Missing covariate data in generalized linear mixed models with distribution-free random effects

Liu, Li; Xiang, Liming

2018

Liu, L., & Xiang, L. (2019). Missing covariate data in generalized linear mixed models with distribution-free random effects. *Computational Statistics & Data Analysis*, 134, 1-16.
doi:10.1016/j.csda.2018.10.011

<https://hdl.handle.net/10356/105576>

<https://doi.org/10.1016/j.csda.2018.10.011>

© 2018 Elsevier B.V. All rights reserved. This paper was published in *Computational Statistics & Data Analysis* and is made available with permission of Elsevier B.V.

Downloaded on 26 Apr 2023 20:33:35 SGT

Missing Covariate Data in Generalized Linear Mixed Models with Distribution-Free Random Effects

Li Liu¹ and Liming Xiang ^{*2}

¹*School of Mathematics and Statistics, Wuhan University, Hubei 430072, P.R.China*

²*School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371*

Abstract

We consider generalized linear mixed models in which random effects are free of parametric distributions and missing at random data are present in some covariates. To overcome the problem of missing data, we propose two novel methods relying on auxiliary variables: a penalized conditional likelihood method when covariates are independent of random effects, and a two-step procedure consisting of a pairwise likelihood for estimating fixed effects in the first step and a penalized conditional likelihood for estimating random effects in the second step while covariates can be related to random effects. Our methods allow [a nonparametric structure for the missing covariate data](#) and do not rely on distribution assumptions for random effects, which are not observed in the data, thus providing great flexibility in capturing a board range of [the missingness mechanism and behaviors of random effects](#). We show that the proposed estimators enjoy desirable theoretical properties by relaxing the conditions for a finite number of clusters or finite cluster size imposed in the literature. The finite sample performance of the estimators is assessed through extensive simulations. We illustrate the application of the methods using a longitudinal data set on forest health monitoring.

*Corresponding Author: lmxiang@ntu.edu.sg

Keywords: Auxiliary variable; Generalized linear mixed model; Missing at random data; Pairwise likelihood; Penalized conditional likelihood

1 Introduction

In many application studies, the covariate data are partly observed to form a validation set because of expensive cost to measure or some sensitive questions in a questionnaire. There are often some auxiliary variables, which may not be of substantial interest but may be associated with the missing data, available in such studies and fully measured to provide surrogate information for the subjects. Thus these missing values are assumed as missing at random (MAR) in the sense of Little and Rubin (2002), that is, conditional on the observed data, the failure to observe a value does not depend on the data that are unobserved. In general, without any further assumptions, MAR cannot be refuted from the data as there could always be some variables cannot be included in the underlying model but help explain the missingness. In this paper, we rely on auxiliary variables to analysis MAR covariate data.

Conventionally, there are two naive approaches, deletion and mean imputation, to handle such missing data. The former one refers to as the complete case, which uses the validation set only, while the latter is to replace the missing covariates with an average of the observed data for the corresponding variable. It is well known that both approaches yield inefficient and inconsistent estimates (see Little and Rubin 2002, p43). As such, there have been a variety of improved approaches to account for missing data problems, especially for missing covariates. Rubin (1987) studied multiple imputation (MI) methods for missing covariates in generalized linear models (GLMs). Robins et al. (1994) discussed a general semiparametric approach using inverse probability weighted estimating equations to obtained consistent regression parameters for MAR data in GLMs and the missingness probabilities are either known or can be parametrically modeled.

Correlated data arise commonly in longitudinal studies, which include repeated measurements of individuals over time, or multicenter studies in which individuals are clustered within natural units. For example, in a recent study of forest health monitoring (Belitz et al, 2015), defoliation levels of trees in the same spatial location are correlated since those trees were exposed to the same environmental risk factors. The GLMM is one of the most widely used methods to account for both inter-cluster heterogeneity and intra-cluster association among typically non-normal data arising from longitudinal or multicenter studies. In the presence of missing values, Ibrahim et al (1999) proposed a method for estimation in GLMMs with missing covariates via Monte Carlo EM algorithm. Their method has been applied to longitudinal data by Sabry et al. (2016) more recently. The MI method was also studied in the analysis of clustered data subject to MAR by Aloisio et al. (2014). Li and Yi (2013) investigated a pairwise likelihood method for longitudinal data when both covariate and response are subject to missingness. All these works require strict parametric assumptions on the missingness mechanism so that the EM algorithm, imputation or pairwise models can be applied. Thus, they are sensitive to misspecification of missing data process. However, the true missing mechanism is unknown in practice and such assumptions for missing data are often non-verifiable.

With the assumption of normally distributed random effects, Zhou et al. (2002) studied the statistical inference problem for a random effects logistic regression model with discrete auxiliary variables through the conditional joint distribution of the outcome variable and random effects given covariates, while Chen and Lin (2010) considered the GLMMs with continuous auxiliary covariates using a kernel smoother. Since random effects are unobservable, it is difficult to verify a parametric distribution assumption in practice. In addition, Agresti et al. (2004) showed that misspecification of the random-effects distribution may result in biased estimators of regression parameters. Wang et al. (2012) proposed a conditional quadratic inference function approach to estimate GLMMs for longitudinal data with unspecified random effects distribution. Furthermore,

as showed in Liu and Xiang (2014), the random effects may depend on covariates, and ignoring the relationship between random effects and covariates would lead to inconsistent estimators. This encourages us to explore novel estimation methods for inference of GLMMs with missingness and no distribution assumptions about random effects while allowing dependence between random effects and covariates.

Throughout the paper, we leave the random-effects distribution unspecified and treat [the conditional model of the missing covariate](#) nonparametrically. In the presence of missing at random in covariates and auxiliary variables being available, following Jiang (1999)’s idea of conditional inference, we propose a penalized conditional likelihood for estimation in the situation where covariates are independent of random effects. We further develop a two-step estimation procedure based on both pairwise likelihood and penalized conditional likelihood to estimate parameters when covariates are correlated with random effects. In particular, we construct a semiparametric pairwise likelihood to estimate regression coefficients in the first step, and then obtain the estimator of the intercept through maximizing a profile penalized conditional likelihood. Pairwise likelihood is a kind of pseudolikelihood constructed by multiplying the likelihood contributions from all pairs of observations within clusters and incorporating information on the non-validation set. For various applications of pairwise likelihood, we refer the readers to Kalbfleisch (1978), Besag (1974), Kuk (2007), Chatelain et al. (2009) and Huang et al. (2010).

The main contribution of this paper is threefold. First, both proposed methods avoid full likelihood analysis and reduce computational cost significantly. Second, unlike the most existing work on GLMMs, our methods are free of distribution assumptions for random effects and [the missing covariate](#), thus yielding robust estimators against misspecification of the [missingness mechanism](#) and/or random effects structures. Third, our methods allow the exposure variable to be associated with the random effects and yield consistent estimators in the circumstances that the number of clusters or cluster size or both diverge to infinity. This advantage makes the proposed methods preferable in both

longitudinal and clustered studies. As such, the proposed methods provide an extension to the works in Wang et al. (2012) and Liu and Xiang (2014), where a finite number of longitudinal observations for each subject and a finite number of clusters were required, respectively. We need to tackle the problems allowing the cluster size or the number of clusters to be divergent, hence the derivation of the subsequent theoretical properties is much more challenging than those in Wang et al. (2012) and Liu and Xiang (2014).

The rest of the paper is organized as follows. We first introduce the notation and model setting up in Section 2. In Section 3, we develop the penalized conditional likelihood and two-step estimation methods. Asymptotic properties are investigated in Section 4. Section 5 presents some simulation results, and Section 6 applies the proposed methods to a forest health dataset. In section 7, we conclude the paper with a discussion.

2 Model setting-up

We assume that there are K clusters and each cluster includes n_k individuals, where $k = 1, \dots, K$ and $N = \sum_{k=1}^K n_k$ is the total number of individuals in the study. Let (X_{ki}, Z_{ki}) be the covariates associated with fixed effects for individual i in cluster k , where X 's are incomplete covariates subject to missingness, and Z 's are complete covariates and always observed. For simplicity's sake and without loss of generality, we consider analysis of outcome Y_{ki} in relation to a scalar incomplete covariate X_{ki} and complete covariate Z_{ki} which can be a vector, though other cases with multiple incomplete covariates are possible. We rely on an auxiliary variable, denoted by W_{ki} , for the incomplete covariate X_{ki} , which may depend on variable Z_{ki} . In addition, we suppose that there are a set of i.i.d. random effects $u_1, \dots, u_K$ with mean 0 and variance σ_T^2 , associated with the cluster-specific structure. Given covariates and random effects, we assume that the distribution of outcome variable Y_{ki} falls into the following canonical exponential family:

$$P_{\theta}(Y_{ki} \mid X_{ki}, Z_{ki}, W_{ki}, u_k) = \exp\{Y_{ki}\eta_{ki} - b(\eta_{ki})/a(\phi) + c(Y_{ki}, \phi)\}, \quad (1)$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions, ϕ is a dispersion parameter, and η is defined by

$$\eta_{ki} = \alpha + u_k + \beta_1' X_{ki} + \beta_2' Z_{ki}$$

with $\beta = (\beta_1, \beta_2)$ a vector of unknown fixed effects. Let $\theta = (\alpha, \beta_1, \beta_2)$, vector of random effects $u = (u_1, \dots, u_K)$ and $\theta_u = (\theta, u)$. The model defined here indicates that W_{ki} does not provide additional information if covariates X_{ki}, Z_{ki} and random effects u_k are given.

