Within-cluster resampling for multilevel models under informative cluster size

Danhyang Lee   Jae-kwang Kim *   Chris J. Skinner †

June 25, 2019

Summary A within-cluster resampling method is proposed for fitting a multilevel model in the presence of informative cluster size. Our method is based on the idea of removing the information in the cluster sizes by drawing bootstrap samples which contain a fixed number of observations from each cluster. We then estimate the parameters by maximising an average, over the bootstrap samples, of a suitable composite log-likelihood. The consistency of the proposed estimator is shown and does not require that the correct model for cluster size is specified. We give an estimator of the covariance matrix of the proposed estimator, and a test for the non-informativeness of the cluster sizes. A simulation study shows, as in Neuhaus and McCulloch (2011), that the standard maximum likelihood estimator exhibits little bias for some regression coefficients. However, for those parameters which exhibit non-negligible bias, the proposed method is successful in correcting for this bias.

*Department of Statistics, Iowa State University, Ames, Iowa 50011, U.S.A.
†Department of Statistics, London School of Economics and Political Sciences, London

Some key words: bootstrap; composite likelihood; generalized linear mixed model; model misspecification; parametric fractional imputation.

1 Introduction

Multilevel models, such as generalized linear mixed models (McCulloch et al., 2008), are widely used in the analysis of clustered data. In this setting, cluster size is said to be informative if it is associated with cluster-level random effects, conditional on cluster-level covariates. Although Neuhaus and McCulloch (2011) have shown that standard maximum likelihood estimators can often exhibit little bias under informative cluster size for some key covariate effects, there remains a need for methods which provide consistent estimation of all the model parameters in this setting. See Seaman et al. (2014) for a review.

One simple approach to controlling for informative cluster size is by including cluster size as a covariate in the model, but the resulting modified model may not be scientifically relevant (e.g. Dunson et al., 2003). Another approach is to incorporate cluster size into the model as a joint outcome alongside the random effects (Dunson et al., 2003; Gueorguieva, 2005; Chen et al. 2011). This depends, however, on the specification of the conditional distribution of the cluster size given the random effects and it is often preferred to treat this part of the model as a nuisance and to avoid such specification.

Hoffman et al. (2001) proposed a within-cluster resampling approach to the related problem of estimating marginal regression models under informative cluster size. Their method involves repeated estimation of the model from resampled datasets of one observation per cluster. Williamson et al. (2003) and Benhin et al. (2005) show how the method converges to a simple weighted estimation method. The method cannot be applied directly to multilevel models, however, since these models are generally inestimable when there is only one observation per cluster. In this paper we show how this approach can be extended to multilevel models by resampling datasets containing a fixed number of at least two observations per cluster. Consistent estimation is achieved without specifying a model for the cluster sizes. A score test for non-informative cluster size is also developed. Chiang & Lee(2008) also proposed resampling at least two observations per cluster, but for the different problem of improving estimation efficiency in a marginal regression model using information on within-cluster correlation. Pavlou et al. (2011) discussed assumptions needed for this method to provide unbiased inference.

2 Basic Setup

We consider clustered data consisting of pairs of values $(y_{ij}, x_{ij}), j = 1, \ldots, n_i$, of a response variable y and a vector of individual-level covariates x for n_i elements in cluster $i = 1, \ldots, K$, together with values z_i of a vector of cluster-level covariates z for these K clusters. We model the data by introducing cluster-specific random effects a_i, which may be vector valued, and factoring the distribution of the data and a_i in cluster i as $f(y_{i1}, \ldots, y_{in_i} \mid x_{i1}, \ldots, x_{in_i}, n_i, a_i, z_i)\ f(x_{i1}, \ldots, x_{in_i} \mid n_i, a_i, z_i)\ f(n_i \mid a_i, z_i)$ $f(a_i \mid z_i)\ f(z_i)$. We assume independence between clusters and conditional independence of the y_{ij} given $x_{i1}, \ldots, x_{in_i}, n_i, a_i$ and z_i with $f(y_{i1}, \ldots, y_{in_i} \mid$

