Fast two-stage estimator for clustered count data with overdispersion

ARTICLE HISTORY
Compiled May 17, 2019

Word count: 6338

ABSTRACT
Clustered count data are commonly analysed by the generalized linear mixed model (GLMM). Here, the correlation due to clustering and some overdispersion is captured by the inclusion of cluster-specific normally distributed random effects. In some cases, the model does not capture the variability completely. Therefore, the GLMM can be extended by including a set of gamma random effects. Routinely, the GLMM is fitted by maximising the marginal likelihood. However, the whole maximisation process is computationally intensive. Although feasible with medium to large data, it can be too time-consuming or computationally intractable with very large data (overall sample and/or cluster size). Therefore, a less computationally intensive two-stage estimator for correlated, overdispersed count data is proposed. It is rooted in the pseudo-likelihood split-sample methodology. Based on a simulation study, it shows good statistical properties. Furthermore, it is computationally much faster than the full maximum likelihood estimator. The approach is illustrated using a large dataset belonging to a network of Belgian general practices.

KEYWORDS
Generalized linear mixed model; Hierarchical data; Negative binomial model; Poisson model; Random effects

1. Introduction

The analysis of count data, also referred to Poisson data, is commonly encountered in many fields, e.g., in a medical study, one may be interested in the weekly number of seizures of epileptic patients. Furthermore, the observations may be collected from structured units, e.g., repeated measures of the same individual or patients nested in hospitals, leading to clustered data. The Poisson as distribution belongs to the exponential family, the analysis of clustered count data is frequently done using generalized linear mixed models (GLMM; [1]; [2]), which is a direct extension of the generalized linear model (GLM; [3]) and the linear mixed model (LMM; [4]; [5]). In the GLMM framework, we assume that conditionally on the normally distributed subject-specific random effects, the outcomes are independent and their distribution belongs to the exponential family. The main idea of including these random effects is to address correlation and some variability due to clustering. Nevertheless, in practice, the model can be too restrictive and may not completely capture the variability.

In the GLM framework, the variance is a deterministic function of the mean. In particular, for the Poisson model, the variance is equal to the mean. However, the variance of observed count data is often larger (overdispersion) and occasionally smaller than the mean. One approach to accommodate overdispersion is to include gamma distributed random effects, leading to the negative-binomial model [6]. Further, the

GLMM for count data can be extended by combining normal and gamma random effects to account for association and overdispersion simultaneously [7].

A GLMM is commonly fit by maximising the marginal likelihood. With Gaussian outcomes, both the conditional and marginal distribution are multivariate normal, simplifying the whole maximisation process. However, in the non-Gaussian case, the derivation of the marginal joint distribution can be complicated, or even not possible in analytical form, although some progress was made by [8], among others. Therefore, marginalisation is routinely done numerically, at the cost of requiring more computing resources. Of course, full likelihood estimation is still computationally tractable with medium to large data. However, when the number of clusters and/or cluster sizes become very large, the fitting process can be too time-consuming or even computationally infeasible.

To facilitate the estimation procedure with large datasets, [9] proposed a pseudo-likelihood-based split-sample methodology. In this approach, the sample is partitioned into K sub-samples, which are analysed separately and afterwards the results are combined to obtain overall inferences. Depending on the model and the data size, the sub-samples can be independent or dependent. The method is not only fast, but it has also exhibited high efficiency with different clustering settings and types of outcomes [10–12].

Based on complete sufficient statistics, findings by [13] suggest that a convenient way to split the sample is by selecting balanced clusters in each sub-sample, i.e., equally distributed clusters. Although feasible in many situations, it is difficult to achieve with very unbalanced clusters, e.g., in meta-analysis or longitudinal studies. In the most extreme case, each sub-sample contains only a single cluster, leading to the so-called cluster-by-cluster (CbC) estimator. Nevertheless, the covariance matrix of the random effects cannot be estimated using a single cluster. Consequently, an estimator based on the cluster-specific estimates is needed. In the LMM, the CbC estimator is unbiased, closed-form, and therefore computationally fast. Furthermore, it is efficient when cluster sizes and the number of clusters grow large at appropriate rates [14]. One interesting finding is that the CbC estimator is equivalent to the restricted maximum likelihood (REML) estimator when analysing balanced clusters.

In the present paper, we introduce the cluster-by-cluster estimator to the GLMM to clustered count data. It is motivated by [8,15]'s finding on marginalised GLMM's. In particular, the Poisson model allows explicit expressions for the marginal joint distribution, including marginal means, variances, and covariances. The CbC estimator requires two stages. At first, a GLM is fitted to each cluster separately. Next, a global estimate for the fixed effects and overdispersion parameter (if needed) is computed using weighted averages. Given the mean-variance relationship and the complex marginal joint distribution in the Poisson model, the estimator of the covariance matrix of the random effects differs from the one in the LMM. Particularly, does not allow for an analytical expression and relies on approximations. Note that, contrary to the implementation in the LMM, the CbC for the GLMM is no longer closed-form even from the first step. Nevertheless, it reduces the computation time considerably compared to the full MLE.

The paper is organised as follows. Section 2 presents a motivating case study based on a large database belonging to a network of Belgian general practices. In Sections 3 and 4, the generalized linear mixed model, with focus on models for count data, and the split-sample method are briefly described. The CbC estimator for count data, with and without overdispersion, is proposed in Section 5. In Section 6, the simulation study to evaluate the cluster-by-cluster estimator is presented. The case study is analysed

in Section 7. Finally, Section 8 is reserved for concluding remarks.

