The aim of this paper is to address this fundamental 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, nonnegativity, monotonicity and additivity.

(a.2) We use stochastic ordering to quantify the relationship between the degree of the perturbation and the magnitude 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 in order to compare Cook's distance for deleted subsets with different numbers of observations.

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. In Section 3, we analyze simulated data and a real dataset using the 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 significantly across all i. Cook's distance and many other deletion diagnostics measure the distance between the maximum likelihood estimators of $\boldsymbol{\theta}$ with and without Y_i [8, 10]. 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

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

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

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

where $G_{n\theta}$ is chosen to be a positive definite matrix. The matrix $G_{n\theta}$ is not changed or re-estimated when a subset of the data is deleted. 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 $\boldsymbol{\theta}$. For clustered data, the observations within the same cluster are correlated. A sensible model $p(\mathbf{Y}|\boldsymbol{\theta})$ should explicitly model the correlation structure in the clustered data and thus $-\partial_\theta^2 \log p(\mathbf{Y}|\hat{\boldsymbol{\theta}})$ implicitly incorporates such a correlation structure.

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 $\operatorname{rank}(\mathbf{L}) = q_1$ [4, 10]. 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

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

For notational simplicity, even though we may focus on a subset of $\boldsymbol{\theta}$, we do not distinguish between $\mathrm{CD}(I|\mathbf{L})$ and $\mathrm{CD}(I)$ 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 the goodness of fit of $\mathcal{M}$ to $\mathbf{Y}$ for I, denoted by $G(I|\mathbf{Y}, \mathcal{M})$, 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:

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

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

We may use the value of CD(I) to assess the influential level of the subset I. We may regard a subset I as influential if either the value of CD(I) is relatively large, compared with other Cook's distances, or the magnitude of CD(I) is greater than the critical points of the χ^2 distribution [10]. 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 developing a measure to quantify the degree of the perturbation to $\mathcal{M}$ introduced by deleting the subset I, regardless of the observed data **Y**. We emphasize here that the degree of perturbation is a property of the model, unlike Cook's distance which is also a property of **Y**. 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 such 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 [11]. Specifically, even though [11] placed the Euclidean geometry on the perturbation space for one-sample problems, such a geometrical structure cannot be easily generalizable to general data structures (e.g., correlated data) and related parametric models. For instance, for correlated data, a sensible model $\mathcal{M}$ should model the correlation structure, and a good measure $\mathcal{P}(I|\mathcal{M})$ should explicitly incorporate the correlation structure specified in $\mathcal{M}$ and the subset I. However, the Euclidean geometry proposed by [11] cannot incorporate the correlation structure in the correlated data.

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

- (P.a) (nonnegativity) For any subset I, $\mathcal{P}(I|\mathcal{M})$ is always nonnegative.
- (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 $I_2 \subset I_1$, $I_{1\cdot2} = I_1 - I_2$ and $p(\mathbf{Y}_{I_{1\cdot2}}|\mathbf{Y}_{[I_1]}, \boldsymbol{\theta}) = p(\mathbf{Y}_{I_{1\cdot2}}|\mathbf{Y}_{[I_{1\cdot2}]}, \boldsymbol{\theta})$ for all $\boldsymbol{\theta}$, then we have $\mathcal{P}(I_1|\mathcal{M}) = \mathcal{P}(I_2|\mathcal{M}) + \mathcal{P}(I_{1\cdot2}|\mathcal{M})$.
- (P.e) $\mathcal{P}(I|\mathcal{M})$ should naturally arise from the current model $\mathcal{M}$, the data **Y** and the subset I.

Principles (P.a) and (P.b) indicate 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 $\mathbf{Y}_{[I_{1\cdot2}]}$ is the union of $\mathbf{Y}_{[I_1]}$ and $\mathbf{Y}_{I_2}$, $p(\mathbf{Y}_{I_{1\cdot2}}|\mathbf{Y}_{[I_1]}, \boldsymbol{\theta}) = p(\mathbf{Y}_{I_{1\cdot2}}|\mathbf{Y}_{[I_{1\cdot2}]}, \boldsymbol{\theta})$ is equivalent to that of $\mathbf{Y}_{I_{1\cdot2}}$ 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.

Principle (P.e) requires that $\mathcal{P}(I|\mathcal{M})$ depend on the triple $(\mathcal{M}, \mathbf{Y}, I)$. We propose $\mathcal{P}(I|\mathcal{M})$ based on the Kullback–Leibler divergence between the fitted probability function $p(\mathbf{Y}|\boldsymbol{\theta})$ and the probability function of a model for characterizing the deletion of $\mathbf{Y}_I$, denoted by $p(\mathbf{Y}|\boldsymbol{\theta}, I)$. 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 $\mathbf{Y}_{[I]}$. Let $\boldsymbol{\theta}_*$ be the true value of $\boldsymbol{\theta}$ under $\mathcal{M}$ [24, 25]. We define $p(\mathbf{Y}|\boldsymbol{\theta}, I)$ as follows:

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

in which $p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \boldsymbol{\theta}_*)$ is independent of $\boldsymbol{\theta}$. In (2.5), by fixing $\boldsymbol{\theta} = \boldsymbol{\theta}_*$ in $p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \boldsymbol{\theta})$, we essentially drop the information contained in $\mathbf{Y}_I$ as we estimate $\boldsymbol{\theta}$. Specifically, $\hat{\boldsymbol{\theta}}_{[I]}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ under $p(\mathbf{Y}|\boldsymbol{\theta}, I)$. 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 $\mathbf{Y}_{[I]}$. The Kullback–Leibler distance between $p(\mathbf{Y}|\boldsymbol{\theta})$ and $p(\mathbf{Y}|\boldsymbol{\theta}, I)$, denoted by $\mathrm{KL}(\mathbf{Y}, \boldsymbol{\theta}|\boldsymbol{\theta}_*, I)$, is given by

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

We use $\mathrm{KL}(\mathbf{Y}, \boldsymbol{\theta}|\boldsymbol{\theta}_*, I)$ to measure the effect of deleting $\mathbf{Y}_I$ on estimating $\boldsymbol{\theta}$ without knowing that the true value of $\boldsymbol{\theta}$ is $\boldsymbol{\theta}_*$. If $\mathbf{Y}_I$ is independent of $\mathbf{Y}_{[I]}$, then we have

$$\mathrm{KL}(\mathbf{Y}, \boldsymbol{\theta}|\boldsymbol{\theta}_*, I) = \int p(\mathbf{Y}_I|\boldsymbol{\theta}) \log\left(\frac{p(\mathbf{Y}_I|\boldsymbol{\theta})}{p(\mathbf{Y}_I|\boldsymbol{\theta}_*)}\right) d\mathbf{Y}_I,$$

which is independent of $\mathbf{Y}_{[I]}$. In this case, the effect of deleting $\mathbf{Y}_I$ on estimating $\boldsymbol{\theta}$ only depends on $\{p(\mathbf{Y}_I|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$.

A conceptual difficulty associated with $\mathrm{KL}(\mathbf{Y}, \boldsymbol{\theta}|\boldsymbol{\theta}_*, I)$ is that both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_*$ are unknown. Although $\boldsymbol{\theta}_*$ is unknown, it can be assumed to be a fixed value from a frequentist viewpoint. For the unknown $\boldsymbol{\theta}$, we can always use the data $\mathbf{Y}$ and the current model $\mathcal{M}$ to calculate an estimator $\hat{\boldsymbol{\theta}}$ in a neighborhood of $\boldsymbol{\theta}_*$. Under some mild conditions [24, 25], one can show that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*)$ is asymptotically normal, and thus $\hat{\boldsymbol{\theta}}$ should be centered around $\boldsymbol{\theta}_*$. Moreover, since Cook's distance is to quantify the change of the parameter estimates after deleting a subset, we need to consider all possible $\boldsymbol{\theta}$ around $\boldsymbol{\theta}_*$, instead of focusing on a single $\boldsymbol{\theta}$. Specifically, we consider $\boldsymbol{\theta}$ 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:

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

This quantity $\mathcal{P}(I|\mathcal{M})$ can also be interpreted as the average effect of deleting $\mathbf{Y}_I$ on estimating $\boldsymbol{\theta}$ with the prior information that the estimate of $\boldsymbol{\theta}$ is centered

around $\boldsymbol{\theta}_*$. Since $\mathcal{P}(I|\mathcal{M})$ is directly calculated from the model $\mathcal{M}$ and the subset I, it can naturally account for any structure specified in $\mathcal{M}$. 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 at their true value.

