Journal of Statistical Research *ISSN 0256 - 422 X*
2017, Vol. 51, No. 1, pp. 1-16

INFLUENCE DIAGNOSTICS FOR MULTIVARIATE GROWTH CURVE MODELS

YUN LING

Department of Biostatistics, University of Pittsburgh Graduate School of Public Health
UPMC Eye Center, Eye and Ear Institute, Ophthalmology and Visual Science Research Center, Department of Ophthalmology, University of Pittsburgh School of Medicine

Email: lingyun0716@gmail.com

STEWART J. ANDERSON*

Department of Biostatistics. University of Pittsburgh Graduate School of Public Health
7130 Parran Hall, 130 Desoto St, Pittsburgh, PA 15261

Email: sja@pitt.edu

RICHARD A. BILONICK

UPMC Eye Center, Eye and Ear Institute, Ophthalmology and Visual Science Research Center, Department of Ophthalmology, University of Pittsburgh School of Medicine
Department of Biostatistics, University of Pittsburgh Graduate School of Public Health

Email: bilonickra@upmc.edu

GADI WOLLSTEIN AND JOEL S. SCHUMAN

NYU Langone Eye Center, New York University School of Medicine
240 E 38th St, New York, NY 10016

Email: {Gadi.Wollstein, Joel.Schuman}@nyumc.org

SUMMARY

Research has shown that in mixed effect longitudinal models, influential observations can have a large effect on the estimates of subject-specific parameters. Furthermore, they cannot always be detected by the classical Cook's distance due to potentially large between-subject variation. Thus, influential observations should be approached by conditioning on the subjects. However, no rigorous approach has been developed for influential observation detection for multivariate longitudinal mixed models where more than one response is measured for each subject at each time point. We propose a multivariate conditional Cook's distance for this more general situation. Examples are given to illustrate how the influential observation in one characteristic changes the effects of both characteristics.

Keywords and phrases: Multivariate; Longitudinal; Influential observations; Mixed effect models

* Corresponding author

1 Introduction

In many medical and epidemiological researches and clinical trials, individuals are measured not only repeatedly, but also with respect to several response variables. Hence, multivariate longitudinal data allow one to study and analyze the joint evolution of multiple response variables over time. The use of multivariate mixed effect models allows one to model a longitudinal process where multiple outcomes are repeatedly measured. These multivariate mixed effect models can accommodate (1) variances of residuals that may be different for different variables; (2) residuals that may be correlated for the same characteristic measured at different time points (within-characteristic correlation); (3) residuals that are also correlated among different characteristics measured at a given time point (inter-characteristic correlation); (4) variables that are often not measured at all time points; and (5) assessments that are not always equally spaced for one or more subjects.

Cook's distance (Cook, 1977) is one of the most important diagnostic tools for detecting influential observations in linear regression for univariate cross-sectional data. Since then, considerable research has been done to extend Cook's distance to detect influential observations in more complex data structures under various kinds of models. For multivariate data, Barrett and Ling (1992) proposed general classes of influence measures for multivariate regression based on analogous forms of univariate Cook's distance. Diaz-Garcia and Gonzalez-Farias (2004) proposed a generalized Cook's distance for elliptical distributions. Hossain and Naik (1989) and Naik (2003) extended deletion of single observation in univariate regression models to the multivariate case. Srivastava and von Rosen (1998) developed a formal test for detecting a single influential observation for a multivariate linear regression model. For longitudinal data, in mixed effect models, these statistics may fail to or incorrectly detect observations influential due to their omission of variances and covariances of associated random effects Tan et al. (2001). Banerjee and Frees (1997) and Banerjee (1998) noticed that the effectiveness of Cook's distance is limited in longitudinal data analysis because it was designed for independent observations and hence, cannot be directly used in the longitudinal setting. Tan et al. (2001) and Ouwens et al. (1999) showed the advantage of using observation-oriented influence measures instead of subject-oriented influence measures because the subject-oriented influence measures may fail to or incorrectly detect influential subjects or influential observations, owing to the relative position of the observations within and across subjects. Tan et al. (2001) proposed a conditional version of Cook's distance by conditioning on the subjects. However, for the detection of influential observations in multivariate mixed effect model, especially multilevel multivariate longitudinal data, no rigorous approach has been developed. Zhu et al. (2012) developed a Bayesian local influence measure method for joint models for longitudinal and survival data. Although assessments of the influence of a model perturbation are generally regarded being useful, a practical and well established approach to influence analysis in statistical modeling is still based on case deletion methods, as pointed out by Lawrance (1990).

The rest of this article is organized as follows. In section 2, we derive the multivariate extension of the conditional Cook's distance. In section 3, we do a simulation study to compare the multivariate conditional Cook's distance and multivariate unconditional Cook's distance, then apply our method to a glaucoma clinical dataset, and present a detailed analysis of the composition of the multivariate conditional Cook's distance.

2 Multivariate Conditional Cook's Distance

Cook's distance is based on the concept of an influence function introduced by Hampel (1974). This concept was applied in a regression setting to the distance measurement between a fitted model and the data by Cook (1977). The idea behind Cook's distance is described as follows. Suppose there is a probability density function. $p(\mathbf{Y}|\boldsymbol{\theta})$, of a random vector $\mathbf{Y}$, where $\boldsymbol{\theta}$ is the vector of the parameters of the probability density function. Cook's distance measures the distance between the maximum likelihood estimators (MLE) of $\boldsymbol{\theta}$ with and without the subset of the data. Let A denote the subset of data to be removed. The new probability density function with A removed is denoted by $p(\mathbf{Y}_{(A)}|\boldsymbol{\theta})$, the MLE of $\boldsymbol{\theta}$ based on the full dataset $\mathbf{Y}$ is $\hat{\boldsymbol{\theta}}$, and the MLE of $\boldsymbol{\theta}$ based on the subsample dataset with A removed, $\mathbf{Y}_{(A)}$, is $\hat{\boldsymbol{\theta}}_{(A)}$, respectively. Hence, Cook's distance for the subset A, denoted by CD(A), is defined as $\text{CD}(A) = (\hat{\boldsymbol{\theta}}_{(A)} - \boldsymbol{\theta})^T \mathbf{B}(\hat{\boldsymbol{\theta}}_{(A)} - \boldsymbol{\theta})$, where $\mathbf{B}$ is a positive definite matrix to be estimated but does not change when the sub dataset is removed.

