A diagnostic of influential cases based on the information complexity criteria in generalized linear mixed models

Junfeng Shang
Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, Ohio

ABSTRACT
Modeling diagnostics assess models by means of a variety of criteria. Each criterion typically performs its evaluation upon a specific inferential objective. For instance, the well-known DFBETAS in linear regression models are a modeling diagnostic which is applied to discover the influential cases in fitting a model. To facilitate the evaluation of generalized linear mixed models (GLMM), we develop a diagnostic for detecting influential cases based on the information complexity (ICOMP) criteria for detecting influential cases which substantially affect the model selection criterion ICOMP. In a given model, the diagnostic compares the ICOMP criterion between the full dataset and a case-deleted dataset. The computational formula of the ICOMP criterion is evaluated using the Fisher information matrix. A simulation study is accomplished and a real data set of cancer cells is analyzed using the logistic linear mixed model for illustrating the effectiveness of the proposed diagnostic in detecting the influential cases.

1. Introduction

In statistical modeling, numerous modeling diagnostics have been employed to identify influential cases for various inferential objectives. For instance, in linear regression, measures such as COOK’s distance, DFBETAS, DFFITS, studentized residuals, and COVRATIO are popularly applied to reveal influential cases which substantially impact the fitted model and associated results (see Belsley et al., 1980; Cook and Weisberg, 1982).

However, little attention has been put on generalized linear mixed modeling diagnostics. The generalized linear models (GLMs) are an extension of the linear modeling process that allows models to be fit to the data that follow probability distributions other than the normal distribution, such as the Poisson, binomial, and multinomial distributions. The generalized linear mixed model (GLMM) is named when random effects are involved in GLMs, and the application of GLMMs can be even more extensively carried out than GLMs, therefore detecting influential cases in generalized linear mixed models becomes quite crucial.

The identification of influential cases involves the problem of model selection since the detected influential cases may be caused by the simplicity of the model. Bozdogan and Bearse
(2003) developed a modeling diagnostic using the information complexity (ICOMP) criteria in dynamic multivariate linear models. In their work, influential case detection and model selection have been addressed jointly. Shang (2008) developed a modeling diagnostic using the ICOMP criterion in the linear mixed modeling setting, which also addressed the model selection along with the proposed diagnostic.

We develop a diagnostic for detecting influential cases based on the ICOMP criteria in the generalized linear mixed modeling framework. The diagnostic compares the information complexity criterion between the full dataset and a case-deleted dataset. The ICOMP criterion is computed from the Fisher information matrix. A simulation study is completed and a real data set of cancer cells is analyzed using the logistic linear mixed model for illustrating the effectiveness of the proposed diagnostic.

2. The Information Complexity (ICOMP) criterion in generalized linear mixed models

2.1. Generalized linear models with random effects

A generalized linear model with random effects is defined in what follows. Let \( Y = (y_1, \ldots, y_N)' \) be a vector of \( N \) observations with the mean \( \mu \) and variance \( V \), and let \( \epsilon \) be a vector of random errors with zero expectation. Let further \( g(\cdot) \) be the link function, which is monotone, such that \( g(\mu) \) can be written as the linear model \( g(\mu) = \eta = X\beta + U\xi \), where \( X_{N \times p} \) is a known design matrix, the \( \beta \) is a vector of fixed effects, the \( U \) is an \( N \times q \) known matrix, and the \( \xi \) is a \( q \times 1 \) vector of random effects.

For the purpose of further analysis, the \( Y \) data can be linearized (McCullagh and Nelder, 1989, p. 31), and then the link function \( g(\cdot) \) is re-written by providing the first order as

\[
g(Y) = g(\mu) + (Y - \mu)g'(\mu) = Z = X\beta + U\xi + \epsilon g'(\mu),
\]

and therefore \( Z \) is called the adjusted dependent variable. Correspondingly, let \( Z = (z_1, \ldots, z_N)' \) be a vector of \( N \) observations. From now on, instead of using the \( Y \), we will utilize the \( Z \) to propose the diagnostic of influential cases in modeling. We know that \( E(Z) = X\beta \) and \( \text{cov}(\xi) = D \). Therefore, \( \text{cov}(\epsilon g'(\mu)) = V(\partial \eta/\partial \mu)^2 = W \), \( \text{cov}(Z) = W + UDU' = \Sigma \), and \( W = V(\partial \eta/\partial \mu)^2 = V \text{diag}(\partial \eta_i/\partial \mu_i)^2, i = 1, \ldots, N \).

