Case-deletion Diagnostics for Linear Mixed Models

Jianxin Pan\textsuperscript{a}, Yu Fei\textsuperscript{b} & Peter Foster\textsuperscript{a}

\textsuperscript{a} School of Mathematics, The University of Manchester, UK
\textsuperscript{b} School of Statistics and Mathematics, Yunnan University of Finance and Economics, China

Accepted author version posted online: 12 Jul 2013.

To cite this article: Technometrics (2013): Case-deletion Diagnostics for Linear Mixed Models, Technometrics, DOI: 10.1080/00401706.2013.810173

To link to this article: http://dx.doi.org/10.1080/00401706.2013.810173

Disclaimer: This is a version of an unedited manuscript that has been accepted for publication. As a service to authors and researchers we are providing this version of the accepted manuscript (AM). Copyediting, typesetting, and review of the resulting proof will be undertaken on this manuscript before final publication of the Version of Record (VoR). During production and pre-press, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal relate to this version also.

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the “Content”) contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions
Case-deletion Diagnostics for Linear Mixed Models

Jianxin Pan*
School of Mathematics, The University of Manchester, UK
(jianxin.pan@manchester.ac.uk)

and

Yu Fei
School of Statistics and Mathematics,
Yunnan University of Finance and Economics, China
(feiyukm@aliyun.com)

and

Peter Foster
School of Mathematics, The University of Manchester, UK
(peter.foster@manchester.ac.uk)

Abstract

Based on the $Q$-function, the conditional expectation of the logarithm of the joint-likelihood between responses and random effects, we propose a case-deletion approach to identify influential subjects and influential observations in linear mixed models. The models considered here are very broad in the sense that any covariance structures can be specified in the covariance matrices of the random effects and random errors. Analytically explicit forms of diagnostic measures for the fixed effects and variance components are provided. Comparisons with existing methods, including likelihood-based case-deletion and local influence methods, are made. Numerical results, including real data analysis and simulation studies, are presented for both illustration and comparison. This paper has supplementary material online.

Keywords: Covariance structures; Generalized Cook distance; Influence analysis; Linear mixed models; $Q$-function.

*Corresponding author
1 Introduction

We consider case-deletion diagnostics for linear mixed models (LMM) (Laid and Ware, 1982)

\[ Y_i = X_i \beta + Z_i u_i + \epsilon_i, \]

where for the \( i \)th subject, \( Y_i \) is the \((n_i \times 1)\) stacked vector of \( n_i \) observations made over time or location, \( X_i \) is an \((n_i \times p)\) design matrix of covariates, \( \beta \) is an \((p \times 1)\) vector of fixed effects, \( Z_i \) is an \((n_i \times q)\) design matrix for the \((q \times 1)\) vector of between-subject random effects \( u_i \), and \( \epsilon_i \) is an \((n_i \times 1)\) vector of random errors \((i = 1, \ldots, m)\). It is common to assume \( u_i \sim N(0, G) \) and \( \epsilon_i | u_i \sim N(0, R_i) \), where \( G = G(\alpha) \) is the \((q \times q)\) between-subject covariance matrix, which may depend on an \((r \times 1)\) vector \( \alpha \) of variance components, and \( R_i = R_i(\gamma) \) is the \((n_i \times n_i)\) within-subject covariance matrix for repeated measurements, which may depend on an \((s \times 1)\) vector \( \gamma \) of variance components. Throughout this paper, we denote \( \theta = (\beta', \alpha', \gamma')' \) as the vector of full parameters for the LMM.

The LMM is widely used in longitudinal and spatial data analysis (Verbeke and Molenberghs, 2000). It also plays a fundamental role in generalized linear mixed models for modeling correlated discrete data (Breslow and Clayton, 1993) and nonparametric smoothing using penalized splines (Eilers and Marx, 1996). Statistical inference for the LMM has been well developed and one may refer to Searle et al (1992) and Verbeke and Molenberghs (2000) for a comprehensive review. However, statistical diagnostics for the LMM have not been fully addressed yet. By assuming that the marginal covariance matrices \( V_i = Z_i G Z_i' + R_i \) are given, Banerjee and Frees (1997) and Hodges (1998) considered case-deletion diagnostics for the fixed effects \( \beta \). Similar considerations were given by Tan et al (2001), Ouwens et al. (2001) and Demidenko and Stukel (2005). When \( V_i \) are unknown, it was suggested to use a consistent estimate \( \hat{V}_i \) to replace \( V_i \) (Banerjee and Frees, 1997; Hodges, 1998) but the resulting statistical diagnostics for the fixed effects \( \beta \) may not be efficient (Atkinson, 1998; Pan and Thompson, 1998). On the other hand, case-deletion diagnostics for the variance components \((\alpha, \gamma)\) are not fully clear yet (Demidenko and Stukel, 2005). When \( G = \sigma_u^2 I_q \) and \( R_i = \sigma_\epsilon^2 I_{n_i} \), leading to \( V_i = \sigma_u^2 Z_i Z_i' + \sigma_\epsilon^2 I_{n_i} \), Christensen et al. (1992) proposed a one-step approximation to Cook’s (1982) distance for the variance components in \( V_i \). However, this approximation may not be computationally efficient, as pointed out by Haslett and Dilane (2004). When either the covariance matrix \( G \) or \( R_i \) has more complicated structures such as AR(1)
or antedependence, it is not clear yet how the one-step approximation approach works in this case. As an alternative, Beckman et al. (1987) and Lesaffre and Verbeke (1998) proposed to use the computationally attractive local influence approach (Cook, 1986) to study influence analysis for the LMM. Beckman et al. (1987) considered the models with \( V_i = \sigma^2_u Z_i Z'_i + \sigma^2_\epsilon I_i \) and examined a variety of perturbation schemes. Lesaffre and Verbeke (1998) studied local influence for the LMM with unstructured covariance for \( G \) and independent structure for \( R_i \), that is, \( R_i = \sigma^2_\epsilon I_i \). Restricted by these special structures for the covariance matrices, their local influence measures can not be directly applied to the models with other covariance structures for either \( G \) or \( R_i \). In fact, Lesaffre and Verbeke (1998) recommended that their local influence measures be used only to flag apparently influential cases for subsequent refit with specifying a particular structure for \( G \).

Based on the conditional expectation of the complete-data log-likelihood at the E-step in the EM-algorithm, namely the \( Q \)-function, Zhu et al. (2001) developed a unified approach to obtain case-deletion diagnostic measures for general models with missing data. Further, Lee and Xu (2004) and Xu et al. (2006) applied similar ideas to study case-deletion influence analysis for generalized linear mixed models and nonlinear mixed-effects models, respectively. Based on the \( Q \)-function, Zhu and Lee (2003) proposed a procedure to detect influential observations in generalized linear mixed models using a local influence approach. The case-deletion diagnostic measures considered by Zhu et al. (2001), Lee and Xu (2004) and Xu et al. (2006) are all based on \( \tilde{Q} \), the second-order derivative of the \( Q \)-function with respect to \( \theta \), which works well for the models they considered but may be difficult to extend to other complicated models due to the complexity of calculating \( \tilde{Q} \). In this paper, we propose to use \( E(\tilde{Q}) \), the expectation of \( \tilde{Q} \) with respect to the responses \( Y \), to replace \( \tilde{Q} \) when constructing case-deletion diagnostic measures. It turns out that the resulting structure of \( E(\tilde{Q}) \) is block-diagonal with respect to the fixed effects \( \beta \) and the variance components \( \alpha \) and \( \gamma \) in the LMM, irrespective of the covariance structures for \( G \) and \( R_i \). It implies that the case-deletion diagnostic measures based on \( E(\tilde{Q}) \) for \( \beta \), \( \alpha \) and \( \gamma \) are mutually independent, which substantially saves computational effort for calculating the case-deletion diagnostic measure of \( \theta \). We also show that the diagnostic measure based on \( E(\tilde{Q}) \) is as efficient as the one based on \( \tilde{Q} \). The LMM considered here is very broad in the sense that any covariance structures can be specified in the covariance matrices of the random effects and random errors. Analytically explicit forms of
diagnostic measures for the fixed effects and variance components in various covariance structures are provided. Comparisons with existing methods, including likelihood-based case-deletion and local influence methods, are made under certain circumstances. Numerical results, including real data analysis and simulation studies, are presented for both illustration and comparison.

The rest of the paper is organized as follows. In Section 2, we study likelihood-based case-deletion diagnostics and develop new diagnostic approaches based on the $Q$-function. In Section 3 we compare the $Q$-function-based case-deletion diagnostic measure to the likelihood-based approach within the framework of the LMM with independent random effects and independent random errors, in which both diagnostic approaches have analytically explicit forms. In Section 4 the $Q$-function-based case-deletion diagnostic measures are considered for the LMM with non-independent covariance structures. In Section 5 the $Q$-function-based diagnostic measures at observation level are discussed. In Section 6, the proposed diagnostic methods are demonstrated with the analysis of data on aerosols. In Section 7 we carry out intensive simulation studies to assess the performance of the proposed diagnostic methods and compare to the existing diagnostic approaches for the LMM, including local influence method by Beckman et al. (1987) and one-step case-deletion method by Christensen et al. (1992). In Section 8, we provide a further discussion on the proposed approaches. Technical details are provided in the Supplementary Material.

2 Case-deletion diagnostics

The case-deletion approach is one of the most commonly used techniques in statistical diagnostics (Cook and Weisberg, 1982). Assume that $\theta$ is the parameter vector of interest in a model which is under consideration and $\hat{\theta}$ is the estimate of $\theta$ using certain principles, such as maximum likelihood. When the $i$th observation $Y_i$ is deleted from the full observational set $Y = \{Y_1, Y_2, ..., Y_n\}$, a refitting of the model using the same estimation principle yields the estimate $\hat{\theta}_{[i]}$ of $\theta$. Appropriate measures of difference between $\hat{\theta}_{[i]}$ and $\hat{\theta}$ can be used to quantify how the $i$th observation influences the estimate $\hat{\theta}$ and statistical inferences. A typical measure is Cook’s distance defined by

$$C_i(\theta) = (\hat{\theta}_{[i]} - \hat{\theta})'M(\hat{\theta}_{[i]} - \hat{\theta}),$$

(2)
where $M$ is an appropriately chosen positive definite matrix. The matrix $M$ is typically the inverse of the covariance of the estimates $\hat{\theta}$ or $\hat{\theta}_{[i]}$ (Cook and Weisberg, 1982). For the ordinary linear regression model $E(Y) = X\beta$ with $\text{Var}(Y) = \sigma^2 V$ where $V$ is a known positive definite matrix, the case-deletion diagnostic measure for the regression coefficients $\beta$ can be formed by taking $M = X'V^{-1}X$ or $M = [X_{[i]}'V_{[i]}^{-1}X_{[i]}]$, leading to the so-called Cook’s statistic or DFFITS, where $X_{[i]}$ and $V_{[i]}$ are the matrices $X$ and $V$ after deleting the $i$th row (and also the $i$th column for $V$). It can be shown that in this case, calculating $\hat{\beta}_{[i]}$ for all $i$’s is not necessary because there is an updating formula for $\hat{\beta}_{[i]} - \hat{\beta}$ which can avoid direct calculation of $\hat{\beta}_{[i]}$ for each $i$ (Cook and Weisberg, 1982). Moreover, Cook’s statistic and DFFITS can be calibrated with reference to the $\chi^2$-distribution with appropriate degrees of freedom so that influential cases can be flagged.

