Supplementary Document of "Perturbation and Scaled Cook's Distances"

Hongtu Zhu, Joseph G. Ibrahim and Hyunsoon Cho

Abstract

We investigate two theoretical examples on generalized linear models and linear mixed models to illustrate the calculation of scaled Cook's distances. We also include additional results obtained from the Monte Carlo simulation studies and real data analysis.

1 Theoretical Examples

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

Example S1. We consider Cook's distance in generalized linear models (McCullagh and Nelder, 1989) as follows. Suppose that the components of $\mathbf{y} = (y_1, \ldots, y_n)^T$ given $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$ are mutually independent, and the conditional density of y_i given $\mathbf{x}_i$ is given by

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

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$ are known functions, $\eta_i = \eta(\mu_i)$ and $\mu_i = \mu_i(\boldsymbol{\beta}) = g(\mathbf{x}_i^T\boldsymbol{\beta})$, in which $g(\cdot)$ is a known monotonic function and twice continuously differentiable and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. Throughout this example, the parameter of interest is $\boldsymbol{\beta}$ and τ is a nuisance parameter and is fixed at $\hat{\tau}$. We define

$$V(\boldsymbol{\beta}) = \text{diag}(\ddot{b}(\eta(\mu_1(\boldsymbol{\beta}))), \ldots, \ddot{b}(\eta(\mu_n(\boldsymbol{\beta})))) \quad \text{and} \quad D(\boldsymbol{\beta})^T = (\partial_\beta\mu_1(\boldsymbol{\beta}), \ldots, \partial_\beta\mu_n(\boldsymbol{\beta})),$$

where ∂_β denotes differentiation with respect to $\boldsymbol{\beta}$ and $\ddot{b}(\eta)$ denotes the second derivative of $b(\eta)$ with respect to η.

We first compute the degree of the perturbation for deleting each $(y_i, \mathbf{x}_i)$ for the case of fixed covariates. Since $\mathcal{M}$ assumes (1). it follows from the Taylor's series expansion that

$$\begin{aligned}\mathcal{P}(\{i\}|\mathcal{M}) &= E_\theta\{g(\mathbf{x}_i^T\boldsymbol{\beta})[a(\tau)^{-1}\eta_i(\boldsymbol{\beta}) - a(\tau_*)^{-1}\eta_i(\boldsymbol{\beta}_*)]\} + E_\theta E_{y_i,\theta}[c(y_i,\tau) - c(y_i,\tau_*)] \\ &\approx \frac{1}{2}\ddot{b}(\eta(\mu_n(\boldsymbol{\beta}_*)))\partial_\beta\eta(\mu_i(\boldsymbol{\beta}_*))^T[\sum_{i=1}^n \ddot{b}(\eta(\mu_n(\boldsymbol{\beta}_*)))\partial_\beta\eta(\mu_i(\boldsymbol{\beta}_*))^{\otimes 2}]^{-1}\partial_\beta\eta(\mu_i(\boldsymbol{\beta}_*)) \\ &+ \frac{1}{2}K_i(\boldsymbol{\theta}_*)[\sum_{i=1}^n K_i(\boldsymbol{\theta}_*)]^{-1}, \qquad (2)\end{aligned}$$

where E_θ is taken with respect to $p(\boldsymbol{\theta}|\boldsymbol{\theta}_*, G_{n\theta}^{-1})$ and $E_{y_i,\theta}$ is taken with respect to $p(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \tau)$ in (1). Moreover, $K_i(\boldsymbol{\theta})$ is defined as

$$[2\dot{a}(\tau)a(\tau)^{-1} - \ddot{a}(\tau)\dot{a}(\tau)^{-1}]E_{y_i,\theta}[\dot{c}(y_i,\tau)] + E_{y_i,\theta}[\ddot{c}(y_i,\tau)].$$

If we are only interested in $\boldsymbol{\beta}$ and treat τ as a nuisance parameter, $0.5K_i(\boldsymbol{\theta}_*)[\sum_{i=1}^n K_i(\boldsymbol{\theta}_*)]^{-1}$ can be dropped from $\mathcal{P}(\{i\}|\mathcal{M})$ in (2).

Following the derivations in (Williams, 1987; Wei, 1998), we can show that Cook's distance for deleting subset I with size$(I) = n(I)$ can be approximated by

$$\widetilde{\mathrm{CD}}(I) = \frac{1}{a(\hat{\tau})}\hat{\mathbf{e}}^T\hat{V}^{-1/2}U_I(\mathbf{I}_{n(I)} - \hat{H}_I)^{-1}\hat{H}_I(\mathbf{I}_{n(I)} - \hat{H}_I)^{-1}U_I^T\hat{V}^{-1/2}\hat{\mathbf{e}}, \qquad (3)$$

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

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

Theorem S1. *Suppose that Assumptions A2-A5 in the appendix hold for the generalized linear model (1). We have the following results:*

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

$$\mathbf{W}_* = V_*^{-1/2}(\mathbf{I}_n - H_*)U_I(\mathbf{I}_{n(I)} - H_{*,I})^{-1}H_{*,I}(\mathbf{I}_{n(I)} - H_{*,I})^{-1}U_I^T(\mathbf{I}_n - H_*)V_*^{-1/2}, \quad (4)$$

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

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

$$\begin{aligned} E[CD_*(I)|\mathcal{M}] &= E\{tr[(\mathbf{I}_{n(I)} - H_{*,I})^{-1}]|\mathcal{M}\} - n(I) = \sum_{j=1}^{n(I)} E[(1-\lambda_{I,j})^{-1}|\mathcal{M}] - n(I), \\ Var[CD_*(I)|\mathcal{M}] &= a(\tau_*)\sum_{i=1}^{n} w_{ii*}b^{(4)}(\eta_{i*}) + Var\{tr[(\mathbf{I}_{n(I)} - H_{*I})^{-1}]|\mathcal{M}\} \quad (5) \\ &+ 2E\{tr[(\mathbf{I}_{n(I)} - H_{*,I})^{-2}]|\mathcal{M}\} - 4E\{tr[(\mathbf{I}_{n(I)} - H_{*,I})^{-1}]|\mathcal{M}\} + 2n(I), \end{aligned}$$

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

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

Proof of Theorem S1. (a). Let $\mu(\boldsymbol{\beta}) = (\mu_1(\boldsymbol{\beta}), \ldots, \mu_n(\boldsymbol{\beta}))^T$. If the model $\mathcal{M}$ is true, then $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*) = (D_*^T V_*^{-1} D_*)^{-1} D_*^T V_*^{-1}\mathbf{e}_* + o_p(n^{-1/2})$. Thus, under Assumptions A2-A5, we have

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

where $\mathbf{e}_* = \mathbf{y} - \mu(\boldsymbol{\beta}_*)$. This yields Theorem S1 (a).

(b). We consider two scenarios including both random and fixed covariates. For the case of random covariate, the current model $\mathcal{M}$ includes the specifications of the distribution on $\mathbf{X}$ and the conditional distribution of $\mathbf{y}$ given $\mathbf{X}$, which are, respectively,

represented as $\mathcal{M}_{\mathbf{X}}$ and $\mathcal{M}_{\mathbf{y}|\mathbf{X}}$. Since $E[\mathbf{e}_*^{\otimes 2}|\mathcal{M}] = a(\tau_*)E[V_*|\mathcal{M}_{\mathbf{X}}]$, we have

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

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

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

With some algebraic calculation, it can be shown that

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

Furthermore, we have can be expressed as

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

which lead to (5). In addition, since H_* only has p non-zero eigenvalues and $H_{*,I}$ is a submatrix of H_*, it follows from Wielandt's eigenvalue inequality that $\lambda_{I,1} \geq \ldots \geq \lambda_{I,p} \geq 0 = \lambda_{I,p+1} = \ldots = \lambda_{I,n(I)}$ for $n(I) \geq p$. This yields Theorem S1 (b).