$x_{i1}, \ldots, x_{in_i}, n_i, a_i, z_i) = \prod_j f_1(y_{ij} \mid x_{ij}, a_i; \theta_1)$. It is further assumed that y_{ij} is conditionally independent of n_i and z_i given x_{ij} and a_i:

$$y_{ij} \mid x_{ij}, n_i, z_i, a_i \sim f_1(y_{ij} \mid x_{ij}, a_i; \theta_1), \tag{1}$$

where $f_1(. \mid .; \theta_1)$ is a fully specified parametric model. We assume conditional independence of the x_{ij} and a_i given n_i and z_i so that $f(x_{i1}, \ldots, x_{in_i} \mid n_i, a_i, z_i) = f(x_{i1}, \ldots, x_{in_i} \mid n_i, z_i)$, that is there is no confounding by cluster. We further assume that

$$a_i \mid z_i \sim f_2(a_i \mid z_i; \theta_2), \tag{2}$$

$$n_i \mid a_i, z_i \sim g(n_i \mid a_i, z_i), \tag{3}$$

where $f_2(. \mid .; \theta_2)$is a fully specified parametric model and $g(\cdot \mid \cdot, \cdot)$ is completely unspecified. The parameters θ_1 and θ_2 along with $f_1(. \mid .; \theta_1)$ and $f_2(. \mid .; \theta_2)$ define the parts of this two-level model of interest, whereas $f(x_{i1}, \ldots, x_{in_i} \mid n_i, z_i)$, $g(n_i \mid a_i, z_i)$ and $f(z_i)$ represent nuisance parts of the model. The cluster size n_i is said to be informative if $g(n_i \mid a_i, z_i) \neq g(n_i \mid z_i)$, that is n_i and a_i are not conditionally independent given z_i. These and alternative sets of assumptions are discussed by Seaman et al. (2014). For likelihood-based inference, we assume that any parameters of the nuisance parts of the model are not functionally related to θ_1 or θ_2. The standard maximum likelihood method based on (1) and (2), ignoring the relation between n_i and a_i in (3), assumes the log-likelihood for (θ_1, θ_2) is

$$\ell(\theta_1, \theta_2) = \sum_{i=1}^{K} \log \int \prod_{j=1}^{n_i} f_1(y_{ij} \mid x_{ij}, a_i; \theta_1) f_2(a_i \mid z_i; \theta_2) da_i.$$

This can lead to biased estimation unless n_i is included in z_i, because the correctly specified log-likelihood function, up to an additive constant, is given by

$$\sum_{i=1}^{K} \log \int \prod_{j=1}^{n_i} f_1(y_{ij} \mid x_{ij}, a_i; \theta_1) g(n_i \mid a_i, z_i) f_2(a_i \mid z_i; \theta_2) da_i. \tag{4}$$

Thus, as pointed out by Neuhaus and McCulloch (2011), the informative cluster size problem is essentially a model misspecification problem. As noted in the Introduction, the approach of incorporating n_i in z_i as a covariate is often unsatisfactory since it can lead to a model which is not scientifically relevant. Moreover, the alternative approach of incorporating n_i into the model as a joint outcome may suffer from the effects of misspecification of the cluster size model $g(n_i \mid a_i, z_i)$ in (3).

3 Proposed method

To estimate the parameters under informative cluster size, we note that if the y_{ij} were generated from (1) for just a fixed number m of elements j for each cluster i, then the sample would be free of the informative cluster size problem. We shall show that we can use a within-cluster resampling method to construct such a data set which overcomes the informative cluster size problem, provided we make the additional assumption that $f(x_{i1}, \ldots, x_{in_i} \mid n_i, z_i) = \prod_{j=1}^{n_i} f(x_{ij} \mid z_i)$. Our proposed resampling method consists of selecting a bootstrap subsample of $m \leq \min_i n_i$ elements from each cluster by simple random sampling without replacement. Let $\{(x_{ij}^*, y_{ij}^*), j = 1, \ldots, m\}$ be the realized element-level data for the bootstrap subsample in cluster i, drawn from $\{(x_{ij}, y_{ij}); j = 1, \ldots, n_i\}$. We assume $m \geq 2$ and that $\theta = (\theta_1, \theta_2)$ remains identified for such a subsample. The observed log-likelihood function for θ constructed from the b-th bootstrap subsample is

$$\ell^{*(b)}(\theta) = \sum_{i=1}^{K} \log \int \prod_{j=1}^{m} f_1(y_{ij}^{*(b)} \mid x_{ij}^{*(b)}, a_i; \theta_1) f_2(a_i \mid z_i; \theta_2) da_i.$$