For the LMM (1), when the marginal covariances $V_i = Z_iGZ_i + R_i$ are known, an updating formula for the fixed effects $\beta$ exists (Banerjee and Frees, 1997). When the covariances $V_i$ are unknown, however, the updating formula for $\beta$ is no longer available, except for balanced longitudinal data (Pan and Fang, 2002). It is common in the literature that the maximum likelihood estimates or restricted maximum likelihood estimates $\hat{V}_i$ are used to replace $V_i$ in the updating formula as if $V_i$ were given (Banerjee and Frees, 1997; Hodges, 1998; Demidenko and Stukel, 2005). The resulting diagnostics for the fixed effects $\beta$, however, remain problematic because $\hat{V}_i$ are also influenced by the deletion of the $i$th subject (Atkinson, 1998; Pan, 1998). Regarding the case-deletion diagnostic measures for the variance components, Christensen et al. (1992) considered a special case with $G = \sigma^2 u_1 I_{q_1} + \ldots + \sigma^2 u_k I_{q_k}$ and $R_i = \sigma^2 \epsilon_i I_{n_i}$ using a one-step approximation. For other covariance structures for $G(\alpha)$ and $R_i(\gamma)$, it is not clear yet how to construct case-deletion diagnostic measures for variance components $\alpha$ and $\gamma$ (Demidenko and Stukel, 2005).

### 2.1 Likelihood-based diagnostics

The log-likelihood function for the LMM (1) is given by

$$\ell(\theta) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{m} \log |V_i| - \frac{1}{2} \sum_{i=1}^{m} (Y_i - X_i\beta)'V_i^{-1}(Y_i - X_i\beta),$$  

(3)

where $V_i = Z_iGZ_i + R_i$ and $n = \sum_{i=1}^{m} n_i$. Denote $\ell_{[i]}(\theta)$ as the log-likelihood with the $i$th subject deleted. Assume $\hat{\theta}$ and $\hat{\theta}_{[i]}$ are the maximum likelihood estimates of $\theta$ under $\ell(\theta)$ and $\ell_{[i]}(\theta)$, respec-
tively. To calculate $\hat{\theta}_{[i]}$, we may use a one-step Newton-Raphson approximation at $\hat{\theta}$, in other words, $\hat{\theta}_{[i]} = \hat{\theta} + [-\hat{\ell}^\prime_{[i]}(\hat{\theta})]^{-1}\hat{\ell}_{[i]}(\hat{\theta})$, where dots over the functions denote derivatives with respect to $\theta$. The one-step approximation to Cook’s distance is thus of the form $C^1_i(\theta) = [\hat{\ell}_{[i]}(\hat{\theta})]'[-\hat{\ell}_{[i]}(\hat{\theta})]^{-1}[\hat{\ell}_{[i]}(\hat{\theta})]$, where we choose $M = [-\hat{\ell}_{[i]}(\hat{\theta})]$, the inverse of the asymptotic covariance matrix of $\hat{\theta}_{[i]}$. Note the computation of $[-\hat{\ell}_{[i]}(\hat{\theta})]^{-1}$ may become cumbersome when $m$ is large. To overcome this difficulty, we propose to use $\hat{\ell}(\hat{\theta})$ to replace $\hat{\ell}_{[i]}(\hat{\theta})$ so that the approximated Cook’s distance becomes

$$C_i(\theta) = [\hat{\ell}_{[i]}(\hat{\theta})]'[-\hat{\ell}(\hat{\theta})]^{-1}[\hat{\ell}_{[i]}(\hat{\theta})].$$ (4)

Compared to $C^1_i(\theta)$, it is obvious that $C_i(\theta)$ largely reduces computational loads. For the purpose of diagnostics, $C_i(\theta)$ may be quite adequate for assessing influences, even if the corresponding parameter estimates are poor (Cook and Weisberg, 1982; Kass et al., 1989).

In general, the asymptotic covariance matrix $[-\hat{\ell}(\hat{\theta})]^{-1}$ is not block-diagonal. It may be difficult to decompose the approximated Cook’s statistic (4) into different components of interest, for example, the diagnostic measure for the fixed effects $\beta$. As an alternative, we propose to use $E(\hat{\ell})$, the expectation of $\hat{\ell}$ with respect to the responses $Y$, to replace $\hat{\ell}$, so that we obtain

$$C^*_i(\theta) = [\hat{\ell}_{[i]}(\hat{\theta})]'[-E(\hat{\ell}(\hat{\theta}))]^{-1}[\hat{\ell}_{[i]}(\hat{\theta})].$$ (5)

Note $[-E(\hat{\ell}(\hat{\theta}))]^{-1}$ is actually the Fisher-information matrix. For the LMM (1), it can be shown that $E[\partial^2\ell(\theta)/\partial\beta\partial\eta']_{\beta=0} = 0$ where $\eta = (\alpha', \gamma')'$. Accordingly, Cook’s statistic (5) can be decomposed into $C^*_i(\theta) = C^*_i(\beta) + C^*_i(\eta)$. In other words, the diagnostic measure of $\theta$ is the sum of the diagnostic measures of the fixed effects $\beta$ and the variance components $\eta$ in terms of $C^*_i$.

We can show that the above approximation to the estimate $\hat{\theta}_{[i]}$ can be actually characterized by $O_p(m^{-2})$ under certain regularity conditions. In fact, we have

**Theorem 1.** Suppose that $E\{\hat{\ell}(\hat{\theta})\} = O_p(m)$, $\bar{\ell}(\hat{\theta}) = O_p(m)$, $\hat{\ell}_{[i]}(\hat{\theta}) = O_p(1)$, $\hat{\ell}_{[i]}(\hat{\theta}) - \bar{\ell}(\hat{\theta}) = O_p(1)$ and $\bar{\ell}(\hat{\theta}) - E\{\bar{\ell}(\hat{\theta})\} = O_p(1)$. We then must have

$$[-\hat{\ell}_{[i]}(\hat{\theta})]^{-1}\hat{\ell}_{[i]}(\hat{\theta}) = [-\bar{\ell}(\hat{\theta})]^{-1}\hat{\ell}_{[i]}(\hat{\theta}) + O_p(m^{-2}) = [-E\{\bar{\ell}(\hat{\theta})\}]^{-1}\hat{\ell}_{[i]}(\hat{\theta}) + O_p(m^{-2}).$$ (6)

The proof of Theorem 1 is sketched in Supplementary Section S1.
2.2 \textit{Q}-function-based diagnostics

The Cook’s statistics (4)-(5) may have an analytically explicit form for the LMM with some simple covariance structures. For example, when the model has independent random effects and independent random errors, it is not difficult to obtain the explicit forms of $C_i^*(\beta)$ and $C_i^*(\sigma^2, \sigma^2_e)$, see Section 3 below. However, for other complicated covariance structures it may be difficult to obtain the formulae for $C_i^*(\beta)$ and $C_i^*(\eta)$ where $\eta = (\alpha', \gamma')'$, because the calculation of the derivatives is analytically intractable. On the other hand, it is difficult to extend the likelihood-based case-deletion diagnostics to other models if the associated likelihood function has no analytical form, for example, the generalized linear mixed models. Accordingly, an alternative to the likelihood-based diagnostics needs to be sought.

It is well-known that the EM-algorithm calculates the maximum likelihood estimates in an iterative manner (Dempster et al., 1977). It is natural to use the \textit{Q}-function, the conditional expectation of the logarithm of the joint likelihood function between the responses $Y$ and the random effects $u$, involved in the EM-algorithm, to replace the log-likelihood $\ell(\theta)$ in statistical diagnostics.

Within the framework of incomplete data, Zhu et al. (2001) showed that the resulting one-step approximation is as good as the one based on the log-likelihood in terms of the error bound $O(p(m^{-2})$.

For the LMM (1), the \textit{Q}-function $Q(\theta|\theta^{(t)}) = E[\log f(Y, u|Y, \theta^{(t)})]$ is of the form

$$Q(\theta|\theta^{(t)}) = -\frac{1}{2} \sum_{i=1}^{m} \log |R_i| - \frac{1}{2} \sum_{i=1}^{m} [\hat{r}_i^{(t)}]R_i^{-1}[\hat{r}_i^{(t)}] - \frac{1}{2} \sum_{i=1}^{m} \text{tr}[R_i^{-1}Z_i[\Phi_i^{(t)}]^{-1}Z_i']$$

$$-\frac{m}{2} \log |G| - \frac{1}{2} \sum_{i=1}^{m} \text{tr}[G^{-1}([\Phi_i^{(t)}]^{-1} + [\hat{\phi}_i^{(t)}][\hat{\phi}_i^{(t)}])],$$

where $\hat{r}_i^{(t)} = Y_i - X_i \beta - Z_i \hat{\alpha}_i^{(t)}$, $\hat{\alpha}_i^{(t)} = G(i)Z_i[V(i)^{-1}](Y_i - X_i \beta^{(t)})$, $\Phi_i^{(t)} = [G(i)^{-1} + Z_i[R_i^{(t)}]^{-1}Z_i]$, $V_i^{(t)} = Z_iG(i)Z_i' + R_i^{(t)}$, $G(i) = G(\theta^{(t)})$, $R_i^{(t)} = R_i(\gamma^{(t)})$, and $\theta^{(t)} = (\beta^{(t)'}, \alpha^{(t)'}, \gamma^{(t)'})'$ is the parameter estimate of $\theta = (\beta', \alpha', \gamma')'$ in the $t$-th iteration of the algorithm (see, e.g., Pan and MacKenzie, 2007).