To construct some likelihood of θ using $[Y|X, Z]$, we first need to check the identifiability of θ . As shown in Liu and Xiang (2014), the parameters in the above model are identifiable when 1, X_{ki} and Z_{ki} are linearly independent. Now we formulate the definition of data structure under the missingness mechanism. For a given cluster k , we denote V_k as a simple random validation set with sample size n_{V_k} . For the individuals belonging to V_k , the covariates (X, W) are always observed. On the other hand, we let $\bar{V}_k$ be the set including the remaining individuals of the k -th cluster and $n_{\bar{V}_k} = n_k - n_{V_k}$ be the size of $\bar{V}_k$. For individuals in the set of $\bar{V}_k$, only their auxiliary variable W can be observed. Thus the observed data for the i -th individual in cluster k take the following form:

$$\begin{cases} \{Y_{ki}, Z_{ki}, W_{ki}, X_{ki}\} & \text{if } i \in V_k \\ \{Y_{ki}, Z_{ki}, W_{ki}\} & \text{if } i \in \bar{V}_k, \end{cases} \quad (2)$$

where $k = 1, 2, \dots, K$, $i = 1, 2, \dots, n_k$. Based on the data in (2), the conditional likelihood function of parameters is given by

$$L(\theta_u) = \prod_{k=1}^K \left\{ \prod_{i \in V_k} P_\theta(Y_{ki} | X_{ki}, Z_{ki}, u_k) \cdot \prod_{i \in \bar{V}_k} P_\theta(Y_{ki} | W_{ki}, Z_{ki}, u_k) \right\}.$$

Furthermore, we denote by r the observable indicator of X , that is, $r = 1$ if X is observed and 0 otherwise. Under the assumption that the data are missing at random,

$$P_\theta(r = 1 | Y, Z, W, X) = P_\theta(r = 1 | Y, W),$$

This assumption allows the missingness mechanism to be dependent on the response variable Y and the auxiliary variable W , implying that the auxiliary variable and the response variable offer sufficient information for the missingness mechanism.

3 Estimation Procedure

3.1 Penalized conditional likelihood

Without loss of generality, we consider the model (1) with $a(\phi) = 1$. To avoid any parametric assumption on $[Z, W]$, we first consider estimating the distribution of W empirically using data in the validation set as in Pepe and Fleming (1991), leading to

$$\hat{P}_\theta(Y_{ki} | Z_{ki}, W_{ki}, u_k) = \frac{\sum_{j \in V_k} P_\theta(Y_{ki} | Z_{ki}, X_{kj}, u_k) I_{[A_{ki}=A_{kj}]}}{\sum_{j \in V_k} I_{[A_{ki}=A_{kj}]}}, \quad (3)$$

where $I_{[\cdot]}$ denotes an indicator function and $A = (Y, W)$. The estimated conditional likelihood function of θ_u is then written as

$$EL(\theta_u) = \prod_{k=1}^K \left\{ \prod_{i \in V_k} P_\theta(Y_{ki} | Z_{ki}, X_{ki}, u_k) \cdot \prod_{j \in \bar{V}_k} \hat{P}_\theta(Y_{kj} | Z_{kj}, W_{kj}, u_k) \right\},$$

which contains unobservable random effects u_k . Following the same line of the conditional inference approach proposed by Jiang (1999), we propose to estimate θ_u by maximizing the penalized conditional log-likelihood (PCL) function

$$U(\theta_u) = \log EL(\theta_u) - \lambda \sum_{k=1}^K u_k^2 / 2, \quad (4)$$

where λ is a tuning parameter and the random effects u_k are treated as unknown fixed parameters to be estimated. As argued by Jiang (1999), such estimation is reasonable when the number of clusters and cluster size are both large enough, i.e., there is sufficient

information about all random effects and there will be no much information loss since we treat the random effects nonparametrically.

We denote the resulting estimator by $\hat{\theta}_u$, which can be numerically obtained by maximizing the PCL in Eq. (4) using the Newton-Raphson algorithm. We refer the corresponding estimation method as the penalized conditional likelihood estimation (PCLE).

3.2 Two-step method

Note that the PCLE in the previous subsection actually requires an assumption that covariates, for example X 's, are independent of the random effects. In the situation when these covariates are associated with the random effects, we propose a two-step estimation procedure. In the first step, we consider estimating the fixed effects β through the pairwise likelihood approach suggested by Liu and Xiang (2014). Specifically, multiplying the likelihood contributions from all pairs of observations (Y_{ki}, Y_{kj}) within clusters, it leads to the following conditional pseudo-likelihood of β :

$$\begin{aligned} & \prod_{k=1}^K \prod_{i < j} \frac{P_{\theta}(Y_{ki}|X_{ki}, Z_{ki}, u_k) P_{\theta}(Y_{kj}|X_{kj}, Z_{kj}, u_k)}{P_{\theta}(Y_{ki}|X_{ki}, Z_{ki}, u_k) P_{\theta}(Y_{kj}|X_{kj}, Z_{kj}, u_k) + P_{\theta}(Y_{kj}|X_{ki}, Z_{ki}, u_k) P_{\theta}(Y_{ki}|X_{kj}, Z_{kj}, u_k)} \\ &= \prod_{k=1}^K \prod_{i < j} \left(1 + e^{-[\beta_1(X_{ki}-X_{kj})+\beta_2(Z_{ki}-Z_{kj})](Y_{ki}-Y_{kj})} \right)^{-1}. \end{aligned}$$

Let $f_{\beta}^{ij}(X_{ki}, X_{kj}) = \left(1 + e^{-[\beta_1(X_{ki}-X_{kj})+\beta_2(Z_{ki}-Z_{kj})](Y_{ki}-Y_{kj})} \right)^{-1}$. When missing covariate values are encountered, $f_{\beta}^{ij}(W_{ki}, W_{kj})$ is adopted instead relying on the auxiliary covariate W . Based on observations from the validation set, its consistent nonparametric estimator can be obtained empirically by

$$\hat{f}_{\beta}^{ij}(W_{ki}, W_{kj}) = \frac{\sum_{i^V, j^V \in V_k, i^V < j^V} f_{\beta}^{ij}(X_{ki^V}, X_{kj^V}) I_{(A_{ki^V}=A_{ki}, A_{kj^V}=A_{kj})}}{\sum_{i^V, j^V \in V_k, i^V < j^V} I_{(A_{ki^V}=A_{ki}, A_{kj^V}=A_{kj})}}. \quad (5)$$

Substituting the estimator given in Eq. (5) for $f_{\beta}^{ij}(W_{ki}, W_{kj})$, it leads to the conditional

pairwise log-likelihood

$$\hat{R}(\beta) = \sum_{k=1}^K \left(\sum_{i,j \in V_k, i < j} \log f_{\beta}^{ij}(X_{ki}, X_{kj}) + \sum_{i,j \in \bar{V}_k, i < j} \log \hat{f}_{\beta}^{ij}(W_{ki}, W_{kj}) \right).$$

As suggested by Kuk and Nott (2000), to obtain more reasonable estimates of the parameters, we take averages of the contributions in $\hat{R}(\beta)$ over the validation and non-validation sets, respectively. A weighted conditional pairwise log-likelihood function of β can be consequently formed by

$$\hat{R}_W(\beta) = \frac{1}{N} \sum_{k=1}^K \left(\frac{1}{n_{V_k} - 1} \sum_{i,j \in V_k, i < j} \log f_{\beta}^{ij}(X_{ki}, X_{kj}) + \frac{1}{n_{\bar{V}_k} - 1} \sum_{i,j \in \bar{V}_k, i < j} \log \hat{f}_{\beta}^{ij}(W_{ki}, W_{kj}) \right).$$

The estimator for fixed-effects, denoted by $\tilde{\beta}$, is in turn obtained by maximizing $\hat{R}_W(\beta)$ with respect to β .

In the second step of the procedure, we estimate the intercept using the PCLE method developed in Subsection 3.1. Given estimator $\tilde{\beta}$, the estimator $(\tilde{\alpha}, \tilde{u})$ of the intercept and random effects can be obtained through maximizing the profile PCL. Equivalently, we solve the following equation:

$$\dot{U}(\tilde{\beta}, \alpha, u) = 0, \tag{6}$$

where $U(\cdot)$ is the PCL function defined in Eq. (4), and $\dot{U}(\cdot)$ represents the first order derivative of function $U(\theta_u)$ with respect to θ_u . Through the whole paper, we also denote the first order partial derivative of function $U(\cdot)$ with respect to variable α by $\dot{U}_{\alpha}(\cdot)$, and the second order partial derivative with respect to θ_u by $\ddot{U}(\cdot)$ and with respect to variables α and β by $\ddot{U}_{\alpha\beta}(\cdot)$.

3.3 Selection of tuning parameter

As with all penalized conditional likelihood methods, the practical performance of the proposed methods depends essentially on whether the choice of the tuning parameter is

appropriate. In general, the preferred estimator is determined corresponding to the tuning parameter λ , which optimizes some criterion, such as the AIC, BIC or Cross Validation. For GLMMs with binary responses considered in the numerical studies of this paper, the criterion based on the ROC curve and its associated metric, the area under ROC curve (AUC), is widely used in selecting λ .

An important issue to assess the performance of a mixed-effects logistic regression model is the model's ability to discriminate between subjects with and without the event (i.e., response=1 or 0). If a fitted probability of the event occurrence is greater than a given threshold, subjects are classified into the positive class, and the negative class otherwise. Then, we assess the sensitivity and specificity of the classification using the ROC curve, the plot of sensitivity against one minus specificity overall possible thresholds. The AUC is used for measuring the overall performance of such discrimination with large AUC indicating better performance. The AUC ranges from 1 to 0.5 corresponding to a model from having a perfect discrimination to having no discrimination ability. We refer to Pepe (2003) for more details. The preferred tuning parameter estimate is obtained by maximizing the AUC over a range of possible λ . We illustrate this selection method in our application example in Section 6.

4 Asymptotic theory

In this section, we investigate the asymptotic properties of the proposed PCLE and two-step estimators under some regularity conditions. We denote α^0 and β^0 as the true values of parameters α and β , and u^0 as the vector of true realizations of random effects u , respectively. Accordingly, we have true values $\theta^0 = (\alpha^0, \beta^0)$ and $\theta_u^0 = (\theta^0, u^0)$.

We first establish the $\sqrt{N}$ -consistency and asymptotically normality for the PCL estimator developed in Subsection 3.1. We denote the expected information of the penalized

conditional log-likelihood by

$$I(\theta_u) = -\frac{1}{N}E\left(\ddot{U}(\theta_u)\right),$$

and suppose that the matrix $I(\theta_u)$ can be partitioned as $\begin{pmatrix} I_{11}(\theta_u) & I_{12}(\theta_u) \\ I_{21}(\theta_u) & I_{22}(\theta_u) \end{pmatrix}$ according to $\theta_u = (\theta, u)$. In what follows, we omit the variable in the information matrix for simplicity when evaluating it at $\theta_u = \theta_u^0$. For example, we write $I_{11} \equiv I_{11}(\theta_u^0)$. In addition, we need the following assumptions on the matrix I , tuning parameter λ and cluster size n_k for theoretical justification of the results.