For multivariate data, longitudinal data, or multivariate longitudinal data, the within subject observations are correlated. Hence, the likelihood function $p(\mathbf{Y} \mid \boldsymbol{\theta})$ has to account for the correlation structure. Thus we set

$$\mathbf{B} = I(\boldsymbol{\theta}) = -\frac{\partial^2 \log(p(\mathbf{Y} \mid \boldsymbol{\theta}))}{\partial\boldsymbol{\theta}\, \partial\boldsymbol{\theta}^T} \tag{2.1}$$

which incorporates the correlation structure of Zhu et al. (2012). Here, $I(\boldsymbol{\theta})$ denotes the Fisher information for $\boldsymbol{\theta}$. In this model, $\boldsymbol{\theta}$ is the vector of the parameters of the probability density function, including both fixed effects and random effects. Multivariate influence measures for models with and without random effects will be developed and compared.

2.1 Model Specification

We now introduce notation for a the multivariate mixed effect model. The model for the observations at j^{th} time point of i^{th} individual is:

$$\underbrace{\mathbf{y}_{ij}}_{m\times 1} = \underbrace{\mathbf{X}_{ij}}_{m\times p} \underbrace{\boldsymbol{\beta}}_{p\times 1} + \underbrace{\mathbf{Z}_{ij}}_{m\times q} \underbrace{\mathbf{b}_i}_{q\times 1} + \underbrace{\boldsymbol{\epsilon}_{ij}}_{m\times 1}, \tag{2.2}$$

where i indicates the number of subjects, $i = 1, 2, \ldots, N$; j indicates the j^{th} measurements for the i^{th} subject, $j = 1, 2, \ldots, n_i$; m indicates the number of characteristics measured for each individual; p is the total number of fixed effects parameters, and q is the total number of random effects parameters.

We assume that $\boldsymbol{\epsilon}_{ij} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{m\times m})$, where $\boldsymbol{\Sigma}_{m\times m}$ is an unstructured variance-covariance matrix. If we stack observations for each individual over time, then the "stacked" error vector for the i^{th} individual has the property that $\boldsymbol{\epsilon}_i \sim N\big(\mathbf{0},\ \mathbf{I}_{n_i} \otimes \boldsymbol{\Sigma}\big)$. The random effects are distributed as $\mathbf{b}_i \sim N_q(\mathbf{0}, \mathbf{G})$ independently for $i = 1, \ldots, N$. Depending on the application, we may allow $\mathbf{G}$ to be either unstructured or block diagonal with m non-zero blocks of size $q_k \times q_k$ corresponding to the m characteristics.

Here, we want to estimate $\boldsymbol{\beta}$, $\mathbf{b}_i$, $\boldsymbol{\Sigma}$ and $\mathbf{G}$. The multivariate mixed effect model with missing value and correlated error term can be fitted using SAS PROC MIXED (version 9.2 or later), the following repeated statement allows one to fit the desired error structure:

```
random int_b1 int_b2 /subject=id type=un g gcorr;
repeated var_type /subject=id*visit_order type=un r rcorr;
```

If we want to fit the multivariate mixed effect model with independent errors structure ($\boldsymbol{\Sigma}$ is diagonal), we simply change the option to `type=vc` in the repeated statement above.

2.2 Multivariate Conditional Cook's Distance

Using a concept similar to Cook's distance and the conditional Cook's distance of Tan et al. (2001), we propose a multivariate longitudinal extension. Conditioning on all of the individuals and each characteristic of the individuals, we have the following log-likelihood:

$$l(\boldsymbol{\Phi}) = -\frac{1}{2}\log|\mathbf{S}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})^T\mathbf{S}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})$$

where $\mathbf{S} = \operatorname{diag}[\mathbf{I}_{n_1} \otimes \boldsymbol{\Sigma}, \ldots, \mathbf{I}_{n_N} \otimes \boldsymbol{\Sigma}]$, and $\boldsymbol{\Phi}$ is the vector containing all the fixed and random effects parameters to be estimated, that is, $\boldsymbol{\Phi} = \{\boldsymbol{\beta}^T, \mathbf{b}^T\}^T$.

The corresponding Fisher Information is given as:

$$\mathbf{B} = I(\boldsymbol{\Phi}) = -\frac{\partial^2\, l(\boldsymbol{\Phi})}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}^T} = \begin{pmatrix} \mathbf{X}^T\mathbf{S}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{S}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{S}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{S}^{-1}\mathbf{Z} \end{pmatrix}$$

and so, the conditional Cook's distance can be written as:

$$\begin{aligned} CD(A) &= (\hat{\boldsymbol{\Phi}}_{(A)} - \hat{\boldsymbol{\Phi}})^T\mathbf{B}(\hat{\boldsymbol{\Phi}}_{(A)} - \hat{\boldsymbol{\Phi}})/c \\ &= (\hat{\boldsymbol{\beta}}_{(A)} - \hat{\boldsymbol{\beta}})^T\mathbf{X}^T\mathbf{S}^{-1}\mathbf{X}(\hat{\boldsymbol{\beta}}_{(A)} - \hat{\boldsymbol{\beta}})/c + (\hat{\mathbf{b}}_{(A)} - \hat{\mathbf{b}})^T\mathbf{Z}^T\mathbf{S}^{-1}\mathbf{Z}(\hat{\mathbf{b}}_{(A)} - \hat{\mathbf{b}})/c \\ &\quad + 2(\hat{\boldsymbol{\beta}}_{(A)} - \hat{\boldsymbol{\beta}})^T\mathbf{X}^T\mathbf{S}^{-1}\mathbf{Z}(\hat{\mathbf{b}}_{(A)} - \hat{\mathbf{b}})/c \\ &= C_{A1} + C_{A2} + C_{A3}, \end{aligned} \tag{2.3}$$

where $c = (Nm - 1)\,q + p$.