To calculate the case-deletion estimate $\hat{\theta}_{i|\hat{\theta}}$ of $\theta$, Zhu et al (2001) proposed use of the following one-step approximation based on the \textit{Q}-function,

$$\hat{\theta}_{i|\hat{\theta}} = \hat{\theta} + [-\hat{Q}(\hat{\theta}|-\hat{\theta})]^{-1} \hat{Q}_{i|\hat{\theta}}(\hat{\theta}|-\hat{\theta}),$$

where $Q_{i|\hat{\theta}}(\hat{\theta}|-\hat{\theta}) = Q_{i|\hat{\theta}}(\hat{\theta})$ is the \textit{Q}-function formed without the $i$th subject but evaluated at
the maximum likelihood estimate $\hat{\theta}$, and $\dot{Q}_{[i]}(\hat{\theta}|\hat{\theta})$ is the first-order derivative of $Q_{[i]}(\theta|\hat{\theta})$ with respect to $\theta$, also evaluated at $\theta = \hat{\theta}$. In other words, $\dot{Q}_{[i]}(\hat{\theta}|\hat{\theta}) = [\partial Q_{[i]}(\theta|\hat{\theta})/\partial \theta]_{\theta=\hat{\theta}}$. Similarly, $\ddot{Q}(\hat{\theta}|\hat{\theta}) = [\partial^2 Q(\theta|\hat{\theta})/\partial \theta \partial \theta^\prime]_{\theta=\hat{\theta}}$. It is noted that in (8) the matrix $-\ddot{Q}_{[i]}(\theta|\hat{\theta})$ in the Newton-Raphson algorithm is already replaced by $-\ddot{Q}(\theta|\hat{\theta}) = -\ddot{Q}(\theta|\hat{\theta})|_{\theta=\hat{\theta}}$. Accordingly, in the same manner as (4) the $Q$-function-based Cook’s statistic is given by

$$D_i(\theta) = [\dot{Q}_{[i]}(\hat{\theta}|\hat{\theta})]^{-1}[\ddot{Q}_{[i]}(\hat{\theta}|\hat{\theta})],$$

(9)

with the choice of $M = -\ddot{Q}(\hat{\theta}|\hat{\theta})$, which is the same as Equation (10) of Zhu et al. (2001).

When applied to the LMM (1), it can be shown that (9) is the sum of the $Q$-function-based Cook’s statistics for the fixed effects $\beta$ and the variance components $\eta = (\alpha', \gamma')'$. In other words, $D_i(\theta) = D_i(\beta) + D_i(\eta)$. However, for the variance components $\alpha$ and $\gamma$, the diagnostic measures $D_i(\alpha)$ and $D_i(\gamma)$ cannot be obtained directly from $D_i(\eta)$ as the off-diagonal block $[\partial^2 Q(\theta|\hat{\theta})/\partial \alpha \partial \gamma']_{\theta=\hat{\theta}}$ may not be equal to zero. In the spirit of (5), we propose to use $-E[\ddot{Q}(\hat{\theta}|\hat{\theta})]$, the expectation of $-\ddot{Q}(\hat{\theta}|\hat{\theta})$ with respect to the responses $Y$, to replace $-\ddot{Q}(\hat{\theta}|\hat{\theta})$ in (9), leading to the following modified $Q$-function-based Cook’s statistic

$$D'_i(\theta) = [\dot{Q}_{[i]}(\hat{\theta}|\hat{\theta})]^{-1}[E[\ddot{Q}(\hat{\theta}|\hat{\theta})]]^{-1}[\ddot{Q}_{[i]}(\hat{\theta}|\hat{\theta})].$$

(10)

For the LMM (1), it can be shown that the matrix $E[\ddot{Q}(\hat{\theta}|\hat{\theta})]$ is always block-diagonal with respect to the parameters $\beta$, $\alpha$ and $\gamma$, irrespective of the covariance structures for $G$ and $R_i$. Technical details can be found in Supplementary Section S2. The modified $Q$-function-based Cook’s statistic $D'_i(\theta)$ in (10) can thus be decomposed into the three components

$$D'_i(\theta) = D'_i(\beta) + D'_i(\alpha) + D'_i(\gamma)$$

(11)

where $D'_i(\beta)$, $D'_i(\alpha)$ and $D'_i(\gamma)$ are the modified $Q$-function-based Cook’s statistics corresponding to the fixed effects $\beta$, the between-subject covariance components $\alpha$ and the within-subject covariance components $\gamma$, respectively. In other words, the modified $Q$-function-based diagnostic measures for those three sets of parameters are mutually independent in this sense.

The following theorem gives the accuracy of the $Q$-function-based one-step approximation of the estimates $\hat{\theta}_{[i]}$ in (8) and the alternative in which $\ddot{Q}(\hat{\theta}|\hat{\theta})$ is replaced by $E[\ddot{Q}(\hat{\theta}|\hat{\theta})].$
Theorem 2. Assume that $E[\hat{Q}(\hat{\theta} | \hat{\theta})] = O_p(m)$, $\hat{Q}_{ij}(\hat{\theta} | \hat{\theta}) = O_p(1)$ and $\hat{Q}(\hat{\theta} | \hat{\theta}) - E[\hat{Q}(\hat{\theta} | \hat{\theta})] = O_p(1)$. Then we have

$$[-\hat{Q}(\hat{\theta} | \hat{\theta})]^{-1}[\hat{Q}_{ij}(\hat{\theta} | \hat{\theta})] = [-E[\hat{Q}(\hat{\theta} | \hat{\theta})]]^{-1}[\hat{Q}_{ij}(\hat{\theta} | \hat{\theta})] + O_p(m^{-2}).$$

(12)

The proof of Theorem 2 is provided in Supplementary Section S1. Theorem 2 implies that, when replacing the log-likelihood with the Q-function in the EM-algorithm, the one-step approximation to the maximum likelihood estimate $\hat{\theta}_{[i]}$ maintains the same accuracy. Zhu et al. (2001) showed that the assumption in Theorem 2 holds at least for the so-called independent-type-incomplete-data models, in which linear mixed models and generalized linear mixed models are included as special cases.

2.3 Cutoff value for influential cases

As indicated by Cook and Weisberg (1982, p113), Cook’s distance $C_i$ in (4) can be compared to a $\chi^2$-distribution with an appropriate degree of freedom for calibration. Theorem 1 above shows that $C^*_i$ in (5) is a good approximation to $C_i$, so that the $\chi^2$ calibration can also be used as a reference to Cook’s distance $C^*_i$ in (5).

For general models with missing data, under mild conditions Zhu et al. (2001, p732) suggested to use $r/m$ as a rough cutoff value for calibrating the Q-function-based Cook’s statistic $D_i$, where $r$ is the dimension of parameter vector $\theta$ and $m$ is the sample size. Note that Theorem 2 indicates the diagnostic measure $D^*_i$ is a good approximation to $D_i$. Following Zhu et al. (2001), we may choose $r/m$ as a rough guideline for calibrating the diagnostic measure $D^*_i$, where $m$ is the number of subjects. Attention should be paid to these subjects with values of $D^*_i$ or $D_i$ that are significantly larger than $r/m$, as suggested by Zhu et al. (2001).

3 Diagnostics for independent covariances

In Section 2 we propose to use the Q-function-based Cook’s statistics to identify influential subjects for the LMM. In particular, the modified Q-function-based Cook’s statistic $D^*_i(\theta)$ in (10) is highly preferable because it has a transparent statistical interpretation in terms of orthogonal de-
composition of the parameters of interest, as shown in (11). Since the Q-function-based diagnostic measure is an approximation to the likelihood-based diagnostic measure, it is natural to assess how good the approximation is. To make a comparison between these two approaches, we need to obtain the computational formulas for both the Q-function-based and likelihood-based case-deletion diagnostic measures for the LMM. The latter approach, unfortunately, does not always have an analytically closed form and in fact it is the reason that the Q-function-based diagnostic measure is proposed as an alternative. An exception is the LMM with independent random effects and independent random errors. In other words, \( G = \sigma_u^2 I_q \) and \( R_i = \sigma_e^2 I_n_i \), leading to the marginal covariance structure \( V_i = \sigma_u^2 Z_i Z_i' + \sigma_e^2 I_n_i \), where \( \sigma_u^2 > 0 \) and \( \sigma_e^2 > 0 \) are unknown variance components. We study these two approaches within the framework of the LMM with this particular covariance structure in this section.

3.1 Likelihood-based diagnostics

When calculating the maximum likelihood estimates \( \hat{\theta} = (\hat{\beta}', \hat{\sigma}_u^2, \hat{\sigma}_e^2)' \) of the parameters \( \theta = (\beta', \sigma_u^2, \sigma_e^2)' \) in the LMM (1) with \( G = \sigma_u^2 I_q \) and \( R_i = \sigma_e^2 I_n_i \), it is common to take the ratio of variance components \( \lambda = \sigma_u^2 / \sigma_e^2 \) as a new parameter for computational reasons (Laid and Ware, 1982). In statistical diagnostics however, we may be interested in diagnostics for the variance components \( \sigma_u^2 \) and \( \sigma_e^2 \), separately. We thus keep the parameters \( \theta = (\beta', \sigma_u^2, \sigma_e^2)' \) unchanged throughout this paper.

Based on the log-likelihood function (3), it is easy to obtain the score equations satisfied by the maximum likelihood estimates \( \hat{\theta} \). Also, after some algebra we obtain the first-order derivative of \( \ell_{[i]}(\theta) \) evaluated at \( \theta = \hat{\theta} \), that is, \( \hat{\ell}_{[i]}(\hat{\theta}) = \ell_{[i]}(\theta)|_{\theta = \hat{\theta}} \) as

\[
\hat{\ell}_{[i]}(\hat{\theta}) = - \begin{bmatrix}
X_i' \hat{V}_i^{-1} \hat{e}_i \\
\frac{1}{2} (\hat{e}_i' \hat{V}_i^{-1} Z_i Z_i' \hat{V}_i^{-1} \hat{e}_i - \text{tr}(\hat{V}_i^{-1} Z_i Z_i')) \\
\frac{1}{2} (\hat{e}_i' \hat{V}_i^{-2} \hat{e}_i - \text{tr}(\hat{V}_i^{-1}))
\end{bmatrix},
\]

where \( \hat{V}_i = \hat{\sigma}_u^2 Z_i Z_i' + \hat{\sigma}_e^2 I_n_i \), \( \hat{e}_i = Y_i - X_i \hat{\beta} \), and \( \hat{\beta} \), \( \hat{\sigma}_u^2 \) and \( \hat{\sigma}_e^2 \) are the maximum likelihood estimates of \( \beta \), \( \sigma_u^2 \) and \( \sigma_e^2 \), respectively.