(c). Note that the matrices $H_{*,I}$ and $(\mathbf{X}_*^T\mathbf{X}_*)^{-1}\mathbf{X}_{*,I}^T\mathbf{X}_{*,I}$ have the same set of nonzero eigenvalues. Since $n^{-1}\mathbf{X}_*^T\mathbf{X}_*$ and $n(I)^{-1}\mathbf{X}_{*,I}^T\mathbf{X}_{*,I}$ converge to the same matrix

almost surely, $n(I)n^{-1}[(n^{-1}\mathbf{X}_*^T\mathbf{X}_*)^{-1}n(I)^{-1}\mathbf{X}_{*,I}^T\mathbf{X}_{*,I} - \mathbf{I}_p]$ converges to $\mathbf{0}$ almost surely as $n, n(I) \to \infty$. This completes the proof of Theorem S1 (c).

Theorem S1 (a) characterizes the stochastic behavior of $\widetilde{\mathrm{CD}}(I)$, which depends on both the responses and the covariates in the set I. To ensure that $E[\mathrm{CD}(I)|\mathcal{M}]$ and $Q_{CD(I)}(0.5|\mathcal{M})$ depend only on the size of the perturbation, not the set I itself, we need to bootstrap the randomness in both the responses and the covariates. Specifically, we can generate a new set of responses from the fitted model and draw an I_s at random from the original covariate data without (or with) replacement, where size(I_s)=size(I). Then, we calculate the CD(I_s) based on the bootstrapped data for $s = 1, \ldots, S$ and use their sample median to approximate $Q_{CD(I)}(0.5|\mathcal{M})$. Theorem S1 (b) gives an approximation of $E[\widetilde{\mathrm{CD}}(I)|\mathcal{M}]$ and $\mathrm{Var}[\widetilde{\mathrm{CD}}(I)|\mathcal{M}]$. We can draw a sample of sets $\{I_s : s = 1, \ldots, S\}$ of size(I) at random from the original covariate data without (or with) replacement and approximate them. Moreover, it should be noted that $\sum_{j=1}^{n(I)} E[(1-\lambda_{I,j})^{-1}|\mathcal{M}] - n(I)$ increases with the size of I even for $n(I) \geq p$. Theorem S1 (c) shows the asymptotic consistency of $\lambda_{I,j}$ for $j \leq p$. As $n(I)/n \to \gamma \in [0, 1)$, $\sum_{j=1}^{p} E[(1-\lambda_{I,j})^{-1}|\mathcal{M}] - p$ converges to $p\gamma/(1-\gamma)$.

Example S1 (continued). For generalized linear models, we fix all covariates, that is $\mathbf{Z} = \mathbf{X}$, and then calculate the CSCDs as follows. First, we can show that

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

Then, similar to the derivations of Theorem S1 (a) and (b), we can show that the conditionally scaled Cook's distance $\mathrm{CSCD}_1(I, \mathbf{X})$ can be approximated by

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

To approximate $\mathrm{CSCD}_2(I, \mathbf{X})$, we can generate responses from the model fitted to the

data and then substitute them into Theorem S1 (a) to obtain a sample of simulated $\widetilde{\mathrm{CD}}(I)$'s given the covariates. Finally, we can use the empirical median and median standard deviation of the simulated $\widetilde{\mathrm{CD}}(I)$'s to approximate $\mathrm{CSCD}_2(I, \mathbf{Z})$.

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

Example S2. Consider the LMCE given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{R}). \tag{6}$$

By choosing various $\mathbf{R}$'s, LMCE includes the linear model with independent data, the multivariate linear model, time series models, geostatistical models, and mixed effects models as special cases (Haslett, 1999; Haslett and Haslett, 2007). Similar to Haslett (1999), we fix $\mathbf{R}$ at an appropriate estimate $\hat{\mathbf{R}}$ throughout the example. We can calculate the generalized least squares estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{Y} = \mathbf{B}\mathbf{Y} \text{ and } \hat{\sigma}^2 = \mathbf{Y}^T\mathbf{Q}\mathbf{Y}/(n-p) = \hat{\mathbf{e}}^T\mathbf{R}^{-1}\hat{\mathbf{e}}/(n-p),$$

where $\mathbf{Q} = \mathbf{R}^{-1} - \mathbf{H}$, $\hat{\mathbf{e}} = \mathbf{R}\mathbf{Q}\mathbf{Y}$, and $\mathbf{H} = \mathbf{R}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{R}^{-1}$. Moreover, we have $\mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X})^{-1}$. It has been shown in Haslett (1999) that Cook's distance for deleting the subset I is given by

$$\begin{aligned} \mathrm{CD}(I) &= \frac{1}{\hat{\sigma}^2}\boldsymbol{\epsilon}^T\mathbf{Q}U_I\mathbf{Q}_{II}^{-1}(\mathbf{R}^{II} - \mathbf{Q}_{II})\mathbf{Q}_{II}^{-1}U_I^T\mathbf{Q}\boldsymbol{\epsilon} \approx \\ \widetilde{\mathrm{CD}}(I) &= \frac{1}{\sigma^2}\boldsymbol{\epsilon}^T\mathbf{Q}U_I\mathbf{Q}_{II}^{-1}(\mathbf{R}^{II} - \mathbf{Q}_{II})\mathbf{Q}_{II}^{-1}U_I^T\mathbf{Q}\boldsymbol{\epsilon}, \end{aligned} \tag{7}$$

where $\mathbf{Q}_{II}$ is the (I, I) subset of $\mathbf{Q}$ and $\mathbf{R}^{II}$ is the (I, I) subset of $\mathbf{R}^{-1}$.

Since $\widetilde{\mathrm{CD}}(I)$ is a quadratic form of $\boldsymbol{\epsilon}$, it follows from the well known results in (Mathai and Provost, 1992) that

$$\begin{aligned} &\mathrm{E}[\mathrm{CD}(I)|\mathcal{M}] \approx \mathrm{E}[\mathrm{tr}(\mathbf{Q}_{II}^{-1}\mathbf{R}^{II})|\mathcal{M}] - n(I) = \sum_{j=1}^{n(I)} \mathrm{E}[(1-\lambda_{I,j})^{-1}|\mathcal{M}] - n(I), \\ &\mathrm{Var}[\mathrm{CD}(I)|\mathcal{M}] \approx 2\mathrm{E}(\mathrm{tr}\{[\mathbf{Q}_{II}^{-1}\mathbf{R}^{II} - \mathbf{I}_{n(I)}]^2\}|\mathcal{M}) + \mathrm{Var}[\mathrm{tr}(\mathbf{Q}_{II}^{-1}\mathbf{R}^{II})|\mathcal{M}_X], \end{aligned} \tag{8}$$

where $\mathcal{M}_X$ represents the distribution of $\mathbf{X}$ and $\lambda_{I,1} \geq \ldots \geq \lambda_{I,n(I)}$ are the ordered eigenvalues of $(\mathbf{R}^{II})^{-1/2}\mathbf{H}_{II}(\mathbf{R}^{II})^{-1/2}$, in which $\mathbf{H}_{II}$ is the (I, I) subset of $\mathbf{H}$. Therefore, the scaled Cook's distance $\mathrm{SCD}_1(I)$ can be approximated by

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

Similar to Theorem S1 (b), when $n(I) \geq p$, the right-hand side of (8) reduces to $\sum_{j=1}^{p} E[(1-\lambda_{I,j})^{-1}|\mathcal{M}] - p$. In many scenarios such as the multivariate linear model, we can follow the strategies in Example S1 to approximate $\mathrm{E}[\mathrm{CD}(I)|\mathcal{M}]$ and $\mathrm{Var}[\mathrm{CD}(I)|\mathcal{M}]$. However, for time series data, since the elements in $\mathbf{X}$ are responses in an autoregressive model, such as the AR(1) model, we can use the parametric bootstrap to generate random samples from the fitted model and then approximate $\mathrm{E}[\mathrm{CD}(I)|\mathcal{M}]$ and $\mathrm{Var}[\mathrm{CD}(I)|\mathcal{M}]$.