Model (2.1) is a linear random effects model with the adjusted dependent variable \( Z \) instead of the \( Y \) and therefore is considered as a linear mixed model. However, the covariance matrix \( \Sigma \) here is not as simple as that in a linear mixed model because it depends on a function of the fixed effects \( \beta \).

Note that model (2.1) is derived by the first order Taylor expansion. For the parameter estimation in the GLMs, the similar idea has been adopted (see McCullagh and Nelder, 1989), and then based on it, the maximum likelihood estimators can be calculated. Analogously, for the parameter estimation in GLMMs, model (2.1) is also employed to conduct the estimation (see Schall, 1991). From such applications, it can be concluded that the linearized model (2.1) is proper for quite a few inferences. Therefore, we assume the linearized model (2.1), even though it is a special case, this assumption is adequate for proposing a modeling diagnostic.

2.2. The Information Complexity (ICOMP) criterion in generalized linear mixed models

For the purpose of comparison with the ICOMP criterion, we first present and comment on the most well-known model selection criterion, the Akaike Information Criterion (AIC,
Akaike, 1973, 1974). The AIC is given by

$$\text{AIC} = -2 \log L(\hat{\theta} | Z) + 2k,$$

where $L(\hat{\theta} | Z)$ is the maximized likelihood function, and $k$ represents the dimension of estimated parameter $\hat{\theta}$ under the given model. The structure of AIC solidly reflects an underlying principle for model selection criteria, that is, a model selection criterion involves both a goodness of fit term gauging how well the model fits the data and a penalty term measuring the model complexity. AIC penalizes the complexity of model by two times of the number of estimated parameters.

Similarly, the information complexity (ICOMP) criterion (Bozdogan, 1988, 1990, 1993, 1994) is also built up by combining a goodness-of-fit term with a term for measuring the complexity of model. As the result of this formation, instead of penalizing the number of estimated parameters, the ICOMP criterion penalizes the covariance complexity of the model.

Based upon the covariance complexity index of van Emdan (1971), the ICOMP criterion is defined as

$$\text{ICOMP} = -2 \log L(\hat{\theta} | Z) + 2 C(\hat{Q}),$$

(2.2)

where $L(\hat{\theta} | Z)$ represents the maximized likelihood function, $\hat{\theta}$ represents the maximum likelihood estimator of the unknown parameter $\theta$, $C$ represents a complexity measure, $Q$ represents the covariance matrix of the estimated parameters for the model, and correspondingly $\hat{Q}$ represents the estimated covariance matrix of $Q$. Here, we address that in the original definition of the ICOMP criterion, the $\hat{\theta}$ could be any estimator of $\theta$. In this article, similar to the utility in Shang (2008), we utilize the maximum likelihood estimator (MLE) of $\theta$.

Obviously, the ICOMP criterion and the AIC possess the similarity in containing two terms, one is the goodness of fit term, $-2 \log L(\hat{\theta} | Z)$; the other one is the penalty term. However, they penalize the model complexity with different quantities. The penalty term of AIC is $2k$, two times of the number of estimated parameters, whereas the penalty term of the ICOMP criterion is the measure of the covariance complexity for the model.

In the expression of (2.2), the complexity measure of the ICOMP criterion needs to be estimated. Bozdogan (1988, 1990, 1993, 1994) proposed a maximal information complexity measure which is expressed as

$$C_m(Q) = \frac{m_k}{2} \log \frac{\text{tr}(Q)}{m_k} - \frac{1}{2} \log |Q|,$$

(2.3)

where $m_k$ is the dimension of $Q$. Because this measure is invariant with respect to scaler multiplication and orthonormal transformation and is also a monotonically increasing function of the dimension $m_k$ of $Q$, it is quite optimal in applications (see Bozdogan, 1988, 1990 for details).