We show in the Supplementary Materials that this is a valid log-likelihood, free of the informative cluster size problem, for any such subsample, under the assumptions in section 2 and the additional assumption above. Combining the B bootstrap subsamples, we seek the maximizer of

$$\ell_B(\theta) = \frac{1}{B} \sum_{b=1}^{B} \ell^{*(b)}(\theta). \tag{5}$$

Computational aspects of maximizing $\ell_B(\theta)$ are discussed in §4. We now establish some asymptotic properties of the proposed estimator that maximizes (5). The score function derived from (5), viewed as a likelihood function, is

$$S_B(\theta) = \frac{\partial}{\partial \theta} \ell_B(\theta) = \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{K} S_i^{*(b)}(\theta), \tag{6}$$

where

$$S_i^{*(b)}(\theta) = \frac{\partial}{\partial \theta} \log \int \prod_{j=1}^{m} f_1(y_{ij}^{*(b)} \mid x_{ij}^{*(b)}, a_i; \theta_1) f_2(a_i \mid z_i; \theta_2) da_i. \tag{7}$$

Our proposed method is based on B replications of a resampling prodedure in which one subsample $\{j_1, \ldots, j_m\}$, is drawn from the $\binom{n_i}{m}$ possible subsamples of size m within cluster i with equal probability, for each cluster $i = 1, \ldots, K$. Thus, given the original sample, $S_B(\theta)$ converges to

$$S_C(\theta) = \sum_{i=1}^{K} \frac{1}{\binom{n_i}{m}} \sum_{1 \le j_1 < \cdots < j_m \le n_i} S_i(\theta; y_{ij_1}, \ldots, y_{ij_m}), \tag{8}$$

as $B \to \infty$, where

$$S_i(\theta; y_{i1}, \ldots, y_{im}) = \frac{\partial}{\partial \theta} \log f_i(y_{i1}, \ldots, y_{im}; \theta),$$
$$f_i(y_{i1}, \ldots, y_{im}; \theta) = \int \prod_{j=1}^{m} f_1(y_{ij} \mid x_{ij}, a_i; \theta_1) f_2(a_i \mid z_i; \theta_2) da_i.$$

Note that $S_C(\theta)$ is a composite score function (Varin et al., 2011). Since $S_C(\theta)$ is a sum of K independent random variables, under suitable moment conditions, we can obtain the asymptotic normality of $S_C(\theta)$ and, hence, the asymptotic normality of the proposed estimator, denoted by $\hat{\theta}_B$.

Theorem 1 *Let $\hat{\theta}_B$ be the maximizer of $\ell_B(\theta)$. Under some regularity conditions stated in the Supplementary Materials, (i) $\hat{\theta}_B \xrightarrow{p} \theta_0$ and (ii) $\sqrt{K}(\hat{\theta}_B - \theta_0) \xrightarrow{d} N(0, V_m(\theta_0))$, as $B \to \infty$ and $K \to \infty$, where θ_0 is the true parameter value. Here, $V_m(\theta_0)$ is a nonzero finite limit given by*

$$V_m(\theta) = \lim_{K \to \infty} H_m(\theta)^{-1} J_m(\theta) H_m(\theta)^{-1}, \tag{9}$$

where

$$H_m(\theta) = -\frac{1}{K} \sum_{i=1}^{K} E\left\{ \binom{n_i}{m}^{-1} \sum_{1 \le j_1 < \ldots, j_m \le n_i} \frac{\partial}{\partial \theta'} S_i(\theta; y_{ij_1}, \ldots, y_{ij_m}) \right\},$$
$$J_m(\theta) = \frac{1}{K} \sum_{i=1}^{K} var\left\{ \binom{n_i}{m}^{-1} \sum_{1 \le j_1 < \ldots, j_m \le n_i} S_i(\theta; y_{ij_1}, \ldots, y_{ij_m}) \right\}.$$

A specific expression for $V_m(\theta)$ is given in the Supplementary Materials for a linear mixed model where θ contains β_1, the coefficient of x_{ij} in the within-cluster model. The expression indicates that the asymptotic variance of the proposed estimator of β_1 is reduced by using a larger bootstrap subsample size. In this sense, $\min_i n_i$ is the preferred choice of m.