2. Intego dataset

These data come from the Intego database, a Belgian general practice-based morbidity registration network at the Department of General Practice of the University of Leuven. It consists of a continuous recording of patient information, diagnoses, drug prescriptions, laboratory results and vaccinations from general practitioners evenly spread throughout Flanders, Belgium [16]. We consider the Intego database from 2011. In total, the sample contains information about 151,971 patients from 65 practices. Particularly, patients with diagnosed chronic diseases, such as cancer, diabetes and heart diseases, are considered reducing the sample size to 54,967. We are interested in evaluating the effect of age, gender, body mass index (BMI), diabetes, cholesterol (mean value per year), and systolic blood pressure on the number of additional diagnosed chronic diseases suffered by patients, as they might be risk factors. Although the covariates age and gender are fully observed, a large amount of missing data is encountered in the other covariates. The percentage of missingness for these range between 55% and 78%.

To handle the missing covariates, we implemented a multiple imputation (MI) procedure. The MI framework embraces an extensive collection of techniques to deal with missing values. Nevertheless, all these techniques follow the same three phases. Firstly, each missing value is replaced by a set of M plausible values. Later, the multiply imputed datasets are analysed by using the standard method for complete data. Finally, the set of estimates from these analyses are combined to obtained overall estimates and their standard error. More details on MI can be found in [17], [19], [20], among others.

3. Generalized linear mixed model

Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ be the n_i-dimensional vector of measurements of cluster i, with $i = 1, \ldots, N$. The GLMM assumes that, conditionally on a q-dimensional vector of random effects $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{D})$, **where D is an unstructured covariance matrix**, the elements of $\boldsymbol{Y}_i$ are independent and follow a distribution that belongs to the exponential family, that is:

$$f\left(y_{ij}|\boldsymbol{b}_i\right) = \exp\left\{\phi^{-1}\left[y_{ij}\theta_{ij} - \psi\left(\theta_{ij}\right)\right] + c\left(y_{ij}, \phi\right)\right\}, \tag{1}$$

where θ_{ij} and ϕ are called natural and scale parameter, respectively; $\psi(\cdot)$ and $c(\cdot, \cdot)$ are known functions. Here, the conditional mean vector $\boldsymbol{\mu}_i^c$ is modeled by a known link function, $\eta\left(\boldsymbol{\mu}_i^c\right) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$, where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are $(n_i \times p)$ and $(n_i \times q)$ matrices of known covariates, and $\boldsymbol{\beta}$ is an unknown p-dimensional vector of fixed-effects coefficients. Furthermore, the conditional covariance matrix $\boldsymbol{V}_i^c$ is a diagonal matrix with the $(j,j)^{\text{th}}$ element equals to $v_{i,jj}^c = \phi\nu\left(\mu_{ij}^c\right)$, where $\nu(\cdot)$ is a function that describes the mean-variance relationship.

Although (1) is expressed hierarchically, the GLMM is commonly fitted through

maximising the marginal likelihood:

$$L(\boldsymbol{\beta}, \boldsymbol{D}, \phi) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}, \phi) f(\boldsymbol{b}_i|\boldsymbol{D}) d\boldsymbol{b}_i, \tag{2}$$

As it can be seen from (2), maximisation of the likelihood involves N integrals over $\boldsymbol{b}_i$. Except for some special cases, e.g., linear mixed models, there is no easy analytic solution for these integrals, and a numerical approximation is required [8].

There are several approaches to solve these intractable integrals, most of them based on an approximation of the integrand (Laplace's method), linearization of the data (Penalized quasi-likelihood-PQL or marginal quasi-likelihood-MQL), or an approximation of the integral (using Gaussian or adaptive Gaussian quadrature) [2]. The estimation based on the latter performs better, but it is computationally more intensive. The Laplace method, PQL, and MQL perform poorly when the number of measurements per cluster is small, and the outcomes are far from a normally distributed, e.g., binary data [21]. More details can be found in [2, chap. 14].

3.1. Poisson-Normal model

Assuming that, conditionally on the random effects, Y_{ij} follows a Poisson distribution, we obtain the Poisson-Normal (PN) model with log link function:

$$Y_{ij}|\boldsymbol{b}_i \sim \text{Poisson}\left(\lambda_{ij}\right), \text{ where } \lambda_{ij} = \exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i\right), \tag{3}$$

where $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ are the jth row of $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$, respectively. This means that, conditionally on $\boldsymbol{b}_i$, the elements of $\boldsymbol{Y}_i$ are independent with expected value and variance equal to λ_{ij}.

[8,15] derived an analytical expression for the marginal distribution of $\boldsymbol{Y}_i$. Particularly, the marginal mean and variance are:

$$E\left(\boldsymbol{Y}_i\right) = \exp\left[\boldsymbol{X}_i\boldsymbol{\beta} + \frac{1}{2}\text{diag}\left(\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}'_i\right)\right],$$

and

$$V\left(\boldsymbol{Y}_i\right) = \boldsymbol{M}_i + \boldsymbol{M}_i\left[\exp\left(\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}'_i\right) - \boldsymbol{J}_{n_i}\right]\boldsymbol{M}_i,$$

respectively, where $\boldsymbol{J}_{n_i}$ is a $(n_i \times n_i)$ matrix of ones, and $\boldsymbol{M}_i$ is a diagonal matrix with the vector $E\left(\boldsymbol{Y}_i\right)$ along the diagonal.