To compute $\mathcal{P}(I|\mathcal{M})$ in a real data analysis, we only need to specify $\mathcal{M}$ and $(\boldsymbol{\theta}_*, \Sigma_*)$. Then we may use some numerical integration methods to compute $\mathcal{P}(I|\mathcal{M})$. Although $(\boldsymbol{\theta}_*, \Sigma_*)$ are unknown, we suggest substituting $\boldsymbol{\theta}_*$ by an estimator of $\boldsymbol{\theta}$, denoted by $\tilde{\boldsymbol{\theta}}$, and Σ_* by the covariance matrix of $\tilde{\boldsymbol{\theta}}$. Throughout the paper, since $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_*$ [24, 25], we set $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$ and Σ_* as the covariance matrix of $\hat{\boldsymbol{\theta}}$.

We obtain the following theorems, whose detailed assumptions and proofs can be found in the Appendix.

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).

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 $y_i = \mathbf{x}_i^T \boldsymbol{\beta}_* + \varepsilon_i$, 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 ith 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

$$(2.8) \quad 0.5E_\theta[\log(\sigma_*^2/\sigma^2)] + 0.5\frac{\mathbf{x}_i^T E_\theta[(\boldsymbol{\beta} - \boldsymbol{\beta}_*)(\boldsymbol{\beta} - \boldsymbol{\beta}_*)^T]\mathbf{x}_i}{\sigma_*^2} \approx \frac{1}{2n} + \frac{1}{2}h_{ii},$$

where E_θ is taken with respect to $p(\boldsymbol{\theta}|\boldsymbol{\theta}_*, G_{n\theta}^{-1})$. Moreover, the right-hand side of (2.8) contains only terms involving n and $\mathbf{X}$, since perturbation is defined only in terms of the underlying model $\mathcal{M}$. This is also at the core of why only stochastic ordering is possible for Cook's distance, which is a function of both the perturbation and the data. See Section 2.3 for detailed discussions. Furthermore, if $\boldsymbol{\beta}$ is the parameter of interest in $\boldsymbol{\theta}$ and σ^2 is a nuisance parameter, then $0.5E_\theta[\log(\sigma_*^2/\sigma^2)]$, and $1/(2n)$ can be dropped from $\mathcal{P}(\{i\}|\mathcal{M})$ in (2.8).

Furthermore, for the case of random covariates, we assume that the $\mathbf{x}_i$'s are independently and identically distributed with mean μ_x and covariance matrix Σ_x.

It can be shown that $\mathcal{P}(\{i\}|\mathcal{M})$ equals

$$(2.9) \quad 0.5E_\theta[\log(\sigma_*^2/\sigma^2)] + 0.5\sigma_*^{-2}\mathrm{tr}\{\Sigma_x E_\theta[(\boldsymbol{\beta} - \boldsymbol{\beta}_*)(\boldsymbol{\beta} - \boldsymbol{\beta}_*)^T]\} \approx \frac{1}{2n} + \frac{p}{2n}.$$

If $\boldsymbol{\beta}$ is the parameter of interest in $\boldsymbol{\theta}$, and σ^2 is a nuisance parameter, then $\mathcal{P}(\{i\}|\mathcal{M})$ reduces to $p/(2n)$. Furthermore, consider deleting a subset of observations $\{(y_{i_k}, \mathbf{x}_{i_k}): k = 1, \ldots, n(I)\}$ and $I = \{i_1, \ldots, i_{n(I)}\}$. It follows from Theorem 1 that $\mathcal{P}(\{i_1, \ldots, i_{n(I)}\}|\mathcal{M}) = \sum_{k=1}^{n(I)} \mathcal{P}(\{i_k\}|\mathcal{M})$. 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. *Cook's distance and degree of perturbation.* To understand the relationship between $\mathcal{P}(I|\mathcal{M})$ and $\mathrm{CD}(I)$ in (2.4), we temporarily assume that the fitted model $\mathcal{M}$ is the true data generator of $\mathbf{Y}$. 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 [8] for case i, $(y_i, \mathbf{x}_i)$, is given by

$$(2.10) \quad \mathrm{CD}(\{i\}) = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[i]})^T \mathbf{X}^T \mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[i]})}{\hat{\sigma}^2} = \frac{\sigma^2}{\hat{\sigma}^2} t_i^2 \frac{h_{ii}}{1 - h_{ii}},$$

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, $\mathrm{CD}(\{i\})$ is almost the same as the original Cook's distance (Cook [8]). As shown in (2.8) and (2.9), regardless of the exact value of $(y_i, \mathbf{x}_i)$, deleting any $(y_i, \mathbf{x}_i)$ has approximately the same degree of perturbation to $\mathcal{M}$. Moreover, the $\mathrm{CD}(\{i\})$ are comparable regardless of i. Specifically, if $\varepsilon_i \sim N(0, \sigma^2)$, 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 $\mathrm{CD}(\{i\})$ 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

$$(2.11) \quad \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[I]})^T \mathbf{X}^T \mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[I]})}{\hat{\sigma}^2} = \frac{1}{\hat{\sigma}^2}\hat{\mathbf{e}}_I^T(\mathbf{I}_{n(I)} - H_I)^{-1}H_I(\mathbf{I}_{n(I)} - H_I)^{-1}\hat{\mathbf{e}}_I,$$

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 $\text{CD}(I)$ 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 $\text{CD}(I_1)$ and $\text{CD}(I_2)$ 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 a 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 $\text{CD}(I)$ 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 $\text{CD}(I)$, as follows.

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

(a) *For any* $I_2 \subset I_1$, $\text{CD}(I_1)$ *is stochastically larger than* $\text{CD}(I_2)$ *for any* $\mathbf{X}$, *that is*, $\text{P}(\text{CD}(I_1) > t|\mathcal{M}) \geq \text{P}(\text{CD}(I_2) > t|\mathcal{M})$ *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*, $\text{CD}(I)$ *and* $\text{CD}(I')$ *follow the same distribution when* $n(I) = n(I')$ *and* $\text{CD}(I_1)$ *is stochastically larger than* $\text{CD}(I_2)$ *for any two subsets* I_2 *and* I_1 *with* $n(I_1) > n(I_2)$.

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 magnitude 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 $\mathbf{Y}$, $\text{CD}(I_1)$, as a random variable, should be *stochastically larger* than $\text{CD}(I_2)$ if $\mathcal{M}$ is the true model for $\mathbf{Y}$. We make the following assumption.

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

$$(2.12) \qquad \text{P}(\text{CD}(I_1) > t|\mathcal{M}) \geq \text{P}(\text{CD}(I_2) > t|\mathcal{M})$$

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 $\text{CD}(I_1)$ stochastically dominates $\text{CD}(I_2)$ whenever $\mathcal{P}(I_1|\mathcal{M}) > \mathcal{P}(I_2|\mathcal{M})$. According to the definition of stochastic ordering [20], 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*

$$(2.13) \qquad E[h(\text{CD}(I_1))|\mathcal{M}] \geq E[h(\text{CD}(I_2))|\mathcal{M}]$$

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

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

2.4. *Scaled Cook's distances.* We focus on developing several scaled Cook's distances for I, denoted by $\text{SCD}(I)$, to detect relatively influential subsets, while accounting for the degree of perturbation $\mathcal{P}(I|\mathcal{M})$. Since we have characterized the stochastic relationship between $\mathcal{P}(I|\mathcal{M})$ and $\text{CD}(I)$ when $\mathcal{M}$ is the true data generator, we are interested in reducing the effect of the difference among $\mathcal{P}(I|\mathcal{M})$ for different subsets I on the magnitude of $\text{CD}(I)$. A simple solution is to calculate several features (e.g., mean, median, or quantiles) of $\text{CD}(I)$ and match them across different subsets under the assumption that $\mathcal{M}$ is the true data generator. Throughout the paper, 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 when $\mathcal{M}$ is the true data generator.

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

$$\text{SCD}_1(I) = \frac{\text{CD}(I) - E[\text{CD}(I)|\mathcal{M}]}{\text{Std}[\text{CD}(I)|\mathcal{M}]} \quad \text{and}$$

$$\text{SCD}_2(I) = \frac{\text{CD}(I) - Q_{\text{CD}(I)}(0.5|\mathcal{M})}{\text{Mstd}[\text{CD}(I)|\mathcal{M}]},$$

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

We can use $SCD_1(I)$ and $SCD_2(I)$ to evaluate the relatively influential level for different subsets I. A large value of $SCD_1(I)$ [or $SCD_2(I)$] indicates that the subset I is relatively influential. Therefore, for any two subsets I_1 and I_2, the probability of observing the event $SCD(I_1) > SCD(I_2)$ and that of the event $SCD(I_1) < SCD(I_2)$ should be reasonably close to each other. Thus, the $SCD(I)$ are roughly comparable. 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. Moreover, the standardization in Definition 1 still implies that higher than average values of $CD(I)$ still correspond with high positive values of $SCD(I)$, even though for some deletions, it is possible for $SCD(I)$ to be negative unlike $CD(I)$.