In the linearized mixed model (2.1), the covariance matrix of estimated parameters plays the role of $Q$ in the expression of (2.3). If the $Q$ is estimable in an approach, the ICOMP criterion can be accessed. However, in model (2.1), the covariance matrix of the estimated parameters $Q$ is unknown in closed form. Alternatively, we will make use of the estimated inverse-Fisher information matrix to estimate $Q$, and then with the estimated $Q$, we can compute the model complexity. Let $F$ represents the Fisher information matrix for the model, then let $F^{-1}$ denote the inverse of $F$. The estimated inverse-Fisher information matrix $\hat{F}^{-1}$ is obtained with $\hat{\theta}$ in place of $\theta$ in the matrix $F^{-1}$.

For model (2.1), we have that $\text{cov}(\epsilon g'(\mu)) = V(\partial \eta / \partial \mu)^2 = W$, $\text{cov}(Z) = W + U D U' = \Sigma$, and $W = V(\partial \eta / \partial \mu)^2 = V \text{diag}((\partial \eta / \partial \mu_i)^2)$, $i = 1, \ldots, N$. $V$ is the covariance matrix of
$Y$ conditional on $\mu$. Write $U = [U_1, \ldots, U_r]$ and $\xi' = [\xi'_1, \ldots, \xi'_r]$ with $\text{Cov}(\xi_i) = \sigma^2_i I_{q_i}$ and $\text{Cov}(\xi_i, \xi_j) = 0$. Let $q_i$ denote the number of columns in $\xi_i$ and then $I_{q_i}$ is a $q_i \times q_i$ identity matrix. Therefore, the covariance matrix of $\xi$ is a block diagonal matrix with blocks $\sigma^2_i I_{q_i}$. In model (2.1), the covariance of $Z$ can be re-written as

$$\Sigma = \sum_{i=1}^{r} \sigma^2_i U_i U'_i + W.$$ 

Using this partitioned covariance matrix is to facilitate the derivation of matrix $F$. In the derivation of matrix $F$, the second derivative of the log-likelihood is desirable, and yet its derivation is not trivial. Also, since the link functions in the GLMMs are different, the $F$-formats are different. In this article, the example of a logistic regression model is utilized, so the $F$ matrix for the logistic regression model is derived and the derivation is shown in the supplementary materials.

For the estimation of parameters, we utilize Schall’s method (1991) to estimate the MLE’s. With respect to the parameters, the unknown parameter vector $\theta$ consists of the elements of the vector $\beta$ and the scalars $\sigma^2_1, \ldots, \sigma^2_r$. Instead of estimating the matrix $D$, we need to estimate scalars $\sigma^2_1, \ldots, \sigma^2_r$. We have $p + r$ parameters to evaluate. Let $\hat{\theta}$ denote the MLE of $\theta$, and $\hat{\theta} = (\hat{\beta}', \hat{\sigma}^2_1, \ldots, \hat{\sigma}^2_r)$.

Based on the derivation of inverse-Fisher information matrix $F$ and by the expressions (2.2) and (2.3), we therefore re-write the ICOMP criterion for model (2.1) as

$$ICOMP = -2 \log L(\hat{\theta} \mid Z) + m_k \log \frac{\text{tr}(\hat{F}^{-1})}{m_k} - \log |\hat{F}^{-1}|$$

$$= N \log(2\pi) + \log |\hat{\Sigma}| + (Z - X\hat{\beta})'(\hat{\Sigma}^{-1})(Z - X\hat{\beta}) + m_k \log \frac{\text{tr}(\hat{F}^{-1})}{m_k} - \log |\hat{F}^{-1}|.$$ 

With this computational formula for the ICOMP criterion, we then can evaluate its value.

### 2.3. A diagnostic of influential cases based on the ICOMP criterion

Regarding the diagnostic of influential cases, we adopt the idea of leave-one-out method, which is extensively and typically utilized to develop measures for identifying influential cases. The leave-one-out method aims to compare inferential quantities such as the regression prediction, regression parameter estimates, and estimated variances based on a fitted model to the full data set with those based on fitting a model to the data set with a case deleted. For instance, in the linear regression modeling framework, Cook (1977, 1979) successfully applied the leave-one-out method and has developed numerous measures for the detection of influential observations.