Let \( X = (X_1', ..., X_m')' \), \( \hat{V} = \text{diag}(\hat{V}_1, ..., \hat{V}_m) \), \( Z = \text{diag}(Z_1, ..., Z_m) \) and \( \hat{e} = (\hat{e}_1', ..., \hat{e}_m')' \) be the
(n × p), (n × n), (n × mq) and (n × 1) matrices, respectively, where n = \sum_{i=1}^{m} n_i. We can show that the negative of the second-order derivative of the log-likelihood ℓ(θ) with respect to θ, evaluated at θ = \hat{θ}, that is, \(-\ddot{\ell}(\hat{θ}) = -\ddot{\ell}(θ)\)_{\theta=\hat{θ}} must be of the form

\[
-\ddot{\ell}(\hat{θ}) = \begin{bmatrix}
X'\hat{V}^{-1}X & X'\hat{V}^{-1}ZZ'\hat{V}^{-1}\hat{ε} & X'\hat{V}^{-2}\hat{ε} \\
\hat{ε}'\hat{V}^{-1}ZZ'\hat{V}^{-1}X & \frac{1}{2}\text{tr}(\hat{V}^{-1}ZZ'\hat{V}^{-1}ZZ') & \frac{1}{2}\text{tr}(\hat{V}^{-2}ZZ') \\
\hat{ε}'\hat{V}^{-2}X & \frac{1}{2}\text{tr}(\hat{V}^{-2}ZZ') & \frac{1}{2}\text{tr}(\hat{V}^{-2})
\end{bmatrix}
\]

(14)

where \(Ω = 2\hat{V}^{-1}\hat{ε}\hat{ε}' - I_n\). Similarly, we can obtain the Fisher information matrix evaluated at the maximum likelihood estimates \(\hat{θ}\), that is, \(E[-\ddot{\ell}(\hat{θ})] = E[-\ddot{\ell}(θ)]\) as follows

\[
E[-\ddot{\ell}(\hat{θ})] = \begin{bmatrix}
X'\hat{V}^{-1}X & 0 & 0 \\
0 & \frac{1}{2}\text{tr}(\hat{V}^{-1}ZZ'\hat{V}^{-1}ZZ') & \frac{1}{2}\text{tr}(\hat{V}^{-2}ZZ') \\
0 & \frac{1}{2}\text{tr}(\hat{V}^{-2}ZZ') & \frac{1}{2}\text{tr}(\hat{V}^{-2})
\end{bmatrix}
\]

(15)

Accordingly, inserting (13)-(15) into (4) and (5) gives the analytical formulas to calculate the likelihood-based Cook’s statistics \(C_i(θ)\) and \(C_i^*(θ)\) for the LMM with independent random effects and independent random errors. The latter, obviously, can be decomposed into \(C_i^*(θ) = C_i^*(β) + C_i^*(σ_u^2, σ_ε^2)\) due to the block structure of \(E[-\ddot{\ell}(\hat{θ})]\) in (15), where

\[
C_i^*(β) = \hat{ε}_i'\hat{V}_i^{-1}X_i(X'\hat{V}^{-1}X)^{-1}X_i'\hat{V}_i^{-1}\hat{ε}_i.
\]

(16)

In other words, the case-deletion diagnostic measures of the fixed effects \(β\) and the variance components \((σ_u^2, σ_ε^2)\) are mutually independent in the sense of \(C_i^*\). This may not be true in the sense of the statistic \(C_i\) in (4), however. It is noted that (16) has the same form as the one developed by Christensen et al. (1992) and Demidenko and Stukel (2005), where the covariance matrices \(V_i\) are assumed to be given first and then replaced by their maximum likelihood estimates.

The case-deletion diagnostic measure \(C_i^*(σ_u^2, σ_ε^2)\) has an analytically closed form which can be obtained using (15), although it is difficult to separate the diagnostic information for \(σ_u^2\) and \(σ_ε^2\) from \(C_i^*(σ_u^2, σ_ε^2)\). We therefore consider the Q-function-based case-deletion diagnostic analysis below.
3.2 \(Q\)-function-based diagnostics

When the \(i\)th subject is deleted, the \(Q\)-function \(Q_{[i]}(\theta|\hat{\theta})\) can be formed in the same manner as (7). It can be shown that its first-derivative \(\dot{Q}_{[i]}(\theta|\hat{\theta})\) with respect to \(\theta\), evaluated at \(\theta = \hat{\theta}\), must have the form

\[
\frac{\partial}{\partial \theta} \left[ Q_{[i]}(\theta|\hat{\theta}) \right]_{\theta=\hat{\theta}} = - \begin{pmatrix} X_i'\hat{r}_i/\hat{\sigma}_\epsilon^2 \\ \hat{u}_i' + tr(\hat{\Phi}_i^{-1}) - q\hat{\sigma}_u^2/2\hat{\sigma}_u^4 \\ \hat{\sigma}_\epsilon^2 \end{pmatrix},
\]

where \(\hat{r}_i = Y_i - X_i\hat{\beta} - Z_i\hat{u}_i, \hat{u}_i = \hat{\sigma}_u^2Z_i'\hat{V}_i^{-1}(Y_i - X_i\hat{\beta}), \hat{V}_i = \hat{\sigma}_\epsilon^2Z_iZ_i' + \hat{\sigma}_u^2I_n, \hat{\Phi}_i = \hat{\sigma}_u^2I_q + \hat{\sigma}_\epsilon^2Z_i'Z_i, \) and \(\hat{\beta}, \hat{\sigma}_u^2 \) and \(\hat{\sigma}_\epsilon^2\) are the maximum likelihood estimates of \(\beta, \sigma_u^2 \) and \(\sigma_\epsilon^2\), respectively.

On the other hand, based on (7) we can show that the negative second-derivative of the \(Q\)-function \(Q(\theta|\hat{\theta})\) for the LMM with \(G = \sigma_u^2I_q\) and \(R_i = \sigma_\epsilon^2I_n\), evaluated at \(\theta = \hat{\theta}\), is given by

\[
-\ddot{Q}(\theta|\hat{\theta}) = \begin{pmatrix} X'X/\hat{\sigma}_\epsilon^2 & 0 & 0 \\ 0 & tr(\hat{\Phi}^{-1}) - \hat{u}\hat{u}'/2\hat{\sigma}_u^6 & 0 \\ 0 & 0 & \{\hat{r}' + tr(\hat{Z}\hat{\Phi}^{-1}\hat{Z}')}\}/2\hat{\sigma}_\epsilon^6 \end{pmatrix},
\]

where \(\hat{\Phi} = \text{diag}(\hat{\Phi}_1, ..., \hat{\Phi}_m)\) is an \((mq \times mq)\) block-diagonal matrix, and \(\hat{u} = (\hat{u}_1', ..., \hat{u}_m')'\) and \(\hat{r} = (\hat{r}_1', ..., \hat{r}_m')'\) are \((n \times 1)\) vectors of the predicted random effects and conditional residuals, respectively. Taking the expectation in (18), we obtain

\[
-E[\ddot{Q}(\theta|\hat{\theta})] = -E[\ddot{Q}(\theta|\hat{\theta})]_{\theta=\hat{\theta}},
\]

which is equal to

\[
-\mathbb{E}[\ddot{Q}(\theta|\hat{\theta})] = \begin{pmatrix} X'X/\hat{\sigma}_\epsilon^2 & 0 & 0 \\ 0 & tr(\hat{\Phi}^{-1}) - \hat{\sigma}_u^2Z'\hat{V}^{-1}\hat{Z}/2\hat{\sigma}_u^6 & 0 \\ 0 & 0 & \{4tr(\hat{Z}\hat{\Phi}^{-1}\hat{Z}') + n\hat{\sigma}_\epsilon^2\}/2\hat{\sigma}_\epsilon^6 \end{pmatrix}.
\]

Technical details for (17)-(19) are provided in Supplementary Section S3. Accordingly, the \(Q\)-function-based case-deletion diagnostic measures \(D_i\) and \(D_i'\) defined in (9) and (10) both can be decomposed into three components, corresponding to the fixed effects \(\beta\) and the variance components \(\sigma_u^2\) and \(\sigma_\epsilon^2\), respectively. In other words, \(D_i(\theta) = D_i(\beta) + D_i'(\sigma_u^2) + D_i'(\sigma_\epsilon^2)\) and \(D_i'(\theta) = D_i'(\beta) + D_i'(\sigma_u^2) + D_i'(\sigma_\epsilon^2)\), implying that the case-deletion diagnostics for the fixed effects \(\beta\), the between-subject variance \(\sigma_u^2\) and the within-subject variance \(\sigma_\epsilon^2\) are mutually independent in terms of \(D_i\) and \(D_i'\). In particular, the modified \(Q\)-function-based Cook’s statistics \(D_i'\) for \(\beta, \sigma_u^2\) and

\[\text{ACCEPTED MANUSCRIPT} \hfill 12\]
\( \sigma^2 \) and \( \epsilon^2 \) are given by, respectively,

\[
D^*_i(\beta) = \hat{r}_i'X_i(X'X)^{-1}X_i\hat{r}_i/\hat{\sigma}^2, \\
D^*_i(\sigma^2_u) = (\hat{u}_i'\hat{u}_i + \text{tr}(\hat{\Phi}_i^{-1}) - q\hat{\sigma}^2_u)^2/2\hat{\sigma}^2_u(2\text{tr}(\hat{\Phi}^{-1} + \hat{\sigma}^4_uZ'\hat{V}^{-1}Z) - mq\hat{\sigma}^2_u), \\
D^*_i(\sigma^2_\epsilon) = (\hat{r}_i'\hat{r}_i + \text{tr}(Z_i\hat{\Phi}_i^{-1}Z_i') - n_i\hat{\sigma}^2_\epsilon)^2/2\hat{\sigma}^2_\epsilon(4\text{tr}(Z\hat{\Phi}^{-1}Z') + n\hat{\sigma}^2_\epsilon). 
\]

(20)

For the fixed effects \( \beta \), in the literature many authors, including Christensen et al. (1992) and Demidenko and Stukel (2005), obtained the same function form as that given above using different methods. However, none of those methods have closed-form expressions for case-deletion diagnostic measures for the variance components \( \sigma^2_u \) and \( \sigma^2_\epsilon \), and all have computational difficulties when the sample size \( m \) is large. In the numerical examples reported later, we show that the proposed analytical forms of the diagnostic measures for \( \beta \), \( \sigma^2_u \) and \( \sigma^2_\epsilon \) work reasonably well.