From equation (2.3), we can see that $CD(A)$ can be decomposed into three parts: C_{A1}, C_{A2}, and C_{A3}. The term, C_{A1}, is written as

$$\begin{aligned} C_{A1} &= (\hat{\boldsymbol{\beta}}_{(A)} - \hat{\boldsymbol{\beta}})^T\mathbf{X}^T\mathbf{S}^{-1}\mathbf{X}(\hat{\boldsymbol{\beta}}_{(A)} - \hat{\boldsymbol{\beta}})/c \\ &= \sum_{i=1}^{N}\sum_{j=1}^{n_i}(\hat{\boldsymbol{\beta}}_{(A)} - \hat{\boldsymbol{\beta}})^T\mathbf{X}_{ij}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_{ij}(\hat{\boldsymbol{\beta}}_{(A)} - \hat{\boldsymbol{\beta}})/c\,, \end{aligned}$$

and is the total distance measurement for the fixed (marginal) effect between the complete dataset and the data with subset A removed. The expression, $(\hat{\boldsymbol{\beta}}_{(A)} - \hat{\boldsymbol{\beta}})^T\mathbf{X}_{ij}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_{ij}(\hat{\boldsymbol{\beta}}_{(A)} - \hat{\boldsymbol{\beta}})/c$, in

C_{A1} is actually the overall marginal Cook's distance for the i^{th} subject at the j^{th} time point. It is the total distance measurement of m characteristics, but only normalizes the residual variance-covariance matrix, without normalizing the random variance-covariance matrices (Tan, 2001). If we assume the residual covariance matrix is diagonal, that is, $\boldsymbol{\Sigma} = \text{diag}\left[\sigma_1^2, \ldots, \sigma_m^2\right]$, then the expression becomes $\sum_{k=1}^{m}\left[(\hat{\boldsymbol{\beta}}_{(A)} - \hat{\boldsymbol{\beta}})^T \mathbf{x}_{ijk}^T \mathbf{x}_{ijk} (\hat{\boldsymbol{\beta}}_{(A)} - \hat{\boldsymbol{\beta}}) / (c\sigma_k^2)\right]$, which is the simple summation of the distance measurements for all the characteristics, and C_{A1} then becomes

$$C_{A1} = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{m} (\hat{\boldsymbol{\beta}}_{(A)} - \hat{\boldsymbol{\beta}})^T \mathbf{x}_{ijk}^T \mathbf{x}_{ijk} (\hat{\boldsymbol{\beta}}_{(A)} - \hat{\boldsymbol{\beta}}) / c\sigma_k^2 \,.$$

When $\boldsymbol{\Sigma}$ is <u>not</u> diagonal, the total distance measurement for the fixed (marginal) effect also takes into account the correlations among all the m characteristics. Similarly, the second term

$$C_{A2} = (\hat{\mathbf{b}}_{(A)} - \hat{\mathbf{b}})^T \mathbf{Z}^T \mathbf{S}^{-1} \mathbf{Z} (\hat{\mathbf{b}}_{(A)} - \hat{\mathbf{b}})/c = \sum_{i=1}^{N}\sum_{j=1}^{n_i} (\hat{\mathbf{b}}_{i(A)} - \hat{\mathbf{b}}_i)^T \mathbf{Z}_{ij}^T \boldsymbol{\Sigma}^{-1} \mathbf{Z}_{ij} (\hat{\mathbf{b}}_{i(A)} - \hat{\mathbf{b}}_i)/c$$

is the total distance measurement for the random effect parameters between the complete dataset and the data with subset A removed. and the third term

$$\begin{aligned} C_{A3} &= 2(\hat{\boldsymbol{\beta}}_{(A)} - \hat{\boldsymbol{\beta}})^T \mathbf{X}^T \mathbf{S}^{-1} \mathbf{Z} (\hat{\mathbf{b}}_{(A)} - \hat{\mathbf{b}})/c \\ &= 2\sum_{i=1}^{N}\sum_{j=1}^{n_i} (\hat{\boldsymbol{\beta}}_{(A)} - \hat{\boldsymbol{\beta}})^T \mathbf{X}_{ij}^T \boldsymbol{\Sigma}^{-1} \mathbf{Z}_{ij} (\hat{\mathbf{b}}_{i(A)} - \hat{\mathbf{b}}_i)/c \end{aligned}$$

is the distance measure of covariation between the change in the population average profile and the change in the subject-specific profile relative to the population average profile. If we assume the residual covariance matrix is diagonal, that is, $\boldsymbol{\Sigma} = \text{diag}\left[\sigma_1^2, \ldots, \sigma_m^2\right]$, then the terms C_{A2} and C_{A3} can be reduced to simple summations of the distance measurements of all the characteristics for subject-specific effects and covariances, that is,

$$\begin{aligned} C_{A2} &= \sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{m} (\hat{\mathbf{b}}_{i(A)} - \hat{\mathbf{b}}_i)^T \mathbf{z}_{ijk}^T \mathbf{z}_{ijk} (\hat{\mathbf{b}}_{i(A)} - \hat{\mathbf{b}}_i)/c\sigma_k^2 \text{ , and} \\ C_{A3} &= \sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{m} (\hat{\boldsymbol{\beta}}_{(A)} - \hat{\boldsymbol{\beta}})^T \mathbf{x}_{ijk}^T \mathbf{z}_{ijk} (\hat{\mathbf{b}}_{i(A)} - \hat{\mathbf{b}}_i)/c\sigma_k^2. \end{aligned}$$

When $\boldsymbol{\Sigma}$ is not diagonal, the total distance measurements take into account the correlations among all the m characteristics, for subject-specific effects and covariance, respectively.