3.2. Poisson-Normal-Gamma model

The traditional GLMM can be extended to address for overdispersion by including gamma distributed random effects. Hence, combining ideas of overdispersion models [6] and PN model (3), we obtain the Poisson-Normal-Gamma (PNG) model:

$$Y_{ij}|\boldsymbol{b}_i, \vartheta_{ij} \sim \text{Poisson}\left(\vartheta_{ij}\lambda_{ij}\right), \text{ where } \lambda_{ij} = \exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i\right), \tag{4}$$

where ϑ_{ij} follows a gamma distribution with $E(\vartheta_{ij}) = 1$, $V(\vartheta_{ij}) = \alpha$ and, $\text{Cov}(\vartheta_{ij}, \vartheta_{ik}) = 0$, for all $j \neq k$. In this model, $\boldsymbol{b}_i$ induces association between observations of the same cluster, and ϑ_{ij} capturing additional overdispersion.

4. Split-sample methodology for clustered data

In this approach, the sample is divided, according to some appropriate rule, into K sub-samples, with c_k independent clusters in sub-sample k, for $k = 1, \ldots, K$. After partitioning, the method is implemented in two stages:

(1) Estimate $\boldsymbol{\theta}$ in each of the K sub-samples. Let us denote these estimates as $\hat{\boldsymbol{\theta}}_k$, with their corresponding variance $V\left(\hat{\boldsymbol{\theta}}_k\right)$;

(2) Calculate a weighted average of $\hat{\boldsymbol{\theta}}_k$, to obtain an overall estimate:

$$\tilde{\boldsymbol{\theta}} = \sum_{k=1}^{K} \boldsymbol{A}_k \hat{\boldsymbol{\theta}}_k, \text{ with } V\left(\tilde{\boldsymbol{\theta}}\right) = \sum_{k=1}^{K} \boldsymbol{A}_k V\left(\hat{\boldsymbol{\theta}}_k\right) \boldsymbol{A}_k',$$

where $\boldsymbol{A}_k$ is a weighting matrix for sub-sample k. There are several ways to determine the weights, but the $\sum_{k=1}^{K} \boldsymbol{A}_k = \boldsymbol{I}$ constraint is needed to retain asymptotic unbiasedness.

The most convenient split is that each sub-sample consists of balanced clusters; i.e., clusters with the same distribution for $\boldsymbol{Y}_i$. In the LMM, it facilitates the estimation process, because with balanced clusters there are complete sufficient statistics and a closed-form MLE exists [13]. For more details on the split-sample methodology, we refer to [12].

5. Cluster-by-cluster estimator

The cluster-by-cluster estimator is referred to as the split-sample method restricted to the most extreme partitioning: a single cluster per stratum, i.e., $c_k = 1$ for $k = 1, \ldots, N$. Nevertheless, it follows the same two steps presented in Section 4. Considering model (1), and with enough information per cluster, an individual analysis of each cluster allows estimating the fixed effects ($\boldsymbol{\beta}$) and overdispersion parameter (α), but not the variance of the random effects ($\boldsymbol{D}$). Its estimation requires information on more than one cluster. Therefore, we propose a method-of-moments estimator based on the cluster-specific estimates, i.e., $\hat{\boldsymbol{\theta}}_i = \hat{\boldsymbol{\beta}}_i$ for the PN model and $\hat{\boldsymbol{\theta}}_i = (\hat{\boldsymbol{\beta}}_i', \hat{\alpha}_i)'$ for the PNG model.

To describe the estimator, $\eta\left(\boldsymbol{\mu}_i^c\right)$ is re-expressed as:

$$\eta\left(\boldsymbol{\mu}_i^c\right) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i = \boldsymbol{T}_i\left[\boldsymbol{K}_i\boldsymbol{\beta} + \begin{pmatrix}\boldsymbol{b}_i \\ \boldsymbol{0}_r\end{pmatrix}\right] = \boldsymbol{T}_i\boldsymbol{\beta}_i, \tag{5}$$

where $\boldsymbol{T}_i = (\boldsymbol{Z}_i\ \boldsymbol{Z}_{ci})$, $\boldsymbol{Z}_{ci}$ is a $(n_i \times r)$ matrix of within-cluster covariates not associated with random effects, and $\boldsymbol{K}_i$ is a $(q + r \times p)$ matrix of known cluster-specific covariates satisfying $\boldsymbol{X}_i = \boldsymbol{T}_i\boldsymbol{K}_i$. **The augmented vector of random effects ($\boldsymbol{b}_i$) with $\boldsymbol{0}_r$ in (5) implies that some, but not all, cluster-specific parameters are necessarily associated with random effects.**

To facilitate the derivation of the estimator, we assume that $\boldsymbol{T}_i = \boldsymbol{Z}_i$. Therefore, for now on $\eta\left(\boldsymbol{\mu}_i^c\right) = \boldsymbol{Z}_i\boldsymbol{\beta}_i$, with $\boldsymbol{\beta}_i = \boldsymbol{K}_i\boldsymbol{\beta} + \boldsymbol{b}_i$. Nevertheless, the general expression requires some further but straightforward algebra.