### 4 Diagnostics for non-independent covariances

The main strength of the \( Q \)-function-based diagnostic measure \( D^* \) is that it can deal with non-standard or non-independent covariance structures for the LMM. We then turn to consider the case-deletion diagnostic measure for the LMM with more complicated covariance structures for \( G \) and \( R_i \). In addition to the case in Section 3, we study five other structures for the covariance matrices \( G \) and \( R_i \), including (a) unstructured covariance for \( G \) and independent structure for \( R_i = \sigma^2_\epsilon I_{ni} \); (b) unstructured covariance for \( G \) and AR(1) structure for \( R_i = \sigma^2_\epsilon (\rho^{j-k}_\epsilon) \); (c) independent structure for \( G = \sigma^2_u I_q \) and AR(1) structure for \( R_i = \sigma^2_\epsilon (\rho^{j-k}_\epsilon) \); (d) AR(1) structure for \( G = \sigma^2_u (\rho^{j-k}_u) \) and independent structure for \( R_i = \sigma^2_\epsilon I_{ni} \); and (e) AR(1) structure for \( G = \sigma^2_u (\rho^{j-k}_u) \) and AR(1) structure for \( R_i = \sigma^2_\epsilon (\rho^{j-k}_\epsilon) \). To save space, we only present the diagnostic measures for the cases (a) and (b) above and include the other cases in Supplementary Sections S3-S5. It is noted that the same principle can be applied to other structures of covariances that may be more complicated than these in (a)-(e) above.
4.1 Case-deletion measures for $G = G(\alpha)$ and $R_i = \sigma^2_\epsilon I_{n_i}$

For the case (a), denote $G = G(\alpha)$ with $\alpha = (g_{11}; g_{21}, g_{22}; \cdots; g_{q1}, g_{q2}, \cdots, g_{qq})'$. Let $\theta = (\beta', \alpha', \sigma^2_\epsilon)'$. Based on (7), in Supplementary Section S4 we show that $\hat{Q}_{ij}(\hat{\theta}|\hat{\theta})$ must be of the form

$$
\hat{Q}_{ij}(\hat{\theta}|\hat{\theta}) = (-1) \begin{bmatrix}
X_i'\hat{r}_i / \hat{\sigma}^2_\epsilon \\
\hat{a}_i / 2 \\
\{\hat{r}_i'\hat{r}_i + tr(Z_i\hat{\Phi}^{-1}Z_i') - n_i\hat{\sigma}^2_\epsilon / 2\hat{\sigma}^4_\epsilon\}
\end{bmatrix},
$$

(21)

where $\hat{r}_i = Y_i - X_i\hat{\beta} - Z_i\hat{\alpha}_i$, $\hat{\alpha}_i = \hat{\Phi}Z_i'\hat{\gamma}_i^{-1}(Y_i - X_i\hat{\beta})$, $\hat{\Phi} = \hat{G}^1 + \hat{\sigma}^2_\epsilon Z_i'Z_i$, $\hat{\alpha}_i = (\hat{\alpha}_{i1}, \cdots, \hat{\alpha}_{ip})'$ with $\hat{\alpha}_{ik} = tr((\hat{\Phi}^{-1} + \hat{u}_i\hat{\alpha}'_i - \hat{G})\hat{G}^{-1}\hat{G}_k\hat{G}^{-1})$ ($k = 1, \cdots, r$), $\hat{G}_k = (1 - \frac{1}{2}\delta_{k,k}) (E_{k1}k2 + E_{k2}k1)$ with $\delta_{k1,k2} = 1$ for $k_1 = k_2$ and 0 for $k_1 \neq k_2$, $k = k_1 + \frac{1}{2}(2q - k_2)(k_2 - 1)$, and $E_{k1}k2$ is $q \times q$ matrix with $(k_1, k_2)$ entry 1, the others 0. $\hat{\beta}$, $\hat{\alpha}$ and $\hat{\sigma}^2_\epsilon$ are the maximum likelihood estimates of $\beta$, $\alpha$ and $\sigma^2_\epsilon$, respectively.

On the other hand, $-E\{\hat{Q}(\hat{\theta}|\hat{\theta})\}$ is block-diagonal as follows

$$
-E\{\hat{Q}(\hat{\theta}|\hat{\theta})\} = \begin{bmatrix}
X'X / \hat{\sigma}^2_\epsilon & 0 & 0 \\
0 & \hat{S} / 2 & 0 \\
0 & 0 & 4tr(Z\hat{\Phi}^{-1}Z') + n\hat{\sigma}^2_\epsilon / 2\hat{\sigma}^4_\epsilon
\end{bmatrix},
$$

(22)

where $\hat{S} = (\hat{s}_{jk})_{r \times r}$ with $\hat{s}_{jk} = tr(\hat{G}_1\hat{G}_j [2\hat{G}_i^{-1} \sum_{i=1}^m (\hat{\Phi}^{-1} + \hat{G}Z_i'\hat{\gamma}_i^{-1}Z_i\hat{G})\hat{G}^{-1} - m\hat{G}^{-1}\hat{G}_k])$, $\hat{\Phi} = \text{diag}(\hat{\Phi}_1, \cdots, \hat{\Phi}_m)$, and $r = \frac{1}{2}q(q + 1)$. Substituting (21) and (22) into (10), we obtain

$$
D_i'(\alpha) = \hat{\alpha}_i'\hat{S}^{-1}\hat{a}_i / 2,
$$

(23)

while $D_i'(\beta)$ and $D_i'(\sigma^2_\epsilon)$ have the same forms as these given in (20).

4.2 Case-deletion measures for $G = G(\alpha)$ and $R_i = \sigma^2_\epsilon (\rho^{|j-k|}_\epsilon)$

For the case (b), denote $G = G(\alpha)$ with $\alpha = (g_{11}; g_{21}, g_{22}; \cdots; g_{q1}, g_{q2}, \cdots, g_{qq})'$ and $P_i = (\rho^{|j-k|}_\epsilon)$. Then $R_i = \sigma^2_\epsilon P_i$. The vector of parameters in the LMM is $\theta = (\beta', \alpha', \sigma^2_\epsilon, \rho_\epsilon)'$. Denote the maximum likelihood estimates by $\hat{\theta} = (\hat{\beta}, \hat{\alpha}, \hat{\sigma}^2_\epsilon, \hat{\rho}_\epsilon)$. It can be shown that in this case the diagnostic measure $D_i'(\alpha)$ for the variance components $\alpha$ has the same form as (23), and the measures for the fixed
effects $\beta$ and variance components $(\sigma^2_\epsilon, \rho_\epsilon)$ are given by
\[
D_i'(\beta) = \hat{\rho}_\epsilon \hat{P}_i^{-1} X_i' \hat{P}_i^{-1} X_i' \hat{P}_i^{-1} \hat{r}_i / \hat{\sigma}_\epsilon^2,
\]
\[
D_i'(\sigma^2_\epsilon, \rho_\epsilon) = [\hat{c}_i^2 \hat{t}_2 + 2 \hat{c}_i \hat{d} \hat{t} + \hat{d}_i^2 \hat{t}_0] / 2 \hat{\sigma}_\epsilon^2 (\hat{t}_0 - \hat{r}_i),
\]
(24)

where
\[
\hat{r}_i = Y_i - X_i \hat{\beta} - Z_i \hat{u}_i,
\]
\[
\hat{u}_i = \hat{G} Z_i' \hat{V}_i^{-1} (Y_i - X_i \hat{\beta}),
\]
\[
\hat{V}_i = Z_i \hat{G} Z_i' + \hat{\sigma}_\epsilon^2 \hat{P}_i,
\]
\[
\hat{\Phi}_i = \hat{G}^{-1} + \hat{\sigma}_\epsilon^{-2} Z_i' \hat{P}_i^{-1} Z_i,
\]
\[
\hat{c}_i = \hat{r}_i' \hat{P}_i^{-1} \hat{r}_i + tr(\hat{P}_i^{-1} Z_i \hat{\Phi}_i^{-1} Z_i') - n \hat{\sigma}_\epsilon^2,
\]
\[
\hat{d}_i = 2 \hat{\rho}_\epsilon (n_i - 1) \hat{\sigma}_\epsilon^2 - (1 - \hat{\rho}_\epsilon^2) (\hat{r}_i' \hat{P}_i^{-1} \hat{r}_i + tr(\hat{P}_i^{-1} Z_i \hat{\Phi}_i^{-1} Z_i')), \quad (25)
\]
\[
\hat{t}_0 = 4tr(\hat{P}_i^{-1} Z \hat{\Phi}_i^{-1} Z') + n \hat{\sigma}_\epsilon^2,
\]
\[
\hat{t}_1 = (1 - \hat{\rho}_\epsilon^2) [2tr(\hat{P}_i^{-1} Z \hat{\Phi}_i^{-1} Z') + \hat{\sigma}_\epsilon^2 tr(\hat{P}_i^{-1} \hat{P})],
\]
\[
\hat{t}_2 = (1 - \hat{\rho}_\epsilon^2) [2tr(\hat{P}_i^{-1} Z \hat{\Phi}_i^{-1} Z') + \hat{\sigma}_\epsilon^2 tr(\hat{P}_i^{-1} \hat{P})] - 2(1 + \hat{\rho}_\epsilon^2) (n - m) \hat{\sigma}_\epsilon^2
\]

with $\hat{G} = G(\hat{\alpha}), \hat{P}_i = P_i(\hat{\rho}_\epsilon), P = \text{diag}(P_1, \cdots, P_m), \hat{P} = \partial P / \partial \rho_\epsilon |_{\psi_\epsilon = \hat{\rho}_\epsilon}$ and $\hat{P} = \partial^2 P / \partial \rho_\epsilon^2 |_{\psi_\epsilon = \hat{\rho}_\epsilon}$.

5 Diagnostics at observation level

So far, we have discussed the identification of influential subjects for the LMM with various covariance structures of random effects and random errors. For longitudinal data, we have two levels of responses, namely, subjects and repeated measures/observations. In many practical problems, it is also important to identify influential observations. Intuitively, an influential subject may or may not contain influential observations. In this section, we address this issue within the framework of the LMM with independent random effects and independent random errors, as it is the simplest case of the LMM. Similar steps would lead to results for other structures.