Using (9), the covariance matrix of $\hat{\theta}^*$ can be estimated by

$$\hat{V}^* = K^{-1} H(\hat{\theta}^*)^{-1} J(\hat{\theta}^*) H(\hat{\theta}^*)^{-1'},$$

where

$$\begin{aligned} J(\theta) &= \frac{1}{B(K-1)} \sum_{b=1}^{B} \sum_{i=1}^{K} \left\{ S_i^{*(b)}(\theta) - \bar{S}^{*(b)}(\theta) \right\} \left\{ S_i^{*(b)}(\theta) - \bar{S}^{*(b)}(\theta) \right\}', \\ H(\theta) &= -\frac{1}{BK} \sum_{b=1}^{B} \sum_{i=1}^{K} \frac{\partial S_i^{*(b)}(\theta)}{\partial \theta'}, \end{aligned}$$

$S_i^{*(b)}(\theta)$ is defined in (7) and $\bar{S}^{*(b)}(\theta) = K^{-1} \sum_{i=1}^{K} S_i^{*(b)}(\theta)$.

4 Computation

To find the maximizer of $l_B(\theta)$ in (5), we can use the Expectation Maximization algorithm of Dempster et al. (1977), treating the a_i as the missing data. Details of the algorithm are given in the Supplementary Materials. For some models, it is possible to obtain closed form expressions for each step of the algorithm. This is illustrated in the Supplementary Materials for a linear mixed model. In general, the E-step of the algorithm involves Monte Carlo methods to compute expectations. Fast computation can be achieved using the parametric fractional imputation of Kim (2011), which introduces fractional weights. In this method, M Monte Carlo imputed values of a_i are obtained from a proposal distribution once, and the fractional weights are assigned to the M Monte Carlo values. In each iteration of the algorithm, there is no need to repeat the Monte Carlo imputation. Only the fractional weights are updated. This approach is illustrated in the Supplementary Materials for a generalized linear mixed model (McCulloch et al., 2008).

5 Test for non-informativeness of the cluster sizes

It is often of interest to test for non-informativeness of the cluster sizes. Previous approaches have focussed on a marginal model (Benhin et al., 2005; Nevalainen et al., 2011). We propose a test in our multilevel model framework, which is essentially a test of model misspecification. Let $\mathcal{M}_1$ be the class of two-level models and let $\mathcal{M}_2 \subset \mathcal{M}_1$ be the subclass of these models with non-informative cluster sizes. We are interested in testing the null hypothesis $H_0 : F_0 \in \mathcal{M}_2$, where F_0 is the true data generating model. For the true parameter of the two level model, θ_0, let $\hat{\theta}_2$ denote the maximum likelihood estimator under $\mathcal{M}_2$ and $\hat{\theta}_1$ denote the proposed estimator under $\mathcal{M}_1$. Under the null hypothesis of model $\mathcal{M}_2$, the two estimators converge in probability to the same limit, θ_0. Otherwise, $\hat{\theta}_2$ does not converge to the true value. Thus, we can consider a score test for testing $H_0 : E\{S_1(\theta_0)\} = E\{S_2(\theta_0)\}$, where $S_1(\theta)$ and $S_2(\theta)$ are the proposed score function and the usual score function of θ under $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. Since $E\{S_1(\theta_0)\} = 0$ always holds, the null hypothesis reduces to $H_0 : E\{S_2(\theta_0)\} = 0$. Thus, the score test statistic is given by

$$Q = \{S_2(\hat{\theta}_1)\}' \left[\hat{V}\{S_2(\hat{\theta}_1)\}\right]^{-1} S_2(\hat{\theta}_1), \tag{10}$$