5.1. First stage

Conditionally on the random effects, the measurements of the same cluster are independent and follow a distribution that belongs to the exponential family, with $\eta\left(\boldsymbol{\mu}_i^c\right) = \boldsymbol{Z}_i\boldsymbol{\beta}_i$. Hence, we can fit a GLM within each cluster. Of course, this requires that all covariates in $\boldsymbol{Z}_i$ change within clusters, allowing estimation of $\boldsymbol{\beta}_i$. Therefore, at the first stage, we use the iteratively re-weighted least squares (IRLS) estimator to obtain a set of estimates $(\hat{\boldsymbol{\beta}}_i, i = 1, \ldots, N)$. So, the asymptotic conditional expectation and variance of $\hat{\boldsymbol{\beta}}_i$ are:

$$E(\hat{\boldsymbol{\beta}}_i|\boldsymbol{b}_i) = \boldsymbol{K}_i\boldsymbol{\beta} + \boldsymbol{b}_i \text{ and } V\left(\hat{\boldsymbol{\beta}}_i|\boldsymbol{b}_i\right) = \left(\boldsymbol{Z}_i'\boldsymbol{W}_i\boldsymbol{Z}_i\right)^{-1},$$

respectively; where $\boldsymbol{W}_i$ is a diagonal matrix with the $(j,j)^{\text{th}}$ element equal to $w_{i,jj} = \left(\frac{\partial\mu_{ij}^c}{\partial\eta_{ij}}\right)^2/\nu_{ij}^c$. Note that $\boldsymbol{W}_i$ depends on the conditional distribution of $\boldsymbol{Y}_i$ in Model (1).

Asymptotically, the marginal mean and covariance matrix of $\hat{\boldsymbol{\beta}}_i$ are $E\left(\hat{\boldsymbol{\beta}}_i\right) = \boldsymbol{K}_i\boldsymbol{\beta}$ and

$$V\left(\hat{\boldsymbol{\beta}}_i\right) = E\left[V\left(\hat{\boldsymbol{\beta}}_i|\boldsymbol{b}_i\right)\right] + V\left[E\left(\hat{\boldsymbol{\beta}}_i|\boldsymbol{b}_i\right)\right] = E\left[(\boldsymbol{Z}_i'\boldsymbol{W}_i\boldsymbol{Z}_i)^{-1}|\boldsymbol{b}_i\right] + \boldsymbol{D},$$

respectively. Generally, there is no closed-form expression for $E\left[(\boldsymbol{Z}_i'\boldsymbol{W}_i\boldsymbol{Z}_i)^{-1}|\boldsymbol{b}_i\right]$. Nevertheless, it can be approximated using the (first-order) delta method. So, the marginal variance is approximately equal to

$$V\left(\hat{\boldsymbol{\beta}}_i\right) \approx \left[\boldsymbol{Z}_i'E\left(\boldsymbol{W}_i|\boldsymbol{b}_i\right)\boldsymbol{Z}_i\right]^{-1} + \boldsymbol{D}. \tag{6}$$

Poisson-Normal model

In the PN model (3), we have that $w_{i,jj} = \lambda_{ij}$. Then,

$$V\left(\hat{\boldsymbol{\beta}}_i\right) \approx \left(\boldsymbol{Z}_i'\boldsymbol{M}_i\boldsymbol{Z}_i\right)^{-1} + \boldsymbol{D}. \tag{7}$$

In the case of a single random effect b_i, i.e., random intercept PN model, there is an analytic expression for $V\left(\hat{\boldsymbol{\beta}}_i\right)$ leading to an unbiased estimator of the variance of b_i (see Section B of the Supplemental Materials).

Poisson-Normal-Gamma model
In the Poisson-Normal-Gamma model (4), we have that:

$$w_{i,jj} = \frac{\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{b}_i\right)}{1 + \alpha\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{b}_i\right)}. \tag{8}$$

There is no analytic expression for $E(w_{i,jj})$, but it can be approximated using a Taylor series expansion around $\boldsymbol{b}_i = \mathbf{0}$:

$$E\left(w_{i,jj}\right) \approx h\left(\boldsymbol{x}'_{ij}\boldsymbol{\beta}\right) + \frac{1}{2}h''\left(\boldsymbol{x}'_{ij}\boldsymbol{\beta}\right)\boldsymbol{z}'_{ij}\boldsymbol{D}\boldsymbol{z}_{ij}, \tag{9}$$

where $h\left(\boldsymbol{x}'_{ij}\boldsymbol{\beta}\right)$ and $h''\left(\boldsymbol{x}'_{ij}\boldsymbol{\beta}\right)$ refers to function (8) and its second derivative with respect to $\boldsymbol{b}_i$, both evaluated at $\boldsymbol{b}_i = \mathbf{0}$, respectively. Therefore, the approximation is:

$$V\left(\hat{\boldsymbol{\beta}}_i\right) \approx \left(\boldsymbol{Z}'_i\boldsymbol{G}_i\boldsymbol{Z}_i\right)^{-1} + \boldsymbol{D}, \tag{10}$$

where $\boldsymbol{G}_i$ is a diagonal matrix with the $(j,j)^{\text{th}}$ element equal to (9).

A wide range of methods are available in the literature, such as likelihood- or moments-based, to estimate α in each cluster. For a detailed description of them, see [6] and, [22]. The marginal variance of $\tilde{\alpha}$ is:

$$V\left(\hat{\alpha}_i\right) = E\left[V\left(\hat{\alpha}_i|\boldsymbol{b}_i\right)\right], \tag{11}$$

which depends on the actual estimator used. Generally, there is no closed-form expression for the expected value (11). As before, Taylor series expansions and delta methods can be implemented to find an approximation.