We calculate the conditionally scaled Cook's distances as $\mathbf{Z} = \mathbf{X}$. Since $\widetilde{\mathrm{CD}}(I)$ is a quadratic form of $\boldsymbol{\epsilon}$, it follows from the well known results in (Mathai and Provost, 1992) that

$$\begin{aligned} &\mathrm{E}[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}] \approx \mathrm{tr}(\mathbf{Q}_{II}^{-1}\mathbf{R}^{II}) - n(I) = \sum_{j=1}^{n(I)}(1-\lambda_{I,j})^{-1} - n(I), \\ &\mathrm{Var}[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}] \approx 2\mathrm{tr}\{[\mathbf{Q}_{II}^{-1}\mathbf{R}^{II} - \mathbf{I}_{n(I)}]^2\}. \end{aligned} \tag{9}$$

Thus, the conditionally scaled Cook's distance $\mathrm{CSCD}_1(I, \mathbf{Z})$ can be approximated by

$$\frac{\mathrm{CD}(I) - \mathrm{tr}(\mathbf{Q}_{II}^{-1}\mathbf{R}^{II}) + n(I)}{(2\mathrm{tr}\{[\mathbf{Q}_{II}^{-1}\mathbf{R}^{II} - \mathbf{I}_{n(I)}]^2\})^{1/2}}.$$

2 Simulation Studies and A Real Data Example

In this section, we include an additional simulation study and some detailed results obtained from the simulated studies and the real data analysis in the paper.

2.1 Simulated Study I

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

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

extremely unbalanced.

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

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

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

The third experiment was to examine the finite sample performance of Cook's distance and the scaled Cook's distances for detecting influential cluster in longitudinal data. We considered two scenarios. In the first scenario, for each of the 100 simulated

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

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

datasets, we deleted all the observations in cluster n and then reset $m_n = 1$ and varied b_n from 0.4 to 8.0 to generate $y_{n,1}$ according to the above random effects model. The second scenario is almost the same as the first scenario except that we reset $m_n = 10$.

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

Finally, across all 100 datasets, we calculated the averages and standard errors of all diagnostic measures for the nth cluster for each scenario. Figures S1 and S2 present some selected results. Comparing the two scenarios, we observed that deleting the n-th cluster with 10 observations causes larger effect than that with 1 observation (Fig S1 (a) and Fig S2 (a)). For the first scenario, $\mathrm{CD}(\{n\})$ is relatively smaller than the other $\mathrm{CD}(I)$ (Fig. S1 (d)), whereas for the second scenario, $\mathrm{CD}(\{n\})$ is relatively larger than other $\mathrm{CD}(I)$ (Fig. S2 (d)). Furthermore, in the two scenarios, $P_A(\{n\}, \mathbf{Z})$ and $P_B(\{n\}, \mathbf{Z})$ for the scaled Cook's distances increase with b_n as expected, while they are quite close to each other across all values of b_n (Fig. S1 (d) and Fig. S2 (d)). It may indicate that all scaled Cook's distances are consistent with each other.

2.2 Simulation Study II

We included some detailed results for the first two experiments of the simulation studies in the paper. The first experiment was to evaluate the accuracy of $\widetilde{\mathrm{CD}}(I)$ 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 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.

For each dataset, we deleted each cluster one at a time and then calculated $\mathrm{CD}(I)$ and its first-order approximation $\widetilde{\mathrm{CD}}(I)$ for each cluster. Moreover, we computed the average $\mathrm{CD}(I)$, and the biases and standard errors of the differences $\mathrm{CD}(I) - \widetilde{\mathrm{CD}}(I)$ for each cluster. When no influential cluster is present in the first scenario, the distribution of $\mathrm{CD}(I)$ shifts up as $\mathcal{P}(I|\mathcal{M})$ increases (Fig. 3(a)). This result indicates that Assumption A1 may be reasonable. In the second scenario, the distribution of $\mathrm{CD}(I)$ for the true 'good' clusters shifts up as $\mathcal{P}(I|\mathcal{M})$ increases, while that for the two influential clusters are associated with both $\mathcal{P}(I|\mathcal{M})$ and the amount of influence that we introduced (Fig. 4 (a)).

For the true 'good' clusters, the first-order approximation is very accurate and leads to relatively small average biases and standard errors (Figs. 3 (d) and 4 (d)). Moreover, the degree of accuracy decreases as $\mathcal{P}(I|\mathcal{M})$ increases (Figs. 3 (d) and 4 (d)). Even for the influential clusters, $\widetilde{\mathrm{CD}}(I)$ is relatively close to $\mathrm{CD}(I)$ (Fig. 4 (d)). Specifically, the bias of $\widetilde{\mathrm{CD}}(I)$ relative to $\mathrm{CD}(I)$ equals 0.01 for cluster $\{1\}$ and 0.19 for cluster $\{n\}$. Even for cluster $\{n\}$, the bias of 0.19 is relatively small compared with 0.78, the mean of $\mathrm{CD}(\{n\})$. Moreover, such bias may be negligible for diagnostic purposes.

In the second experiment, we considered the same two scenarios as the first experiment in order to examine the finite sample performance of $E[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}]$ and $\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}]$ and their first-order approximations. Specifically, for each dataset, we set $S = 200$ and simulated $S = 200$ random samples from the fitted linear mixed model.

Then, we approximated $E[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$ and $\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$ by using their empirical ones, and calculated their first approximations $\widehat{M}[\widetilde{\mathrm{CD}}(I)]$ and $\widehat{\mathrm{Std}}[\widetilde{\mathrm{CD}}(I)]$.

In both scenarios, the distribution of $E[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$ shifts up as $\mathcal{P}(I|\mathcal{M})$ increases (Fig. 3 (b) and Fig. 4(b)). This is in agreement with the results of Proposition 1. For all clusters, the first-order approximations of $E[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$ and $\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$ are very accurate and lead to small average biases and standard errors (Fig. 3 (e) and (f), Fig. 4 (e) and (f)). Moreover, the degree of accuracy decreases as $\mathcal{P}(I|\mathcal{M})$ increases (Figs. 3 (d) and 4 (d)).

Table 2 presents the degrees of perturbation and the means and standard deviations of four conditionally scaled Cook's distances including $\mathrm{CSCD}_1(I,\mathbf{Z})$ and $\mathrm{CSCD}_2(I,\mathbf{Z})$ and their first-order approximations for all 12 clusters. In both scenarios, the four CSCDs are weakly associated with $\mathcal{P}(I|\mathcal{M})$. Figure 5 presents the box plots of the four CSCDs under the two scenarios. Inspecting Figure 5 (a)-(d) does not reveal any obvious relationship between the distributions of the four CSCDs and the degree of perturbation in the first scenario. Moreover, in the second scenario, we did not observe obvious relationship between the distributions of the four CSCDs and the degree of perturbation for these 'good' clusters. We also observed from Figure 5 (e)-(h) that for the two influential clusters, the distributions of the four CSCDs are associated with the influence level that we introduced.