(A1) Assume that I_{11} is positive definite with the smallest eigenvalue $\Lambda_{\min}(I_{11}) = \rho_1 > 0$ and the tuning parameter λ satisfies that $\lambda = o(\sqrt{N/K})$;

(A2) Let $\rho_k^V = \lim_{n_k \rightarrow \infty} n_{V_k}/n_k, k = 1, \dots, K$. There exists finite constants γ and δ such that

$$\lim_{N \rightarrow \infty} \frac{\sum_{k=1}^K n_{V_k}}{N} = \gamma \quad \text{and} \quad \lim_{N \rightarrow \infty} \frac{\sum_{k=1}^K \frac{(1-\rho_k^V)n_{V_k}}{\rho_k^{V^2}}}{N} = \delta.$$

We summarize asymptotic results of the proposed PCL estimator $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ in the following two theorems. Their proofs are sketched in the Appendix.

Theorem 4.1. ($\sqrt{N}$ -consistency) *Under some regular conditions and Assumptions (A1)-(A2), the solution $\hat{\theta}_u$ of the scoring equation $\dot{U}(\theta_u) = 0$ is $\sqrt{N}$ -consistent, i.e., $\sqrt{N}\|\hat{\theta}_u - \theta_u^0\|_2 = O_p(1)$ in some neighbor of the true θ_u .*

Theorem 4.2. (Asymptotic normality) *Under some regular conditions and Assumptions (A1)-(A2), $\sqrt{N}(\hat{\theta} - \theta^0)$ is distributed asymptotically by a normal distribution with mean 0 and variance-covariance matrix Σ , where Σ is defined as Eqs. (15) and (16) in Appendix.*

Next, we infer the consistency and asymptotic normality properties in Theorem 4.3 for the proposed two-step estimator given in Subsection 3.2, while the proof of the theorem is provided in the Appendix.

Theorem 4.3. (i) *The estimator $\tilde{\beta}$ is consistent and asymptotically normal,*

$$\sqrt{N}(\tilde{\beta} - \beta^0) \xrightarrow{d} N(0, 4(I_1^{-1} + I_1^{-1}\Sigma I_1^{-1}))$$

as $N \rightarrow \infty$, where I_1 and Σ are defined as in Theorem 1 of Liu and Xiang (2014);

(ii) *For the estimator of $\alpha_u = (\alpha, u')'$, we have $\sqrt{N}\kappa'\Sigma_\alpha^{-1}(\tilde{\alpha}_u - \alpha_u^0) \xrightarrow{d} N(0, 1)$ for any K -dimensional vector κ with norm $\|\kappa\|_2 = 1$, where Σ_α is specified by Eq. (18) in the proof of this theorem in the Appendix.*

All these three theorems show that the two proposed estimators are consistent and asymptotically normal as either the number of clusters K or the cluster size n_k , or both go to infinity. These theoretical properties relax the constraints of finite n_i in Wang et al. (2012) or finite K required in the earlier work of Liu and Xiang (2014), and make the proposed methods appealing for both longitudinal and clustered data. The result of Theorem 4.1 infers that $\sqrt{N}\|\hat{u} - u^0\|_2 = O_p(1)$, which is obviously different from the cases with finite numbers of clusters or finite cluster sizes. The reason inducing to this order can be explained intuitively by the fact that the difference $\hat{u}_k - u_k^0$ is related to $\hat{\theta} - \theta^0$, which is in the order of $O_p(N^{-1/2})$, and then each $\hat{u}_k - u_k^0$ can not be of the order $O_p(n_k^{-1/2})$ any more as $n_k \rightarrow \infty$ and $K \rightarrow \infty$. Especially when n_k is finite, we obtain that $\|\hat{\theta}_u - \theta_u^0\|_2 = O_p(K^{-1/2})$, which coincides with the result in Wang et al. (2012). In addition, Theorems 4.2 and 4.3 provide closed-form expressions for asymptotic variances of the estimators, making our estimation approaches more accessible.

5 Simulation studies

To evaluate the finite sample properties of the proposed methods, we conduct simulation studies in this section. In particular, we aim to 1) demonstrate the competitiveness of the proposed estimators under the assumption that random effects are normally distributed; 2) show the performance of the proposed methods in capturing non-normal random ef-

fects; and 3) evaluate the efficiency of the proposed methods in comparison with the existing counterparts.

Since the GLMM for binary data is most challenging in the presence of missing data, in the simulations below we only consider binary response data, which are generated according to the following mixed-effects logistic regression model:

$$P_{\theta}(Y_{ki} = 1|X_{ki}, u_k) = \frac{\exp(\alpha + \beta X_{ki} + u_k)}{1 + \exp(\alpha + \beta X_{ki} + u_k)}$$

for $i = 1, \dots, n_k$ and $k = 1, \dots, K$, where parameters $\theta = (\alpha, \beta) = (\log(3), \log(2))$ or $(\log(3), -\log(2))$, the number of clusters $K = 30, 50$ or 100 and the cluster size $n_k = 100$ for $k = 1, \dots, K$. Covariates X_{ki} and random effects u_k is generated from two different sampling designs: i) X_{ki} follows the standard normal distribution and is independent of random effects u_k , where $u_k \sim N(0, \sigma_T^2)$; ii) X_{ki} follows the standard normal distribution and is independent of random effects u_k , where $u_k \sim LN(-\sigma_T^2/2, \sigma_T^2) - 1$; and iii) $X_{ki} = u_k + X_{ki}^b$, where X_{ki}^b follows the standard normal distribution and u_k are distributed according to the mixture $1/6N(0.2, \sigma_T^2) + 1/3N(0.5, \sigma_T^2) + 1/2N(-0.4, \sigma_T^2)$. The second design relaxes the normality assumption for u_k and the third one further allows dependence between u_k and X_{ki} , leading to a more general and practical situation. We set the variance component $\sigma_T = 0.5, 1$ or 1.5 in the first and the third designs representing a weak, medium or strong correlation, respectively, while we take $\sigma_T = 1.2$ in the second design.

The auxiliary variable W_{ki} are defined as

$$W_{ki} = -3I_{(X_{ki} + \epsilon_{ki} \leq -3)} - 2I_{(-3 < X_{ki} + \epsilon_{ki} \leq -1)} + 2I_{(1 < X_{ki} + \epsilon_{ki} \leq 3)} + 3I_{(X_{ki} + \epsilon_{ki} > 3)},$$

where the random errors $\epsilon_{ki} \sim N(0, \sigma_{ki}^2)$ are independent, noting that $\sigma_{ki}^2 \sim U(0.2, 1)$ can measure the correlation between variables X and W for the i -th cluster. The indicator

variable r of the missing data is also assumed to be distributed by

$$P_{\theta}(r_{ki} = 1|W_{ki}) = \frac{\exp(0.2 + 0.5W_{ki})}{1 + \exp(0.2 + 0.5W_{ki})},$$

which results in around 46% of non-validation fraction.

For each parameter setting, 400 simulation runs are conducted. To gauge the effectiveness of the proposed methods, we estimate parameters for each simulated sample dataset using the proposed PCLE and two-step methods, existing method (ZCC) proposed by Zhou et al. (2002) and a naive version of Zhou et al's method based on the validation set only, respectively. The naive method accounts for measurement errors but ignores the auxiliary covariate information provided in the nonvalidation set. The variance component is estimated by $\hat{\sigma}_T^2 = 1/K \sum_{k=1}^K \hat{u}_k^2$.

Simulation results are summarized in terms of estimated bias (Bias), average of estimated standard errors (SE) obtained using the square root of the asymptotic variance, sample standard deviation (SD) and coverage percentage (CP) of the 95% confidence intervals.

Tables 1-2 about here

When the random effects are normally distributed indeed, results in Tables 1 and 2 show that both the proposed PCLE and two-step estimators of regression coefficients θ perform comparably with the ZCC estimator. They are all approximately unbiased and their SEs work well, while the naive likelihood method produces biased estimates for the intercept and slightly greater SEs than the proposed methods. It is reasonable that both proposed estimators exhibit slightly less efficient than the ZCC estimator since the ZCC is a likelihood-based method which is often more efficient if the random effects distribution assumption is true. In addition, the proposed methods yield more accurate estimates of the variance component σ_T^2 than the ZCC method and the naive one.

Tables 3-4 about here

When the random effects distribution deviates from the normality, Tables 3 and 4 show that both the PCLE and two-step estimates are unbiased and outperform the Naive and ZCC estimates, especially for the intercept α . Furthermore, when u_k are covariate-dependent, we see from Tables 5 and 6 that the two-step method yields unbiased and more efficient estimates for regression coefficients β , while the PCLE, naive and ZCC estimators are biased and their biases become severe for cases with a large variance component σ_T^2 , thus leading to poor CPs.

Tables 5-6 about here

Overall, it is clear that in general the proposed methods provide significant efficiency gains over the naive method under both simulation designs and all parameter settings. The two-step method and PCLE method **perform** as well as the ZCC method in the case when random effects are normally distributed, but **outperform** the ZCC and naive methods when the normality assumption of random effects is violated. In the case that random effects are correlated with covariates, the two-step method performs the best. These results confirm the validity of the proposed estimators and their improved efficiency due to the incorporation of the auxiliary covariate information in the inference procedure. Moreover, the estimated standard errors for the proposed estimators agree well with the sample standard deviations. The proposed methods also produce coverage percentages of the 95% confidence intervals very close to the nominal level.

6 Application to forest health data

We apply the proposed methods to analysis of data from a longitudinal study of the health status of beech trees at plots located in a northern Bavarian forest district, Germany. The main purpose of this study is to determine the risk factors affecting beech trees health.

The data set is available in R package *R2BayesX* (Belitz et al 2015). In the study, the sample consists of $K = 78$ beech trees (clusters) with $n_k = 22$ observations for each tree k ($k = 1, \dots, K$) over the period of 1983 to 2004. Originally, the health status was characterized by the percentage of defoliation on a 9-point ordinal scale. Following previous works of Kneib et al. (2009) and Groll and Tutz (2014), we consider the response variable as whether a tree is healthy or not, that is, the binary variable $Y = 1$ (unhealthy) if defoliation $\geq 12.5\%$ and $Y = 0$ (healthy) otherwise.