Note that this is an extension of the work presented by Tan et al. (2001). We have multiple characteristics per individual at each time point whereas Tan had only one characteristic per individual at each time point.

2.3 Simulation Study

The purpose of our simulation study is two fold: (1) To demonstrate the conditional and naive multivariate Cook's distance for a single realization; and (2) To investigate the ability of each method to detect a "known" influential observation.

2.3.1 The Model

For both purposes described above, we generated a bivariate longitudinal dataset for our simulation study. The dataset contains n individuals and each individual has two characteristics, y_{ij1} and y_{ij2}, which are repeatedly measured. The bivariate mixed effect model is:

$$\begin{aligned} y_{ij1} &= \beta_{10} + \beta_{11}u_{i1} + \beta_{12}t_{ij} + b_{1i} + \epsilon_{ij1} \\ y_{ij2} &= \beta_{20} + \beta_{21}u_{i2} + \beta_{22}t_{ij} + b_{2i} + \epsilon_{ij2} \end{aligned} \tag{2.4}$$

where i indicates the individual, $i = 1, \dots, N$; j indicates the time point, $j = 1, \dots, n_i$, which is randomly sampled from $\{1, 2, \dots, 9\}$. The random effects $\mathbf{b}_i = [b_{1i}, b_{2i}]^T$, are generated from a bivariate normal distribution

$$\mathbf{b}_i = \begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} 1 & .2 \\ .2 & 1 \end{pmatrix}\right).$$

The fixed effects design matrix

$$\mathbf{X}_{ij} = \begin{pmatrix} 1 & u_{i1} & t_{ij} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & u_{i2} & t_{ij} \end{pmatrix},$$

where t_{ij} is the j^{th} time point for i^{th} individual. u_{i1} and u_{i2} denote baseline covariates for the two characteristics. The random variables, u_{i1} and u_{i2}, were generated from a bivariate normal distribution

$$\mathbf{u}_i = \begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}\right), t_{ij} = \log(j), \boldsymbol{\epsilon}_{ij} = \begin{pmatrix} \epsilon_{ij1} \\ \epsilon_{ij2} \end{pmatrix} \sim N\left((\mathbf{0}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right).$$

The true components of $\boldsymbol{\beta}$ are $[\beta_{10}, \beta_{11}, \beta_{12}, \beta_{20}, \beta_{21}, \beta_{22}]^T = [1, 1, 1, 1, 1, 1]^T$.

One goal of the simulation is to compare the multivariate conditional and the naive Cook's distances. To do so, we first generated 50 individuals ($N = 50$). Without loss of generality, we set the number of measurements of the 50^{th} individual to be 9. Then we reset $b_{1,50} = 6$ for time point 5. Thus, the observation of y_{ij1} at time point 5 of the 50^{th} individual has a strong influence due to the extreme values of b_{1i}. One thousand datasets using this process were generated.

2.3.2 Demonstration of Methods for One Simulated Dataset

First, we demonstrate a single realization of the 1,000 datasets generated. Figure 1 shows the scattergram of the relationship between the response $[y_{ij1}, y_{ij2}]^T$ and the time points. Note that the diamonds indicate the y_1's and the small circles indicate the y_2's. It can be seen that the fifth observation of the y_1's of individual 50 (diamond), that is, $y_{50,5,1}$, is extremely high. In this dataset, all individuals have at least one measurement and at most nine measurements.

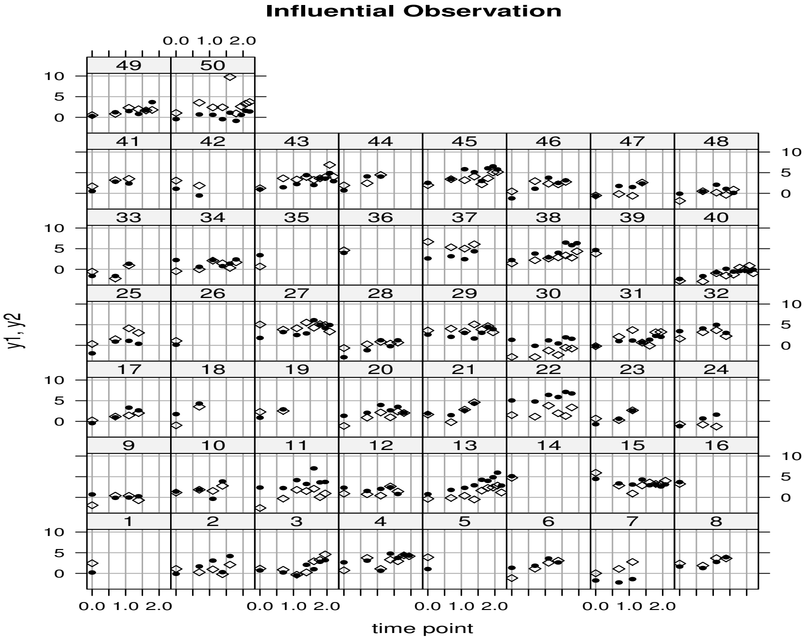

Figure 1: Scattergram of one simulated dataset

Figure 2 shows the multivariate conditional Cook's distance for all observations. Clearly the y_1 value of the fifth measurement of individual 50 was detected. Figure 3 shows the multivariate naive Cook's distance for all observations. Clearly the y_1 value of the fifth measurement of individual 50 was *not* detected.