2.3 Yale Infant Growth Data

Under each model, we calculated $\mathrm{CD}(I)$ for each child, which relates more to the detection of influential clusters (Banerjee and Frees, 1997). We treated $\boldsymbol{\beta}$ as parameters of interest and all elements of $\boldsymbol{\alpha}$ as nuisance parameters. For models M_1 and M_2, inspecting Figures 6 (a) and 7 (a) reveals that $\mathcal{P}(I|\mathcal{M})$ is positively associated with m_i. This

Table 2: The two conditionally scaled Cook's distances of $\mathrm{CSCD}_1(I)$ and $\mathrm{CSCD}_2(I)$ and their first-order approximations $\widetilde{\mathrm{CSCD}}_1(I)$ and $\widetilde{\mathrm{CSCD}}_2(I)$ from simulation studies for $n = 12$ and the two scenarios: M denotes the mean; and SD denotes the standard deviation. In the first scenario, all observations were generated from the linear mixed model, while in the second scenario, two influential clusters were highlighted in bold. For each case, 100 simulated datasets were used. Results were sorted according to the degree of perturbation.

	Scenario I							
	$\mathrm{CSCD}_1(I)$		$\mathrm{CSCD}_2(I)$		$\widetilde{\mathrm{CSCD}}_1(I)$		$\widetilde{\mathrm{CSCD}}_2(I)$	
$\mathcal{P}(I\|\mathcal{M})$	M	SD	M	SD	M	SD	M	SD
0.102	-0.016	0.739	1.457	2.854	-0.008	0.789	1.366	2.746
0.108	0.039	0.862	0.976	2.160	0.032	0.844	0.838	1.977
0.110	0.178	1.078	1.599	3.058	0.151	1.010	1.305	2.429
0.128	0.123	1.098	1.436	3.356	0.102	1.120	1.237	2.990
0.147	0.351	1.264	2.210	3.946	0.364	1.315	2.076	3.806
0.159	0.019	0.928	0.854	2.384	-0.001	0.921	0.732	2.088
0.188	0.237	1.037	1.956	3.532	0.240	1.119	1.798	3.478
0.224	0.109	0.832	1.227	2.376	0.118	0.875	1.084	2.215
0.264	0.128	0.974	1.138	2.580	0.114	0.922	1.077	2.272
0.403	0.301	1.390	1.407	3.153	0.297	1.339	1.408	3.008
0.569	0.151	1.023	1.568	3.308	0.131	0.971	1.566	3.244
0.599	0.346	1.583	1.808	3.862	0.312	1.380	1.920	3.825
	Scenario II							
$\mathcal{P}(I\|\mathcal{M})$	M	SD	M	SD	M	SD	M	SD
0.079	**2.486**	**1.477**	**9.654**	**5.280**	**2.583**	**1.489**	**9.721**	**5.215**
0.109	-0.186	0.724	0.455	2.015	-0.175	0.783	0.412	1.969
0.114	-0.116	0.674	0.992	2.345	-0.116	0.705	0.934	2.305
0.131	-0.294	0.544	0.166	1.541	-0.293	0.583	0.157	1.577
0.155	-0.273	0.553	0.304	1.689	-0.285	0.571	0.234	1.653
0.199	-0.284	0.540	0.190	1.522	-0.288	0.565	0.142	1.514
0.226	-0.222	0.619	0.278	1.497	-0.243	0.630	0.224	1.465
0.245	-0.134	0.824	0.328	1.741	-0.106	0.883	0.380	1.764
0.279	**1.061**	**1.695**	**3.516**	**4.620**	**0.644**	**0.962**	**2.234**	**2.189**
0.368	-0.096	0.773	0.475	1.738	-0.072	0.841	0.520	1.817
0.534	-0.176	0.862	0.408	2.098	-0.148	0.847	0.505	2.186
0.555	-0.275	0.608	0.309	2.018	-0.261	0.619	0.353	2.064

indicates that the bigger the cluster size, the larger the degree of perturbation.

Under model M_1, we used CD(I) to select the top five influential subjects 269, 217, 294, 289, and 274 (Fig. 6 (b)), while we used $\text{CSCD}_1(I)$ to select the top eight influential subjects 274, 90, 217, 109, 294, 289, 246, and 149 (Fig. 6 (c)). Although we observed some difference between CD(I) and $\text{CSCD}_1(I)$ in detecting highly influential subjects, the value of CD(I) and that of $\text{CSCD}_1(I)$ are positively correlated with each other across all subjects (Figure 6 (d)). Moreover, Table 3 presents the top twelve influential subjects detected by CD(I), $\text{CSCD}_1(I|\mathcal{M}, \mathbf{Z})$, and $\text{CSCD}_2(I|\mathcal{M}, \mathbf{Z})$. We used $P_B(I, \mathbf{Z})$ to quantify whether a specific subject is influential relative to the fitted model for all subjects (Figure 6 (e)). Inspecting Figure 6 (f) reveals that there are large number of influential subjects, and thus it may indicate the potential model misspecification in model $\mathcal{M}_1$.

Under model M_2, we used CD(I) to select the top five influential subjects 269, 285, 280, 246, and 58 (Fig. 7 (b)), while we used $\text{CSCD}_1(I)$ to select the top eight influential subjects 246, 141, 109, 31, and 193 (Fig. 7 (c)). Although we observed some difference between CD(I) and $\text{CSCD}_1(I)$ in detecting highly influential subjects, the value of CD(I) and that of $\text{CSCD}_1(I)$ are positively correlated with each other across all subjects (Figure 7 (d)). We used $P_B(I, \mathbf{Z})$ to quantify whether a specific subject is influential relative to the fitted model for all subjects (Figure 7 (e)). Inspecting Figure 7 (f) reveals that the number of influential subjects has dramatically reduced, and thus it may indicate that model M_2 outperforms model M_1 in fitting the Yale infant growth data.

References

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

Banerjee, M. and Frees, E. W. (1997), "Influence Diagnostics for Linear Longitudinal

Table 3: *Yale infant growth data.* Top 12 influential subjects for single case deletion with the compound symmetry model.

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

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

Models," *Journal of the American Statistical Association*, 92, 999–1005.

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

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

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

Mathai, A. M. and Provost, S. (1992), *Quadratic Forms in random Variables: Theory and Applications,*, New York: Marcel Dekker.

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

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

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.

Wasserman, D. and Leventhal, J. (1993), "Maltreatment of Children Born to Cocaine-Dependent Mothers," *American Journal of Diseases of Children*, 147, 1324–1328.

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

Williams, D. A. (1987), "Generalized linear model diagnostics using the deviance and single case deletions," *Applied Statistics*, 36, 181–191.