There were a number of covariates available in the data set and their brief descriptions can be found in Kneib et al. (2009). Groll and Tutz (2014) and Hui et al (2016) have selected important variables, which have great impacts on defoliation of trees, using the glmmLasso and regularized PQL methods, respectively. In this application, we focus on investigating the relationship between the probability of defoliation and those selected important covariates, which include *pH* (soil pH at 0-2 cm depth valued at the mid points of intervals $(3, 3.5]$, $(3.5, 4]$, $(4, 4.5]$, $(4.5, 5]$, $(5, 5.5]$ and $(5.5, 6.05]$ for five categories, respectively), *age* (age of the tree in years, ranging from 7 to 234), *canopy* (density of forest canopy in percent), *humus* (thickness of humus layer in 5 ordinal categories) and *stand* (type of stand, 0= mixed forest and 1= deciduous forest). We consider the case that the *pH* values of some trees were missing at random depending on the following missing indicator:

$$P(r = 1 | alkali) = \frac{\exp(-1 + 4 \cdot alkali)}{1 + \exp(-1 + 4 \cdot alkali)},$$

which results in a missing rate of 53.96%. In other words, the variable *pH* is taken as covariate X . Note that soils with *pH* values higher than 7.3 are increasingly more alkaline due to the presence of alkali salts. Thus we further take the variable *alkali* (proportion of base alkali-ions with categories 0= low or very low and 1= high or very high) as the auxiliary variable of the missing covariate *pH*.

We now fit the data with a mixed-effects logistic regression model. Estimation results obtained by the proposed PCLE and two-step methods in comparison with the ZCC and

naive methods are reported in Table 7. In general, all methods except the naive one produce estimates of the variance component σ_T^2 greater than 1, quite different from 0. It is interesting to test whether or not such estimated variance components imply significant cluster effects in the model, equivalently to test $H_0: \sigma_T^2 = 0$ vs $H_1: \sigma_T^2 > 0$. There are methods proposed in the literature for testing the significance of random effects in GLMMs. Detailed information can be found in a review article by Zhang and Lin (2008). The most commonly used test for such hypotheses is the likelihood ratio test (LRT) due to its desirable theoretical properties. Given that the null hypothesis lies on the boundary of the parameter space, Self and Liang (1987) showed that the LRT statistic is asymptotically distributed as a 50 : 50 mixture of χ_0^2 and χ_1^2 under H_0 . We apply the LRT to the model under both proposed methods and the ZCC method, respectively. Our resulting p-values are all less than 0.01, indicating strong evidence against H_0 regardless of which estimation method is used.

It is seen that both the proposed methods yield similar estimates of regression coefficients and reveal that covariates *age*, *canopy* and *stand* have significant effects on the defoliation of trees. Similar findings can be obtained by using the ZCC method. On the other hand, the naive method claims no cluster effects in the forest health data. This may lead to a misleading conclusion as the naive estimator makes use of the complete cases only and ignores the information provided by the nonvalidation set.

In this analysis, we use the area under the ROC curve to estimate the optimal tuning parameter λ . Plots (a) and (b) in Figure 1 show the AUC values obtained from the proposed PCLE and two-step methods, respectively, over an excerpt (from 0.03 to 2.03 by step 0.1) of possible λ values. We observe that for each method there exists a well-separated and unique maximizer $\lambda = 0.13$, at which the AUC reaches up to 0.90 or above. This indicates that the fitted models obtained from both methods with the optimal value of λ can effectively discriminate healthy and unhealthy trees. We also present a comparison between the ROC curves of the two methods corresponding to the optimum

AUC in Figure 2, which again verifies similar fitting results obtained from proposed two methods and the ZCC method for this data set.

Table 7 about here

Figure 1 about here

Figure 2 about here

7 Discussion

In practice, since no one knows the true missing mechanism and the true distribution of random effects, assumptions for missing data and the random effects are often non-verifiable. We propose two estimation methods based on a penalized conditional likelihood and a two-step procedure, respectively, to analyze clustered/longitudinal data with covariates missing at random and auxiliary information being readily available, especially when usual parametric distribution assumption for the random effects is suspicious.

Using an empirical density function to estimate the conditional distribution of missing data, we are able to develop estimating equations, which result in efficient estimators. Both theoretical and numerical results show favorable performance of the proposed methods. In particular, our methods relax the restraints in the early works of Wang et al. (2012) and Liu and Xiang (2014) in which either the number of observations within clusters or the number of clusters has to be finite, and allow both the number of clusters and cluster size to be divergent simultaneously. Such a flexibility facilitates the applicability of the methods for both clustered and longitudinal studies, thus extending the existing work to a more general situation, while preserving the robustness to misspecification of the random effects distribution and the missingness.

It is noted that selection of the auxiliary variable plays important role in the proposed methods. Since the distribution of the missing covariate in the non-validation set is estimated by that of the auxiliary variable, the auxiliary variable is required to be at least moderately associated with the covariate with missing values. In practice, a potential auxiliary variable can be extraneous to the model of our main interest but beneficial in predicting the covariate with missing data. Although there is still a lot to be learned about the selection of auxiliary variables in analysis of missing data, the following three common scenarios may provide auxiliary information according to the work of Horton and Laird (2001). First, there exist administrative record data possibly from previous studies that can be matched to subjects. Second, proxy informants are available in addition to the primary exposure of interest in a study. Third, in cases where the assessment of a covariate is either expensive or invasive, only a subset of subjects might be included in a validation sample, while more error-prone auxiliary data might be collected on all subjects.

In this work, we consider discrete response variables only, which are common types of data handled by GLMMs typically, such as the binary forest health data in our motivating example. On the other hand, the proposed estimating equations rely on the empirical distribution of missing data conditional on the auxiliary variable, which is also required to be discrete. Based on our current work, it is straightforward to extend the methods to incorporating a continuous auxiliary covariate by either discretizing the continuous variable into categories or introducing a kernel smoother as suggested by Chen and Lin (2010) to replace the empirical distribution estimator used in the proposed estimation procedures and thus to enhance the efficiency of the estimator. An in-depth discussion on this issue deserves further investigation.

Acknowledgements

This research was supported in part by the Natural Science Foundation of China (11671310) and the Singapore Ministry of Education Academic Research Fund Tier 2 Grant (MOE2013-T2-2-118).

Appendix: Proofs of theorems

Proof of Theorem 4.1. Define $B(C) = \{\theta_u : \|\theta - \theta^0\|_2 = C/\sqrt{N}, \|u - u^0\|_2 = C/\sqrt{N}\}$.

It is suffice to prove that for every $\epsilon > 0$,

$$Pr\left(\sup_{\theta_u \in B(C)} (\theta_u - \theta_u^0)' \dot{U}(\theta_u) < 0\right) > 1 - \epsilon.$$

By Taylor's expansion, we have

$$\begin{aligned} & (\theta_u - \theta_u^0)' \dot{U}(\theta_u) \\ &= (\theta_u - \theta_u^0)' \dot{U}(\theta_u^0) + (\theta_u - \theta_u^0)' \ddot{U}(\theta_u^0)(\theta_u - \theta_u^0) + o_p(1) \\ &= J_1 + J_2 + o_p(1) \quad (\text{say}). \end{aligned} \tag{7}$$

To determine the order of J_1 , we first note that

$$\begin{aligned} \frac{1}{\sqrt{N}} \dot{U}(\theta_u^0) &= \frac{1}{\sqrt{N}} \frac{d}{d\theta_u} \log L(\theta_u^0) + \frac{1}{\sqrt{N}} \sum_{k=1}^K \sum_{i \in \bar{V}_k} \left(\frac{d}{d\theta_u} \log \hat{P}_{\theta^0}(Y_{ki}|Z_{ki}, W_{ki}, u_k^0) \right. \\ &\quad \left. - \frac{d}{d\theta_u} \log P_{\theta^0}(Y_{ki}|Z_{ki}, W_{ki}, u_k^0) \right) - \frac{\lambda}{\sqrt{N}} \frac{d}{d\theta_u} \sum_{k=1}^K u_k^2/2 \\ &= B_1 + B_2 + B_3 \quad (\text{say}). \end{aligned} \tag{8}$$

Let

$$M_{ki,kj} = \frac{dP_{\theta^0}(Y_{ki}|Z_{ki}, X_{kj}, u_k^0)/d\theta_u}{P_{\theta^0}(Y_{ki}|Z_{ki}, W_{ki}, u_k^0)} - \frac{dP_{\theta^0}(Y_{ki}|Z_{ki}, W_{ki}, u_k^0)/d\theta_u}{(P_{\theta^0}(Y_{ki}|Z_{ki}, W_{ki}, u_k^0))^2} P_{\theta^0}(Y_{ki}|Z_{ki}, X_{kj}, u_k^0).$$

By Eq. (3), the term B_2 can be rewritten as

$$\begin{aligned}
B_2 &= \frac{1}{\sqrt{N}} \sum_{k=1}^K \sum_{i \in \bar{V}_k} \sum_{j \in V_k} \frac{M_{ki,kj} I_{[A_{ki}=A_{kj}]}}{\sum_{j \in V_k} I_{[A_{ki}=A_{kj}]}} + o_p(1) \\
&= \frac{1}{\sqrt{N}} \sum_{k=1}^K \sum_{j \in V_k} \frac{n_{\bar{V}_k}}{n_{V_k}} \sum_{i \in \bar{V}_k} \frac{M_{ki,kj} I_{[A_{ki}=A_{kj}]}}{\sum_{i \in \bar{V}_k} I_{[A_{ki}=A_{kj}]}} + o_p(1) \\
&\stackrel{a.s.}{=} \frac{1}{\sqrt{N}} \sum_{k=1}^K \left\{ \frac{1 - \rho_k^V}{\rho_k^V} \sum_{j \in V_k} M_{X_{kj}, A_{kj}}(\theta_u^0) \right\}, \tag{9}
\end{aligned}$$

where

$$M_{X_{kj}, A_{kj}}(\theta_u^0) = \sum_{i \in \bar{V}_k} \frac{M_{ki,kj} I_{[A_{ki}=A_{kj}]}}{\sum_{i \in \bar{V}_k} I_{[A_{ki}=A_{kj}]}}.$$