where $\hat{V}\{S_2(\hat{\theta}_1)\}$ denotes the variance estimator of $S_2(\hat{\theta}_1)$. Under the null hypothesis, the limiting distribution of Q is χ^2_p, where $p = \dim(\theta)$. In our setup, we have

$$S_1(\theta) = \frac{1}{BK} \sum_{b=1}^{B} \sum_{i=1}^{K} S_i^{*(b)}(\theta), \quad S_2(\theta) = \frac{1}{K} \sum_{i=1}^{K} S_i(\theta),$$

where $S_i^{*(b)}(\theta)$ is defined in (7) and $S_i(\theta) = \frac{\partial}{\partial\theta}\log\int\prod_{j=1}^{n_i} f_1(y_{ij} \mid x_{ij}, a_i; \theta_1) f_2(a_i; \theta_2) da_i$. To compute $\hat{V}\{S_2(\hat{\theta}_1)\}$ in (10), we can use a Taylor expansion to obtain

$$
\begin{aligned}
S_2(\hat{\theta}_1) &\approx S_2(\theta_0) - E\left\{\frac{\partial}{\partial\theta'}S_2(\theta_0)\right\} E\left\{\frac{\partial}{\partial\theta'}S_1(\theta_0)\right\}^{-1} S_1(\theta_0), \\
&= \frac{1}{K}\sum_{i=1}^{K}\left\{S_i(\theta_0) - \kappa(\theta_0)B^{-1}\sum_{b=1}^{B} S_i^{*(b)}(\theta_0)\right\} := \frac{1}{K}\sum_{i=1}^{K} u_i(\theta_0),
\end{aligned}
$$

where $\kappa(\theta_0) = E\left\{\partial S_2(\theta_0)/\partial\theta'\right\} E\left\{\partial S_1(\theta_0)/\partial\theta'\right\}^{-1}$. A consistent estimator of $\text{var}\{S_2(\hat{\theta}_1)\}$ is

$$
\hat{V}\{S_2(\hat{\theta}_1)\} = \frac{1}{K(K-1)}\sum_{i=1}^{K}\{\hat{u}_i(\hat{\theta}_1) - \bar{u}(\hat{\theta}_1)\}\{\hat{u}_i(\hat{\theta}_1) - \bar{u}(\hat{\theta}_1)\}',
$$

where $\bar{u}(\theta) = K^{-1}\sum_{i=1}^{K}\hat{u}_i(\theta)$, $\hat{u}_i(\theta) = S_i(\theta) - \hat{\kappa}(\theta)B^{-1}\sum_{b=1}^{B} S_i^{*(b)}(\theta)$, and $\hat{\kappa}(\theta) = \left\{K^{-1}\sum_{i=1}^{K}\partial S_i(\theta)/\partial\theta'\right\}\left\{(BK)^{-1}\sum_{b=1}^{B}\sum_{i=1}^{K}\partial S_i^{*(b)}(\theta)/\partial\theta'\right\}^{-1}$.

6 Simulation Study

We conduct a simulation study to compare the performance of the proposed method with the usual maximum likelihood method, which ignores the informative cluster size problem. The study has a 2×2 factorial design: (1) a linear mixed model and a generalized linear mixed model with logit link; (2) informative and non-informative cluster sizes.

We first generate data from a linear mixed model, where $y_{ij} = \beta_0 + \beta_1 x_{ij} + a_i + e_{ij}$, $a_i \sim N(0, \sigma_a^2)$, $e_{ij} \sim N(0, \sigma_e^2)$, $x_{ij} \sim N(1, 1)$ for $j = 1, \ldots, n_i$ and $i = 1, \ldots, K$. We set $\beta_0 = 0{\cdot}5, \beta_1 = 1, \sigma_e^2 = 1, \sigma_a^2 = 0{\cdot}25$, and $K = 50$ and 100.