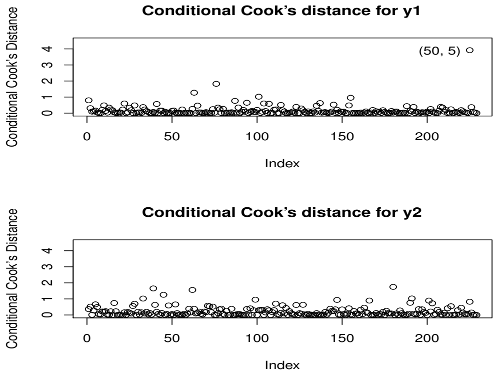

Figure 2: Conditional Cook's Distance for all observations

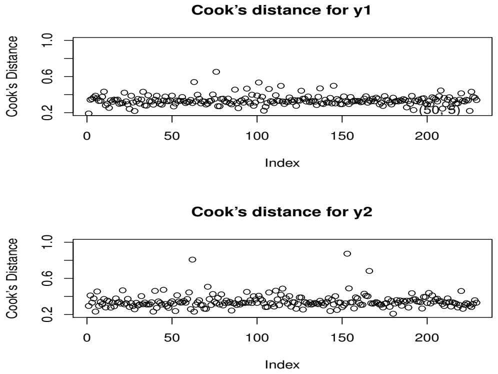

Figure 3: (Unconditional) Cook's Distance for all observations

Figure 4 shows, for each observation, the percentage changes relative to their values with all observations included in the estimated fixed effects of y_1, $\hat{\beta}_{10}$, $\hat{\beta}_{11}$, $\hat{\beta}_{12}$, the estimated fixed effects of y_2, $\hat{\beta}_{20}$, $\hat{\beta}_{21}$, $\hat{\beta}_{22}$, and the estimated random intercept of the 50^{th} individual, $\hat{b}_{50,1}$ and $\hat{b}_{50,2}$. Note that the percentage change for $\hat{b}_{50,1}$ and $\hat{b}_{50,2}$ were divided by 10 so it can be shown more clearly in the plot. The relative change of $\hat{b}_{50,1,y_2}$ is around 51%, not 5%. The diamonds indicate the y_1's and the small circles indicate the y_2's.

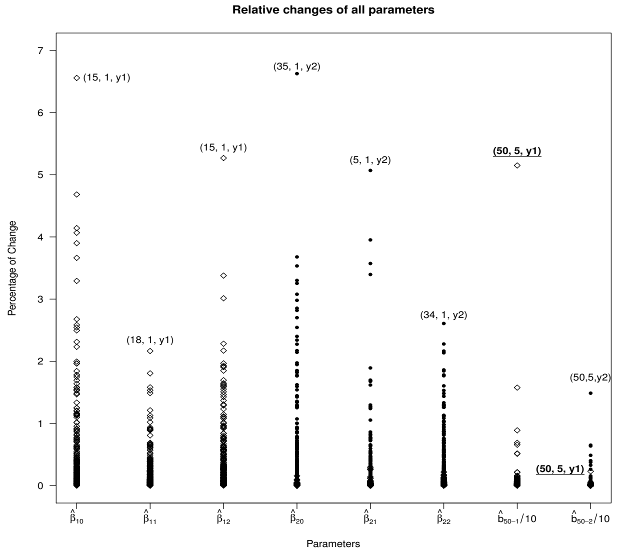

Figure 4: Relative changes for all parameters estimated (in percentage)

Figure 4 indicates that the fifth observation of y_1 of the 50^{th} individual (the extreme observation we made) does not have the largest effect on any of the six fixed effects parameters. But Figure 4 shows that the value of $\hat{b}_{50,1}$ (the random intercept of y_1 of the 50^{th} individual) is strongly influenced by the fifth observation of y_1 of the 50^{th} individual. Figure 4 also shows that observations of y_1 have much stronger influence on y_1's parameters (both fixed effects and random effects) than those of y_2, and similar for y_2. This is understandable.

For the random intercept of y_2 of the 50^{th} individual, of course, one of the observations of y_2 has the largest influence. But it is noticeable that, among the observations of y_1, the fifth observation of the 50^{th} individual (the extreme observation we made) has the largest effect. That is because the two characteristics y_1 and y_2 are correlated (estimated correlation coefficient is 0.3904).

2.3.3 Comparing Performance of Methods

In order to compare our extended conditional Cook's distance to that of the unconditional (original) Cook's distance, we repeated the simulation 1,000 times. Accordingly, we generated 1,000 datasets using the model in Equation 2.4, and then use our method to detect a "known" influential observation in the 1,000 datasets. For each of the 1,000 datasets, a bivariate linear mixed effect model was fitted, and the model parameters and variance-covariance matrices were calculated. The average of the 1,000 estimated model parameters, random effect variance-covariance matrix ($\mathbf{G}$ matrix) and the residual variance-covariance matrix ($\mathbf{\Sigma}$ matrix) are listed below. Table 1 shows the average estimates of the fixed effect parameters and the standard deviations for 1000 simulations.

Table 1: The average estimates of the fixed effect parameters for 1000 simulations

Parameters	Estimated value	Standard Deviation
$\hat{\beta}_{10}$	1.0021	0.0341
$\hat{\beta}_{11}$	0.9992	0.0257
$\hat{\beta}_{12}$	1.0164	0.0100
$\hat{\beta}_{20}$	0.9922	0.0448
$\hat{\beta}_{21}$	0.8018	0.0326
$\hat{\beta}_{22}$	0.9998	0.0110

The estimated $\mathbf{G}$ matrix and its associated correlation matrix, $\mathbf{G}$ are:

$$\overline{\hat{\mathbf{G}}} = \begin{array}{c} \\ b_1 \\ b_2 \end{array} \begin{array}{c} \begin{array}{cc} \;\; b_1 \;\;\;\; & \;\;\;\; b_2 \;\; \end{array} \\ \begin{pmatrix} 0.8996 & 0.1925 \\ 0.1925 & 1.3020 \end{pmatrix} \end{array} \qquad \overline{\hat{\mathcal{G}}} = \begin{array}{c} \\ b_1 \\ b_2 \end{array} \begin{array}{c} \begin{array}{cc} \;\; b_1 \;\;\;\; & \;\;\;\; b_2 \;\; \end{array} \\ \begin{pmatrix} 1.0000 & 0.1778 \\ 0.1778 & 1.0000 \end{pmatrix} \end{array}$$

We can see that the estimated $\overline{\hat{\rho}}_G = 0.1778$ and the true value $\rho_G = 0.20$. The model fits well for the 1,000 simulations.