It follows from the strong law of large numbers that

$$M_{X_{kj}, A_{kj}}(\theta_u^0) \stackrel{a.s.}{\rightarrow} -E \left[\frac{d}{d\theta_u} \log P_{\theta^0}(Y_{k1} | Z_{k1}, W_{k1}, u_k^0) \middle| A_{kj}, X_{kj} \right]. \tag{10}$$

Combining Eqs. (8)–(10), we have

$$\begin{aligned}
\frac{1}{\sqrt{N}} \dot{U}(\theta_u^0) &= \frac{1}{\sqrt{N}} \frac{d}{d\theta_u} \log L(\theta_u^0) - \frac{1}{\sqrt{N}} \sum_{k=1}^K \frac{1 - \rho_k^V}{\rho_k^V} \sum_{i \in V_k} E \left[\frac{d}{d\theta_u} \log P_{\theta^0}(Y_{k1} | Z_{k1}, W_{k1}, u_k^0) \middle| A_{ki}, X_{ki} \right] \\
&\quad - \frac{\lambda}{\sqrt{N}} \frac{d}{d\theta_u} \sum_{k=1}^K u_k^2 / 2 + o_p(1). \tag{11}
\end{aligned}$$

Thus,

$$\begin{aligned}
J_1 &= (\theta - \theta^0)' \dot{U}_\theta(\theta_u^0) + (u - u^0)' \dot{U}_u(\theta_u^0) \\
&= O_p(C/\sqrt{N} \cdot \sqrt{N}) + O_p\left(\sum_{k=1}^K \sqrt{n_k}(u_k - u_k^0)\right) - \lambda \sum_{k=1}^K u_k(u_k - u_k^0).
\end{aligned}$$

Since $\sum_{k=1}^K \sqrt{n_k}(u_k - u_k^0) \leq \sqrt{\left(\sum_{k=1}^K n_k\right) \sum_{k=1}^K (u_k - u_k^0)^2} = \sqrt{N \cdot C^2/N} = C$ and $\lambda \sum_{k=1}^K u_k(u_k - u_k^0) \leq \lambda \sqrt{\sum_{k=1}^K (u_k - u_k^0)^2 \sum_{k=1}^K u_k^2} = \lambda O_p(C\sqrt{K/N})$ by Schwarz inequality, we have $J_1 =$

$O_p(C)$ by the Assumption (A1) that $\lambda = o(\sqrt{N/K})$.

Then we consider the order of J_2 . From Eq. (11) we can see that

$$\begin{aligned} \frac{1}{N} \ddot{U}(\theta_u^0) &= \frac{1}{N} \frac{d^2}{d\theta_u^2} \log L(\theta_u^0) - \frac{1}{N} \sum_{k=1}^K \frac{1 - \rho_k^V}{\rho_k^V} \sum_{i \in V_k} E \left[\frac{d^2}{d\theta_u^2} \log P_{\theta^0}(Y_{k1} | Z_{k1}, W_{k1}, u_k^0) \middle| A_{ki}, X_{ki} \right] \\ &\quad - \frac{\lambda}{N} \frac{d^2}{d\theta_u^2} \sum_{k=1}^K u_k^2/2 + o_p(1). \end{aligned}$$

Let

$$\begin{aligned} s_0 &= -E \left[\frac{\partial^2}{\partial \theta \partial \theta'} \log P_{\theta^0}(Y_{k1} | Z_{k1}, X_{k1}, u_k^0) \right], \\ s_1 &= -E \left[\frac{\partial^2}{\partial \theta \partial u_k} \log P_{\theta^0}(Y_{k1} | Z_{k1}, X_{k1}, u_k^0) \right], \\ s_2 &= -E \left[\frac{\partial^2}{\partial u_k^2} \log P_{\theta^0}(Y_{k1} | Z_{k1}, X_{k1}, u_k^0) \right]. \end{aligned}$$

We note that s_0 is a $p \times p$ dimensional matrix, s_1 is a $p \times 1$ vector, s_2 is a scalar number, and p is the dimension of variable θ . Then $I_{11} = s_0$, $I_{12} = \frac{1}{N}(n_1 s_1, \dots, n_K s_1)$ and $I_{22} = \frac{1}{N} \text{diag}(n_1 s_2, \dots, n_K s_2)$. Applying Chebyshev inequality we get that $\|\frac{1}{N} \ddot{U}(\theta_u^0) + I(\theta_u^0)\|_F = o_p(1)$, where $\|\cdot\|_F$ means the Frobenius norm for a matrix. It follows that

$$\begin{aligned} J_2 &= -N(\theta_u - \theta_u^0)' I(\theta_u^0)(\theta_u - \theta_u^0) + o_p(1) \\ &= - \left(N(\theta - \theta^0)' I_{11}(\theta - \theta^0) + 2N(\theta - \theta^0)' I_{12}(u - u^0) \right. \\ &\quad \left. + N(u - u^0)' I_{22}(u - u^0) \right) + o_p(1) \\ &= -(D_1 + D_2 + D_3) + o_p(1). \quad (\text{say}) \end{aligned}$$

Since

$$\begin{aligned}
D_1 &\geq N\rho_1 \|\theta - \theta^0\|_2^2 = \rho_1 C^2 \\
D_2 &= 2(\theta - \theta^0)' s_1 \cdot \sum_{k=1}^K n_k (u_k - u_k^0) \leq 2\|s_1\|_F \|\theta - \theta^0\|_2 \cdot \sqrt{\sum_{k=1}^K n_k^2} \|u - u^0\|_2 \\
&= o_p\left(\frac{KC^2}{N}\right) = o_p(C^2) \\
D_3 &= s_2 \sum_{k=1}^K n_k (u_k - u_k^0)^2 \leq s_2 \cdot \sup n_k \cdot \|u - u^0\|_2^2 = O_p(C^2),
\end{aligned}$$

it is obviously that the term $D_1 > 0$ dominates D_2 and $D_3 > 0$ in addition, then we have $J_2 = -O_p(C^2)$.

The above concludes show that for sufficiently large C , J_2 dominates J_1 uniformly in $B(C)$. This completes the proof of Theorem 4.1. $\square$

Proof of Theorem 4.2. By Theorem 4.1, the Taylor expansion of $\frac{1}{\sqrt{N}}\dot{U}(\hat{\theta}_u)$ at θ_u^0 is given by

$$\frac{1}{\sqrt{N}}\dot{U}(\hat{\theta}_u) = \frac{1}{\sqrt{N}}\dot{U}(\theta_u^0) - I \cdot \sqrt{N}(\hat{\theta}_u - \theta_u^0) + o_p(1).$$

Therefore,

$$\sqrt{N}I(\hat{\theta}_u - \theta_u^0) = \frac{1}{\sqrt{N}}\dot{U}(\theta_u^0) + o_p(1),$$

which is equivalent to the equations that

$$\sqrt{N}s_0(\hat{\theta} - \theta^0) + \frac{1}{\sqrt{N}}s_1 \sum_{k=1}^K n_k (\hat{u}_k - u_k^0) = \frac{1}{\sqrt{N}}\dot{U}_\theta(\theta_u^0) + o_p(1), \quad (12)$$

$$\frac{n_k}{\sqrt{N}}(s_1'(\hat{\theta} - \theta^0) + s_2(\hat{u}_k - u_k^0)) = \frac{1}{\sqrt{N}}\dot{U}_{u_k}(\theta_u^0) + o_p(1), \quad k = 1, \dots, K. \quad (13)$$

Eq. (13) gives that for $k = 1, \dots, K$,

$$\frac{1}{\sqrt{N}}(\hat{u}_k - u_k^0) = \frac{1}{\sqrt{N}n_k s_2}[\dot{U}_{u_k}(\theta_u^0) - n_k s_1'(\hat{\theta} - \theta^0)] + o_p(1).$$

Substituting this equation into Eq. (12), we obtain that

$$\sqrt{N}(\hat{\theta} - \theta^0) = \frac{1}{\sqrt{N}} \left(s_0 - \frac{1}{s_2} s_1 s_1' \right)^{-1} \left(\dot{U}_\theta(\theta_u^0) - \frac{1}{s_2} s_1 \sum_{k=1}^K \dot{U}_{u_k}(\theta_u^0) \right). \quad (14)$$

Eq. (11) implies that $E(\dot{U}(\theta_u^0)/\sqrt{N}) = 0$ when $N \rightarrow \infty$. Given the missing covariate X and variable A , $\{M_{X_{ki}, A_{ki}}(\theta_u), i \in V_k\}$ is independent and uncorrelated to the score function in the validation set. Consequently, the covariance matrix can be obtained as

$$\begin{aligned} & Var \left[\frac{1}{\sqrt{N}} \dot{U}(\theta_u^0) \right] \\ &= (1 - \gamma) E \left[- \frac{d^2}{d\theta_u^2} \log P_{\theta^0}(Y|Z, W, u^0) \right] + \gamma E \left[- \frac{d^2}{d\theta_u^2} \log P_{\theta^0}(Y|Z, X, u^0) \right] \\ &\quad + \delta Var \left\{ E \left[\frac{d}{d\theta_u} \log P_{\theta^0}(Y|Z, W, u^0) | X, A \right] \right\} + o(1) \\ &= I + \Sigma_1 + o(1). \end{aligned}$$

Letting $\Sigma_2 = I + \Sigma_1 = \begin{pmatrix} \Sigma_2^{11} & \Sigma_2^{12} \\ \Sigma_2^{12'} & \Sigma_2^{22} \end{pmatrix}$ according to $\theta_u = (\theta, u)$, it follows from the central limit theorem that

$$\frac{1}{\sqrt{N}} \left(\dot{U}_\theta(\theta_u^0) - \frac{1}{s_2} s_1 \sum_{k=1}^K \dot{U}_{u_k}(\theta_u^0) \right) \xrightarrow{d} N(0, \Sigma_3), \quad (15)$$

where $\Sigma_3 = \Sigma_2^{11} + \frac{\|\Sigma_2^{22}\|_S}{s_2^2} s_1 s_1' - \frac{1}{s_2} (s_1 \|\Sigma_2^{12}\|_{RS}' + \|\Sigma_2^{12}\|_{RS} s_1')$ with $\|\cdot\|_S$ meaning to the element sum of a matrix and $\|\cdot\|_{RS}$ meaning to the row sum of a matrix. And then by using the relation (14) it gives that

$$\sqrt{N}(\hat{\theta} - \theta^0) \xrightarrow{d} N(0, \Sigma) \quad \text{and} \quad \Sigma = \left(s_0 - \frac{1}{s_2} s_1 s_1' \right)^{-1} \Sigma_3 \left(s_0 - \frac{1}{s_2} s_1 s_1' \right)^{-1}. \quad (16)$$

This concludes the theorem. $\square$

Proof of Theorem 4.3. The result in part (i) can be derived directly following Theorem 1 in the work of Liu and Xiang (2014). Thus, here we prove part (ii) only.