When the $j$th observation of the $i$th subject is deleted from the data set, we use the subscript $[i,j]$ denote the corresponding vectors and matrices. For the LMM with independent random effects
and independent random errors, we obtain

$$
\dot{Q}_{i,j}(\hat{\theta}) = (-1) \left[ \frac{[X_i^r \hat{r}_i - X_{i,j}^r \hat{r}_{i,j}]/\hat{\sigma}_e^2}{([\hat{u}_i^r \hat{u}_i + \text{tr}(\hat{\Phi}_i^{-1})) - [\hat{u}_{i,j}^r \hat{u}_{i,j} + \text{tr}(\hat{\Phi}_{i,j}^{-1})]}/2\hat{\sigma}_u^4}
\right],
$$

(26)

in a manner similar to (17), where \( Y_{i,j} \) is the \( i \)th subject’s response vector with the \( j \)th observation \( Y_{i,j} \) deleted, i.e., \( Y_{i,j} = (Y_{i,1}, \ldots, Y_{i,j-1}, Y_{i,j+1}, \ldots, Y_{i,n})' \). Similarly, \( X_{i,j} = (X_{i,j}', \ldots, X_{i,j-1}', X_{i,j+1}', \ldots, X_{i,m}')' \) is the \((n_i - 1) \times p\) design matrix for subject \( i \) in which the \( j \)th row \( X_{i,j} \) is deleted; \( Z_{i,j} = (Z_{i,1}', \ldots, Z_{i,j-1}', Z_{i,j+1}', \ldots, Z_{i,m}')' \) is the \((n_i - 1) \times q\) design matrix for subject \( i \) with the \( j \)th row \( Z_{i,j} \) deleted; \( \hat{r}_{i,j} = Y_{i,j} - X_{i,j} \hat{\beta} - Z_{i,j} \hat{u}_{i,j} \), \( \hat{u}_{i,j} = \hat{\sigma}_u Z_{i,j} \hat{V}_{i,j}^{-1} (Y_{i,j} - X_{i,j} \hat{\beta}) \), \( \hat{V}_{i,j} = \hat{\sigma}_u^2 Z_{i,j} Z_{i,j} + \hat{\sigma}_e^2 I_{n_i}, \) and \( \hat{\Phi}_{i,j} = \hat{\sigma}_u^{-2} I_{q_i} + \hat{\sigma}_e^{-2} Z_{i,j} Z_{i,j} \).

Note that \(-E[\dot{Q}(\hat{\theta} \theta)]\) in (19) is the same for both subject-level and observation-level diagnostic measures. It can be shown that the observation-level Cook’s statistic must be of the form

$$
D_{i,j}^*(\theta) = D_{i,j}^*(\beta) + D_{i,j}^*(\sigma_u^2) + D_{i,j}^*(\sigma_e^2),
$$

(27)

where

$$
D_{i,j}^*(\beta) = (X_i^r \hat{r}_i - X_{i,j}^r \hat{r}_{i,j})'(X_i^r \hat{r}_i - X_{i,j}^r \hat{r}_{i,j})/\hat{\sigma}_e^2,
$$

$$
D_{i,j}^*(\sigma_u^2) = ([\hat{u}_i^r \hat{u}_i + \text{tr}(\hat{\Phi}_i^{-1})) - [\hat{u}_{i,j}^r \hat{u}_{i,j} + \text{tr}(\hat{\Phi}_{i,j}^{-1})]}/2\hat{\sigma}_u^4[2\text{tr}(\hat{\Phi}_i^{-1} + \hat{\sigma}_u^2 Z_i \hat{V}_i^{-1} Z) - mq\hat{\sigma}_e^2],
$$

$$
D_{i,j}^*(\sigma_e^2) = ([\hat{r}_i^r \hat{r}_i - \hat{r}_{i,j}^r \hat{r}_{i,j}) + [\text{tr}(Z_i \hat{\Phi}_i^{-1} Z_i') - \text{tr}(Z_{i,j} \hat{\Phi}_{i,j}^{-1} Z_{i,j}')] - \hat{\sigma}_e^2)}/2\hat{\sigma}_e^2[4\text{tr}(Z_i \hat{\Phi}_i^{-1} Z_i) + n\hat{\sigma}_e^2].
$$

(28)

In other words, the orthogonality of the fixed effects \( \beta \) and the variance components \( \sigma_u^2 \) and \( \sigma_e^2 \) in terms of \( D^* \) still holds when studying observation-level influences.

## 6 Real data analysis: Aerosol data

We illustrate the use of the proposed case-deletion diagnostic measures through a real example in this section. We also compare the proposed \( Q \)-function based case-deletion diagnostics to these based on the likelihood function. All these approaches are then compared to the exact Cook’s
The aerosol data were originally presented by Kerschner et al. (1984). The study involves high-efficiency particulate air filter cartridges that are used in commercial respirators to prevent or reduce the respiration of toxic fumes, radionuclides, dusts, mists, and other particulate matter. The analysis aims to determine factors that contribute most to the variability in penetration of filters and to determine whether the standard aerosol can be replaced by an alternative aerosol in quality assurance testing. Filters fail a quality test if the percentage of penetration of an aerosol is too great. Two aerosols were crossed with two filter manufacturers. Within each manufacturer three filters were used to evaluate the penetration of the two aerosols, so that there were six filters in total. By taking a filter nested within the two manufacturers as a case/subject, Beckman et al. (1987) proposed to use the following LMM with independent random effects and independent random errors

\[ y_{tjkl} = \mu + A_t + M_j + F_{jk} + \epsilon_{tjkl} \]  

(30)

To model the data, where \( y_{tjkl} \) is the percent penetration, \( \mu \) is an intercept, \( A_t \) is a fixed effect for the \( t \)th aerosol type \( (t = 1, 2) \), \( M_j \) is a fixed effect for the \( j \)th filter manufacturer \( (j = 1, 2) \), \( F_{jk} \sim N(0, \sigma_u^2) \) is a random effect for the \( k \)th filter nested within the \( j \)th manufacturer \( (k = 1, 2, 3) \), and \( \epsilon_{tjkl} \sim N(0, \sigma^2) \) is the error associated with the \( l \)th replication in the \( tjk \)th cell \( (l = 1, 2, 3) \), subject to the usual restrictions: \( \sum_t A_t = \sum_j M_j = 0 \). Beckman et al., 1987) developed local influence diagnostics for identifying influential filters and influential observations. Based on the same model as Beckman et al (1987), Christensen et al. (1992) considered likelihood-based case-deletion diagnostics for the fixed effects \( \beta = (\mu, A_1, M_1)' \) and variance components \( (\sigma_u^2, \sigma^2) \).

We now apply the likelihood-based approximated diagnostic measure \( C_i^* \) and the \( Q \)-function-based measures \( D_i \) and \( D_i^* \), developed in Section 3, to the aerosol penetration data analysis. We consider such diagnostic measures for the full parameter vector \( \theta = (\beta, \sigma_u^2, \sigma^2)' \), and the individual parameters \( \beta, \sigma_u^2 \) and \( \sigma^2 \) in the model (30), respectively. We also compare the proposed methods
to the case-deletion diagnostic approach by Christensen et al. (1992). The numerical results are summarized in Table 1, where $C_{\beta i}$ is the diagnostic measure of Christensen et al. (1992) for the fixed effects $\beta$, and $CD_i(\beta)$ is the exact Cook’s statistic for the fixed effects $\beta$ which has a form similar to (29).

Table 1 shows that the proposed case-deletion diagnostic measures $C^*_i$, $D_i$ and $D^*_i$ perform as well as the exact Cook’s statistic in terms of correctly identifying influential cases. In fact, all the diagnostic measures indicate that the 5th subject is the most influential subject for the full parameter vector $\theta$ and the fixed effects $\beta$. This implies that both the proposed likelihood-based measure $C^*_i$ and the $Q$-function based measures $D_i$ and $D^*_i$ are a good alternative to the exact Cook’s statistic $CD_i$. Note that the computation of the exact Cook’s statistic $CD_i$ may be very intensive because it involves the calculation of $\{-\ell_{i|\hat{\theta}}(\hat{\theta})\}^{-1}$ for each $i$, as shown in the simulation studies in Section 7. On the other hand, although Christensen et al.’s (1992) $C_{\beta i}$ indicates that the 5th filter/subject stands out as the most influential, we note that their suggested cutoff value, $\chi^2_{0.05}(3) = 7.8$, seems to be too big to calibrate $C_{\beta i}$. In contrast, the approximated cutoff value $p/m = 3/6 = 0.5$ for $C^*_i(\beta)$ and $D^*_i(\beta)$, as discussed in Section 2.3, seems to be more reasonable and effective than $C_{\beta i}$.

It is noted that the proposed $Q$-function-based case-deletion diagnostic measure $D^*_i(\theta)$ can be orthogonally decomposed into the measures corresponding to the fixed effects $\beta$ and variance components $\sigma_\mu^2$ and $\sigma_\epsilon^2$, respectively. This is the strength of $D^*_i$ that Christensen et al.’s (1992) method does not have. From the values of $D^*_i(\beta)$, $D^*_i(\sigma_\mu^2)$ and $D^*_i(\sigma_\epsilon^2)$ reported in Table 1, it is clear that the 5th filter/subject is the most influential for the full parameter vector $\theta$, and also for the fixed effects $\beta$ and the within-subject variance component $\sigma_\epsilon^2$. However, the 4th filter/subject is the most influential for the between-subject variance component $\sigma_\mu^2$. We note that this information was not revealed by Christensen et al. (1992).

Beckman et al. (1987) considered the local influence analysis for the aerosol penetration data, where various perturbation schemes were studied. By perturbing the random effects variance $\sigma_\mu^2$, they found that the 4th filter/subject is the most influential and commented on that “Surprisingly, the fifth filter, which contains the two anomalous observations identified previously, is not identified as influential. Rather, the fourth filter stands out.” (Beckman et al., 1987; pp423-424). In com-
parison to Beckman’s et al. (1987) local influence analysis, our $Q$-function-based case-deletion measure $D_i^*(\sigma_u^2)$ clearly reports that the fourth filter is influential on the estimate of the random effects variance component $\sigma_u^2$. In addition, the fifth filter is influential on the full model fitting, the estimates of the fixed effects $\beta$ and the random errors variance component $\sigma_\epsilon^2$ as well.

In Section 5, we also discussed the $Q$-function-based case-deletion diagnostic measure $D^*$ at observation level. For the aerosol penetration data, there are six filters and six observations within each filter. We now apply the diagnostic measure $D_{ij}^*$ at the observation level to the data analysis. The index plot of $D_{ij}^*$ for the aerosol data is displayed in Figure 1, where the plots (a) and (b) are the index plots for the fixed effects $\beta$ and the random errors variance component $\sigma_\epsilon^2$, respectively. In Figure 1, we add vertical light lines at indices 6, 12, ..., 36 to indicate the six different filters. Note that in Figure 1 the index order of filters is not natural but is chosen to be the same as that of Beckman et al. (1987), see their Figures 3 and 6, so that comparison can be made directly to their local influence approach. From Figure 1, it is clear that the 13th and 14th observations, which are the first two observations within the 5th filter, are the most influential for the fixed effects $\beta$ and the random errors variance component $\sigma_\epsilon^2$. In particular, the influence of the 14th observation stands out very clearly. This conclusion is in agreement with that by Beckman et al. (1987) and Christensen et al. (1992).