5.2. Second stage

A weighted average of the sets of estimates $(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_N)$ and $(\hat{\alpha}_1, \ldots, \hat{\alpha}_N)$ is computed to obtain global estimates:

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{k=1}^{N}\boldsymbol{K}'_k\boldsymbol{A}_k\boldsymbol{K}_k\right)^{-1}\sum_{i=1}^{N}\boldsymbol{K}'_i\boldsymbol{A}_i\hat{\boldsymbol{\beta}}_i \text{ and } \tilde{\alpha} = \sum_{i=1}^{N}a_i\hat{\alpha}_i.$$

Furthermore, the variances of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\alpha}$ are

$$V\left(\tilde{\boldsymbol{\beta}}\right) = \left(\sum_{k=1}^{N}\boldsymbol{K}'_k\boldsymbol{A}_k\boldsymbol{K}_k\right)^{-1}\left[\sum_{i=1}^{N}\boldsymbol{K}'_i\boldsymbol{A}_iV\left(\hat{\boldsymbol{\beta}}_i\right)\boldsymbol{A}'_i\boldsymbol{K}_i\right]\left(\sum_{k=1}^{N}\boldsymbol{K}'_k\boldsymbol{A}_k\boldsymbol{K}_k\right)^{-1}$$

and

$$V\left(\tilde{\alpha}\right) = \sum_{i=1}^{N}a_i^2V\left(\hat{\alpha}_i\right),$$

respectively.

Estimator of D

The estimator of $\boldsymbol{D}$ is based on the sum of the cross-product of the difference between the cluster-specific estimates ($\hat{\boldsymbol{\beta}}_i$) and the global estimate ($\tilde{\boldsymbol{\beta}}$):

$$\boldsymbol{S}_b = \sum_{i=1}^{N} \left(\hat{\boldsymbol{\beta}}_i - \boldsymbol{K}_i\tilde{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}_i - \boldsymbol{K}_i\tilde{\boldsymbol{\beta}}\right)' = \sum_{i=1}^{N} \tilde{\boldsymbol{b}}_i\tilde{\boldsymbol{b}}_i'.$$

A method-of-moments estimator is found by equating $\boldsymbol{S}_b$ to its expected value and solving for $\boldsymbol{D}$. Since $E\left(\tilde{\boldsymbol{b}}_i\right) \approx \boldsymbol{0}$, the expected value of $\boldsymbol{S}_b$ is approximately:

$$E\left(\boldsymbol{S}_b\right) \approx \sum_{i=1}^{N} V\left(\tilde{\boldsymbol{b}}_i\right) = \sum_{i=1}^{N} (\boldsymbol{I} - \boldsymbol{H}_{ii}) V\left(\hat{\boldsymbol{\beta}}_i\right) (\boldsymbol{I} - \boldsymbol{H}_{ii})' + \sum_{k\neq i} \boldsymbol{H}_{ik} V\left(\hat{\boldsymbol{\beta}}_i\right) \boldsymbol{H}_{ik}', \quad (12)$$

where $\boldsymbol{H}_{ij} = \boldsymbol{K}_i \left(\sum_{k=1}^{N} \boldsymbol{K}_k'\boldsymbol{A}_k\boldsymbol{K}_k\right)^{-1} \boldsymbol{K}_j'\boldsymbol{A}_j$. For the PN and PNG model, $V\left(\hat{\boldsymbol{\beta}}_i\right)$ is equal to (7) and (10), respectively. For the latter, we plug $\tilde{\alpha}$ into expression (9).

Given that (12) is non-linear, an iterative procedure, e.g., Newton-Raphson, is needed to find the solution of $\boldsymbol{D}$. The approximation used in (7) or (10) leads to a biased estimator. Nevertheless, the bias goes to zero as $n_i \to \infty$. An expression for the variance of $\tilde{\boldsymbol{D}}$ can be found using the delta method (see Section A of the Supplemental Materials).

For the random intercept PN model, there is an analytical expression for the marginal variance of $\hat{\boldsymbol{\beta}}$. Therefore, unbiasedness for the variance of the random intercept can be reached (see Section B of the Supplemental Materials).

In $\tilde{\boldsymbol{D}}$, each set of estimates ($\hat{\boldsymbol{\beta}}_i$) contributes equally to the estimation. However, estimates based on relatively small clusters are less precise than the ones obtained from large clusters. Therefore, in the case of small and highly unbalanced clusters, weights can be added to $\boldsymbol{S}_b$ as follows:

$$\boldsymbol{S}_b = \sum_{i=1}^{N} w_i\tilde{\boldsymbol{b}}_i\tilde{\boldsymbol{b}}_i'.$$

Of course, the expected value (12) should be modified accordingly. A comparison of the estimator of $\boldsymbol{D}$ with proportional weights as well as unweighted is shown in Section C of the Supplemental Materials.

5.3. Weighting scheme

The most obvious choice is a constant weight, i.e., $\boldsymbol{A}_i = (1/N)\boldsymbol{I}$. When the clusters vary substantially by size, proportional weights, $\boldsymbol{A}_i = [n_i/(\sum_{k=1}^{N} n_k)]\boldsymbol{I}$, are more advisable. The advantages of these alternatives are their simplicity, and that they are parameter-free.

A more formal weighting scheme is the optimal one [12]. To estimate $\boldsymbol{\beta}$, these take

the form:

$$\boldsymbol{A}_i^{opt} = \left[\sum_{k=1}^{N} V\left(\hat{\boldsymbol{\beta}}_k\right)^{-1}\right]^{-1} V\left(\hat{\boldsymbol{\beta}}_i\right)^{-1}. \tag{13}$$

In our case, the main drawback is that $\boldsymbol{A}_i^{opt}$ depends on unknown parameters, as one can see from expression (6). Furthermore, this expression is an approximation of $V\left(\hat{\boldsymbol{\beta}}_i\right)$. To overcome the former, we can compute the optimal weights iteratively as follows:

(1) Estimate $\boldsymbol{\beta}_i$ and α in each cluster.
(2) Calculate $\tilde{\boldsymbol{\beta}}$, $\tilde{\alpha}$ and $\tilde{\boldsymbol{D}}$ using a simple weighting scheme, e.g., proportional weights.
(3) Using $\tilde{\alpha}$ and the current estimates, $\tilde{\boldsymbol{\beta}}^{(t)}$ and $\tilde{\boldsymbol{D}}^{(t)}$, calculate $V\left(\hat{\boldsymbol{\beta}}_i\right)^{(t+1)}$.
(4) Update $\tilde{\boldsymbol{\beta}}^{(t+1)}$ and $\tilde{\boldsymbol{D}}^{(t+1)}$ using optimal weights (13) and replace $V\left(\hat{\boldsymbol{\beta}}_i\right)$ by $V\left(\hat{\boldsymbol{\beta}}_i\right)^{(t+1)}$.
(5) Repeat steps 3 and 4 until convergence.