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

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

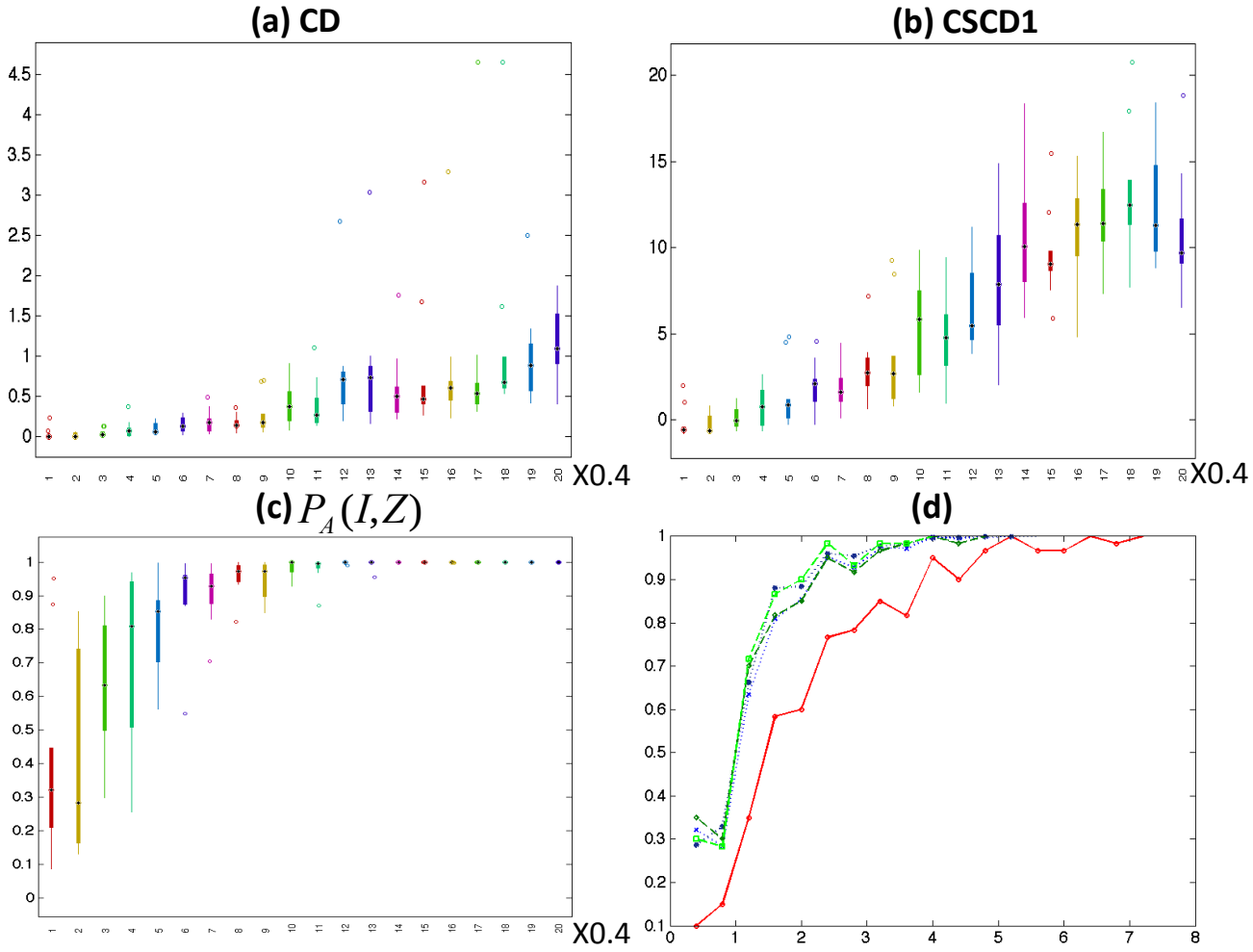

Figure 1: Results from 100 datasets simulated from a linear mixed model, in which $m_{30} = 1$ and b_{30} varies from 0.4 to 8.0. Panel (a) shows the box plots of Cook's distances as a function of b_{30}; panel (b) shows the box plots of $\text{CSCD}_1(I, \mathbf{Z})$ as a function of b_{30}; panel (c) shows the box plots of $P_A(I, \mathbf{Z})$ as a function of b_{30}; panel (d) shows the mean curves of $P_A(I, \mathbf{Z})$ based on the four scaled Cook's distances, in which the green line is for $\text{CSCD}_1(I, \mathbf{Z})$, the dark green line is for $\text{CSCD}_2(I, \mathbf{Z})$, the blue line is for $\widetilde{\text{CSCD}}_1(I, \mathbf{Z})$, and the dark line is for $\widetilde{\text{CSCD}}_1(I, \mathbf{Z})$, and the mean curve of $P_C(I, \mathbf{Z})$ based on $\text{CD}(I)$ (red line) as functions of b_{30}.

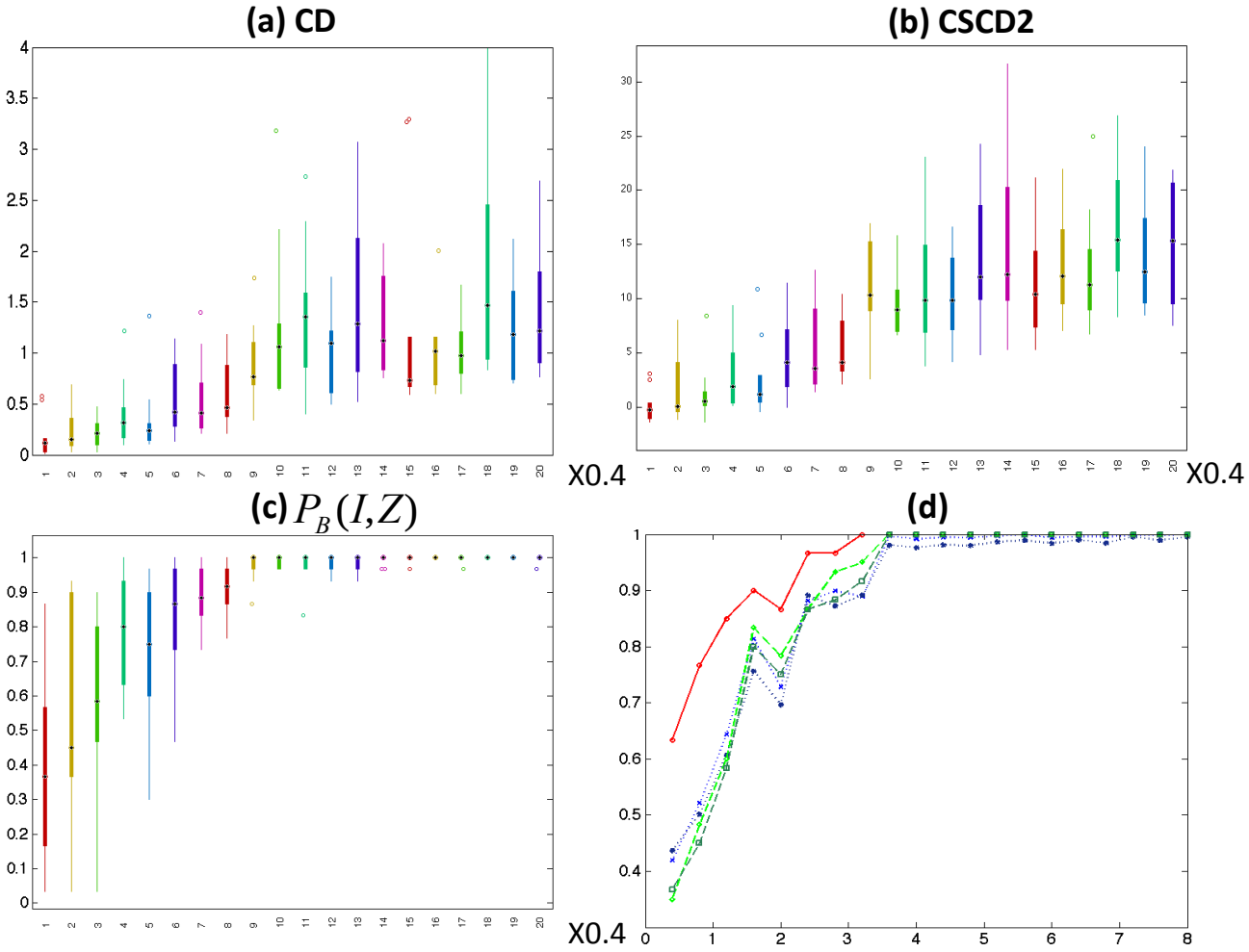

Figure 2: Results from 100 datasets simulated from a linear mixed model, in which $m_{30} = 1$ and b_{30} varies from 0.4 to 8.0. Panel (a) shows the box plots of Cook's distances as a function of b_{30}; panel (b) shows the box plots of $\text{CSCD}_1(I, \mathbf{Z})$ as a function of b_{30}; panel (c) shows the box plots of $P_B(I, \mathbf{Z})$ as a function of b_{30}; panel (d) shows the mean curves of $P_B(I, \mathbf{Z})$ based on the four scaled Cook's distances, in which the green line is for $\text{CSCD}_1(I, \mathbf{Z})$, the dark green line is for $\text{CSCD}_2(I, \mathbf{Z})$, the blue line is for $\widetilde{\text{CSCD}}_1(I, \mathbf{Z})$, and the dark line is for $\widetilde{\text{CSCD}}_1(I, \mathbf{Z})$, and the mean curve of $P_C(I, \mathbf{Z})$ based on $\text{CD}(I)$ (red line) as functions of b_{30}.