### 7 Simulation studies

In Section 6, we have compared the proposed case-deletion diagnostic methods with Christensen et al.’s (1992) likelihood-based diagnostic approach and Beckman et al.’s (1987) local influence analysis within the framework of the LMM with independent random effects and independent random errors. The comparison there was made through real data analysis for the aerosol penetration data. To evaluate the effectiveness of the proposed diagnostic measures and provide more evidence of the good performance of the proposed methods, we carry out intensive simulation studies for the LMM with both independent covariances and non-independent covariances. We consider the following models.

**Model A:** The simulated data are generated from the model (1) with $m = 50, \beta = (1, 0.2, 0.6)'$,
$u_i \sim N(0, 0.5)$ and $\epsilon_i \sim N_6(0, 0.5I_6)$. The design matrices $X_i$ and $Z_i$ take the similar forms as those in (30) for the aerosol penetration data. In other words, the LMM has independent random effects and independent random errors.

**Model B**: The simulated data are generated from the model (1) with $m = 50$, $\beta = (5, -2, 1)'$, $u_i \sim N_3(0, G)$ and $\epsilon_i \sim N_6(0, R_i)$. The design matrices $X_i$ and $Z_i$, and the covariance matrices $G$ and $R_i$ have the following forms,

\[
X_i = Z_i = \begin{bmatrix}
1 & 1 & 1 \\
1 & 2 & 2^2 \\
\vdots & \vdots & \vdots \\
1 & 6 & 6^2
\end{bmatrix}, \quad G = \begin{bmatrix}
1 & 0.5 & 0.2 \\
0.5 & 1.2 & 0.3 \\
0.2 & 0.3 & 0.9
\end{bmatrix}, \quad R_i = \begin{bmatrix}
1 & 0.5 & \cdots & 0.5^5 \\
0.5 & 1 & \cdots & 0.5^4 \\
\cdots & \cdots & \cdots & \cdots \\
0.5^5 & 0.5^4 & \cdots & 1
\end{bmatrix}.
\]

In other words, the LMM has non-independent covariances and actually it is the case in Section 4.2.

For Models A and B, we generated 500 and 100 simulated data sets, respectively. For each simulated data set, one or two influential subjects or observations are artificially generated using certain perturbation schemes, see the details below.

### 7.1 Subject-level diagnostics

Firstly, under Models A and B we consider the following perturbation scheme: (a) **change the fixed effects $\beta$ to $3\beta$ when generating the responses of the 1st subject $Y_1$**. The $Q$-function-based case-deletion measures in (20) and (24) are computed for each simulated data set, and then averaged over the 500 and 100 simulations for Models A and B, respectively. For comparison, we compute the exact Cook’s distance $CD_i$ for the full parameters $\theta$ and the fixed effects $\beta$. See (29) for the definition of the exact Cook’s distance. We also compute Christensen et al.’s (1992) likelihood-based diagnostic measure $C_{\beta i}$ in our simulation studies. Note that Christensen et al.’s (1992) method assumes that the LMM must have independent random effects and independent random errors, so that it can be directly applied to Model A. When applying their method to Model B, however, Christensen et al.’s (1992) method ignores the non-independent structures for $G$ and $R_i$ and incorrectly assumes independent structures for both $G$ and $R_i$. 

**ACCEPTED MANUSCRIPT**
The diagnostic results are reported in Table 2, where the $\chi^2$-calibration is used for $CD_i$, and $r/m$ is chosen as the rough cutoff value for $D^*_i$, see Section 2.3. An exception is for Christensen et al.'s (1992) $C_{\beta i}$, for which they suggested to use the $\chi^2$ calibration. But we find that the value of the $\chi^2$ calibration is too big for $C_{\beta i}$, a phenomenon we saw in the aerosol data analysis. In fact, in our simulation studies no influential subject was identified when using the $\chi^2$ calibration for $C_{\beta i}$. Instead, we look at the largest value of $C_{\beta i}$ that correctly identifies $Y_1$ as the most influential. Table 2 gives the number of simulations for which $Y_1$ is correctly identified as the most influential.

From Table 2, it is not surprising that the exact Cook’s statistic $CD_i$ is the best diagnostic measure and actually it does not cause any misidentification of influential subjects for both Models A and B. But its computation is very time-consuming, as it involves calculating $[-\hat{\ell}_i(\hat{\theta})]^{-1}$ for each $i$. For the $Q$-function-based diagnostic measure $D^*_i$, under Model A only one misidentification in the 500 simulations occurs for the full parameter vector $\theta$ and 19 out of the 500 simulations for the fixed effects $\beta$. As such, the rate of correct identification of influential subjects achieves 96%. Similarly, under Model B there are 12 misidentifications in the 100 simulations that occur for the fixed effects $\beta$ and the rate of correct identification is 88%. In contrast, Christensen et al.'s (1992) $C_{\beta i}$ seems to work reasonably well for Model A. However, it performs very poorly in identification of influential subject for Model B, where the rate of correct identification is reported to be only 37%. The reason is that it misspecifies the covariance structures of the model and incorrectly takes non-independent covariances as independent ones. Note that the identification for $C_{\beta i}$ is made on the basis of its largest value rather than the $\chi^2$ calibration.

It is interesting to comment on the computation time too. We ran all the simulation studies in S-Plus on a PC (Intel CPU@3.40GHz). For the the exact Cook’s statistic $CD_i$, not surprisingly, it is the most effective but took 54 hours for the 500 simulations under Model A and 80 hours for the 100 simulations under Model B. For Christensen et al.'s (1992) $C_{\beta i}$, the computation time is significantly reduced when compared to $CD_i$, taking 3.5 hours for the 500 simulations under Model A and 41 minutes for the 100 simulations under Model B, respectively. In contrast, our proposed $Q$-function-based method saves substantially on computation time. For example, it only took 49 and 32 minutes to achieve 96% and 88% correct identification of influential subjects for the 500 and 100 simulations, under Models A and B, respectively.
To visualize the performance of $CD_i$ and $D_i^*$ in the simulations, we calculate their values for each simulated data set and then take the average value over the simulation runs. For illustration, the index plot of the averaged measure for Model A is displayed in Figure 2. Figure 2(a) shows that the index plot of $D_i^*(\theta)$ (circle with line) is very close to that of $CD_i(\theta)$ (box with line). Clearly, both the diagnostic measures correctly identify the first subject $Y_1$ as the most influential subject for $\theta$. Figure 2(b) reports the same feature for the fixed effects $\beta$.

Secondly, under Model A we also consider the following perturbation schemes: (b) change the random effects variance component $\sigma_u^2$ to $3\sigma_u^2$ when generating the responses of the 2nd subject $Y_2$; and (c) change the random errors variance component $\sigma_\epsilon^2$ to $3\sigma_\epsilon^2$ when generating the responses of the 3rd subject $Y_3$. Note that the computation of the exact Cook’s statistic $CD_i$ is extremely time-consuming, so that we only compute the proposed diagnostic measure $D_i^*$ for these two schemes. The measure $D_i^*$ performs very well, for example, under the perturbation (b) there is only one misidentification of the influential subject out of the 500 simulations. In Figure 3, we present the index plot of the averaged $D_i^*$ over the 500 simulations. Figure 3(a) is for $D_i^*(\sigma_\epsilon^2)$ where $Y_2$ is perturbed through the perturbation (b), and Figure 3(b) is for $D_i^*(\sigma_u^2)$ with $Y_3$ perturbed using the perturbation (c). Figure 3 clearly shows that $Y_2$ and $Y_3$ are the influential subjects for $\sigma_\epsilon^2$ and $\sigma_u^2$, respectively. We also computed the averaged $D_i^*(\theta)$ value for the full parameters $\theta$ and obtained similar diagnostic results.

Thirdly, from the above analysis it is clear that there is a trade-off between accuracy of diagnostics and computation time and so it is interesting to see how much accuracy is sacrificed when the level of induced perturbation is smaller. Hence, we consider the following perturbation schemes under Model B: (d) change the fixed effects $\beta$ to $2\beta$ when generating $Y_1$; and (e) change the fixed effects $\beta$ to $1.5\beta$ when generating $Y_1$. Note that the model which is under consideration is a LMM with non-independent covariances. We compute the diagnostic measures $CD_i$ and $D_i^*$ for 100 simulations and report the numerical results in Table 3.

From Table 3, we can see that when the level of induced perturbation reduces from $3\beta$ to $2\beta$ or $1.5\beta$, the rate of correct identification of influential subject for $\beta$ is down from 88% to 76% or 65%. In other words, the accuracy of the diagnostic measure $D_i^*$ decreases approximately by 10% to 20% percent in this case. We consider this is still acceptable, because the level of induced perturbation
is relatively small.

Fourthly, we consider a null case in which there is no perturbation as this can help to see whether or not the proposed $Q$-function-based diagnostic measure $D_i^*$ suffers from “false positives”. We generated 500 simulated data sets using Model A without perturbation. We find that there are 12 simulations in which at least one subject is identified as being falsely influential by $D_i^*(\theta)$, where the comparison is made to the cutoff value $r/m = 5/50 = 0.1$. In other words, the rate of false positives is about $12/500=2.4\%$. We therefore consider that the diagnostic measure $D_i^*$ has a fairly small chance of generating false positives.

Finally, we consider the case of multiple influential subjects. Under Model B, we look at the following perturbation scheme: (f) change $\beta$ to $3\beta$ and $2\beta$ when generating $Y_1$ and $Y_2$, respectively. In other words, there are two artificially influential subjects, $Y_1$ and $Y_2$. We ran 100 simulations and computed the diagnostic measure $D_i^*(\theta)$ for each simulated data set. We find that there are only 24 out of 100 simulations that correctly identify both $Y_1$ and $Y_2$ as being influential. This low rate implies that, like other case-deletion methods, our proposed diagnostic measure $D_i^*$ may suffer from the so-called masking effects. We have seen that, as is the situation with case-deletion diagnostics for linear regression models, masking and swamping can be a problem. Approaches to addressing this elsewhere in the literature include multiple deletion diagnostic methods. See for example Hadi and Simonoff (1993). This issue is not considered further here.