First, we partition the information submatrix I as $\begin{pmatrix} I_{\alpha_u \alpha_u} & I_{\alpha_u \beta} \\ I_{\beta \alpha_u} & I_{\beta \beta} \end{pmatrix}$ according to $\theta_u = (\alpha_u, \beta)$, where $I_{\alpha_u \alpha_u}$ is the first block entry of I . Similar to the proof of Theorem 4.3, under the regular conditions, we have

$$\sqrt{N}(\tilde{\alpha}_u - \alpha_u^0) = I_{\alpha_u \alpha_u}^{-1} \left(\frac{1}{\sqrt{N}} \dot{U}_{\alpha_u}(\tilde{\beta}, \alpha_u^0) \right) + o_p(1). \quad (17)$$

Let $I_2 = -E(\ddot{R}_W(\beta^0))$. By Taylor's expansion and the result in part (i) of this theorem, we get

$$\begin{aligned} \frac{1}{\sqrt{N}} \dot{U}_{\alpha_u}(\tilde{\beta}, \alpha_u^0) &= \frac{1}{\sqrt{N}} \dot{U}_{\alpha_u}(\theta_u^0) + \frac{1}{N} \ddot{U}_{\alpha_u \beta}(\theta_u^0) \sqrt{N}(\tilde{\beta} - \beta^0) + o_p(1) \\ &= \frac{1}{\sqrt{N}} \dot{U}_{\alpha_u}(\theta_u^0) + \sqrt{N} I_{\alpha_u \beta} I_2^{-1} \left(\frac{1}{N} \dot{R}_W(\beta^0) - \frac{1}{N} \sum_{k=1}^K \frac{n_{k2}}{n_{k1}(n_{k1} - 1)} \right. \\ &\quad \left. \sum_{i^V, j^V \in V_k, i^V < j^V} E \left(\frac{d}{d\beta} \log f_{\beta^0}^{12}(A_{ki^V}, A_{kj^V}) \middle| A_{ki^V}, A_{kj^V}, X_{ki^V}, X_{kj^V} \right) \right) + o_p(1) \\ &= \psi(\theta_u^0) + o_p(1) \quad (\text{say}). \end{aligned}$$

Consequently, we obtain that $\sqrt{N} \kappa' \Sigma_\alpha^{-1}(\tilde{\alpha}_u - \alpha_u^0)$ is asymptotically standard normal with

$$\Sigma_\alpha = I_{\alpha_u \alpha_u}^{-1} \text{Var}(\psi(\theta_u^0)) I_{\alpha_u \alpha_u}^{-1}. \quad (18)$$

This completes the proof. □

References

Agresti, A., Caffo, B. and Ohman-Strickland, P. (2004) Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies.

Computational Statistics and Data Analysis **47**, 639–653.

- Aloisio, et al. (2014) Analysis of partially observed clustered data using generalized estimating equations and multiple imputation. *Stata J* **14**, 863–883.
- Belitz, C., Brezger, A., Klein, N., Kneib, T., Lang, S., and Umlauf, N. (2015) *BayesX - Software for Bayesian Inference in Structured Additive Regression Models*. Version 3.0. Available from <http://www.bayesx.org>.
- Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with Discussion). *Journal of the Royal Statistical Society, Series B* **36**, 192–236.
- Carroll, R.J. and Wand, M.P. (1991) Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society, Series B* **53**, 573–585.
- Chatelain, F., Lambert-Lacroix, S. and Tournet, J.-Y. (2009) Pairwise likelihood estimation for multivariate mixed Poisson models generated by Gamma intensities. *Statistics and Computing*, **19**, 283–301.
- Chen, J. and Lin, C.-D. (2010) Modeling continuous auxiliary covariate data in generalized linear mixed models using the kernel smoother. *Biometrical Journal* **52**, 362–376.
- Groll, A. and Tutz, G. (2014) Variable selection for generalized linear mixed models by L1-penalized estimation. *Statistics and Computing* **24**, 137–154.
- Horton, N.J. and Laird, N.M. (2001) Maximum likelihood analysis of logistic regression models with incomplete covariate data and auxiliary information. *Biometrics* **57**, 34–42.
- Huang, C.Y., Qin, J. and Wang, M.C. (2010) Semiparametric analysis for recurrent event data with time-dependent covariates and informative censoring. *Biometrics* **66**, 39–49.
- Hui, K.C. Müller, S. and Welsh, A.H. (2017) Joint Selection in Mixed Models using Regularized PQL. *Journal of the American Statistical Association* **112**, 1323–1333.

- Ibrahim, J.G., Lipsitz, S.R. and Chen, M-H. (1999) Missing covariates in generalized linear models with the missing data mechanism is non-ignorable. *Journal of the Royal statistical society, Series B* **61**, 173–190.
- Jiang, J.M. (1999) Conditional inference about generalized linear mixed models. *The Annals of Statistics* **27**, 1974–2007.
- Kalbfleisch, J.D. (1978) Likelihood methods and nonparametric tests. *Journal of American Statistical Association* **73**, 167–170.
- Kneib, T., Hothorn, T., Tutz, G. (2009) Variable selection and model choice in geodetic regression models. *Biometrics* **65**, 626–634.
- Kuk, A.Y. (2007) A hybrid pairwise likelihood method. *Biometrika* **94**, 939–952.
- Kuk, A.Y and Nott, D.J. (2000) A pairwise likelihood approach to analyzing correlated binary data. *Statistics & probability letters* **47**, 329–335.
- Li, H.C. and Yi, G.Y. (2013) A pairwise likelihood approach for longitudinal data with missing observations in both response and covariates. *Computational Statistics and Data Analysis* **68**, 66–81.
- Liang, K.Y. and Qin, J. (2000) Regression analysis under non-standard situations: a pairwise pseudolikelihood approach. *Journal of Royal Statistics Society, Series B* **62**, 773–786.
- Liu, L. and Xiang, L.M. (2014) Semiparametric Estimation in Generalized Linear Mixed Models with Auxiliary Covariates: A Pairwise Likelihood Approach. *Biometrics* **70**, 910–919.
- Little, R. and Rubin, D. *Statistical analysis with missing data*. New York: John Wiley, 2002.
- Pepe, M.S. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. UK: Oxford University Press; 2003.

- Pepe, M. and Fleming, T. (1991) A General Nonparametric Method for Dealing with Errors in Missing or Surrogate Data. *Journal of the American Statistical Association* **86**, 108–113.
- Robins, J.M. Rotnitzky, A. and Zhao, L.P. (1994) Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association* **89**, 846–866.
- Rubin, D. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons Inc, 1987.
- Sabry, M.Y., Kholy, R.B. and Gad, A.M. (2016). Generalized linear mixed model for longitudinal data with missing values: a monte carlo EM approach. *International Journal of Probability and Statistics* **5**, 82–88.
- Self, S.G. and Liang, K.Y (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**, 605–610.
- Wang, P., Tsai G.-F. and Qu, A. (2012) Conditional inference functions for mixed-effects models with unspecified random-effects distribution. *Journal of the American Statistical Association* **107**, 725–736.
- Xie, F. and Paik, M.C. (1997) Generalized Estimating Equation Model for Binary Outcomes with Missing Covariates. *Biometrics* **53**, 1458–1466.
- Zhang D. and Lin X. (2008) Variance component testing in generalized linear mixed models for longitudinal clustered data and other related topics. In: Dunson DB, editor. *Random effect and latent variable model selection*. **Chapter 2**, p. 192. Springer Lecture Notes in Statistics.
- Zhou, H., Chen, J. and Cai, J. (2002) Random effects logistic regression analysis with auxiliary covariates. *Biometrics* **58**, 352–360.

Table 1: Simulation results for normally distributed random effects, where $\alpha = \log(3) = 1.099$, $\beta = \log(2) = 0.693$ and $n_k = 100$