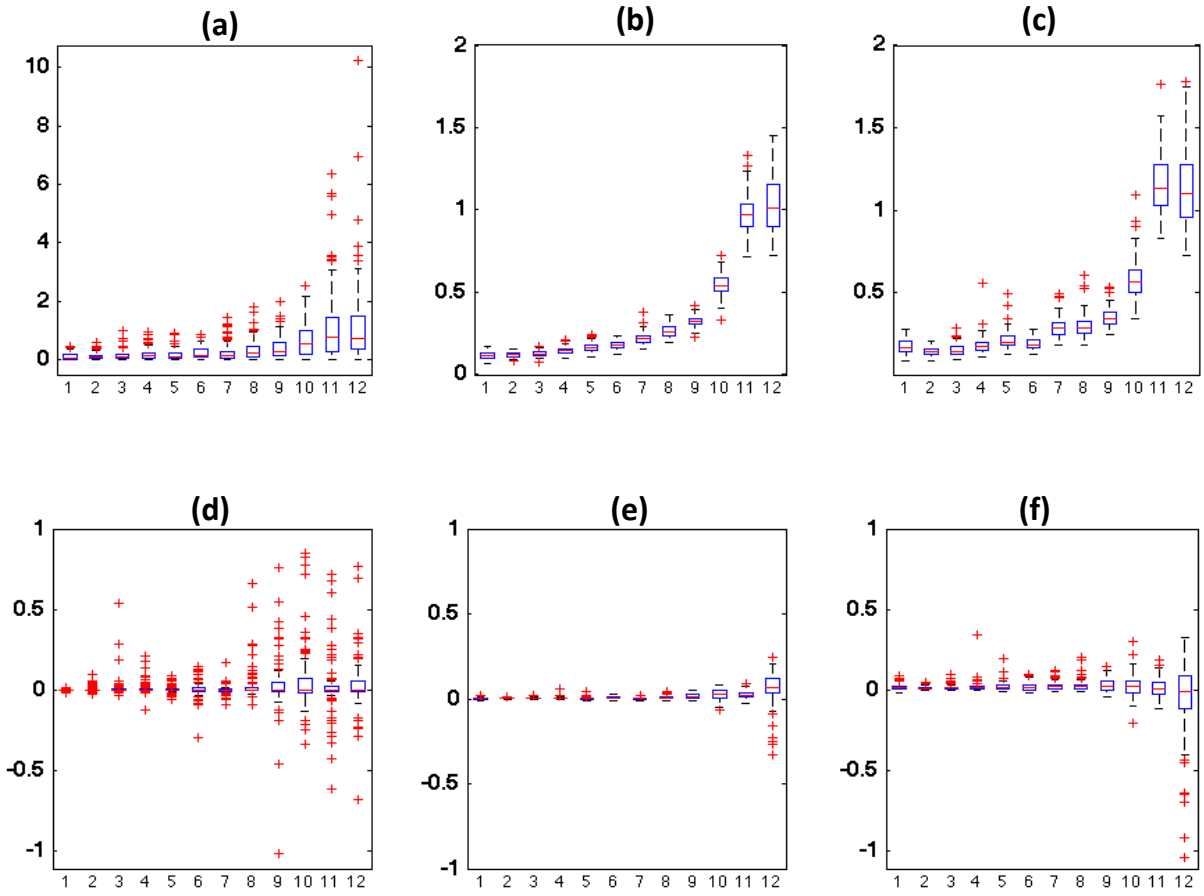

Figure 3: Simulation results from 100 datasets without influential clusters directly simulated from a linear mixed model. The $x-$axis corresponds to the order of the sorted degree of perturbation for all clusters. Panels (a), (b), and (c) show the box plots of $\mathrm{CD}(I)$, $E[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$, and $\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}]$ as a function of $\mathcal{P}(I|M)$; panels (d), (e), and (f) show the box plots of $\mathrm{CD}(I) - \widetilde{\mathrm{CD}}(I)$, $E[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}] - \widehat{M}[\widetilde{\mathrm{CD}}(I)]$, and $\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M},\mathbf{Z}] - \widehat{\mathrm{Std}}[\widetilde{\mathrm{CD}}(I)]$ as a function of $\mathcal{P}(I|M)$.