The iterative procedure involves only calculations based on the estimates; the data is used only once to yield $\hat{\boldsymbol{\beta}}_i$ and $\hat{\alpha}_i$ (in step 1). This is a convenient advantage in cases where the number of elements per cluster is large.

6. Simulation study

6.1. Setting

The simulation study recreates a longitudinal study for count data with an unbalanced number of measurements per individual. The data-generation model is:

$$Y_{ij}|\boldsymbol{b}_i \sim \text{Poisson}\left(\mu_{ij}\right), \tag{14}$$

where $\mu_{ij} = \theta_{ij}\exp\left(\beta_0 + \beta_1 T_i + \beta_2 t_{ij} + \beta_3 t_{ij} T_i + b_{0i} + b_{1i} t_{ij}\right)$; t_{ij} is measurement time j of individual i; T_i is the treatment administrated to individual i (0 for control and 1 for treatment) and $\boldsymbol{b}_i = (b_{0i}, b_{1i})' \sim N(\boldsymbol{0}, \boldsymbol{D})$. **In model (14), β_0 and β_2 represent the intercept and the slope for the control group, respectively. Moreover, β_1 and β_3 are the difference in intercept and slope for the treatment effect with respect to the control group, respectively. In a longitudinal setting, the primary interest lies in β_3, since it measures the treatment effect on the average growth.**

For the simulation, we set:

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)' = (1.5, -0.1, -0.5, -0.2)' \quad \text{and} \quad \boldsymbol{D} = \begin{pmatrix} 0.4 & -0.2 \\ -0.2 & 0.6 \end{pmatrix}. \tag{15}$$

For the PNG model, we assumed that θ_{ij} follows a gamma distribution with $\alpha = 0.5$. On the other hand, for the PN model, we fixed $\theta_{ij} = 1$.

The number of measurements per individual (n_i) was determined using a normal distribution, $n_i \sim N\left[\mu_n, (0.25\mu_n)^2\right]$ (rounded to the nearest integer), with a minimum of 10 observations. The measurement times of each individual (t_{ij}) range uniformly over the interval $(0, 1)$. Finally, the treatment allocation (T_i) follows a Bernoulli distribution with $p = 0.5$, restricting the minimum number of individuals per treatment arm to five.

We varied the number of individuals (N) and mean number of measurements per individual (μ_n), at first fixing $N = 50$ and varying $\mu_n = \{20, 50, 100, 150, 250\}$, and later, fixing $\mu_n = 50$ and increasing $N = \{20, 50, 100, 150, 250\}$. A total of 1,000 datasets were generated for each scenario. Afterwards, each simulated dataset was analysed using the following methods:

- The cluster-by-cluster estimator with proportional and iterated optimal weights.
- The MLE based on adaptive quadrature.

A comparison between both methodologies is undertaken via the relative efficiency (RE), i.e., the mean squared error (MSE) ratio of the cluster-by-cluster estimator over the MLE, and the relative bias (RB), separately for each parameter of model (14). Additionally, the coverage of the 95% confidence interval (CI) for the parameter associated with the treatment effect (β_3) is evaluated.

To evaluate the computational efficiency in large data, we increased N and μ_n as follows: first, fixing $N = 500$ and increasing μ_n ={500, 1000, 1500, 2000}; second, fixing $\mu_n = 500$ and varying N ={500, 1000, 1500, 2000}. Here, we simulated 25 datasets per scenario.

In Section 6.2, we present a comparison between the cluster-by-cluster estimator using iterated optimal weights and the MLE. The use of proportional weights performs somewhat worse than the iterated optimal ones; nevertheless, the difference is small (see Section C of the Supplemental Materials).

6.2. Results

Table 1 exhibits the RB and RE of the CbC estimator of the PN model. When the mean number of measurements per individual is bounded by 50 and the number of individuals increases (Table 1a), the estimator of the fixed effects is unbiased and the efficiency loss seems to be negligible. On the contrary, a constant but small positive bias is observed for each variance component. Regarding efficiency, the loss increases as the number of individuals gets larger. When the number of individuals is fixed at 50 and the mean number of measurements per individual increases (Table 1b), the estimator of the fixed effects remains unbiased and highly efficient. For the variance components, it is asymptotically unbiased and efficient. Its RE goes slowly to one; with $\mu_n = 250$, the MSE of the CbC estimator is only around 10% larger than the MSE of the full MLE.

The RB and RE of the CbC estimator of the PNG model are displayed in Table 2. It shows similar behaviour as before, unbiased and small efficiency loss for the fixed effects. Furthermore, the estimator of the variance components and the overdispersion parameter are asymptotically unbiased and efficient when the μ_n increase faster than N. However, the efficiency loss of $\tilde{\alpha}$ goes to one at a slower rate; with $\mu_n = 250$, its efficiency loss is around 25%. Meanwhile, it is roughly 6% for the variance components.