Motivated by the leave-one-out method, we therefore propose a diagnostic which makes use of the deletion of cases at a time based on the ICOMP criteria. Accordingly, the diagnostic is defined by the discrepancy of the two ICOMP criteria, one is computed based on the full data; the other one is computed based on a case-deleted data set. We define the diagnostic as

$$\delta_{\text{ICOMP}}(i) = ICOMP_{\text{Full-Data}} - ICOMP^{(i)},$$  \hspace{1cm} (2.4)
where $ICOMP_{\text{full-data}}$ is the ICOMP criterion value for a fitted mixed model when the full data set is utilized; $ICOMP^{(i)}$ is the ICOMP criterion value for the same fitted mixed model when the $i$th case is deleted.

We comment that the magnitude of $\delta_{ICOMP}^{(i)}$ reflected on definition (2.4) evaluates the influence of $y_i$ on the ICOMP criterion. Again, we can recall that the ICOMP criterion consists of two terms and essentially takes into account both goodness of fit and model complexity. The magnitude of $\delta_{ICOMP}^{(i)}$ therefore combines the influences of $y_i$ on both goodness of fit and model complexity.

In the evaluation of estimated $\delta_{ICOMP}^{(i)}$ values, we suppose a case is potentially influential, and once this case is removed, the leave-one-out data will make the model better fit. As a result, the value of the leave-one-out ICOMP criterion, i.e., $ICOMP^{(i)}$, will shrink compared to the ICOMP criterion under the full data set. Thus, the $\delta_{ICOMP}^{(i)}$ value is positive. However, positive diagnostics only imply that the corresponding cases are potentially influential. As expected, we hope to have an approach to detect the influential cases. To benefit of the detection of influential cases, among all the evaluated $\delta_{ICOMP}^{(i)}$ for the cases in a data set, the outstanding positive ones can flag the cases which are highly likely to be influential.

To find a benchmark for detecting an influential case, a simple approach is applied. We standardize the $\delta_{ICOMP}^{(i)}$ values, and for the one value which is above the two standard deviations, the corresponding case is influential. Let $S.\delta_{ICOMP}^{(i)}$ denote the standardized $\delta_{ICOMP}^{(i)}$, if $|S.\delta_{ICOMP}^{(i)}| > 2$, then the case for that $S.\delta_{ICOMP}^{(i)}$ value is influential.

Both serving as model selection criteria, however, the AIC and the ICOMP criterion come to be quite different in that the analogous criterion as in expression (2.4) based on the AIC cannot provide a diagnostic of influential cases as effective as $\delta_{ICOMP}^{(i)}$ because the dimension of estimated parameters is identical for both the full data set and a case-deleted data set, and the difference of the model complexity cannot be measured in the changed amount. From this point, the ICOMP criterion can catch more information from candidate models than the AIC.

3. The Information Complexity (ICOMP) criterion in the logistic linear regression with random effects

In what follows, we consider the logistic linear regression with random effects in the setting of generalized linear models as an example to illustrate the performance and effectiveness of the proposed diagnostic for distinguishing the influential cases in the data set for the model.

Let $Y = (y_1, \ldots, y_N)'$ be a vector of $N$ observations, which can be written as

$$Y = \mu + \epsilon,$$

where $\epsilon$ is a vector of random errors with zero expectation and covariance matrix $V$ given the $\mu$, and $V$ is a diagonal matrix with the element $V_i = \frac{\mu_i}{n_i} (n_i - \mu_i)$, $i = 1, \ldots, N$, the $n_i$ is the total number of trials for the binomial distribution. Let the link function be the logit, which is monotone, such that $g(\mu)$ can be written as the linear model