The next task is how to compute $E[CD(I)|\mathcal{M}]$, $Std[CD(I)|\mathcal{M}]$, $Mstd[CD(I)|\mathcal{M}]$ and $Q_{CD(I)}(0.5|\mathcal{M})$ for each subset I under the assumption that $\mathcal{M}$ is the true data generator. Computationally, we suggest using the parametric bootstrap to approximate the four quantities of $CD(I)$ 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 $CD(I)^{(s)} = F_1(I, \hat{\mathcal{M}}, \mathbf{Y}^s)$ for each s and each subset I.

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

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

In this process, we have used $\hat{\mathcal{M}}$ to approximate $\mathcal{M}$ [24] 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 $\overline{CD(I)}$ is a consistent estimate of $E[F_1(I, \mathcal{M}, \mathbf{Y})|\mathcal{M}]$. Similar arguments hold for the other three quantities of $CD(I)$. In Steps 2 and 3, we can use a moderate S, say $S = 100$, in order to accurately approximate all four quantities of $CD(I)$. According to our experience, such an approximation is very accurate, even for small n. See the 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.

As an illustration, we consider how to calculate $SCD_1(I)$ for any subset I in the linear regression model.

EXAMPLE 1 (Continued). In (2.11), since all $CD(I)$ share $\hat{\sigma}^2$, we replace $\hat{\sigma}^2$ by σ_*^2. Thus, we approximate $CD(I)$ by $CD_*(I) = \boldsymbol{\varepsilon}^T \mathbf{W}_* \boldsymbol{\varepsilon}/\sigma_*^2$, where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim N(\mathbf{0}, \sigma_*^2 \mathbf{I}_n)$ and

$$\mathbf{W}_* = (\mathbf{I}_n - H_x) U_I (\mathbf{I}_{n(I)} - H_I)^{-1} H_I (\mathbf{I}_{n(I)} - H_I)^{-1} U_I^T (\mathbf{I}_n - H_x).$$

To compute $\mathrm{SCD}_1(I)$, we just need to calculate the two quantities $E[\mathrm{CD}_*(I)|\mathcal{M}]$ and $\mathrm{Std}[\mathrm{CD}_*(I)|\mathcal{M}]$. Since $\mathrm{CD}_*(I)$ is a quadratic form, it can be shown that

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

$$\begin{aligned}\mathrm{Var}[\mathrm{CD}_*(I)|\mathcal{M}] = {}& \mathrm{Var}\{\mathrm{tr}[(\mathbf{I}_{n(I)} - H_I)^{-1}]|\mathcal{M}_X\} \\ &+ 2E\{\mathrm{tr}[\{(\mathbf{I}_{n(I)} - H_I)^{-1}H_I\}^2]|\mathcal{M}_X\},\end{aligned}$$

where $E[\cdot|\mathcal{M}_X]$ denotes the expectation taken with respect to $\mathbf{X}$.

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 comparing 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 conditionally scaled Cook's distances 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

$$\mathrm{CSCD}_1(I, \mathbf{Z}) = \frac{\mathrm{CD}(I) - E[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}]}{\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}]},$$

$$\mathrm{CSCD}_2(I, \mathbf{Z}) = \frac{\mathrm{CD}(I) - Q_{\mathrm{CD}(I)}(0.5|\mathcal{M}, \mathbf{Z})}{\mathrm{Mstd}[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}]},$$

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 data structure when $\mathcal{M}$ is true and $\mathbf{Z}$ is fixed.

The next problem is how to compute $E[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}]$, $\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}]$, $Q_{\mathrm{CD}(I)}(0.5|\mathcal{M}, \mathbf{Z})$ 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 $\text{CSCD}_1(I, \mathbf{Z})$ and $\text{CSCD}_2(I, \mathbf{Z})$. However, a slight difference occurs in the way that we simulate the data. Specifically, let $\mathbf{Y}_Z$ be the data $\mathbf{Y}$ with $\mathbf{Z}$ fixed. 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 $\text{CD}(I)^{(s)} = F_1(I, \hat{\mathcal{M}}_Z, (\mathbf{Y}_Z^s, \mathbf{Z}))$ for each subset I.

As an illustration, we consider how to calculate $\text{CSCD}_1(I, \mathbf{Z})$ for any subset I in the linear regression model.

EXAMPLE 1 (Continued). We set $\mathbf{Z} = \mathbf{X}$ to calculate $\text{CSCD}_1(I, \mathbf{Z})$. We need to compute $E[\text{CD}_*(I)|\mathcal{M}, \mathbf{Z}]$ and $\text{Std}[\text{CD}_*(I)|\mathcal{M}, \mathbf{Z}]$. Since $\text{CD}_*(I)$ is a quadratic form, it is easy to show $E[\text{CD}_*(I)|\mathcal{M}] = \text{tr}[(\mathbf{I}_{n(I)} - H_I)^{-1}] - n(I)$ and $\text{Var}[\text{CD}_*(I)|\mathcal{M}] = 2\,\text{tr}[\{(\mathbf{I}_{n(I)} - H_I)^{-1}H_I\}^2]$.

2.6. *First-order approximations.* We have focused on developing the scaled Cook's distances and their approximations for the linear regression model. More generally, we are interested in approximating the scaled Cook's distances for a large class of parametric models for both independent and dependent data.

We obtain the following theorem.

THEOREM 3. *If Assumptions [A2]–[A5] in the [Appendix] hold and $n(I)/n \to \gamma \in [0, 1)$, where $n(I)$ denotes the number of observations 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* $\mathbf{s}_I(\boldsymbol{\theta}) = -\partial_\theta^2 \log p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \boldsymbol{\theta})$, $\text{CD}(I)$ *can be approximated by*

$$(2.14) \quad \widetilde{\text{CD}}(I) = \mathbf{f}_I(\hat{\boldsymbol{\theta}})^T[\mathbf{F}_n(\hat{\boldsymbol{\theta}}) - \mathbf{s}_I(\hat{\boldsymbol{\theta}})]^{-1}\mathbf{F}_n(\hat{\boldsymbol{\theta}})[\mathbf{F}_n(\hat{\boldsymbol{\theta}}) - \mathbf{s}_I(\hat{\boldsymbol{\theta}})]^{-1}\mathbf{f}_I(\hat{\boldsymbol{\theta}});$$

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

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

Theorem 3(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 $\text{CD}(I)$. Theorem 3(b) and (c) give an approximation of $E[\text{CD}(I)|\mathcal{M}]$ and $E[\text{CD}(I)|\mathcal{M}, \mathbf{Z}]$, respectively. Generally, it is difficult to give a simple approximation to $\text{Var}[\text{CD}(I)|\mathcal{M}]$ and $\text{Var}[\text{CD}(I)|\mathcal{M}, \mathbf{Z}]$, since it involves the fourth moment of $\mathbf{f}_I(\hat{\boldsymbol{\theta}})$, which does not have a simple form.

Based on Theorem 3, we can approximate the scaled Cook's distance measures as follows.

Step (i). 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 $\widetilde{\mathrm{CD}}(I)^s$ for large S, say $S > 100$. 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 (ii). By repeating Step (i) S times, we can use the empirical quantities of $\{\widetilde{\mathrm{CD}}(I)^s : s = 1, \ldots, S\}$ to approximate $E[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}]$, $\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}]$, $Q_{\mathrm{CD}(I)}(0.5|\mathcal{M}, \mathbf{Z})$ 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

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

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

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's 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:

$$(2.15) \qquad P_A(I, \mathbf{Z}) = \sum_{s=1}^{S} \mathbf{1}(\widetilde{\mathrm{CSCD}}_1(I, \mathbf{Z})^s \leq \widetilde{\mathrm{CSCD}}_1(I, \mathbf{Z}))/S,$$

$$(2.16) \qquad P_B(I, \mathbf{Z}) = \sum_{\tilde{I}} \sum_{s=1}^{S} \frac{\mathbf{1}(\widetilde{\mathrm{CSCD}}_1(\tilde{I}, \mathbf{Z})^s \leq \widetilde{\mathrm{CSCD}}_1(I, \mathbf{Z}))}{S \times \#(\tilde{I})},$$

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 $\widetilde{\text{CD}}(I)$ to the exact $\text{CD}(I)$. For relatively influential subsets, even though the accuracy of the first-order approximation may be relatively low, $\widetilde{\text{CD}}(I)$ 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 $\widetilde{\text{CD}}(I)$ relative to the exact $\text{CD}(I)$. Numerical comparisons are given in Section 3.