### 7.2 Observation-level diagnostics

For the observation-level diagnostic analysis, under Model A we consider the following two perturbation schemes: (g) only the 1st observation of the 1st subject, $Y_{11}$, is perturbed by changing the fixed effects from $\beta$ to $3\beta$, whereas all other observations/subjects are the same as before; and (h) only the 3rd observation of the 3rd subject, $Y_{33}$, is perturbed by inflating the random error variance from $\sigma_\epsilon^2$ to $(3\sigma_\epsilon)^2$, whereas all other observations/subjects are the same as before.

For these two perturbations, we conducted 100 simulations and computed the value of the observation-level diagnostic measure $D_{ij}^*$ for each simulated data set, in order to identify the influential observation for the full parameter vector $\theta$, the fixed effects $\beta$ and the variance components $\sigma_u^2$ and $\sigma_\epsilon^2$. Similar to the previous analysis, we found that the proposed $Q$-function-based diag-
nostic measure $D^*_ij$ performs very well. In fact, the rate of correct identification of the influential observation for the perturbations (g) and (h) is as high as 92%. For illustration, the index plots of the averaged values of $D^*_ij(\beta)$ and $D^*_ij(\sigma^2_\epsilon)$ over 100 simulations are displayed in Figure 4(a) and (b), respectively. Figure 4 shows that $Y_{11}$ and $Y_{33}$ clearly stand out and are hence considered to be the most influential for the fixed effects $\beta$ and the random error variance $\sigma^2_\epsilon$, respectively.

Finally, we compare the proposed case-deletion diagnostic measure $D^*_ij(\theta)$ to the local influence method of Beckman et al (1987), although these two methods are quite different in the range of applicability. Note that Beckman et al.’s (1987) local influence method is only available to the LMM with independent random effects and independent random errors. We therefore choose to use Model A with the perturbation (h) above to generate 100 data sets, so that Beckman et al.’s (1987) local influence method can also be correctly applied. The diagnostic results are reported in Table 4, where $\theta = (\beta', \sigma^2_u, \sigma^2_\epsilon)'$, and $h_{\text{max}}$ represents the local influence diagnostic measure developed by Beckman et al. (1987).

Table 4 shows that these two diagnostic methods perform almost equally as well and both actually correctly identify the influential observation, $Y_{33}$, more than 90% of the time. However, the computation of the local influence method is quite time-consuming, taking 31.25 hours for the 100 simulations. In contrast, the proposed diagnostic measure $D^*_ij$ required only 49 minutes for the 100 simulations, a significant computational savings. Note that the local influence approach of Beckman et al. (1987) is less susceptible to masking effects.

8 Discussion

In this paper, which is based on the $Q$-function in the EM-algorithm, we proposed new case-deletion Cook-type statistics for the LMM, which are shown to correctly identify influential subjects/observations and can be computed efficiently. The major advantage of the proposed diagnostic methods lies in the fact that they can be used for the LMM with very non-standard covariance structures for both random effects and random errors. In contrast, Beckman et al.’s (1987) local influence method and Christensen et al.’s (1992) case-deletion diagnostic approach can only be applied to the LMM with independent random effects and independent random errors. Our inten-
sive simulation studies and real data analysis demonstrate that the proposed diagnostic measures work very well in correctly identifying influential subjects and observations. In terms of computation time, the proposed diagnostic measures are very appealing when compared to existing methods. The reason is that the proposed diagnostic measures have analytical forms even for very non-standard covariance structures. Such analytical forms of the proposed diagnostic measures for a number of combinations of covariance structures specified to random effects and random errors are provided in the paper and the Supplementary Material.

Another major advantage of the proposed $Q$-function-based diagnostic measure $D^*$ is that the diagnostic measure for the full parameter vector $\theta$ can be orthogonally decomposed into terms for the fixed effects $\beta$, the random effects variance components $\alpha$, and the random errors variance components $\gamma$. In other words, diagnostic analysis for these three sets of parameters in the LMM, based on $D^*$, can be made independently. This property is very helpful in the identification of influential subjects and observations when an individual set of parameters is of interest.

Many issues still deserve further research. For example, a fundamental issue is to develop appropriate influence diagnostics by taking into account the number of observations, as it may vary from subject to subject, as pointed out by Zhu et al. (2007). Also, we need to study how to apply the proposed $Q$-function-based diagnostic measures to more complicated models such as generalized linear mixed models, generalized additive mixed models, and time-varying coefficients models, etc. On the other hand, the issue of masking effects needs to be addressed. We are currently working on these topics and will report the results elsewhere.

Acknowledgements

Pan’s research was supported by a grant from the Royal Society of the UK (International Joint Project; 2010/R112497), and Fei’s research was supported by The National Natural Science Foundation of China (11061036). We gratefully acknowledge very helpful and constructive comments and suggestions made by the Editor, Associate Editor and two anonymous referees, which leads to significant improvements to the paper.
Supplementary sections: Supplementary sections containing the proofs of Theorems 1 and 2, the justification of Equations (11), (19)-(25), which are the $Q$-function-based diagnostic measures for the LMM with various covariance structures. (.pdf file)

S-Plus codes: S-Plus codes containing the main codes to compute the Cook-type statistics presented in the paper, for example, AeroCDij is the codes for computing the measures for Figure 1. DY2M3 and DY3 are for Figure 3, and CUY3M3 and DY33M3 are for Table 4, etc. (zip folder containing .ssc files).

Data sets: Data sets used in Sections 6 and 7. (zip folder containing .txt, .csv and .ssd files).

References


Table 1. The exact ($CD_i$), Likelihood-based ($C_i^*$) and $Q$-function-based Cook’s statistics ($D_i$ and $D_i^*$) for aerosol data

<table>
<thead>
<tr>
<th>Subject</th>
<th>$CD_i(\theta)$</th>
<th>$C_i^*(\theta)$</th>
<th>$D_i(\theta)$</th>
<th>$D_i^*(\theta)$</th>
<th>$CD_i(\beta)$</th>
<th>$C_\beta(\beta)$</th>
<th>$C_i^*(\beta)$</th>
<th>$D_i^*(\beta)$</th>
<th>$D_i^*(\sigma_u^2)$</th>
<th>$D_i^*(\sigma_\epsilon^2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.1155</td>
<td>0.8714</td>
<td>0.5765</td>
<td>0.5262</td>
<td>0.386</td>
<td>0.417</td>
<td>0.629</td>
<td>0.3902</td>
<td>0.0020</td>
<td>0.1340</td>
</tr>
<tr>
<td>2</td>
<td>0.8716</td>
<td>0.5570</td>
<td>0.4593</td>
<td>0.3741</td>
<td>0.121</td>
<td>0.116</td>
<td>0.188</td>
<td>0.1338</td>
<td>0.0135</td>
<td>0.2269</td>
</tr>
<tr>
<td>3</td>
<td>0.9003</td>
<td>0.5723</td>
<td>0.4622</td>
<td>0.3744</td>
<td>0.123</td>
<td>0.125</td>
<td>0.195</td>
<td>0.1295</td>
<td>0.0111</td>
<td>0.2338</td>
</tr>
<tr>
<td>4</td>
<td>0.7306</td>
<td>1.4585</td>
<td>0.7293</td>
<td>0.6834</td>
<td>0.558</td>
<td>0.696</td>
<td>0.975</td>
<td>0.4990</td>
<td><strong>0.0624</strong></td>
<td>0.1220</td>
</tr>
<tr>
<td>5</td>
<td><strong>6.8097</strong></td>
<td><strong>8.9948</strong></td>
<td><strong>7.8191</strong></td>
<td><strong>6.4354</strong></td>
<td><strong>2.185</strong></td>
<td><strong>1.568</strong></td>
<td><strong>3.007</strong></td>
<td><strong>2.7483</strong></td>
<td><strong>0.0038</strong></td>
<td><strong>3.6833</strong></td>
</tr>
<tr>
<td>6</td>
<td>0.3361</td>
<td>0.2260</td>
<td>0.1651</td>
<td>0.1427</td>
<td>0.062</td>
<td>0.063</td>
<td>0.098</td>
<td>0.0652</td>
<td>0.0180</td>
<td>0.0595</td>
</tr>
</tbody>
</table>

Table 2. Number of simulations for which $Y_1$ is correctly identified as the most influential

<table>
<thead>
<tr>
<th>Measure</th>
<th>Model A (500 simulations)</th>
<th>Model B (100 simulations)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>For $\theta$</td>
<td>For $\beta$</td>
</tr>
<tr>
<td>$CD_i$</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>$D_i^*$</td>
<td>499</td>
<td>481</td>
</tr>
<tr>
<td>$C_\beta$</td>
<td>–</td>
<td>473</td>
</tr>
</tbody>
</table>

Table 3. Number of simulations for which $Y_1$ is correctly identified as being influential under Model B

<table>
<thead>
<tr>
<th>Measure</th>
<th>Scheme (d) (100 simulations)</th>
<th>Scheme (e) (100 simulations)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>For $\beta$</td>
<td>Running time</td>
</tr>
<tr>
<td>$CD_i$</td>
<td>100</td>
<td>80 hours</td>
</tr>
<tr>
<td>$D_i^*$</td>
<td>76</td>
<td>32 minutes</td>
</tr>
</tbody>
</table>

Table 4. Number of simulations for which $Y_{33}$ is correctly identified as being influential

<table>
<thead>
<tr>
<th>Measure</th>
<th>For $\theta$</th>
<th>Running Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{i,j}^*$</td>
<td>94</td>
<td>49 minutes</td>
</tr>
<tr>
<td>$h_{\text{max}}$</td>
<td>91</td>
<td>31.25 hours</td>
</tr>
</tbody>
</table>
Figure 1: The aerosol data analysis. (a): Index plot of $D^*_ij(\beta)$; (b): Index plot of $D^*_ij(\sigma^2)$. 

Figure 2: Subject-level index plots for simulated data with $Y_1$ perturbed in Model A. (a): Index plot of averaged $CD_i(\theta)$ and $D^*_i(\theta)$ over 500 realizations; (b): Index plot of averaged $CD_i(\beta)$ and $D^*_i(\beta)$ over 500 realizations.
Figure 3: Subject-level index plots for the simulated data where either $Y_2$ or $Y_3$ is perturbed. (a): Index plot of the averaged $D_i^*(\sigma_u^2)$ where $Y_2$ is perturbed; (b): Index plot of the averaged $D_i^*(\sigma_\epsilon^2)$ where $Y_3$ is perturbed.

Figure 4: Observation-level index plots for the simulation data where either $Y_{11}$ or $Y_{33}$ is perturbed. (a): Index plot of $D_{ij}(\beta)$ when $Y_{11}$ is perturbed; (b): Index plot of $D_{ij}(\sigma_\epsilon^2)$ when $Y_{33}$ is perturbed.