$$g(\mu) = \eta = X\beta + U\xi.$$

Here, $g(\mu)$ is an $N \times 1$ vector containing the element $\eta_i = \log \frac{\mu_i}{n_i - \mu_i}$, $i = 1, \ldots, N$. Here, $X_{N \times p}$ is a known design matrix, the $\beta$ is a vector of fixed effects, the $U$ is an $N \times q$ known matrix, and the $\xi$ is $q \times 1$ vector of random effects. Conditionally on $\mu$, the components of $Y$ are independently distributed.
Suppose that random effects \( \xi \) is partitioned as \([\xi_1, \ldots, \xi_r]\) and let \( q_1 + \cdots + q_r = q \) and \( U = [U_1, \ldots, U_r] \). The random vectors \( \xi_1, \ldots, \xi_r \) are assumed to be uncorrelated with zero expectation. The random effects are also assumed to be uncorrelated with \( \epsilon \). Further, \( \text{cov}(\xi) = \sigma_i^2 I_q \), \( i = 1, \ldots, r \) and \( \text{cov}(\xi) = D = \text{diag}(\sigma_1^2 I_1, \ldots, \sigma_r^2 I_r) \), where \( I_1, \ldots, I_r \) are identity matrices of orders \( q_1 \times q_1, \ldots, q_r \times q_r \).

Again, the link function \( g(.) \) is applied to the data \( Y \) (McCullagh and Nelder, 1989, p. 31) is linearized, providing the first order by

\[
g(Y) = g(\mu) + (Y - \mu)g'(\mu) = Z = X\beta + U\xi + \epsilon g'(\mu).
\] (3.3)

\( Z \) is called the adjusted dependent variable. Correspondingly, let \( Z = (Z_1, \ldots, Z_N)' \) be a vector of \( N \) observations. We know that \( E(Z) = X\beta \) and \( \text{cov}(\xi) = D \). Therefore, \( \text{cov}(\epsilon g'(\mu)) = V(\partial \eta/\partial \mu)^2 = W, \text{cov}(Z) = W + UD' = \Sigma, \) and \( W = V(\partial \eta/\partial \mu)^2 \) \( = V \text{diag}((\partial \eta/(\partial \mu_i)^2) = \text{diag}(\frac{\eta_i}{\mu_i(\eta_i - \mu_i)}), i = 1, \ldots, N. \)

### 3.1. The estimation of the parameters in the logistic linear regression model with random effects

For model (2.1), to estimate the maximum likelihood estimation in the normal variance components, we utilize the Schall’s estimation method, and its iteration algorithm is described as follows:

First, given estimates \( \hat{\sigma}^2, \hat{\sigma}_1^2, \ldots, \hat{\sigma}_r^2 \) for \( \beta \) and \( \xi_1, \ldots, \xi_r \) as least-squares solutions to the set of overdetermined linear equations

\[
C \begin{bmatrix} \hat{\beta} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} W^{-\frac{1}{2}}X & W^{-\frac{1}{2}}U \\ 0 & D^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\xi} \end{bmatrix} = \begin{bmatrix} W^{-\frac{1}{2}}Z \\ 0 \end{bmatrix},
\]

where \( W \) and \( D \) are evaluated at the current estimates of variance components.

Then, let \( T \) be the inverse of the matrix formed by the last \( q \) rows and columns of \( CC' \), partitioned conformably with \( D \) as

\[
\begin{bmatrix}
T_{11} & \cdots & T_{1r} \\
\vdots & \ddots & \vdots \\
T_{r1} & \cdots & T_{rr}
\end{bmatrix}
\]

Given estimates \( \hat{\beta} \) and \( \hat{\xi}_1, \ldots, \hat{\xi}_r \), compute estimates \( \hat{\sigma}^2, \hat{\sigma}_1^2, \ldots, \hat{\sigma}_r^2 \) for \( \sigma^2 \) and \( \sigma_1^2, \ldots, \sigma_r^2 \) as

\[
\hat{\sigma}^2 = (Z - X\hat{\beta} - U\hat{\xi})'(Z - X\hat{\beta} - U\hat{\xi})/(N - \sum_{i=1}^r (q_i - v_i)), \quad \hat{\sigma}_i^2 = \frac{\hat{b}_i\hat{b}_i}{q_i - v_i},
\]

where \( v_i = \text{tr}(T_{ii})/\sigma_i^2 \) is evaluated at the current estimates of \( \sigma_i^2 \). Note that \( \sigma^2 \) is the residual variance of the model, and for the logistic linear model and without the extra-binomial variation, its estimate \( \hat{\sigma}^2 \) should be close to 1. Otherwise, this value will be larger than one. In addition, it is mentioned earlier that this algorithm yields maximum likelihood estimates of the parameters.
3.2. A simulation study in the logistic linear regression model with random effects