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

EXAMPLE 2. 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 [18]

$$(2.17) \qquad p(y_{ij}|\mathbf{b}_i) = \exp\{a(\tau)^{-1}[y_{ij}\eta_{ij} - b(\eta_{ij})] + c(y_{ij}, \tau)\},$$

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 $\text{CD}(I_i)$ for deleting the ith cluster is given by

$$(2.18) \quad \widetilde{\text{CD}}(I_i) = \partial_\beta \ell_i(\hat{\boldsymbol{\beta}})^T [\mathbf{F}_n(\hat{\boldsymbol{\beta}}) - \mathbf{f}_i(\hat{\boldsymbol{\beta}})]^{-1} \mathbf{F}_n(\hat{\boldsymbol{\beta}})[\mathbf{F}_n(\hat{\boldsymbol{\beta}}) - \mathbf{f}_i(\hat{\boldsymbol{\beta}})]^{-1} \partial_\beta \ell_i(\hat{\boldsymbol{\beta}}),$$

where $I_i = \{(i, 1), \ldots, (i, m_i)\}$, $\ell_i(\boldsymbol{\beta})$ is the log-likelihood function for the ith 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 $\text{E}[\widetilde{\text{CD}}(I_i)|\mathcal{M}, \mathbf{Z}]$ can be approximated by $\text{tr}(\{\text{E}[\mathbf{F}_n(\hat{\boldsymbol{\beta}})|\mathcal{M}, \mathbf{Z}] - \text{E}[\mathbf{f}_i(\hat{\boldsymbol{\beta}})|\mathcal{M}, \mathbf{Z}]\}^{-1}\text{E}[\mathbf{f}_i(\hat{\boldsymbol{\beta}})|\mathcal{M}, \mathbf{Z}])$ when $\mathcal{M}$ is true. Moreover, we may approximate $\text{Var}[\widetilde{\text{CD}}(I_i)|\mathcal{M}, \mathbf{Z}]$ by using the fourth moment of $\partial_\beta \ell_i(\boldsymbol{\beta}_*)$. It is not straightforward to approximate $Q_{\text{CD}(I_i)}(0.5|\mathcal{M}, \mathbf{Z})$ and $\text{Mstd}[\text{CD}(I_i)|\mathcal{M}, \mathbf{Z}]$. Computationally, we employ the parametric bootstrap method described above to approximate the conditionally scaled Cook's distances $\text{CSCD}_1(I_i, \mathbf{Z})$ and $\text{CSCD}_2(I_i, \mathbf{Z})$.

3. Simulation studies and a real data example. In this section, we illustrate our methodology with simulated data and a real data example. We also include some additional results in the supplemental article [27]. The code along with its documentation for implementing our methodology is available on the first author's website at http://www.bios.unc.edu/research/bias/software.html.

3.1. *Simulation studies.*   The goals of our simulations were to examine the finite sample performance of Cook's distance and the scaled Cook's distances and their first-order approximations 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} $(j = 1, \ldots, m_i; i = 1, \ldots, n)$ were independently generated as $y_{ij} \sim N(\mathbf{x}_{ij}^T \boldsymbol{\beta} + b_i, \sigma_y^2)$ and the m_i were randomly drawn from $\{1, \ldots, 5\}$. 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, the responses were repeatedly simulated, whereas we generated the covariates and cluster sizes only once in order to fix the effect of the covariates and cluster sizes on Cook's distance for each cluster. The true value of $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma_b, \sigma_y)^T$ was fixed at $(1, 1, 1, 1, 1)^T$. The sample size n was set at 12 to represent a small number of clusters.

For each simulated dataset, we considered the detection of influential clusters [4]. We fit the same linear mixed model and used the expectation–maximization (EM) algorithm to calculate $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{[I]}$ for each cluster I. We treated (σ_b, σ_y) as nuisance parameters and $\boldsymbol{\beta}$ as the parameter vector of interest. We calculated the degree of the perturbation $\mathcal{P}(\{i\}|\mathcal{M})$ for deleting each subject $\{i\}$ while fixing the covariates, and then we calculated the conditionally scaled Cook's distances and associated quantities. Let $\mathbf{x}_i$ be an $m_i \times 3$ matrix with the jth row being $\mathbf{x}_{i,j}^T$. It can be shown that for the case of fixed covariates, we have

$$(3.1) \qquad \mathcal{P}(\{i\}|\mathcal{M}) = 0.5 \operatorname{tr}\{\mathbf{x}_i^T R_i(\hat{\boldsymbol{\alpha}})^{-1} \mathbf{x}_i E_\beta[(\boldsymbol{\beta} - \boldsymbol{\beta}_*)(\boldsymbol{\beta} - \boldsymbol{\beta}_*)^T]\},$$

where E_β is taken with respect to $p(\boldsymbol{\beta}|\boldsymbol{\beta}_*, G_{n\beta}^{-1})$ and $R_i(\boldsymbol{\alpha}) = \sigma_y^2 \mathbf{I}_{m_i} + \sigma_b^2 \mathbf{1}_{m_i}^{\otimes 2}$, in which $\boldsymbol{\alpha} = (\sigma_b^2, \sigma_y^2)^T$ and $\mathbf{1}_{m_i}$ is an $m_i \times 1$ vector with all elements equal to one. We set $G_{n\beta}^{-1} = [\sum_{i=1}^n \mathbf{x}_i^T R_i(\hat{\boldsymbol{\alpha}})^{-1} \mathbf{x}_i]^{-1}$ and substituted $\boldsymbol{\beta}_*$ by $\hat{\boldsymbol{\beta}}$.

We carried out three experiments as follows. The first experiment was to evaluate the accuracy of the first-order approximation to $\mathrm{CD}(I)$. The explicit expression of $\widetilde{\mathrm{CD}}(I)$ is given in Example S2 of the supplementary document. We considered two scenarios. In the first scenario, we directly simulated 100 datasets from 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 $(m_1, b_1) = (1, 4)$ and $(m_n, b_n) = (5, 3)$ to generate $y_{i,j}$ for $i = 1, n$ and all j according to the above linear mixed model. Thus, the new first and nth clusters can be regarded as influential clusters due to the extreme values of b_1 and b_n. Moreover, the number of observations in these two clusters is unbalanced. We calculated $\mathrm{CD}(I)$ and $\widetilde{\mathrm{CD}}(I)$, the average $\mathrm{CD}(I)$, and the biases and standard errors of the differences $\mathrm{CD}(I) - \widetilde{\mathrm{CD}}(I)$ for each cluster $\{i\}$ (Table 1).

Inspecting Table 1 reveals three findings as follows. First, when no influential cluster is present in the first scenario, the average $\mathrm{CD}(I)$ is an increasing function

TABLE 1

*Selected results from simulation studies for $n = 12$ and the two scenarios: m_i, $\mathcal{P}(\{i\}|\mathcal{M})$, M, SD, Mdif $(\times 10^{-2})$ and SDdif $(\times 10^{-1})$ of the three quantities CD(I), E[CD(I)|$\mathcal{M}$, **Z**] and Std[CD(I)|$\mathcal{M}$, **Z**]. m_i denotes the cluster size of subset $\{i\}$; $\mathcal{P}(\{i\}|\mathcal{M})$ denotes the degree of perturbation; 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, two clusters were influential clusters and highlighted in bold. For each case, 100 simulated datasets were used. Results were sorted according to the degree of perturbation for each cluster*