The estimated $\mathbf{\Sigma}$ matrix averaged over the 1,000 simulations and the associated correlation matrix are:

$$\overline{\hat{\mathbf{\Sigma}}} = \begin{array}{c} \\ y_1 \\ y_2 \end{array} \begin{array}{c} \begin{array}{cc} \;\; y_1 \;\;\;\; & \;\;\;\; y_2 \;\; \end{array} \\ \begin{pmatrix} 1.2171 & 0.4976 \\ 0.4976 & 0.9962 \end{pmatrix} \end{array} \qquad \overline{\hat{\varSigma}} = \begin{array}{c} \\ y_1 \\ y_2 \end{array} \begin{array}{c} \begin{array}{cc} \;\; y_1 \;\;\;\; & \;\;\;\; y_2 \;\; \end{array} \\ \begin{pmatrix} 1.0000 & 0.4519 \\ 0.4519 & 1.0000 \end{pmatrix} \end{array}$$

We can see the estimated $\hat{\rho}_R = 0.4519$ and the true value $\rho_R = 0.50$. The model fits well.

Table 2: Number of detections for conditional and original Cook's distance

Conditional	Original Cook's D		
Cook's D	No	Yes	Total
No	75	0	75
Yes	663	262	925
Total	738	262	1000

Our multivariate conditional Cook's distance successfully detected the "known" influential observation in 925 of the 1,000 datasets. The original Cook's distance only detected the "known" influential observation in 262 of the 1,000 datasets. In Table 2, a contingency table for the multivariate conditional Cook's distance and the original Cook's distance summarizes the result from the 1,000 simulations.

2.4 Application

We applied our method to clinical data obtained from a study in patients with glaucoma. In this study, the patients' eyes were repeatedly measured resulting in multiple responses. Specifically, we jointly modeled the thicknesses of the retinal nerve fiber layer (RNFL) and the retinal ganglion cells complex (GCC), which had been repeatedly measured, to find out if there were abnormal observations. The dataset is from the Eye Center at the University of Pittsburgh Medical Center (UPMC).

Our dataset was derived from information on 487 eyes from 256 patients. Because the two eyes from each patient are typically correlated, we randomly chose one eye from each of 256 patients (which is a common practice in the opthalmology literature). Hence, for demonstrating our method, we have 256 eyes, each having the two measurements described above at each time point. The follow-up duration varied from 1.3 to 6.4 years. All eyes were divided into three diagnostic groups: healthy(H), glaucoma suspect(GS) and glaucoma(G). In the complete dataset, there were 97 healthy eyes, 279 glaucoma suspect eyes and 111 glaucoma eyes. Patient baseline age (in years) and diagnostic group were included as covariates.

2.4.1 The Model

We fitted the following bivariate linear mixed effect model:

$$\begin{aligned} Y_{RNFL} &= (\beta_{10} + \beta_{11}GS + \beta_{12}G + b_{10}) + (\beta_{13} + \beta_{14}GS + \beta_{15}G + b_{11})\, Fu + \beta_{16}Age + \epsilon_{RNFL} \\ Y_{GCC} &= (\beta_{20} + \beta_{21}GS + \beta_{22}G + b_{20}) + (\beta_{23} + \beta_{24}GS + \beta_{25}G + b_{21})\, Fu + \beta_{26}Age + \epsilon_{GCC} \end{aligned}$$

where Age indicates baseline age; and Fu indicates the time of Follow-up (in years). We assume that

$$\mathbf{b} = [b_{10}, b_{20}, b_{11}, b_{21}]^T \sim N(\mathbf{0}, \mathbf{G}) \text{ and } \boldsymbol{\epsilon} = [\epsilon_{RNFL}, \epsilon_{GCC}]^T \sim N(\mathbf{0}, \mathbf{\Sigma}_{2\times 2})$$

Table 3: The estimated fixed effects parameters

β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}
111.60	−6.32	−0.185	−0.059	−0.616	−0.705	−17.22
β_{20}	β_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}
107.70	−5.22	−0.221	−0.972	0.380	0.497	−12.74

Table 3 shows the estimated parameters of the fixed effects. As can be seen from the parameter estimates, RNFL and GCC both have similar relationships with the age of the patient but demonstrate different trajectories within the groups.

The estimated variance-covariance matrix of the random effects, $(\mathbf{G})$, is

$$\hat{\mathbf{G}} = \begin{array}{c} \\ b_{10} \\ b_{20} \\ b_{11} \\ b_{21} \end{array} \begin{array}{c} \begin{array}{cccc} b_{10} & b_{20} & b_{11} & b_{21} \end{array} \\ \begin{pmatrix} 95.797 & 51.865 & -0.964 & -0.281 \\ 51.865 & 52.309 & -0.0696 & -0.271 \\ -0.964 & -0.0696 & 0.882 & 0.108 \\ -0.281 & -0.271 & 0.108 & 0.753 \end{pmatrix} \end{array}$$

The correlation matrix of $\mathbf{G}$ is:

$$\hat{\mathcal{G}} = \begin{array}{c} \\ b_{10} \\ b_{20} \\ b_{11} \\ b_{21} \end{array} \begin{array}{c} \begin{array}{cccc} b_{10} & b_{20} & b_{11} & b_{21} \end{array} \\ \begin{pmatrix} 1.000 & 0.733 & -0.105 & -0.0331 \\ 0.733 & 1.000 & -0.0103 & -0.0433 \\ -0.105 & -0.0103 & 1.000 & 0.132 \\ -0.0331 & -0.0433 & 0.132 & 1.000 \end{pmatrix} \end{array}$$