The simulated data is generated from models (3.2) and (3.1). In (3.2), we have $\beta = -0.70$, and $X$ is a vector consisting of all 1’s. We also have $\xi = [\xi_1', \xi_2']'$ with $\text{Cov}(\xi_i) = \sigma_i^2 I_{4i}$ and $\text{Cov}(\xi_1, \xi_2) = 0$. The dimensions for $\xi_1$ and $\xi_2$ are 50 and 150, respectively. Then we set $\sigma_1^2 = 1.00$ and $\sigma_2^2 = 0.25$. The trial number of binomial distribution is 500. Then from model (3.2), a sample is generated and $\mu$ is calculated; from model (3.1), we generate 50 cases, each case consisting of 3 observations, total 150 observations from the binomial distribution using the success probability computed from model (3.1).

For the generated sample, we use models (3.2) and (3.1) to do estimations, and $\sigma^2 = 1.13$, close to 1, so there is no overdispersion, $\beta = -0.71$, $\sigma_1^2 = 0.87$, and $\sigma_2^2 = 0.24$ are obtained, indicating the model is a good one for fitting the data.

For the generated data set, we utilize the proposed diagnostic to detect the influential cases for the model selection ICOMP criterion.

In the generated data set, the most values for observations are from 100-250 and some are around 300. For these five possibly influential cases, the observations are (71, 44, 25), (48, 31, 57), (58, 25, 68), (376, 478, 288), and (27, 30, 35), for cases 8, 13, 31, 43, and 49, respectively. Intuitively, we expect that these five cases or some of them will be identified as influential.

As mentioned previously, we utilize a simple method to determine the benchmark for the proposed modeling diagnostic. We can standardize the $\delta_{ICOMP(i)}$ values, and then $S.\delta_{ICOMP(i)}$ values are computed. Figure 1 features the graph displaying the $S.\delta_{ICOMP(i)}$ value vs. the index. Among all $S.\delta_{ICOMP(i)}$ values, only $S.\delta_{ICOMP(8)}$ and $S.\delta_{ICOMP(49)}$ are above 2, and they are 2.11 and 2.77, respectively. We therefore identify cases 8 and 49 as influential. Since $S.\delta_{ICOMP(13)} = 1.99$ and $S.\delta_{ICOMP(31)} = 1.92$, we can judge them as potentially influential. The value of $S.\delta_{ICOMP(43)}$ is 1.76, case 43 is thus just a normal data. The influential and potentially influential cases are all marked in Figure 1.

3.3. The presentation of the application results

For the illustration of the effectiveness of the diagnostic, we apply the diagnostic to a data set which comes from an experiment to measure the mortality of cancer cells under radiation from Schall (1991). For this data set, four hundred cells ($n_i = 400$) were placed on a dish, and three dishes were irradiated at a time, or occasion. After the cells were irradiated, the surviving cells were counted. Since cells would also die naturally, dishes with cells were put in
the radiation chamber without being irradiated, to establish the natural mortality. Taking the difference for the two mortalities will be the one of the cancer. This data set can be described by model (3.1) and (3.2), and the models can be linearized by (3.3). The results will demonstrate that the proposed diagnostic can effectively flag the influential cases in the mixed model which is rewritten from the logistic linear model.