	Scenario I					Scenario II				
m_i	$\mathcal{P}(\{i\}|\mathcal{M})$	M	Mdif	SD	SDdif	m_i	$\mathcal{P}(\{i\}|\mathcal{M})$	M	Mdif	SD	SDdif
					CD(*I*)						
1	0.10	0.11	0.01	0.09	0.03	**1**	**0.08**	**0.37**	**1.01**	**0.18**	**0.18**
2	0.11	0.12	0.32	0.12	0.15	2	0.11	0.10	0.08	0.09	0.12
2	0.11	0.15	1.24	0.18	0.64	1	0.11	0.08	0.02	0.11	0.02
2	0.13	0.18	0.87	0.19	0.36	2	0.13	0.13	0.08	0.12	0.12
2	0.15	0.17	0.25	0.19	0.20	2	0.16	0.13	−0.13	0.12	0.08
3	0.16	0.23	0.55	0.19	0.50	2	0.20	0.20	0.08	0.19	0.12
2	0.19	0.26	−0.02	0.32	0.25	3	0.23	0.21	−0.06	0.18	0.22
3	0.22	0.34	2.97	0.35	0.99	4	0.25	0.23	0.37	0.23	0.26
4	0.27	0.41	3.35	0.38	1.77	**5**	**0.28**	**0.78**	**18.59**	**0.61**	**4.71**
5	0.40	0.70	5.43	0.60	1.90	5	0.37	0.38	0.90	0.32	0.46
4	0.57	1.15	1.57	1.29	1.73	5	0.54	0.70	1.32	0.68	0.82
5	0.60	1.21	3.62	1.49	1.62	4	0.56	0.65	1.06	0.69	0.54
					E[CD(*I*)|$\mathcal{M}$, **Z**]						
1	0.10	0.12	0.22	0.02	0.05	**1**	**0.08**	**0.09**	**0.43**	**0.01**	**0.04**
2	0.11	0.12	0.41	0.01	0.03	2	0.11	0.12	0.45	0.02	0.04
2	0.11	0.13	0.46	0.02	0.04	1	0.11	0.13	0.09	0.02	0.03
2	0.12	0.15	0.40	0.02	0.07	2	0.13	0.15	0.38	0.02	0.04
2	0.15	0.17	0.34	0.03	0.08	2	0.16	0.18	0.26	0.02	0.04
3	0.16	0.18	0.77	0.02	0.08	2	0.20	0.23	0.12	0.03	0.05
2	0.19	0.22	0.21	0.04	0.09	3	0.23	0.27	0.46	0.03	0.07
3	0.22	0.26	0.62	0.04	0.09	4	0.25	0.29	1.13	0.03	0.13
4	0.26	0.32	1.63	0.03	0.15	**5**	**0.28**	**0.36**	**1.94**	**0.04**	**0.18**
5	0.40	0.55	2.58	0.07	0.29	5	0.37	0.48	1.86	0.05	0.18
4	0.57	0.97	2.21	0.12	0.21	5	0.53	0.82	4.26	0.10	0.34
5	0.60	1.03	5.87	0.16	0.99	4	0.56	0.93	1.64	0.11	0.17
					Std[CD(*I*)|$\mathcal{M}$, **Z**]						
1	0.10	0.18	1.48	0.04	0.20	**1**	**0.08**	**0.13**	**1.05**	**0.04**	**0.22**
2	0.11	0.14	1.16	0.03	0.10	2	0.11	0.14	1.18	0.03	0.12
2	0.11	0.15	1.37	0.03	0.16	1	0.11	0.18	0.78	0.04	0.10
2	0.13	0.18	1.72	0.05	0.35	2	0.13	0.18	1.15	0.03	0.13
2	0.15	0.21	2.02	0.05	0.25	2	0.16	0.23	1.28	0.04	0.14
3	0.16	0.19	2.05	0.03	0.25	2	0.20	0.30	1.07	0.06	0.16
2	0.19	0.29	2.36	0.07	0.24	3	0.23	0.31	1.72	0.06	0.22
3	0.22	0.30	2.55	0.07	0.32	4	0.25	0.30	1.96	0.05	0.42
4	0.26	0.35	2.84	0.06	0.39	**5**	**0.28**	**0.39**	**4.06**	**0.09**	**0.66**
5	0.40	0.58	2.13	0.11	0.71	5	0.37	0.50	2.67	0.09	0.52
4	0.57	1.16	1.17	0.18	0.55	5	0.53	0.89	0.60	0.14	0.68
5	0.60	1.14	−4.18	0.25	2.29	4	0.56	1.13	0.94	0.21	0.41

of $\mathcal{P}(I|\mathcal{M})$, whereas it is only positively proportional to the cluster size $n(I)$ with a correlation coefficient of 0.83. This result agrees with the results of Proposition 1. Second, in the second scenario, the average $CD(I)$ for the true "good" clusters is positively proportional to $\mathcal{P}(I|\mathcal{M})$ with a correlation coefficient of 0.76, while that for the influential clusters is associated with both $\mathcal{P}(I|\mathcal{M})$ and the amount of influence that we introduced. Third, 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, $\widetilde{CD}(I)$ is relatively close to $CD(I)$. For instance, for cluster $\{n\}$, the bias of 0.19 is relatively small compared with 0.78, the mean of $CD(\{n\})$.

In the second experiment, we considered the same two scenarios as the first experiment. Specifically, for each dataset, we approximated $E[CD(I)|\mathcal{M}, \mathbf{Z}]$ and $Std[CD(I)|\mathcal{M}, \mathbf{Z}]$ by setting $S = 200$ and using their empirical values, and calculated their first approximations $\widehat{M}[\widetilde{CD}(I)]$ and $\widehat{Std}[\widetilde{CD}(I)]$. Across all 100 data sets, for each cluster I, we computed the averages of $E[CD(I)|\mathcal{M}, \mathbf{Z}]$ and $Std[CD(I)|\mathcal{M}, \mathbf{Z}]$, and the biases and standard errors of the differences $E[CD(I)|\mathcal{M}, \mathbf{Z}] - \widehat{M}[\widetilde{CD}(I)]$ and $Std[CD(I)|\mathcal{M}, \mathbf{Z}] - \widehat{Std}[\widetilde{CD}(I)]$.

Table 1 shows the results for each scenario. First, in both scenarios, the average $E[CD(I)|\mathcal{M}, \mathbf{Z}]$ is an increasing function of $\mathcal{P}(I|\mathcal{M})$, whereas it is only positively proportional to the cluster size $n(I)$ with a correlation coefficient (CC) of 0.80. This is in agreement with the results of Proposition 1. The average of $Std[CD(I)|\mathcal{M}, \mathbf{Z}]$ are positively proportional to m_i (CC $= 0.76$) and $\mathcal{P}(I|\mathcal{M})$ (CC $= 0.99$). Second, for all clusters, the first-order approximations of $E[CD(I)|\mathcal{M}, \mathbf{Z}]$ and $Std[CD(I)|\mathcal{M}, \mathbf{Z}]$ are very accurate and lead to small average biases and standard errors.

The third experiment was to examine the finite sample performance of Cook's distance and the scaled Cook's distances for detecting influential clusters 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 0.6 to 6.0 to generate $y_{n,j}$ according to the above linear mixed model. The second scenario is almost the same as the first scenario, except that we reset $m_n = 10$. Note that when the value of b_n is relatively large, for example, $b_n = 2.5$, the nth cluster is an influential cluster, whereas the nth cluster is not influential for small b_n. A good case-deletion measure should detect the nth cluster as truly influential for large b_n, whereas it does not for small b_n. For each data set, we approximated $CSCD_1(I, \mathbf{Z})$, $CSCD_2(I, \mathbf{Z})$, $\widetilde{CSCD}_1(I, \mathbf{Z})$ and $\widetilde{CSCD}_2(I, \mathbf{Z})$ by setting $S = 100$. Subsequently, we calculated $P_A(I, \mathbf{Z})$ and $P_B(I, \mathbf{Z})$ in (2.15) and $P_C(I, \mathbf{Z}) = \sum_{I \neq \{n\}} \mathbf{1}(CD(I) \leq CD(\{n\}))/(n-1)$. Finally, across all 100 datasets, we calculated the averages and standard errors of all diagnostic measures for the nth cluster for each scenario.