One interesting result is that, in most cases, no more than two iterations were needed to get the iterated optimal weights. Compared to proportional weights (see

Table C.1 of the Supplemental Materials), there is an improvement in the estimation of the fixed effects, but not for the variance components. Taking into account that the iterative procedure to estimate $\boldsymbol{D}$ can be demanding with a large number of clusters, a computationally efficient way to proceed is to estimate initially $\boldsymbol{\beta}$ and $\boldsymbol{D}$ using proportional weights, compute the optimal weights using $\tilde{\boldsymbol{D}}$, and thereafter, update the overall estimate of $\boldsymbol{\beta}$. We call this method approximate optimal weights.

Regarding coverage of the 95% confidence interval of β_3, the proportion of samples for which the parameter is contained in the confidence interval is around 0.95 in all scenarios for both estimators (see Table C.3 of the Supplemental Materials).

The median computation time of the CbC estimator, using approximate and iterated optimal weights, and the full MLE for the PN model is displayed in Table 3a. As expected, the CbC estimator is faster than the full MLE in all scenarios. Furthermore, the latter shows a steeper increase in computation time. With 2,000 clusters of mean size of 500, the CbC estimator can be more than 30 times faster. Regarding the weighting scheme, the use of approximate optimal weights is less demanding than the iterated optimal weights. As one observes in Table 3b, the PNG model is computationally more demanding. Nevertheless, the CbC estimator is still considerably faster than the full MLE in all settings.

The CbC estimator was implemented in R 3.5.1 [23] and the full MLE in SAS software 9.4 [24] using the nlmixed procedure. Both programs have been run on a laptop computer with a Intel(R) Core(TM) i5-6200U CPU 2.30GHz processor and 16GB of RAM memory.

7. Analysis of the Intego database

Defining Y_{ij} as the number of additional chronic diseases by patient i in practice j, a PNG model (4) with:

$$\ln \lambda_{ij} = \beta_{01} + b_i + \text{age}_{ij}\beta_1 + \text{gender}_{ij}\beta_2 + \text{BMI}_{ij}\beta_3 + \text{systolic}_{ij}\beta_4 + \text{cholesterol}_{ij}\beta_5 \quad (16)$$

where $b_i \sim N(0, d)$, is proposed. We also considered the PN model, assuming that $\vartheta_{ij} = 1$.

Before fitting the model, a MI procedure was performed to complete the covariates. We used fully conditional specification (FCS; [25]) implemented in the mice package in R [26]. After creating 20 multiply imputed datasets, we fitted model (16) using the CbC estimator to each one, and the estimates were combined using Rubin's rule [**?**]. Furthermore, all continuous covariates were centred to reduce collinearity and possible convergence issues. The clusters are highly unbalanced, their sizes range between 11 up to 4,240 patients, with a mean of 846. Therefore, we considered proportional weights to the estimator of d and α. Furthermore, iterated optimal weights are implemented for $\boldsymbol{\beta}$. For comparison, the same analysis was performed by MLE. Table 4 displays the estimates and standard error of the estimates of model (16) by both estimators.

For both models, the CbC and ML estimator provide similar estimates for the fixed effects and negative-binomial parameter (for the PNG model), with a slightly larger standard error for the former. For the variance of the random intercept, the ML estimate is somewhat smaller. Based on the Wald test, there is a significant effect of all covariates, excepted for cholesterol. Regarding gender, the number of chronic diseases in men is roughly 1.3 times higher than in women. Regarding computation time, fitting the CbC estimator for th PNG model for all the multiply imputed datasets

took roughly 6 minutes. On the contrary, the MLE was more than 15 times more time-consuming.

8. Final remarks

The so-called cluster-by-cluster (CbC) estimator has been proposed for hierarchical count data with and without overdispersion. Although the estimator is no longer closed-form, it is computationally less intensive than the standard MLE based on adaptive quadrature. Furthermore, it shows good statistical properties. For the fixed effects, it is unbiased and almost as efficient as the MLE. For the variance of the random effects and overdispersion parameter, it is biased. However, it is asymptotically efficient when the number of elements per cluster increases faster than the number of clusters. Particularly in the random intercept Poisson-Normal model, unbiasedness can be attained. Therefore, we suggest that the cluster-by-cluster estimator is an attractive alternative to fit a GLMM **for count data** with several large-size clusters. **Our findings are based on simulations in the context of the random-slope PN and PNG models, with unstructured covariance matrix. For a larger dimension of the random-effects vector, we expect a higher computational efficiency of our proposed method, with similar statistical properties for the fixed effects.** In the case of a large number of clusters, all processes in the first stage can be executed in parallel, reducing the computation time considerably.

Although the estimator still has attractive properties with medium cluster-sizes, its implementation can be problematic, especially during the first stage. With few observations or several zeros in a cluster, the IRLS algorithm may diverge or converge to a spurious solution, leading to unstable overall estimates. Therefore, we suggest performing a sensitivity analysis by excluding any problematic clusters and evaluating the overall estimates. Furthermore, the addition of weights in the estimator of $\boldsymbol{D}$ reduces the influence of small and unstable clusters.

Regarding the weighting scheme, iterated optimal weights lead to relatively more efficient estimates of $\boldsymbol{\beta}$ than proportional weights. However, its implementation is computationally more expensive. Hence, we recommend to estimate $\boldsymbol{D}$ using a simple weighting scheme for $\boldsymbol{\beta}$, and later, estimate $\boldsymbol{\beta}$ using approximated optimal weights based on the foregoing estimation of $\boldsymbol{D}$. In this way, the efficiency loss is negligible, and there is a large gain in computation time.