For the informative cluster size case, we generate n_i from the cluster size model $n_i \sim \text{Poi}(e^{1+\gamma a_i}) + C$, where C is the minimum cluster size. We set $\gamma = 3$ and C=5 in this simulation. For the non-informative cluster size case, we generate data from the same linear mixed model but generate n_i from the model $n_i \sim \text{Poi}(e^{1+\gamma b_i}) + C$, where $b_i \sim N(0, \sigma_a^2)$, that is b_i follows the same distribution of a_i but is independent of a_i.

For the simulation, we compute the proposed estimates using $B = 50$ bootstrap samples. As seen from Table 1, in the informative cluster size case, our proposed method provides almost unbiased estimation, while the maximum likelihood method has significant biases for the regression intercept and variance component of the level two model. In the non-informative cluster size case, both the proposed and maximum likelihood estimators are unbiased for all parameters. As expected, the Monte Carlo standard errors of the proposed estimator tend to be slightly larger than those of the maximum likelihood estimator.

We next consider a generalized linear mixed model, where $y_{ij} \sim \text{Ber}(p_{ij})$, $\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_{ij} + a_i$, $a_i \sim N(0, \sigma_a^2)$, $x_{ij} \sim N(1, 1)$ for $j = 1, \ldots, n_i$ and $i = 1, \ldots, K$. We set $\beta_0 = -1, \beta_1 = 1, \sigma_a^2 = 0{\cdot}25$, and $K = 50$ and 100. Cluster sizes for the informative and non-informative cases are generated from the same models used for the linear mixed model with $\gamma = 3$ and C=10.

Table 2 shows that the proposed method removes the biases due to informative cluster size, in line with the previous simulation study. In the non-informative cluster size case, the proposed estimator is comparable with the maximum likelihood estimator with respect to Monte Carlo bias, but has larger Monte Carlo standard errors.

We have also computed the sizes and powers of the proposed score test of non-informativeness under the linear mixed model with nominal significance levels $\alpha = 0{\cdot}01, 0{\cdot}05$ and $0{\cdot}10$. Here, we set $\gamma = 1, 2$ and 3 in the cluster size models. Table 3 shows that the test performs well with respect to both size and power.

Table 1: Monte Carlo biases, standard errors (SEs) and root mean squared errors (RMSEs) of estimators, based on 1,000 Monte Carlo samples under Linear Mixed Model

	Number of clusters	Parameter	Proposed Bias	Proposed SE	Proposed RMSE	Maximum Likelihood Bias	Maximum Likelihood SE	Maximum Likelihood RMSE
ICS	50	β_0	0·002	0·086	0·086	0·071	0·089	0·114
		β_1	0·003	0·053	0·053	0·000	0·040	0·040
		σ_e^2	-0·010	0·077	0·078	-0·009	0·057	0·058
		σ_a^2	-0·008	0·072	0·073	0·009	0·069	0·070
	100	β_0	0·000	0·059	0·059	0·070	0·061	0·093
		β_1	0·002	0·038	0·038	0·002	0·029	0·029
		σ_e^2	-0·003	0·056	0·056	-0·005	0·042	0·042
		σ_a^2	-0·002	0·055	0·055	0·016	0·053	0·055
Non-ICS	50	β_0	0·001	0·090	0·090	0·001	0·090	0·090
		β_1	0·001	0·055	0·055	0·002	0·052	0·052
		σ_e^2	-0·003	0·079	0·079	-0·002	0·074	0·074
		σ_a^2	-0·010	0·074	0·075	-0·009	0·073	0·074
	100	β_0	-0·000	0·061	0·061	-0·000	0·060	0·060
		β_1	0·001	0·042	0·042	0·001	0·039	0·039
		σ_e^2	-0·002	0·059	0·059	-0·002	0·055	0·055
		σ_a^2	-0·004	0·056	0·056	-0·004	0·055	0·055

Table 2: Monte Carlo biases, standard errors (SEs) and root mean squared errors (RMSEs) of estimators, based on 1,000 Monte Carlo samples under Generalized Linear Mixed Model