The estimated Residual Variance-Covariance Matrix is:

$$\hat{\mathbf{\Sigma}} = \begin{array}{c} \\ Y_{RNFL} \\ Y_{GCC} \end{array} \begin{array}{c} \begin{array}{cc} Y_{RNFL} & Y_{GCC} \end{array} \\ \begin{pmatrix} 13.842 & 0.184 \\ 0.184 & 8.459 \end{pmatrix} \end{array}$$

Its associated correlation matrix is

$$\hat{\mathit{\Sigma}} = \begin{array}{c} \\ Y_{RNFL} \\ Y_{GCC} \end{array} \begin{array}{c} \begin{array}{cc} Y_{RNFL} & Y_{GCC} \end{array} \\ \begin{pmatrix} 1.0000 & 0.0169 \\ 0.0169 & 1.0000 \end{pmatrix} \end{array}$$

, indicating a nearly independent structure in the RNFL and GCC residuals.

2.4.2 The Influential Observations (Observation Level)

Using our method, we calculated the observation level conditional Cook's distance and the decomposed C_{A1}, C_{A2} and C_{A3}. "Observation level" means that the subset A to be removed is the whole observation of the i^{th} subject at j^{th} time point. That is, A contains both RNFL and GCC values measured at the j^{th} time point for the i^{th} subject.

Figure 5 illustrates the bivariate observations for 10 eyes, and Table 4 lists the 10 observations with largest values of the conditional Cook's distance. Note that the diamonds and thick dashed lines indicate the observed RNFL values and individual fitted regression lines for RNFL, respectively. The light dashed lines indicate the marginal fitted regression lines for RNFL. Similarly, the small solid circles, thick solid lines and light solid lines are for the GCC measurements.

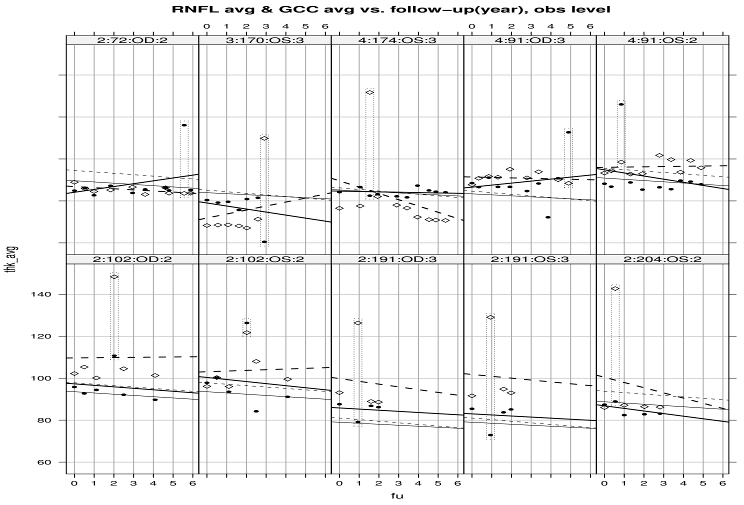

Figure 5: Bivariate observations with largest conditional Cook's Distance in 10 eyes

Table 4 shows the values of the three decomposition terms of the multivariate conditional Cook's distance. We notice that for most observations, the distance measurement of the random effects are much greater than the distance measurement of the fixed effects, that is, $C_{A2} \gg C_{A1}$. This is obvious because the subject-specific effects are much more sensitive to the influential observation than the marginal effects. Also, the covariance between the distance measurements of fixed and random effects, that is, C_{A3}, is very small. This is similar to the conclusion by Tan et al. (2001) made for the univariate case.

2.4.3 The Influential Observations (Component Level)

Using our method, we also calculated the component level conditional Cook's distance and the decomposed C_{A1}, C_{A2} and C_{A3}. "Component level" means that the subset A to be removed is only one component of the whole observation of the i^{th} subject at j^{th} time point. That is, A contains only one of the two components, either the RNFL or GCC value measured at the j^{th} time point for the i^{th} subject. Figure 6 illustrates 10 components in 10 eyes, and Table 5 shows the list of the 10 components with the largest values of conditional Cook's distance.

3 Concluding Remarks

Case-deletion methods are an effective diagnostic algorithm to detect influential observations and outliers. We have developed a method for identifying outliers for multivariate longitudinal observations by extending the conditional Cook's distance proposed by Tan et al. (2001). Our method takes into account both the serial (within-characteristic) correlation and the inter-characteristic correlation. We use random effects to dominate the within-characteristic correlations among different time

Table 4: Decomposition of the Conditional Cook's Distance for the 10 observations with largest conditional cook's distance

Eye ID	Follow-up	Follow-up	Diagnostic	Conditional			
	(in years)	(in days)	group	Cook's D	C_{A1}	C_{A2}	C_{A3}
2:204:OS	0.5448	199	GS	0.003597	0.000022	0.003577	-0.000002
2:103:OD	2.0205	738	GS	0.002645	0.000036	0.002615	-0.000006
2:102:OS	2.0205	738	GS	0.002321	0.000033	0.002294	-0.000006
3:170:OS	2.9103	1063	G	0.002292	0.000101	0.002240	-0.000048
4:174:OS	1.5387	562	G	0.002151	0.000059	0.002102	-0.000010
2:191:OS	0.9391	343	G	0.001932	0.000035	0.001905	-0.000009
4:91:OS	0.8487	310	GS	0.001921	0.000013	0.001908	0.000000
4:92:OD	4.8953	1788	G	0.001848	0.000027	0.001811	0.000010
2:72:OD	5.5633	2032	GS	0.001822	0.000008	0.001807	0.000007
2:194:OD	0.9391	343	G	0.001767	0.000031	0.001743	-0.000007