To describe the cancer cell data and to avoid the presence of extra-binomial variation, the model is written as

\[ \log \frac{\mu_{ij}}{n - \mu_{ij}} = \beta + \xi_{1i} + \xi_{2ij}, \]  

where \( i = 1, \ldots, 9, j = 1, \ldots, 3 \). The number of locations is 9, and \( \xi_{1i} \) are the random effects originating from the location. The number of dishes for each location is 3, and \( \xi_{2ij} \) are the random effects coming from the dish of the location. That is, the random effects are initiated from the location and the error term for each dish. Therefore the total observed random effects coming from the dish of the location. That is, the random effects are initiated from the location and the error term for each dish. Therefore the total observed \( y_i \) is 27, i.e., \( N = 27 \) and \( n_i = 400 \) for the binomial distribution.

For model (3.4), \( \beta \) is the fixed effect, so \( p = 1 \), and the corresponding design matrix \( X \) is a \( 27 \times 1 \) vector consisting of all 1’s. The dimension of random effects is 2, so \( r = 2 \).

Further, \( \text{cov}(\xi_i) = \sigma^2 I_{q_l} (i = 1, 2) \) and \( \text{cov}(\xi) = D = \text{diag}(\sigma^2 I_1, \sigma^2 I_2) \), where \( I_1, I_2 \) are identity matrices of orders \( q_1 \times q_1, q_2 \times q_2 \), and in this data set, \( q_1 = 9 \) and \( q_2 = 27 \). We can write \( U = [U_1, U_2] \), and the \( U_1 \) is a \( 27 \times 9 \) block matrix and the \( U_2 \) is a \( 27 \times 1 \) matrix. Also, we can write \( \xi = [\xi_1', \xi_2']^T \) with \( \text{cov}(\xi_1) = \sigma^2 I_{q_1} \) and \( \text{cov}(\xi_1, \xi_2) = 0 \). Note that \( \xi_1 \) is a vector of \( 9 \times 1 \) and \( \xi_2 \) is a vector of \( 27 \times 1 \).

For the calculation of \( \delta_{\text{ICOMP}}(i) \) in (2.4), we need to find out the Fish information matrix of model (3.3), and its derivation is illustrated in the supplementary materials.

As described previously, the cancer cell data have been used in Schall (1991). The data were collected from the nine locations, and each location contains three dishes. It could be observed that the numbers of cells surviving out of 400 placed are mostly around 110–145, some are about 170–180. Only for cases (locations) 3 and 8, the surviving cell numbers are very far from the other data. Intuitively, these two cases may be influential.

For this cancer cell data set, we may want to see the estimates of the two cases whose \( \delta_{\text{ICOMP}}(i) \) values are outstanding are individually eliminated from the data set. It is easy to see that these estimates are quite away from the others, indicating that when case 3 or 8 is eliminated from the data set, the parameter estimates are significantly changed. As a result, the corresponding \( \delta_{\text{ICOMP}}(i) \) values are very large. The results therefore demonstrate that the diagnostic \( \delta_{\text{ICOMP}}(i) \) can effectively detect the comprehensive change of the parameter estimates when case is removed from the data set and further can successfully evaluate the magnitude of the influence of each case. The graph of \( \delta_{\text{ICOMP}}(i) \) values is shown in the supplementary materials.

Figure 2 features the \( S.\delta_{\text{ICOMP}}(i) \) values, and it shows that only \( S.\delta_{\text{ICOMP}}(3) \) is greater than 2, which is 2.01. Thus, case 3 is influential, however, \( S.\delta_{\text{ICOMP}}(8) \) is only 1.3, indicating that case 3 is not influential by the criterion of our benchmark.

To further examine the effectiveness of the proposed diagnostic, we then artificially changed the values for cases 3 and 8. We change case 3 because \( \delta_{\text{ICOMP}}(3) \) is large. The values for cases 3 and 8 are originally 66, 75, 80 and 88, 76, 90 respectively. For Version 1, we changed them to 104, 105, 116 and 120, 110, 117. The \( S.\delta_{\text{ICOMP}}(i) \) values are shown in Figure 3, none of them is greater than 2, so there is no influential case this time.

From the previous application results, it is exhibited that the proposed diagnostic \( \delta_{\text{ICOMP}}(i) \) performs well in detecting an influential case in the GLMMs.
4. Concluding remarks

We develop a diagnostic for detecting influential cases based on the ICOMP criteria in generalized linear mixed models. The ICOMP is a model selection criterion taking into account of both goodness-of-fit and model complexity. The diagnostic is defined for revealing influential cases as the discrepancy of the ICOMP criteria based on the full data set and a case-deleted data set.