	Number of clusters	Parameter	Proposed Bias	Proposed SE	Proposed RMSE	Maximum Likelihood Bias	Maximum Likelihood SE	Maximum Likelihood RMSE
ICS	50	β_0	-0·008	0·160	0·160	0·092	0·150	0·176
		β_1	0·004	0·111	0·111	0·009	0·093	0·093
		σ_a^2	-0·013	0·141	0·142	0·047	0·121	0·129
	100	β_0	-0·000	0·110	0·110	0·100	0·105	0·145
		β_1	0·002	0·077	0·077	0·005	0·064	0·064
		σ_a^2	-0·007	0·100	0·100	0·055	0·087	0·103
Non-ICS	50	β_0	-0·008	0·153	0·153	-0·008	0·141	0·141
		β_1	0·007	0·104	0·104	0·006	0·092	0·092
		σ_a^2	-0·009	0·137	0·138	-0·009	0·122	0·122
	100	β_0	-0·004	0·104	0·104	-0·002	0·096	0·096
		β_1	0·002	0·074	0·074	0·001	0·063	0·063
		σ_a^2	-0·008	0·102	0·103	-0·005	0·090	0·090

Acknowledgement

We are grateful to the reviewers for comments that have helped to improve this paper. The research of the second author was partially supported by the U.S. National Science Foundation.

Supplementary material

Supplementary material available at *Biometrika* online includes the proof of Theorem 1, expressions for variances under a linear mixed model, and a

Table 3: Sizes and powers of the proposed test based on $2,000$ Monte Carlo samples with pre-determined nominal levels α

Number of clusters	α	$\gamma = 1$		$\gamma = 2$		$\gamma = 3$	
		Size	Power	Size	Power	Size	Power
50	0·01	0·015	0·318	0·010	0·840	0·011	0·942
	0·05	0·067	0·639	0·055	0·970	0·045	0·992
	0·10	0·119	0·780	0·108	0·992	0·010	0·997
100	0·01	0·013	0·823	0·009	0·999	0·012	1·000
	0·05	0·069	0·951	0·052	1·000	0·054	1·000
	0·10	0·126	0·977	0·106	1·000	0·097	1·000

description of the EM algorithm used for computation.

References

Benhin, E., Rao, J.N.K., & Scott, A.J. (2005). Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes, *Biometrika*, **92**, 435–50.

Chen, Z., Zhang, B. & Albert, P.S. (2011). A joint modeling approach to data with informative cluster size: robustness to the cluster size model. *Statistics in Medicine*, **30**, 1825–36.

Chiang, C.-T. & Lee, K.-Y. (2008). Efficient estimation methods for informative cluster size data. *Statistica Sinica* **18**, 121-33.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*,**39**, 1-38.

Dunson, D., Chen, Z., & Harry, J. (2003). A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes, *Biometrics*, **59**, 521–30.

Gueorguieva, R.V. (2005). Comments about joint modelling of cluster size and binary and continuous subunit-specific outcomes. *Biometrics*, **61**, 862–7.

Hoffman, E. B., Sen, P. K., & Weinberg, C. R. (2001). Within-cluster resampling. *Biometrika*, **88**(4), 1121-34.

Kim, J.K. (2011). Parametric fractional imputation for missing data analysis, *Biometrika*, **98**, 119–32.

McCulloch, C.E., Searle, S.R. & Neuhaus, J.M. (2008). *Generalized, Linear and Mixed Models*, 2nd ed. New York: Wiley.

Neuhaus, J. & McCulloch, C. (2011). Estimation of covariate effects in generalized linear mixed models with informative cluster sizes, *Biometrika*, **98**, 147–62.

Nevalainen, J., Oja, H. & Datta, S. (2011). Tests for informative cluster size using a novel balanced bootstrap scheme. *Statistics in Medicine*, **36**, 2630-40.

Pavlou, M., Seaman, S. & Copas, A. (2011). An examination of a method for marginal inference when the cluster size is informative. *Statistica Sinica* **23**, 791-808.

Seaman, S., Pavlou, M., & Copas, A. (2014). Review of methods for handling confounding by cluster and informative cluster size in clustered data, *Statistics in Medicine*, **33**, 5371–87.

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 5-42.

Williamson, J.M., Datta, S., & Satten, G.A. (2003). Marginal analysis of clustered data when cluster size is informative, *Biometrics*, **59**, 36–42.