Inspecting Figure 1 reveals some findings as follows. First, deleting the nth cluster with 10 observations causes a larger effect than that with 1 observation
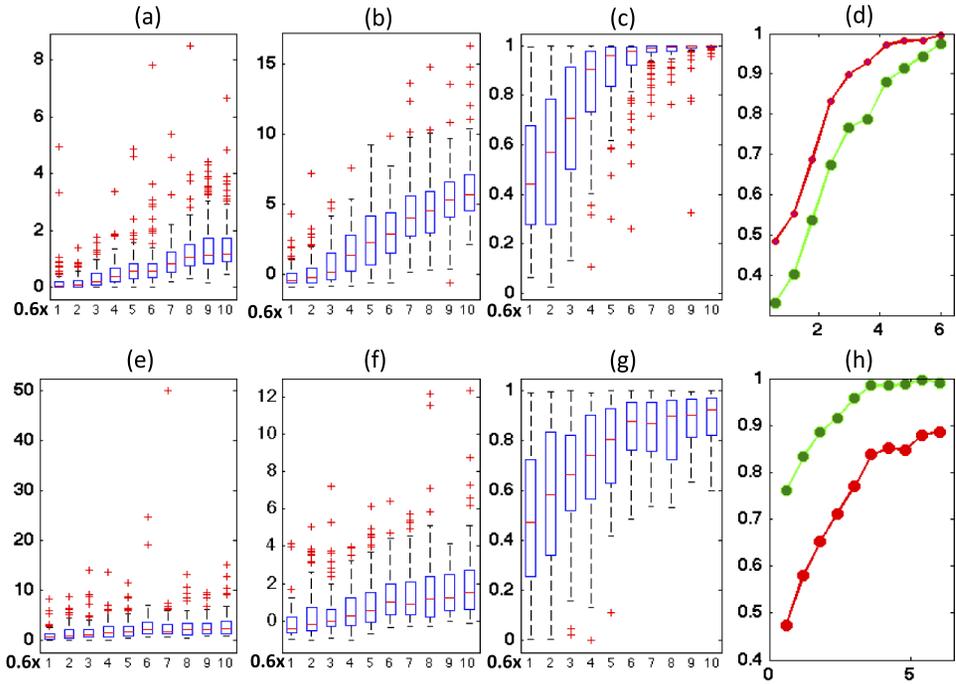
FIG. 1. *Simulation results from* 100 *datasets simulated from a linear mixed model in the two scenarios. The first row corresponds to the first scenario, in which* $m_{12} = 1$ *and* b_{12} *varies from* 0.6 *to* 6.0. *The second row corresponds to the second scenario, in which* $m_{12} = 10$ *and* b_{12} *varies from* 0.6 *to* 6.0. *Panels* (a) *and* (e) *show the box plots of Cook's distances as a function of* b_{12}; *panels* (b) *and* (f) *show the box plots of* $\mathrm{CSCD}_1(I, \mathbf{Z})$ *as a function of* b_{12}; *panels* (c) *and* (g) *show the box plots of* $P_B(I, \mathbf{Z})$ *as a function of* b_{12}; *panels* (d) *and* (h) *show the mean curve of* $P_B(I, \mathbf{Z})$ *based on* $\mathrm{CSCD}_1(I, \mathbf{Z})$ *(red line) and the mean curve of* $P_C(I, \mathbf{Z})$ *based on* $\mathrm{CD}(I)$ *(green line) as functions of* b_{12}.

[Figure 1(a) and (e), (d) and (h)]. As expected, the distributions of $\mathrm{CD}(\{n\})$ and $\widetilde{\mathrm{CSCD}}_1(I, \mathbf{Z})$ shift up as b_n increases [Figure 1(a), (b), (e) and (f)]. Second, in the first scenario, $\mathrm{CD}(\{n\})$ is stochastically smaller than most other $\mathrm{CD}(I)$s, when the value of b_n is relatively small [Figure 1(d)]. However, in the second scenario, $\mathrm{CD}(\{n\})$ is stochastically larger than most other $\mathrm{CD}(I)$s [Figure 1(h)] even for small values of b_n. Specifically, when $m_n = 1$, the average $P_C(\{n\}, \mathbf{Z})$ is smaller than 0.4 as $b_n = 0.6$ and $b_n = 1.2$, whereas when $m_n = 10$, the average $P_C(\{n\}, \mathbf{Z})$ is higher than 0.75 even as $b_n = 0.6$. In contrast, in the two scenarios, the value of $P_B(\{n\}, \mathbf{Z})$ is close to 0.5 as $b_n = 0.6$ [Figure 1(d) and (h)]. It indicates that the cluster size does not have a big effect on the distribution of $\widetilde{\mathrm{CSCD}}_1(I, \mathbf{Z})$ [Figure 1(c) and (g)].

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). One feature of this dataset is that the number of observations per children m_i varies significantly from 2 to 30 [21, 22]. The total number of data points is $\sum_{i=1}^{n} m_i = 3176$. Following Zhang [26], we considered two linear mixed models given by $y_{i,j} = \mathbf{x}_{i,j}^T \boldsymbol{\beta} + \varepsilon_{i,j}$, where $y_{i,j}$ is the weight (in kilograms) of the jth visit from the ith subject, $\mathbf{x}_{i,j} = (1, d_{i,j}, (d_{i,j} - 120)^+, (d_{i,j} - 200)^+, (g_i - 28)^+, d_{i,j}(g_i - 28)^+, (d_{i,j} - 60)^+(g_i - 28)^+, (d_{i,j} - 490)^+(g_i - 28)^+, s_i d_{i,j}, s_i(d_{i,j} - 120)^+)^T$, in which $d_{i,j}$ and g_i (days) are the age of visit and gestational age, respectively, and s_i is the indicator for gender. In addition, we assumed $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,m_i})^T \sim N_{m_i}(\mathbf{0}, R_i(\boldsymbol{\alpha}))$, where $\boldsymbol{\alpha}$ is a vector of unknown parameters in $R_i(\boldsymbol{\alpha})$. We first considered $R_i(\boldsymbol{\alpha}) = \alpha_0 \mathbf{I}_{m_i} + \alpha_1 \mathbf{1}_{m_i}^{\otimes 2}$. We refer to this model as model M_1. Then, it is assumed that variance and autocorrelation parameters are, respectively, given by $V(d) = \exp(\alpha_0 + \alpha_1 d + \alpha_2 d^2 + \alpha_3 d^3)$ and $\rho(l) = \alpha_4 + \alpha_5 l$, where l is the lag between two visits. We refer to this model as model M_2.

We systematically examined the key assumptions of models M_1 and M_2 as follows.

(i) We presented a cumulative residual plot and calculated the cumulative sums of residuals over the age of the visit to test $E[y_{i,j}|\mathbf{x}_{i,j}] = \mathbf{x}_{i,j}^T \boldsymbol{\beta}$ [17], whose p-value is greater than 0.543. It may suggest that the mean structure is reasonable. The cumulative residual plot is given in Figure 2(b).

(ii) For model M_1, inspecting the plot of raw residuals $r_{i,j} = y_{i,j} - \mathbf{x}_{i,j}^T \hat{\boldsymbol{\beta}}$ against age in Figure 2(c) reveals that the variance of the raw residuals appears to increase with the age of visit. As pointed by Zhang [26], it may be more sensible to use model M_2. Let $\tilde{\mathbf{r}}_i = (\tilde{r}_{i,1}, \ldots, \tilde{r}_{i,m_i})^T = R_i(\hat{\boldsymbol{\alpha}})^{-1/2} \mathbf{r}_i$ be the vector of standardized residuals of M_2, where $\mathbf{r}_i = (r_{i,1}, \ldots, r_{i,m_i})^T$. The standardized residuals under M_2 do not have any apparent structure as age increases [Figure 2(d)].

(iii) Under each model, we calculated $\mathrm{CD}(I)$ for each child [4]. We treated $\boldsymbol{\beta}$ as parameters of interest and all elements of $\boldsymbol{\alpha}$ as nuisance parameters. For model M_1, 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 4(b) highlights the top ten influential subjects. Compared with model M_1, we observed similar findings by using $\mathrm{CD}(I)$ under model M_2, which were omitted for space limitations.