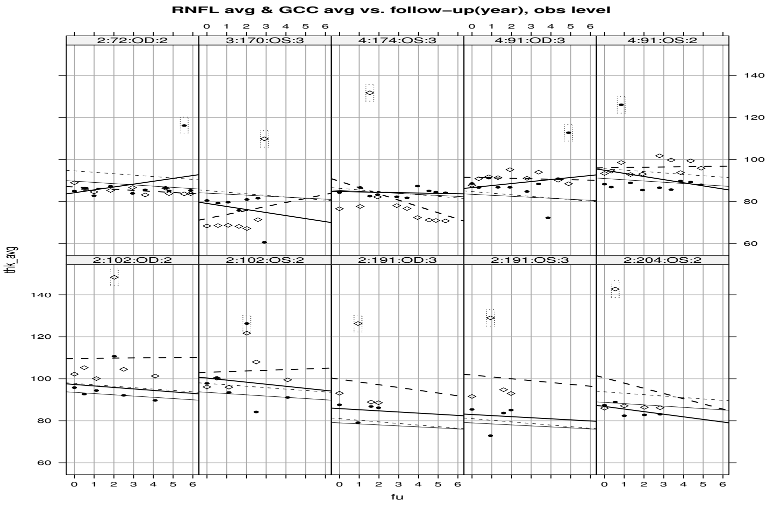

Figure 6: 10 components with the largest conditional cook's distance in 10 eyes

Table 5: Decomposition of Conditional Cook's Distances (CCDs) for the Ten Components with the Largest CCDs

Eye ID	Type	Follow-up (in years)	Follow-up (in days)	Diagnostic group	Conditional Cook's D	C_{A1}	C_{A2}	C_{A3}
2:204:OS	RNFL	0.5448	199	GS	0.003488	0.000022	0.003471	-0.000002
4:174:OD	RNFL	1.5387	562	G	0.002157	0.000059	0.002118	-0.000010
2:103:OD	RNFL	2.0205	738	GS	0.002074	0.000029	0.002054	-0.000005
4:91:OS	GCC	0.8487	310	GS	0.001888	0.000013	0.001875	0.000000
4:92:OD	GCC	4.8953	1788	G	0.001857	0.000027	0.001811	0.000010
2:72:OD	GCC	5.5633	2032	GS	0.001830	0.000008	0.001808	0.000008
2:191:OS	RNFL	0.9391	343	G	0.001828	0.000031	0.001812	-0.000008
2:102:OS	GCC	2.0205	738	GS	0.001803	0.000027	0.001786	-0.000005
3:170:OS	RNFL	2.9103	1063	G	0.001782	0.000084	0.001777	-0.000039
2:194:OD	RNFL	0.9391	343	G	0.001763	0.000029	0.001748	-0.000007

points, and the residual variance-covariance matrix to handle the correlations among different characteristics.

We also explored the three parts of the multivariate conditional Cook's distance : (1) influences on the estimated average profile (fixed effect parameters), (2) on the individual-specific parameters, and (3) on the covariance between the average profile and individual profiles. We show that for each component, the measurement of the influence is a combination of the influence measurements on all characteristics. If the characteristics are independent from each other, then the measurement of the influence is simply the summation of the influence measurements on all characteristics.

Our simulation results show that, similar to the finding by Tan et al. (2001) for one observation per time point, the conditional Cook's distance is superior to the unconditional Cook's distance in a *multivariate* longitudinal data analysis. We also showed in our simulation study that our method successfully detected the influential vector component for a large percentage (92.5%) of datasets whereas the unconditional Cook's distance only detected that component in a relatively small percentage (26.2%) of the datasets.

Acknowledgements

This work was partially supported by the grant, NIMH–P30 MH090330 (Anderson); and NIH contracts, R01 EY013178, P30 EY008098 (Bethesda, MD); Eye and Ear Foundation (Pittsburgh, PA); Research to Prevent Blindness (New York, NY).

References

Banerjee, M. (1998), “Cook’s distance in linear longitudinal models,” *Communs Stat. Theory Meth*, 27, 2973–2983.

Banerjee, M. and Frees, E. W. (1997), “Influence diagnostics for linear longitudinal models,” *Journal of American Statistics Association*, 92, 999–1005.

Barrett, B. E. and Ling, R. F. (1992), “General Classes of influence measures for multivariate regression,” *Journal of American Statistic Association*, 87, 187–191.

Cook, R. D. (1977), “Detection of influential observation in linear regression,” *Econometrics*, 19, 15–18.

Diaz-Garcia, J. A. and Gonzalez-Farias, G. (2004), “A note on the Cook’s Distance,” *Journal of Statistical Planning and Inference*, 120, 119–136.

Hampel, F. R. (1974), “The influence curve and its role in robust estimation,” *Journal of American Statistical Association*, 69, 383–393.

Hossain, A. and Naik, D. N. (1989), “Detection of influential observation in multivariate regression,” *Journal of Applied Statistics*, 16, 25–37.

Lawrance, A. J. (1990), *Directions in Robust Statistics and Diagnostics, Part I*, vol. 29, Springer Verlag.

Naik, D. N. (2003), “Diagnostic methods for univariate and multivariate normal data,” *Handbook of Statistics*, 22, 957–993.

Ouwens, J. N. M., Tan, F. E. S., and Berger, M. P. F. (1999), “Local Influence for repeated measures generalized linear mixed models,” *Proceedings 14th International Workshop Statistical Modeling*.

Srivastava, M. S. and von Rosen, D. (1998), “Outliers in multivariate regression model,” *Journal of Multivariate Analysis*, 65, 195–208.

Tan, F. E. S., Ouwens, M. J. N., and Berger, M. P. F. (2001), “Detection of Influential Observation in Longitudinal Mixed Effects Regression Data,” *Journal of the Royal Statistical Society*, 50, 271–284.

Zhu, H., Ibrahim, J. G., Chi, Y.-Y., and Tang, N. (2012), “Bayesian Influence Measures for Joint Models for Longitudinal and Survival Data,” *Biometrics*, 68, 954–964.

Received: June 30, 2017

Accepted: July 29, 2017