K	σ_T	Methods	$\hat{\sigma}_T$	$\hat{\alpha}$				$\hat{\beta}$			
				Bias	SD	SE	CP	Bias	SD	SE	CP
30	0.5	PCLE	0.530	0.008	0.099	0.108	0.954	-0.004	0.055	0.053	0.924
		Two step	0.538	0.010	0.109	0.108	0.935	-0.005	0.058	0.053	0.910
		Naive	0.064	-0.046	0.105	0.065	0.739	-0.032	0.064	0.066	0.923
		ZCC	0.322	-0.023	0.096	0.088	0.883	-0.022	0.052	0.052	0.930
	1	PCLE	0.985	-0.003	0.182	0.187	0.953	-0.003	0.058	0.055	0.935
		Two step	0.995	0.006	0.193	0.188	0.945	-0.007	0.061	0.055	0.931
		Naive	0.852	-0.039	0.179	0.167	0.920	-0.029	0.071	0.068	0.925
		ZCC	0.924	-0.022	0.175	0.173	0.940	-0.021	0.055	0.054	0.928
	1.5	PCLE	1.448	-0.013	0.269	0.271	0.950	-0.006	0.061	0.058	0.935
		Two step	1.477	0.008	0.284	0.276	0.950	-0.004	0.062	0.057	0.920
		Naive	1.335	-0.051	0.253	0.251	0.945	-0.024	0.073	0.071	0.940
		ZCC	1.401	-0.032	0.252	0.258	0.955	-0.020	0.058	0.057	0.948
	0.5	PCLE	0.539	0.011	0.082	0.084	0.950	0.000	0.042	0.041	0.954
		Two step	0.543	0.013	0.079	0.084	0.965	-0.002	0.043	0.041	0.930
		Naive	0.031	-0.046	0.086	0.048	0.678	-0.026	0.054	0.051	0.934
		ZCC	0.352	-0.020	0.078	0.063	0.890	-0.019	0.042	0.040	0.910
	1	PCLE	1.009	0.008	0.151	0.148	0.942	-0.005	0.042	0.042	0.950
		Two step	1.018	0.016	0.149	0.149	0.955	-0.004	0.042	0.043	0.938
		Naive	0.864	-0.035	0.146	0.131	0.905	-0.018	0.057	0.053	0.930
		ZCC	0.930	-0.015	0.146	0.136	0.933	-0.017	0.041	0.042	0.940
	1.5	PCLE	1.484	0.000	0.228	0.215	0.932	-0.010	0.046	0.045	0.940
		Two step	1.507	0.016	0.223	0.218	0.950	-0.005	0.044	0.045	0.955
		Naive	1.348	-0.047	0.214	0.197	0.920	0.018	0.059	0.055	0.930
		ZCC	1.414	-0.023	0.215	0.203	0.935	-0.016	0.042	0.045	0.955
100	0.5	PCLE	0.550	0.010	0.058	0.061	0.958	0.003	0.030	0.029	0.940
		Two step	0.556	0.009	0.059	0.061	0.965	-0.007	0.031	0.029	0.935
		Naive	0.017	-0.054	0.058	0.033	0.578	-0.030	0.034	0.036	0.870
		ZCC	0.380	-0.023	0.056	0.046	0.870	-0.019	0.029	0.028	0.898
	1	PCLE	1.027	0.009	0.108	0.107	0.948	0.000	0.031	0.030	0.942
		Two step	1.037	0.010	0.107	0.107	0.955	-0.007	0.033	0.030	0.905
		Naive	0.876	-0.040	0.101	0.094	0.913	-0.020	0.038	0.037	0.905
		ZCC	0.937	-0.022	0.103	0.097	0.920	-0.016	0.030	0.030	0.910
	1.5	PCLE	1.532	0.013	0.164	0.157	0.936	-0.002	0.033	0.032	0.940
		Two step	1.540	0.013	0.162	0.157	0.955	-0.009	0.034	0.031	0.920
		Naive	1.357	-0.052	0.149	0.141	0.930	-0.021	0.039	0.039	0.918
		ZCC	1.422	-0.030	0.152	0.145	0.923	-0.016	0.032	0.031	0.928

Note: SD represents the sample standard deviation, SE represents the mean of the estimated standard errors and CP represents the empirical 95% coverage probability.

Table 2: Simulation results for normally distributed random effects, where $\alpha = \log(3) = 1.099$, $\beta = -\log(2) = -0.693$ and $n_k = 100$

K	σ_T	Methods	$\hat{\sigma}_T$	$\hat{\alpha}$				$\hat{\beta}$			
				Bias	SD	SE	CP	Bias	SD	SE	CP
30	0.5	PCLE	0.531	0.011	0.100	0.108	0.960	-0.006	0.049	0.052	0.960
		Two step	0.537	0.011	0.105	0.108	0.950	-0.007	0.046	0.053	0.965
		Naive	0.099	-0.046	0.108	0.071	0.759	0.024	0.061	0.063	0.952
		ZCC	0.324	-0.020	0.096	0.089	0.880	0.012	0.048	0.051	0.958
	1	PCLE	0.995	0.004	0.186	0.189	0.950	0.002	0.053	0.054	0.956
		Two step	1.008	0.012	0.196	0.191	0.945	-0.005	0.053	0.055	0.945
		Naive	0.869	-0.032	0.187	0.170	0.923	0.014	0.063	0.066	0.958
		ZCC	0.927	-0.014	0.178	0.174	0.928	0.012	0.051	0.054	0.963
	1.5	PCLE	1.463	-0.006	0.272	0.274	0.948	0.007	0.058	0.057	0.955
		Two step	1.478	0.006	0.289	0.276	0.925	-0.004	0.059	0.057	0.940
		Naive	1.351	-0.042	0.263	0.254	0.935	0.013	0.065	0.069	0.970
		ZCC	1.405	-0.022	0.259	0.259	0.935	0.013	0.055	0.057	0.963
50	0.5	PCLE	0.539	0.010	0.082	0.084	0.952	-0.001	0.041	0.040	0.944
		Two step	0.541	0.014	0.081	0.084	0.955	0.000	0.043	0.041	0.960
		Naive	0.079	-0.042	0.087	0.053	0.722	0.025	0.048	0.048	0.930
		ZCC	0.346	-0.020	0.079	0.062	0.888	0.019	0.040	0.040	0.915
	1	PCLE	1.004	0.007	0.151	0.148	0.936	0.002	0.042	0.042	0.954
		Two step	1.017	0.018	0.150	0.149	0.965	0.000	0.042	0.042	0.955
		Naive	0.879	-0.028	0.149	0.139	0.923	0.016	0.050	0.051	0.945
		ZCC	0.933	-0.012	0.147	0.136	0.938	0.015	0.041	0.042	0.940
	1.5	PCLE	1.483	-0.021	0.213	0.199	0.930	0.010	0.041	0.044	0.940
		Two step	1.500	0.015	0.225	0.217	0.947	0.004	0.043	0.044	0.955
		Naive	1.362	-0.037	0.215	0.200	0.928	0.016	0.052	0.054	0.928
		ZCC	1.417	-0.014	0.217	0.203	0.943	0.016	0.041	0.044	0.943
100	0.5	PCLE	0.549	0.010	0.059	0.061	0.960	-0.004	0.030	0.029	0.934
		Two step	0.559	0.013	0.060	0.061	0.955	-0.006	0.030	0.029	0.930
		Naive	0.050	-0.051	0.060	0.036	0.582	0.030	0.034	0.034	0.864
		ZCC	0.384	-0.021	0.057	0.046	0.875	0.014	0.027	0.028	0.920
	1	PCLE	1.030	0.010	0.108	0.107	0.938	-0.001	0.031	0.030	0.940
		Two step	1.045	0.015	0.109	0.108	0.955	-0.005	0.031	0.030	0.935
		Naive	0.894	-0.034	0.103	0.096	0.913	0.013	0.035	0.036	0.923
		ZCC	0.941	-0.019	0.104	0.098	0.915	0.011	0.028	0.029	0.945
	1.5	PCLE	1.534	0.013	0.163	0.157	0.936	0.000	0.033	0.032	0.938
		Two step	1.546	0.019	0.163	0.158	0.940	-0.006	0.032	0.031	0.930
		Naive	1.375	-0.043	0.149	0.143	0.918	0.012	0.038	0.038	0.943
		ZCC	1.426	-0.025	0.153	0.145	0.925	0.010	0.030	0.031	0.948

Note: SD represents the sample standard deviation, SE represents the mean of the estimated standard errors and CP represents the empirical 95% coverage probability.

Table 3: Simulation results for log-normally distributed random effects, where $\alpha = \log(3) = 1.099$, $\beta = \log(2) = 0.693$, $n_k = 100$ and $\sigma_T = 1.2$

K	Methods	$\hat{\sigma}_T$	$\hat{\alpha}$				$\hat{\beta}$			
			Bias	SD	SE	CP	Bias	SD	SE	CP
30	PCLE	1.140	-0.079	0.210	0.216	0.945	0.002	0.049	0.052	0.962
	Two step	1.244	-0.044	0.247	0.230	0.937	-0.002	0.052	0.052	0.940
	Naive	0.636	-0.210	0.172	0.135	0.613	-0.018	0.061	0.065	0.939
	ZCC	0.832	-0.168	0.180	0.157	0.749	-0.013	0.048	0.052	0.942
50	PCLE	1.162	-0.067	0.175	0.170	0.924	0.002	0.039	0.041	0.949
	Two step	1.263	-0.054	0.188	0.181	0.947	-0.001	0.041	0.040	0.940
	Naive	0.665	-0.203	0.145	0.107	0.505	-0.013	0.050	0.051	0.931
	ZCC	0.851	-0.159	0.152	0.125	0.689	-0.012	0.038	0.040	0.947
100	PCLE	1.160	-0.077	0.117	0.120	0.899	0.003	0.028	0.029	0.962
	Two step	1.295	-0.044	0.132	0.131	0.942	0.001	0.028	0.028	0.968
	Naive	0.663	-0.216	0.099	0.075	0.221	-0.013	0.036	0.036	0.934
	ZCC	0.838	-0.170	0.103	0.087	0.510	-0.011	0.027	0.028	0.953

Note: SD represents the sample standard deviation, SE represents the mean of the estimated standard errors and CP represents the empirical 95% coverage probability.

Table 4: Simulation results for log-normally distributed random effects, where $\alpha = \log(3) = 1.099$, $\beta = -\log(2) = -0.693$, $n_k = 100$ and $\sigma_T = 1.2$

K	Methods	$\hat{\sigma}_T$	$\hat{\alpha}$				$\hat{\beta}$			
			Bias	SD	SE	CP	Bias	SD	SE	CP
30	PCLE	1.145	-0.078	0.212	0.217	0.940	-0.005	0.052	0.052	0.955
	Two step	1.454	-0.005	0.278	0.275	0.959	-0.003	0.055	0.052	0.953
	Naive	0.718	-0.193	0.180	0.148	0.696	0.009	0.064	0.063	0.944
	ZCC	0.836	-0.168	0.181	0.158	0.774	0.009	0.050	0.051	0.955
50	PCLE	1.169	-0.064	0.180	0.171	0.937	-0.004	0.039	0.040	0.947
	Two step	1.027	-0.004	0.221	0.217	0.960	-0.002	0.041	0.040	0.955
	Naive	0.743	-0.185	0.153	0.117	0.585	0.014	0.047	0.049	0.947
	ZCC	0.857	-0.156	0.156	0.126	0.685	0.010	0.039	0.040	0.948
100	PCLE	1.163	-0.077	0.120	0.120	0.894	-0.003	0.027	0.028	0.960
	Two step	1.322	-0.085	0.110	0.136	0.944	-0.001	0.029	0.028	0.953
	Naive	0.736	-0.199	0.104	0.082	0.353	0.016	0.033	0.036	0.935
	ZCC	0.842	-0.169	0.104	0.088	0.520	0.011	0.027	0.028	0.950

Note: SD represents the sample standard deviation, SE represents the mean of the estimated standard errors and CP represents the empirical 95% coverage probability.