There are several difficulties in using Cook's distance under both models M_1 and M_2 [3, 4, 7, 19]. First, cluster size varies significantly across children, and deleting a larger cluster may have a higher probability of having a larger influence as discussed in Section 2.3. For instance, we observe $(m_{285}, \mathrm{CD}(\{285\})) = (8, 0.738)$ and $(m_{274}, \mathrm{CD}(\{274\})) = (22, 1.163)$. A larger $\mathrm{CD}(\{274\})$ can be caused by a larger $m_{274} = 22$ and/or influential subject 274, among others. Since m_{274} is much larger than m_{285}, it is difficult to claim that subject 274 is more influential than subject 285. Second, there is no rule for determining whether a
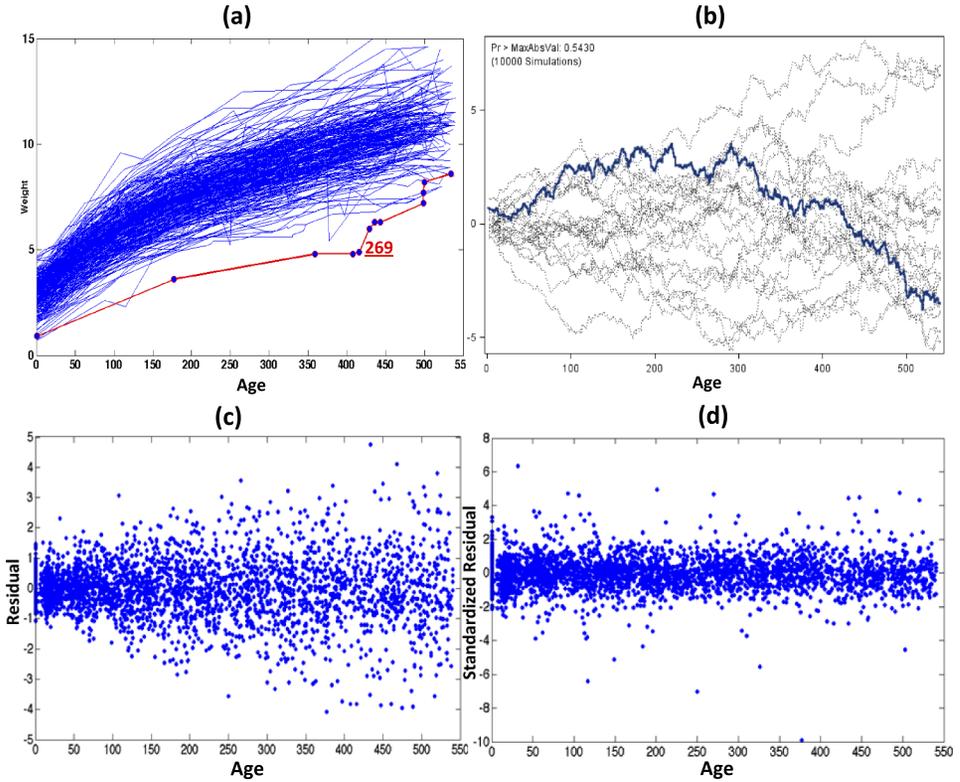
FIG. 2.   *Yale infant growth data. Panel* (a) *presents the line plot of infant weight against age, in which the observations of subject* 269 *are highlighted*; *panel* (b) *shows the cumulative residual curve versus age, in which the observed cumulative residual curve is highlighted in blue*; *and panels* (c) *and* (d), *respectively, present age versus raw residual and age versus standardized residual for cluster deletion.*

specific subject is influential relative to the fitted model. Specifically, it is unclear whether the subjects with larger CD($\{i\}$) are truly influential or not. Third, inspecting Cook's distance solely does not seem to delineate the potential misspecification of the covariance structure under model M_1. We will address these three difficulties by using the new case-deletion measures.

(iv) Under each model, we calculated $\mathcal{P}(\{i\}|\mathcal{M})$ for deleting each subject $\{i\}$ for fixed covariates, and then we calculated the conditionally scaled Cook's distances and associated quantities. We then used 1000 bootstrap samples to approximate $\text{CSCD}_1(I, \mathbf{Z})$ and $\text{CSCD}_2(I, \mathbf{Z})$. Subsequently, we calculated $P_A(I, \mathbf{Z})$ and $P_B(I, \mathbf{Z})$ in (2.15).

We observed several findings. First, under model M_1, we observed a strong positive correlation between $\mathcal{P}(\{i\}|\mathcal{M})$ and m_i [Figure 3(a)]. Second, even though $m_{269} = 12$ is moderate, subject 269 has the largest degree of perturbation. Inspecting the raw data in Figure 2(a) reveals that subject 269 is of older age during visits
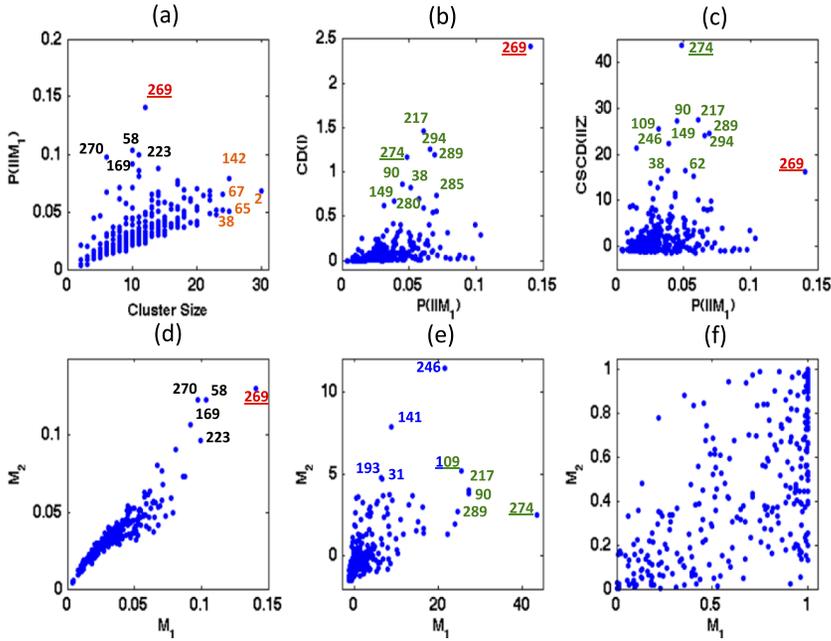
FIG. 3. *Yale infant growth data. Panel* (a) *shows* m_i *versus* $\mathcal{P}(I|\mathcal{M}_1)$, *in which the ten subjects with the largest values of degree of perturbation or cluster size are highlighted; panel* (b) *shows* $\mathcal{P}(I|\mathcal{M}_1)$ *versus* $\mathrm{CD}(I)$, *in which the top ten influential subjects are highlighted; panel* (c) *shows* $\mathcal{P}(I|\mathcal{M}_1)$ *versus* $\mathrm{CSCD}_1(I, \mathbf{Z})$, *in which the top eleven influential subjects are highlighted; and panels* (d), (e) *and* (f), *respectively, show* $\mathcal{P}(I|\mathcal{M})$, $\mathrm{CSCD}_1(I, \mathbf{Z})$ *and* $P_B(I, \mathbf{Z})$ *for models* $\mathcal{M}_1$ *and* $\mathcal{M}_2$.

compared with other subjects. Third, we also observed a strong positive correlation between $\mathcal{P}(\{i\}|\mathcal{M})$ and Cook's distance [Figure 3(b)], which may indicate their stochastic relationship as discussed in Section 2.3. Fourth, we observed a positive correlation between Cook's distance and the conditionally scaled Cook's distance [Figure 3(b) and (c)], but their levels of influence for the same subject are quite different. For instance, the magnitude of $\mathrm{CSCD}_1(\{269\}, \mathbf{Z})$ is only moderate, whereas $\mathrm{CD}_1(\{269\}, \mathbf{Z})$ is the highest one. We observed similar findings under model $\mathcal{M}_2$ and presented some findings in Figure 3(d) and (e).

We used $P_B(I, \mathbf{Z})$ to quantify whether a specific subject is influential relative to the fitted model $\mathcal{M}_1$ [Figure 3(f)]. For instance, since $\mathrm{CD}(\{246\}) = 0.253$, it is unclear whether subject 246 is influential or not according to CD, whereas we have $\mathrm{CSCD}_1(\{246\}, \mathbf{Z}) = 21.443$ and $P_B(\{246\}, \mathbf{Z}) = 1.0$. Thus, subject 246 is really influential after eliminating the effect of the cluster size. Moreover, it is difficult to compare the influential levels of subjects 274 and 285 using CD. All of the conditionally scaled Cook's distances and associated quantities suggest that subject 274 is more influential than subject 285 after eliminating the degree of perturbation difference. We observed similar findings under model $\mathcal{M}_2$ and omitted them due to space limitations. See Figure 3(d) and (e) for details.

We compared the goodness of fit of models $\mathcal{M}_1$ and $\mathcal{M}_2$ to the data by using the proposed case-deletion measures. First, inspecting Figure 3(d) reveals a strong similarity between the degrees of perturbation under models $\mathcal{M}_1$ and $\mathcal{M}_2$ for all subjects. Second, by using the conditionally scaled Cook's distance, we observed different levels of influence for the same subject under $\mathcal{M}_1$ and $\mathcal{M}_2$. For instance, $\text{CSCD}_1(I, \mathbf{Z})$ identifies subjects 246, 141, 109, 193 and 31 as the top five influential subjects under $\mathcal{M}_1$, whereas it identifies subjects 274, 217, 90, 109 and 289 as the top ones under $\mathcal{M}_2$. Finally, examining $P_B(I, \mathbf{Z})$ reveals a large percentage of influential points for model $\mathcal{M}_1$, but a small percentage of influential points for model $\mathcal{M}_2$; see Figure 3(f) for details. This may indicate that model $\mathcal{M}_2$ outperforms model $\mathcal{M}_1$. Furthermore, although we may develop goodness-of-fit statistics based on the scaled Cook's distances and show that model $\mathcal{M}_2$ outperforms model $\mathcal{M}_1$, this will be a topic of our future research.

