Generalized Linear Mixed Models With Crossed Effects and Unit-specific Survey Weights

Jan Pablo Burgard\textsuperscript{a} and Patricia Dörr\textsuperscript{b}

\textsuperscript{a}University of Trier, FB IV—Economics, Trier, Germany; \textsuperscript{b}Universität Trier, Trier, Germany

ABSTRACT
Mixed models are frequently used in social and economic analysis. Typically, the analyzed data are gathered through a survey sample from the population of interest. The underlying survey design may be non-ignorable, leading to possible bias in the regression parameters if not accounted for. To circumvent this problem, survey weights should be included in the estimation process. We propose a survey weighted generalized linear mixed model allowing for unit-specific survey weights and a flexible random effects structure. An estimation procedure is proposed and evaluated in two simulation studies under different scenarios. The first simulation study strongly encourages the use of survey weighted generalized linear mixed models if the survey design is non-ignorable. The second one replicates a previous simulation study in order to study the competitiveness of the proposed method with established approaches and software. Further, the proposed method is used to re-estimate a regression on reading proficiency using US PISA data from 2000. Supplementary files for this article are available online.

ARTICLE HISTORY
Received November 2018
Revised August 2021

KEYWORDS
Complex designs; Generalized linear mixed models; Regression analysis; Survey weighting

1. Introduction
In the recent article, Lumley and Scott (2017) resumed the work of Pfeffermann (1993) and outline the importance of survey-weighting in regression analysis when the empirical research is done on complex survey data. If not accounted for, a non-ignorable sampling design could lead to erroneous inference (Pfeffermann 1993; Lumley and Scott 2017). Consequently, the authors advise incorporating survey weights in the estimation procedure. An overview of the use of survey weights in regression analysis, in general, is given in Pfeffermann (1993). The use of survey weights in the classical generalized linear regression context is widely understood (Kott 2018). However, the inclusion of survey weights in a (generalized) linear mixed model (GLMM) framework is a non-trivial task as Lumley and Scott (2017) noted:

All the design-based inference we discuss is for marginal models. Design-based inference even for linear mixed models is substantially more complicated. In the special case where the clusters in the model are the same as the clusters in the sampling design, there is literature on consistent estimation [...] and some implementations [...].

It is the more general case thereof that we analyze in this work.

The GLMM framework extends the classical generalized linear model by additional unobserved random effects structures. An exemplary application for GLMM is the sampling of pupils in classes in schools, where there are assumed unobserved random effects both on classes and teachers. The teachers give lessons in various classes leading to a crossed random effects structure (e.g., Leckie 2009). To the best of our knowledge, there is up to now no proposal on how to handle such a crossed effects structure considering survey weights. A study on reading proficiency of teenage U.S. students presented by Rabe-Hesketh and Skrondal (2006) will serve as an exemplary application. Besides illustrating the theoretical developments, it allows us to distinguish our approach from the established ones, too.

For the two-level LMM case, Pfeffermann et al. (1998) proposed to use inverse selection probabilities at each stage to reduce the bias due to non-ignorability of the survey design. For the case of nested GLMM, where the unobserved random effects structure is hierarchically ordered and corresponds to the sampling stages, Rabe-Hesketh and Skrondal (2006) propose a pseudo-likelihood estimation method accounting for weights in each sampling stage. However in practice, the data producers only provide a single survey weight for the dataset, and to reduce disclosure risk, often the sampling stages are not communicated. Alternative approaches are found in Rao, Verret, and Hidiroglou (2013) who proposed an estimating equation approach for the LMM scenario and the two-stage cluster sampling scheme only. In the same set-up, Oguz-Alper and Berger (2016) derived survey-weighted estimators through empirical maximum likelihood (EML). Yi, Rao, and Li (2016) elaborated a composite likelihood approach that is applicable to mixed models with the endogenous family belonging to the exponential family. Though Yi, Rao, and Li’s (2016) approach is already a valuable extension from linear mixed models to GLMMs, it is nonetheless necessary that the design mirrors the random effects structure of the model.
Our contribution is to allow for survey weighted GLMM estimation with a flexible random effects structure including both nested and crossed random effects. As already mentioned, the surveys only provide a single survey weight per unit in the sample. Therefore, in contrast to Rabe-Hesketh and Skrondal (2006), only one survey weight per unit must be sufficient in the presented method. We provide the underlying theory and propose an algorithm to fit such GLMMs. Further, we present simulation results supporting our findings.