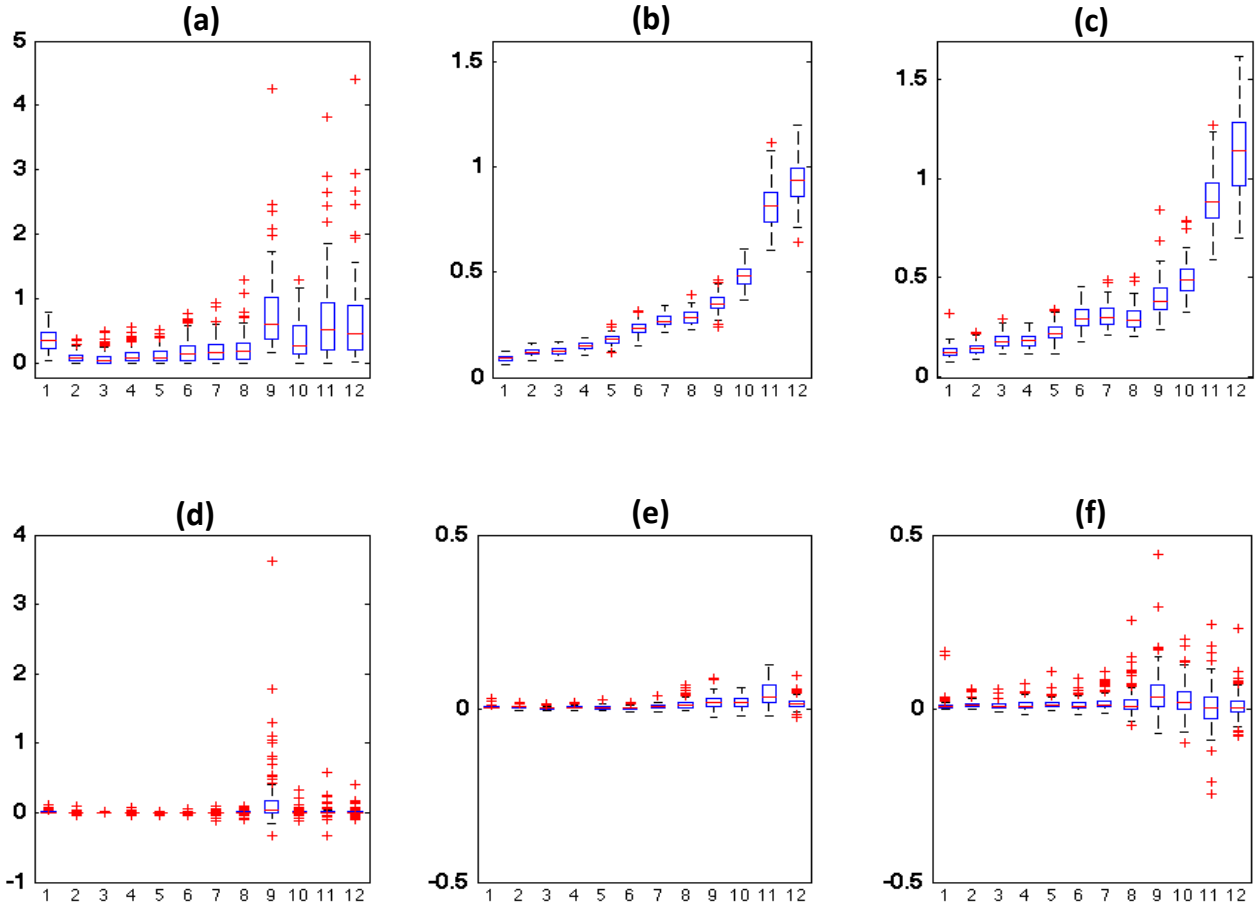

Figure 4: Simulation results from 100 datasets with two influential clusters simulated from a linear mixed model, in which we reset $(m_1, b_1) = (1, 4)$ and $(m_n, b_n) = (5, 3)$ to generate $y_{i,j}$ for $i = 1, n$ and all j according to the same linear mixed model.. The $x-$axis corresponds to the order of the sorted degree of perturbation for all clusters. Panels (a), (b), and (c) show the box plots of $\mathrm{CD}(I)$, $E[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}]$, and $\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}]$ as a function of $\mathcal{P}(I|M)$; panels (d), (e), and (f) show the box plots of $\mathrm{CD}(I) - \widetilde{\mathrm{CD}}(I)$, $E[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}] - \widehat{M}[\widetilde{\mathrm{CD}}(I)]$, and $\mathrm{Std}[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}] - \widehat{\mathrm{Std}}[\widetilde{\mathrm{CD}}(I)]$ as a function of $\mathcal{P}(I|M)$.

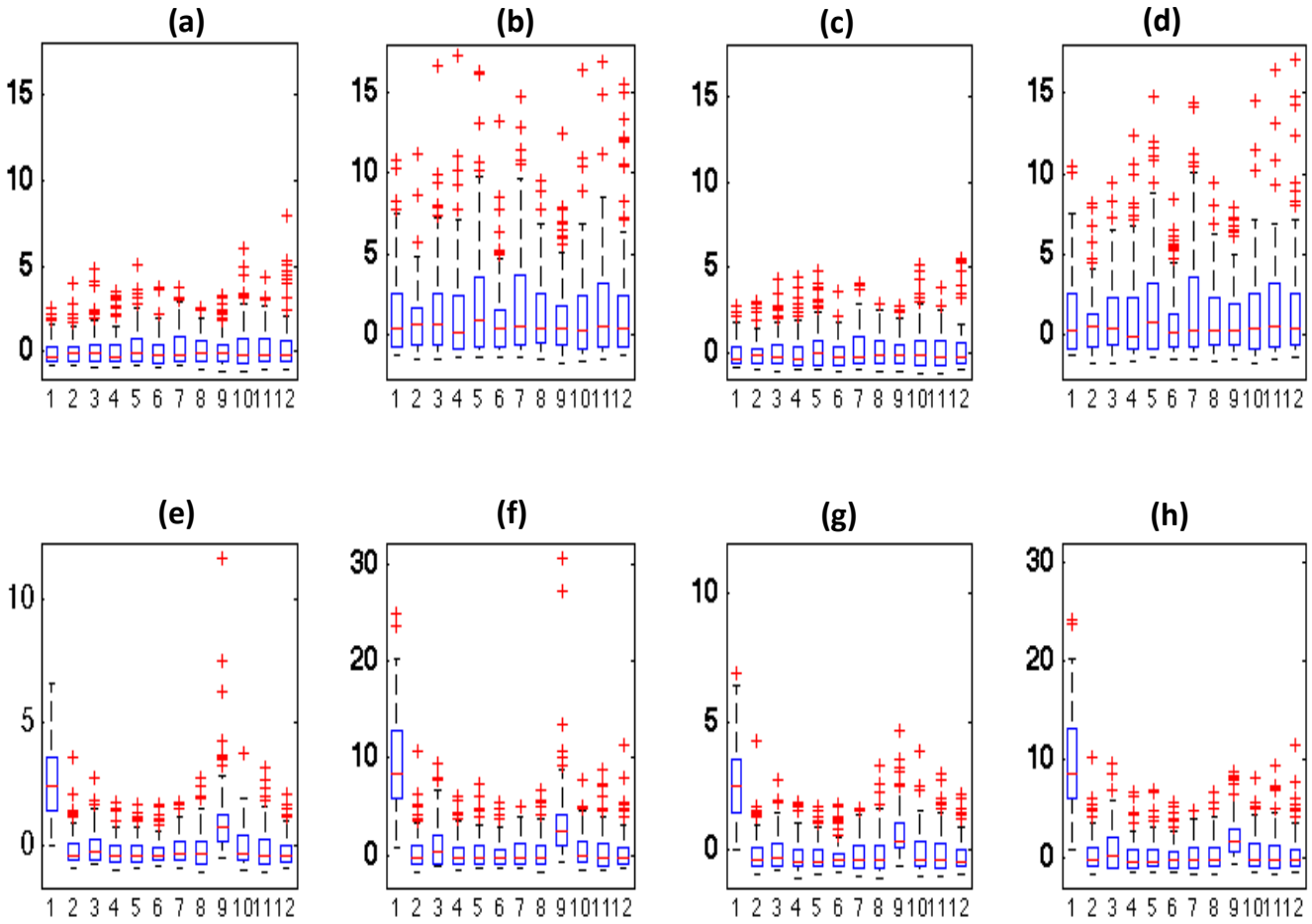

Figure 5: Simulation results from 100 datasets simulated from a linear mixed model in the two scenarios. The first row corresponds to the first scenario with no influential subjects. The second row corresponds to the second scenario, in which we set $(m_1, b_1) = (1, 4)$ and $(m_{12}, b_{12}) = (5, 3)$. The $x-$axis corresponds to the order of the sorted degrees of perturbation. For the second scenario, the 1st and 9th cases are, respectively, the two influential subjects with $(m_1, b_1) = (1, 4)$ and $(m_{12}, b_{12}) = (5, 3)$. Panels (a) and (e) show the box plots of $\text{CSCD}_1(I, \mathbf{Z})$ as a function of $\mathcal{P}(I|\mathcal{M})$; panels (b) and (f) show the box plots of $\text{CSCD}_2(I, \mathbf{Z})$ as a function of $\mathcal{P}(I|\mathcal{M})$; panels (c) and (g) show the box plots of $\widetilde{\text{CSCD}}_1(I, \mathbf{Z})$ as a function of $\mathcal{P}(I|\mathcal{M})$; panels (d) and (h) show the box plots of $\widetilde{\text{CSCD}}_2(I, \mathbf{Z})$ as a function of $\mathcal{P}(I|\mathcal{M})$.