The implementation of the CbC estimator for a broader class of GLMM deserves further work and simulations. To address excess zeros in the outcome, the PNG can be extended by adding a zero-inflated component to the model [22,27]. Here, the CbC estimator proceeds in the same way. However, $V\left(\hat{\boldsymbol{\beta}}\right)$ has to be modified accordingly to estimate $\boldsymbol{D}$. Furthermore, a specific covariance structure for $\boldsymbol{D}$, e.g., compound-symmetry (CS) or autoregressive (AR), can be considered. Each case implies that a different system of equations needs to be solved to find an estimator of $\boldsymbol{D}$. Findings by [12] and [11] for normally distributed hierarchical data with CS and AR structure can be extended to non-Gaussian outcomes, but arguable will be more complicated. In most cases, we may rely on approximations to estimate $\boldsymbol{D}$. Therefore, a biased estimator is expected. However, we believe that it will be negligible as the cluster sizes increases. Most importantly, we expect that there may still be computational advantages of our

methodology over the full maximum likelihood estimator.

We have focused on the analysis of count data, for which an explicit expression of the marginal distribution is available. However, the cluster-by-cluster estimator can be considered for other types of non-Gaussian outcomes, such as binary and time-to-event data.

Disclosure statement

No potential conflict of interest was reported by the authors.

Supplemental materials

Expression for the precision of the estimator of D is presented in Appendix A. The cluster-by-cluster estimator for the random intercept model is introduced in Appendix B. Finally, additional results of the simulation study are shown in Appendix C.

References

[1] Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association. 1993;88(421):9–25.

[2] Molenberghs G, Verbeke G. Models for discrete longitudinal data. Springer, New-York; 2005.

[3] McCullagh P, Nelder JA. Generalized linear models. London, UK: Chapman & Hall / CRC; 1989.

[4] Laird N, Ware J. Random-effects models for longitudinal data. Biometrics. 1983; 38(4):963–974.

[5] Verbeke G, Molenberghs G. Linear mixed models for longitudinal data. New York: Springer; 2000.

[6] Lawless JF. Negative binomial and mixed poisson regression. Canadian Journal of Statistics. 1987;15(3):209–225.

[7] Booth JG, Casella G, Friedl H, et al. Negative binomial loglinear mixed models. Statistical Modelling. 2003;3(3):179–191.

[8] Molenberghs G, Verbeke G, Demétrio CGB, et al. A family of generalized linear models for repeated measures with normal and conjugate random effects. Statistical Science. 2010; 25(3):325–347.

[9] Molenberghs G, Verbeke G, Iddi S. Pseudo-likelihood methodology for partitioned large and complex samples. Statistics & Probability Letters. 2011;81(7):892–901.

[10] Ivanova A, Molenberghs G, Verbeke G. Fast and highly efficient pseudo-likelihood methodology for large and complex ordinal data. Statistical Methods in Medical Research. 2015; 26(6):2758–2779.

[11] Hermans L, Nassiri V, Molenberghs G, et al. Fast, closed-form, and efficient estimators for hierarchical models with AR(1) covariance and unequal cluster sizes. Communications in Statistics - Simulation and Computation. 2018;47(5):1492–1505.

[12] Molenberghs G, Hermans L, Nassiri V, et al. Clusters with random size: maximum likelihood versus weighted estimation. Statistica Sinica. 2018;28(3):1107–1132.

[13] Hermans L, Molenberghs G, Aerts M, et al. A tutorial on the practical use and implication of complete sufficient statistics. International Statistical Review. 2018;86(3):403–414.

[14] Flórez AJ, Molenberghs G, Verbeke G, et al. A closed-form estimator for meta-analysis and surrogate markers evaluation. Journal of Biopharmaceutical Statistics. 2019; 29(2):318–332.

[15] Molenberghs G, Verbeke G, Demétrio CGB. An extended random-effects approach to modeling repeated, overdispersed count data. Lifetime Data Analysis. 2007;13(4):513–531.
[16] Truyers C, Goderis G, Dewitte H, et al. The intego database: background, methods and basic results of a flemish general practice-based continuous morbidity registration project. BMC Medical Informatics and Decision Making. 2014;14(1):48–57.
[17] Rubin DB. Multiple imputation for nonresponse in surveys. Jhon Wiley & Sons; 1987.
[18] Little RJA, Rubin DB. Statistical analysis with missing data . Jhon Wiley & Sons; 2002.
[19] van Buuren S. Flexible imputation of missing data. Boca Ratón, FL: Chapman and Hall/CRC; 2012.
[20] Carpenter J, Kenward M. Multiple imputation and its application. John Wiley and Sons Ltd; 2013.
[21] Tuerlinckx F, Rijmen F, Verbeke G, et al. Statistical inference in generalized linear mixed models: A review. British Journal of Mathematical and Statistical Psychology. 2006; 59(2):225–255.
[22] Cameron AC, Trivedi P. Regression analysis of count data. Cambridge University Press,; 2013.
[23] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2018.
[24] SAS Institute. The SAS system for Windows Release 9.4. Cary, NC: SAS Institute; 2011.
[25] van Buuren S, Brand J, Groothuis-Oudshoorn C, et al. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation. 2006; 76(12):1049–1064.
[26] van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of Statistical Software. 2011;45(3):1–67.
[27] Ridout M, Demétrio CGB, Hinde J. Models for count data with many zeros. In: International Biometric Conference XIX; Cape Town, South Africa; 1998. p. 179–192.