Given the generalized linear mixed model (GLMM), it can be linearized using the Taylor expansion, and then the GLMMs are simplified to the linear mixed models, and correspondingly the focus on the response variable is shifted to the adjusted dependent variable. Based on the linearized mixed model, the diagnostic can be computed for detecting the influential case among the data where the GLMM can be utilized.

Since the covariance matrix of estimated parameters in the linear mixed modeling framework is unknown, the Fisher information matrix is employed to compute the ICOMP criterion. The Fisher information matrix is derived for the logistic linear mixed model in the supplementary materials. Since the covariance of the adjusted dependent variable is a function of the fixed effects in the mixed model, the derivation of the Fisher information matrix is not trivial, yet it is feasible.

To demonstrate the effectiveness of the proposed diagnostic, a simulation study and an application on a cancer cell data are carried out. The generated and the real data are described by the logistic linear mixed model. The simulation and application results verify that the proposed procedure performs effectively in detecting the influential case.

From the derivation of the proposed diagnostic, the second derivative of the loglikelihood function is manageable. Hence, the proposed diagnostic $\delta_{ICOMP}$ serves as an effective approach to conduct modeling diagnosis in GLMMs. However, this approach could be quite complicated in some other modeling settings.
Regarding the determination of the benchmark for the proposed diagnostic, we utilize a simple approach where the diagnostic $\delta_{ICOMP(i)}$ values are standardized. If the standardized value is greater than 2, the corresponding case is diagnosed as influential. The simulation and application results demonstrate that the outstanding $\delta_{ICOMP(i)}$ values may not be influential after they are standardized. However, the outstanding diagnostic values indeed can provide enough indication regarding the cases which could be influential.

**Acknowledgments**

The author would like to express her appreciation to the Associate Editor and anonymous referee for their valuable and insightful comments and suggestions which helped to improve the original version of this article.

**References**

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F., eds. *2nd International Symposium on Information Theory* (pp. 267–281). Budapest: Akademia Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Cont.* AC-19:716–723.

Belsley, D.A., Kuh, E., Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley and Sons.

Bozdogan, H. (1988). ICOMP: A new model selection criterion. In: Bock, H.H., ed. *Classification and Related Methods of Data Analysis* (pp. 599–608). Amsterdam: North-Holland.

Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application on the evaluation of multivariate linear models. *Commun. Statist. Theor. Meth.* 19(1):221–278.

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture model using a new informational complexity criterion of the inverse Fisher information matrix. In: Opitz, O., Lausen, B., Klar, R., eds. *Studies in Classification, Data Analysis, and Knowledge Organization* (pp. 40–54). Heidelberg: Springer.

Bozdogan, H. (1994). Mixture-model cluster analysis using a new informational complexity and model selection criteria. In: Bozdogan, H., ed. *Multivariate Statistical Modeling*, vol 2. *Proc. 1st US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach* (pp. 69–113). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Bozdogan, H., Bearse, P. (2003). Information complexity criteria for detecting influential observations in dynamic multivariate linear models using the genetic algorithm. *J. Statist. Plann. Infer.* 114:31–44.

Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics* 19:15–18.

Cook, R.D. (1979). Influential observations in linear regression. *J. Amer. Statist. Assoc.* 74:169–174.

Cook, R.D., Weisberg, S. (1982). *Residuals and Influence in Regression*. London: Chapman and Hall.

McCullagh, P., Nelder, J.A. (1989). *Generalized Linear Models*. 2nd ed. London: Chapman and Hall.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometr.* 78:719–727.

Shang, J. (2008). A diagnostic of influential cases based on the Information Complexity Criterion in mixed models. *Proc. of the International Conference on Information Theory and Statistical Learning (ITSL)*, pp. 36–41.

van Emden, M. (1971). *An Analysis of Complexity*. Mathematisch Centrum Tracts 35. Amsterdam: Mathematisch Centrum.