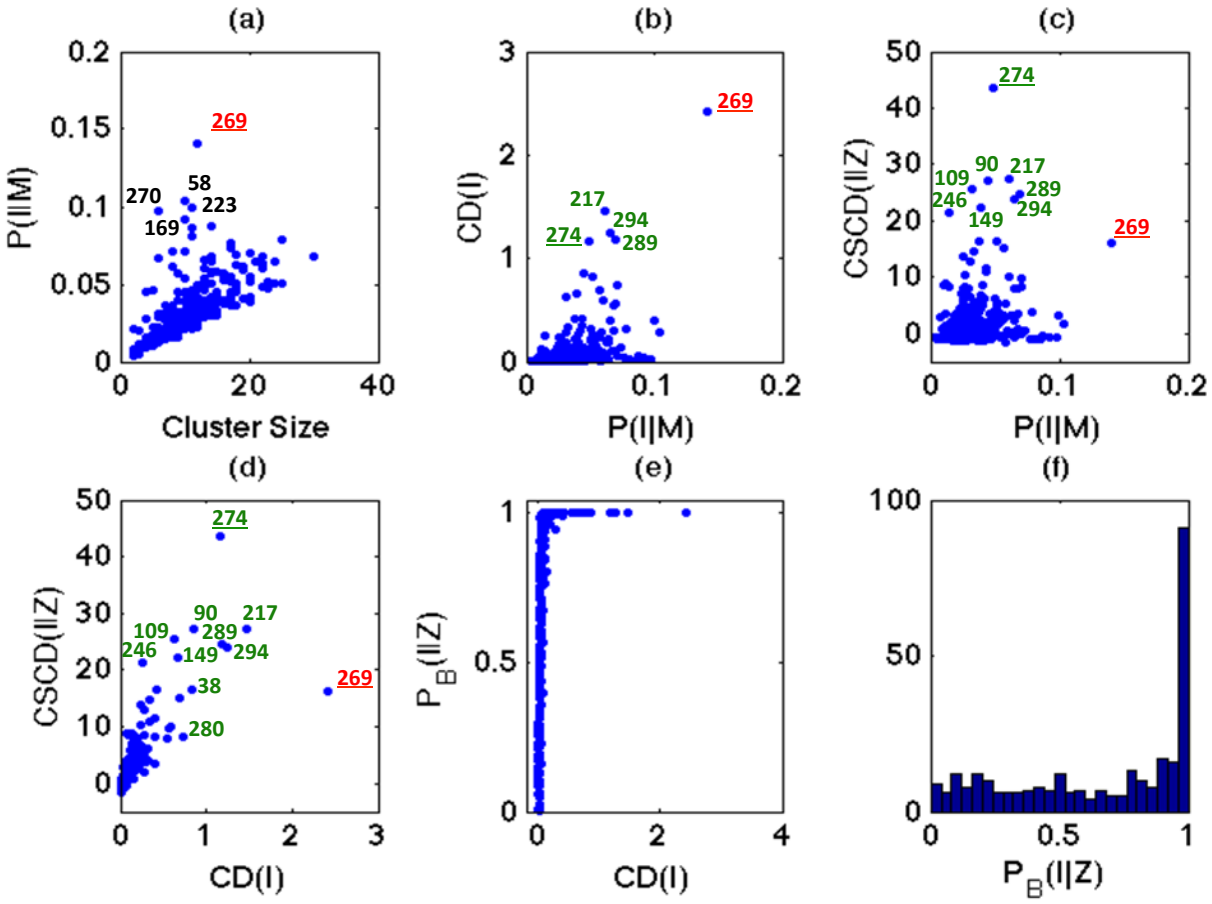

Figure 6: Yale infant growth data analysis based on model $\mathcal{M}_1$. Panel (a) shows $\mathcal{P}(I|M)$ versus m_i; panels (b), (c), (d), and (e), respectively, present CD(I) versus $\mathcal{P}(I|M)$, CSCD($I|\mathbf{Z}$) versus $\mathcal{P}(I|M)$, CSCD($I|\mathbf{Z}$) versus CD(I), and $P_B(I|\mathbf{Z})$ versus CD(I); panel (f) presents the histogram of $P_B(I|\mathbf{Z})$.

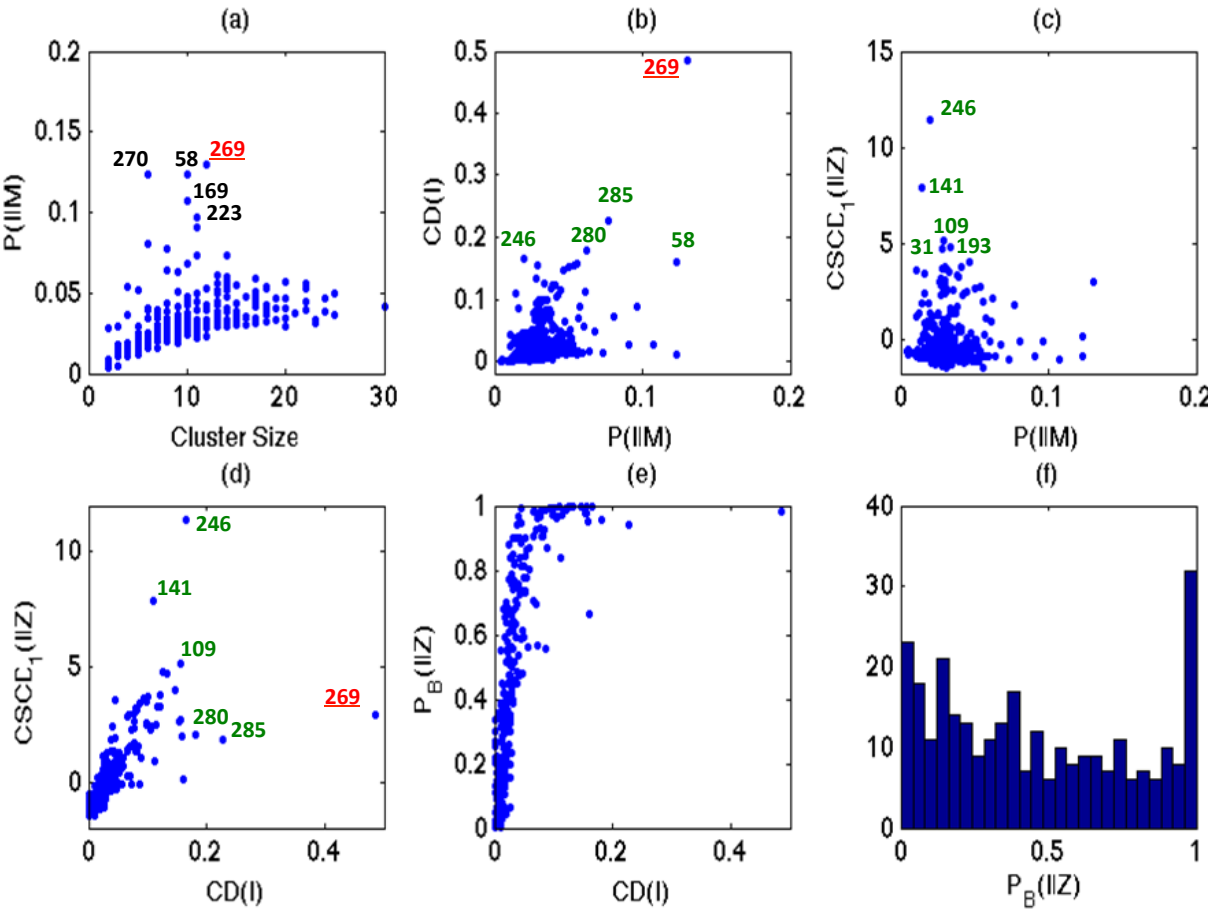

Figure 7: Yale infant growth data analysis based on model $\mathcal{M}_2$. Panel (a) shows $\mathcal{P}(I|M)$ versus m_i; panels (b), (c), (d), and (e), respectively, present CD(I) versus $\mathcal{P}(I|M)$, CSCD($I|\mathbf{Z}$) versus $\mathcal{P}(I|M)$, CSCD($I|\mathbf{Z}$) versus CD(I), and $P_B(I|\mathbf{Z})$ versus CD(I); panel (f) presents the histogram of $P_B(I|\mathbf{Z})$.