Table 5: Simulation results for correlated random effects, where $\alpha = \log(3) = 1.099$, $\beta = \log(2) = 0.693$ and $n_k = 100$

K	σ_T	Methods	$\hat{\sigma}_T$	$\hat{\alpha}$				$\hat{\beta}$			
				Bias	SD	SE	CP	Bias	SD	SE	CP
30	0.5	PCLE	0.688	0.012	0.134	0.136	0.959	0.015	0.059	0.057	0.920
		Two step	0.673	0.003	0.143	0.133	0.951	0.004	0.058	0.057	0.946
		Naive	0.215	-0.038	0.113	0.083	0.823	0.219	0.120	0.054	0.208
		ZCC	0.457	-0.026	0.117	0.099	0.923	0.090	0.086	0.055	0.676
	1	PCLE	1.116	0.012	0.214	0.215	0.948	0.020	0.063	0.064	0.938
		Two step	1.081	-0.015	0.219	0.208	0.929	0.003	0.065	0.061	0.929
		Naive	0.824	-0.028	0.179	0.163	0.919	0.152	0.105	0.057	0.375
		ZCC	0.825	-0.054	0.172	0.161	0.925	0.097	0.083	0.063	0.664
	1.5	PCLE	1.545	-0.005	0.297	0.300	0.959	0.031	0.071	0.076	0.939
		Two step	1.459	-0.006	0.295	0.281	0.937	0.005	0.071	0.067	0.937
		Naive	1.263	-0.026	0.250	0.237	0.940	0.147	0.088	0.059	0.388
		ZCC	1.068	-0.104	0.216	0.207	0.920	0.135	0.094	0.074	0.595
50	0.5	PCLE	0.703	0.023	0.103	0.108	0.957	0.008	0.044	0.044	0.950
		Two step	0.687	0.022	0.100	0.105	0.951	-0.002	0.048	0.044	0.906
		Naive	0.210	-0.045	0.088	0.063	0.810	0.226	0.112	0.042	0.108
		ZCC	0.468	-0.030	0.091	0.077	0.898	0.082	0.065	0.043	0.561
	1	PCLE	1.146	0.038	0.168	0.170	0.950	0.015	0.047	0.049	0.954
		Two step	1.127	0.031	0.165	0.168	0.949	0.264	0.158	0.055	0.139
		Naive	0.831	-0.029	0.136	0.127	0.920	0.146	0.076	0.045	0.185
		ZCC	0.833	-0.052	0.138	0.126	0.915	0.091	0.059	0.048	0.523
	1.5	PCLE	1.598	0.036	0.242	0.240	0.944	0.021	0.053	0.059	0.959
		Two step	1.541	0.016	0.239	0.231	0.931	0.000	0.055	0.052	0.946
		Naive	1.255	-0.026	0.190	0.183	0.933	0.148	0.067	0.045	0.175
		ZCC	1.073	-0.092	0.174	0.162	0.885	0.128	0.082	0.057	0.423
100	0.5	PCLE	0.710	0.018	0.074	0.077	0.956	0.008	0.030	0.031	0.950
		Two step	0.704	0.010	0.076	0.076	0.940	-0.003	0.032	0.031	0.926
		Naive	0.192	-0.046	0.070	0.043	0.673	0.242	0.106	0.029	0.010
		ZCC	0.488	-0.026	0.064	0.056	0.897	0.074	0.046	0.031	0.383
	1	PCLE	1.173	0.033	0.123	0.124	0.947	0.012	0.034	0.035	0.941
		Two step	1.139	0.013	0.124	0.120	0.943	-0.002	0.035	0.033	0.937
		Naive	0.846	-0.024	0.103	0.091	0.902	0.145	0.056	0.032	0.048
		ZCC	0.837	-0.049	0.097	0.090	0.905	0.089	0.044	0.034	0.335
	1.5	PCLE	1.648	0.036	0.171	0.175	0.960	0.017	0.037	0.041	0.945
		Two step	1.586	0.007	0.175	0.168	0.946	0.000	0.039	0.037	0.929
		Naive	1.262	-0.021	0.141	0.131	0.928	0.150	0.063	0.032	0.030
		ZCC	1.074	-0.095	0.123	0.115	0.850	0.126	0.058	0.041	0.223

Note: SD represents the sample standard deviation, SE represents the mean of the estimated standard errors and CP represents the empirical 95% coverage probability.

Table 6: Simulation results for correlated random effects, where $\alpha = \log(3) = 1.099$, $\beta = -\log(2) = -0.693$ and $n_k = 100$

K	σ_T	Methods	$\hat{\sigma}_T$	$\hat{\alpha}$				$\hat{\beta}$			
				Bias	SD	SE	CP	Bias	SD	SE	CP
30	0.5	PCLE	0.657	0.004	0.127	0.128	0.955	0.011	0.055	0.051	0.919
		Two step	0.670	0.006	0.135	0.130	0.946	-0.004	0.053	0.053	0.957
		Naive	0.062	-0.055	0.107	0.067	0.702	0.023	0.063	0.062	0.951
		ZCC	0.511	-0.023	0.113	0.103	0.929	0.098	0.070	0.048	0.513
	1	PCLE	1.011	-0.011	0.195	0.191	0.941	0.045	0.069	0.050	0.765
		Two step	1.081	-0.003	0.213	0.202	0.934	-0.005	0.056	0.053	0.931
		Naive	0.622	-0.135	0.168	0.132	0.751	0.013	0.065	0.063	0.945
		ZCC	0.950	-0.021	0.185	0.177	0.938	0.088	0.058	0.049	0.563
	1.5	PCLE	1.412	-0.021	0.270	0.262	0.950	0.069	0.076	0.050	0.639
		Two step	1.539	-0.013	0.302	0.285	0.940	-0.004	0.056	0.054	0.926
		Naive	0.734	-0.308	0.183	0.149	0.441	0.013	0.062	0.062	0.941
		ZCC	1.389	-0.022	0.263	0.254	0.943	0.090	0.058	0.049	0.568
50	0.5	PCLE	0.667	0.015	0.097	0.101	0.953	0.009	0.043	0.039	0.931
		Two step	0.683	0.017	0.094	0.102	0.954	-0.004	0.045	0.041	0.923
		Naive	0.046	-0.056	0.084	0.051	0.679	0.026	0.047	0.048	0.918
		ZCC	0.505	-0.022	0.090	0.079	0.925	0.095	0.053	0.037	0.334
	1	PCLE	1.032	0.011	0.151	0.150	0.946	0.038	0.054	0.039	0.766
		Two step	1.101	0.015	0.153	0.160	0.951	-0.004	0.044	0.041	0.931
		Naive	0.660	-0.124	0.139	0.106	0.701	0.016	0.047	0.049	0.951
		ZCC	0.945	-0.019	0.143	0.137	0.943	0.087	0.047	0.038	0.378
	1.5	PCLE	1.441	-0.017	0.211	0.207	0.942	0.060	0.063	0.039	0.636
		Two step	1.567	-0.013	0.219	0.224	0.940	-0.003	0.046	0.041	0.923
		Naive	0.769	-0.297	0.160	0.120	0.323	0.015	0.045	0.048	0.955
		ZCC	1.382	-0.018	0.200	0.197	0.943	0.090	0.047	0.038	0.388
100	0.5	PCLE	0.672	0.008	0.072	0.072	0.951	0.010	0.031	0.028	0.906
		Two step	0.681	0.010	0.075	0.072	0.940	0.000	0.031	0.029	0.917
		Naive	0.013	-0.061	0.055	0.034	0.510	0.027	0.035	0.034	0.876
		ZCC	0.522	-0.017	0.066	0.058	0.900	0.093	0.037	0.026	0.115
	1	PCLE	1.035	0.002	0.111	0.107	0.942	0.040	0.046	0.028	0.675
		Two step	1.095	0.013	0.116	0.112	0.943	0.000	0.031	0.029	0.920
		Naive	0.656	-0.137	0.096	0.074	0.559	0.017	0.033	0.035	0.930
		ZCC	0.952	-0.013	0.106	0.098	0.935	0.090	0.036	0.027	0.148
	1.5	PCLE	1.457	-0.001	0.153	0.148	0.949	0.055	0.047	0.028	0.529
		Two step	1.559	0.015	0.162	0.158	0.954	0.001	0.031	0.029	0.940
		Naive	0.771	-0.311	0.108	0.084	0.095	0.015	0.031	0.034	0.936
		ZCC	1.389	-0.012	0.149	0.141	0.933	0.091	0.037	0.027	0.150

Note: SD represents the sample standard deviation, SE represents the mean of the estimated standard errors and CP represents the empirical 95% coverage probability.

Table 7: Results for the analysis of ForestHealth data

Parameter	$\hat{\beta}^P$		$\hat{\beta}^T$		$\hat{\beta}^Z$		$\hat{\beta}^N$	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	−0.848*	0.243	−0.690*	0.269	−0.763*	0.191	−0.648*	0.092
PH	−0.783	0.436	−0.499	0.329	−0.775	0.797	−0.446	0.234
Age	0.024*	0.004	0.012*	0.004	0.021*	0.003	0.017*	0.002
Canopy	−3.979*	0.565	−4.898*	0.616	−3.366*	0.500	−2.317*	0.353
Humus	0.077	0.077	0.123	0.081	−0.048	0.076	−0.065	0.078
Stand	1.155*	0.476	1.644*	0.539	0.937*	0.384	0.459*	0.190
$\hat{\sigma}_T$	1.932 [†]		1.184 [†]		1.509 [†]		0.000	

Note: $\hat{\beta}^P$ means the estimators by using the penalized conditional likelihood method, $\hat{\beta}^T$ means the estimators by using the two-step method, $\hat{\beta}^Z$ means the estimators by using Zhou et al.'s method, and $\hat{\beta}^N$ means the naive estimators; * indicates a significant fixed effect on the health status of beech trees at level 0.05; [†] indicates a significant variance component at level 0.01 based on the LRT.

(a) PCL method

(b) Two-step method

Figure 1 Plots of AUC over λ using the PCL and two-step methods for Forest Health data.

Figure 2 Plot of ROC curves corresponding to the largest AUC using the PCL and two-step methods for Forest Health data.