In summary, the use of the new case-deletion measures provides new insights in real data analysis. First, $\mathcal{P}(I|\mathcal{M})$ explicitly quantifies the degree of perturbation introduced by deleting each subject. Second, $\text{CSCD}_k(I, \mathbf{Z})$ for $k = 1, 2$ explicitly account for the degree of perturbation for each subject. Third, $P_B(I, \mathbf{Z})$ allows us to quantify whether a specific subject is influential relative to the fitted model. Fourth, inspecting $P_B(I, \mathbf{Z})$ and $\text{CSCD}_k(I, \mathbf{Z})$ may delineate the potential misspecification of the covariance structure under model M_1.

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 magnitude of Cook's distance. We have developed several scaled Cook's distances to address the fundamental issue of deletion diagnostics in general parametric models. We have shown that the scaled Cook's distances provide important information about the relative influential level of each subset. Future work includes developing goodness-of-fit statistics based on the scaled Cook's distances, developing Bayesian analogs to the scaled Cook's distances, and developing user-friendly R code for implementing our proposed measures in various 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 $\boldsymbol{\theta}_* \in \Theta$.

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 $\mathbf{F}_{n,[I]}(\boldsymbol{\theta}_*) = \partial_\theta^2 \log p(\mathbf{Y}_{[I]}|\boldsymbol{\theta})$.

ASSUMPTION A4. For any 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 $0 < \inf_{\theta \in B(\theta_*, \delta_0 n^{-1/2})} \lambda_{\min}(n^{-1}\mathbf{F}_{n,[I]}(\boldsymbol{\theta})) \le \sup_{\theta \in B(\theta_*, \delta_0 n^{-1/2})} \lambda_{\max}(n^{-1} \times \mathbf{F}_{n,[I]}(\boldsymbol{\theta})) < \infty$.

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

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

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

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

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 [2]. Assumption A2 assumes that the parameter estimates 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 [2, 30]. Sufficient conditions of Assumptions A2–A5 have been extensively discussed in the literature [2, 30].

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

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

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

(A.1)
$$\begin{aligned} \mathcal{P}(I_1|\mathcal{M}) = {}& \mathcal{P}(I_2|\mathcal{M}) \\ & + \int p(\boldsymbol{\theta}|\boldsymbol{\theta}_*, \Sigma_{n*}) p(\mathbf{Y}|\boldsymbol{\theta}) \log\left(\frac{p(\mathbf{Y}_{[I_2]}|\mathbf{Y}_{[I_1]}, \boldsymbol{\theta})}{p(\mathbf{Y}_{[I_2]}|\mathbf{Y}_{[I_1]}, \boldsymbol{\theta}_*)} \right) d\boldsymbol{\theta} \, d\mathbf{Y}, \end{aligned}$$

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

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

which yields (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 (A.1) reduces to $\mathcal{P}(I_{1 \cdot 2}|\mathcal{M})$, which finishes the proof of (P.d). $\quad\square$

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 $\lambda_{1,1} \geq \cdots \geq \lambda_{1,n(I_1)} \geq 0$ and $\lambda_{2,1} \geq \cdots \geq \lambda_{2,n(I_2)} \geq 0$ be the 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 [13] 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 $\Lambda_k = \mathrm{diag}(\lambda_{k,1}, \ldots, \lambda_{k,n(I_k)})$. It can be shown that for $k = 1, 2$,

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

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 $CD(I) = (\hat{\sigma}^2)^{-1} \sum_{j=1}^{n(I)} \lambda_j (1 - \lambda_j)^{-1} h_j^2$, where the 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 $CD(I)$ and $CD(I')$ 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). $\square$

PROOF OF THEOREM 3. (a) It follows from a Taylor 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

$$\hat{\boldsymbol{\theta}}_{[I]} - \hat{\boldsymbol{\theta}} = \left[-\partial_\theta^2 \log p(\mathbf{Y}_{[I]}|\hat{\boldsymbol{\theta}}) \right]^{-1} \partial_\theta \log p(\mathbf{Y}_{[I]}|\hat{\boldsymbol{\theta}})[1 + o_p(1)]$$

$$(A.2) \qquad = -\left[-\partial_\theta^2 \log p(\mathbf{Y}_{[I]}|\hat{\boldsymbol{\theta}}) \right]^{-1} \partial_\theta \log p(\mathbf{Y}_I|\mathbf{Y}_{[I]}, \hat{\boldsymbol{\theta}})[1 + o_p(1)].$$

Substituting (A.2) into $CD(I) = (\hat{\boldsymbol{\theta}}_{[I]} - \hat{\boldsymbol{\theta}})^T \mathbf{F}_n(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{[I]} - \hat{\boldsymbol{\theta}})$ completes the proof of Theorem 3(a).

(b) It follows from Assumptions A2–A4 that

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_* = \mathbf{F}_n(\boldsymbol{\theta}_*)^{-1} \partial_\theta \log p(\mathbf{Y}|\boldsymbol{\theta}_*)[1 + o_p(1)]$$

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

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

$$J_I(\hat{\boldsymbol{\theta}}) = J_I(\boldsymbol{\theta}_*) - \mathbf{s}_I(\boldsymbol{\theta}_*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*)[1 + o_p(1)]$$

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

$$(A.3) \qquad = \left(\{ \mathbf{I}_p - E[\mathbf{s}_I(\boldsymbol{\theta})|\mathcal{M}] \mathbf{F}_n(\boldsymbol{\theta}_*)^{-1} \} J_I(\boldsymbol{\theta}_*) \right.$$

$$\left. - E[\mathbf{s}_I(\boldsymbol{\theta})|\mathcal{M}] \mathbf{F}_n(\boldsymbol{\theta}_*)^{-1} \partial_\theta \log p(\mathbf{Y}_{[I]}|\boldsymbol{\theta}_*) \right)[1 + o_p(1)].$$

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 3(b).

(c) Similarly to Theorem 3(b), we can prove Theorem 3(c). $\square$

SUPPLEMENTARY MATERIAL

REFERENCES

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

[2] ANDREWS, D. W. K. (1999). Estimation when a parameter is on a boundary. *Econometrica* **67** 1341–1383. MR1720781

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

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

[5] BECKMAN, R. J. and COOK, R. D. (1983). Outlier…s. *Technometrics* **25** 119–163. MR0702168

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

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

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

[9] COOK, R. D. (1986). Assessment of local influence. *J. Roy. Statist. Soc. Ser. B* **48** 133–169. MR0867994

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

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

[12] DAVISON, A. C. and TSAI, C. L. (1992). Regression model diagnostics. *International Statistical Review* **60** 337–353.

[13] 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. *Ann. Statist.* **19** 260–271. MR1091849

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

[15] HASLETT, J. (1999). A simple derivation of deletion diagnostic results for the general linear model with correlated errors. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 603–609. MR1707863

[16] HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York. MR0606374

[17] LIN, D. Y., WEI, L. J. and YING, Z. (2002). Model-checking techniques based on cumulative residuals. *Biometrics* **58** 1–12. MR1891037

[18] MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman & Hall/CRC, Boca Raton.

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

[20] SHAKED, M. and SHANTHIKUMAR, G. J. (2006). *Stochastic Orders*. Springer, New York.

[21] 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.

[22] WASSERMAN, D. R. and LEVENTHAL, J. M. (1993). Maltreatment of children born to cocaine-dependent mothers. *Am. J. Dis. Child.* **147** 1324–1328.

[23] WEI, B.-C. (1998). *Exponential Family Nonlinear Models*. *Lecture Notes in Statist.* **130**. Springer, Singapore. MR1646023

[24] WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163

[25] WHITE, H. (1994). *Estimation*, *Inference and Specification Analysis*. *Econometric Society Monographs* **22**. Cambridge Univ. Press, Cambridge. MR1292251

[26] ZHANG, H. (1999). Analysis of infant growth curves using multivariate adaptive splines. *Biometrics* **55** 452–459.

[27] ZHU, H. and IBRAHIM, J. G. (2011). Supplement to "Perturbation and scaled Cook's distance." DOI:10.1214/12-AOS978SUPP.

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

[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. MR1859405

[30] ZHU, H. and ZHANG, H. (2006). Asymptotics for estimation and testing procedures under loss of identifiability. *J. Multivariate Anal.* **97** 19–45. MR2208842

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
CHAPEL HILL, NORTH CAROLINA 27599-7420
USA
E-MAIL: hzhu@bios.unc.edu
        ibrahim@bios.unc.edu
        hyunsoon.cho@nih.gov