Section 2 gives a brief description of the working example that will be used alongside the theoretical development. Section 3 introduces the estimation algorithm for GLMMs under complex survey designs. In this context, we describe the underlying theoretical requirements for the estimator’s consistency in Section 3.2. Section 4 describes the path from the original lying theoretical requirements for the estimator’s consistency to the presented algorithm. In Section 4.2, we introduce the statistical model to the presented algorithm. In Section 5, we describe the underpinnings of the complete 15-year-old population in the United States. Obviously, the ignorability depends on the chosen model.

Section and can be skipped by practitioners.

3. Weighted Generalized Linear Mixed Models

3.1. Estimation Algorithm

In the following, we assume as a data-generating process (DGP) a generalized linear mixed model (GLMM)

\[
\eta_j = x_j^T \beta + z_j^T G,
\]

\[
\mu_j = g(\eta_j),
\]

\[
Y_j \sim F(\mu_j, \varphi),
\]

\[
G \sim N(0, \Sigma),
\]

where \(x_j \in \mathbb{R}^p\) and \(z_j \in \mathbb{R}^q\) are nonrandom explanatory variables. The likelihood of the model depends on these explanatory variables, too, but for brevity, we omit them in the subsequent formulae. For a binary endogenous variable, this model was first introduced by Stratelli, Laird, and Ware (1984). In the systematic component (3.1), the linear prediction \(\eta_j\) is a function of \(x_j\) with fixed effect vector \(\beta\) and of \(z_j\) interacting with the multivariate Gaussian random effects \(G\) whose realization is denoted by \(\varphi\). The symmetric positive definite covariance matrix \(\Sigma\) is assumed to originate from at most \(q(q+1)/2\) disjoint parameters \(\sigma\), i.e. \(\Sigma = \Sigma(\sigma)\). A distribution from the exponential family with corresponding probability density \(f\), scale parameter \(\varphi\) and inverse link function \(g\). As the scaling parameter \(\varphi\) does not affect the estimation of \(\psi := (\beta^T, \sigma^T)^T\) through maximum likelihood (ML) estimation (Wedderburn 1976), we omit it in the following formulae for brevity too.

In our working example, \(Y_j\) would be a binary variable indicating whether student \(j\) achieved the top two reading proficiency levels defined in the PISA study: \(x_{ij}\) contains explainatories such as gender, indicators for the education level of the parents, indicators for foreign born parents and the international socio-economic index of student \(j\); \(z_{ij}\) is an indicator vector for the school that student \(j\) attends.

Before going into deeper detail of the estimation process, we first introduce survey sampling theory and our estimator of the population log-likelihood as it will be used in Algorithm 3.1. The development of the algorithm is postponed to the next section and can be skipped by practitioners.

Assume that there are \(N\) realizations of the DGP described in Equations (3.1) to (3.4). We denote the index set \(\{1, \ldots, N\} =: U\) as the finite population, for \(j \in U\), we observe realization \(y_j\) of \(Y_j\) and, for the complete finite population, \(\gamma\) of \(G\). The DGP is also named superpopulation model. In our working example, \(U\) is the set of all 15-year-old students in the United States. From \(U\), a survey sample \(s \subset U\) is drawn as a realization of the random Variable \(S\) that follows the random process \(P_S\), the survey design. Accordingly, we can define the inclusion probability of a unit
j, \( P_S(j \in S) =: \pi_j \). Let \( \delta_S \) denote the indicator function whether a unit is included in an arbitrary realization of the random variable \( S \). The inverse inclusion probability \( \pi_j^{-1} \) is the basic survey weight of unit \( j \) and we set \( w_j := \pi_j^{-1} \). The Horvitz-Thompson estimator (Horvitz and Thompson 1952) of the population log-likelihood is then

\[
\hat{L}(\mathbf{y}, \mathbf{y}_S, \mathbf{\psi}, S) = \sum_{j \in U} w_j \cdot \delta_S(j) \cdot \left( \log f(y_j | \mathbf{y}_S, \mathbf{\psi}) \right)
\]

\[
= A + \log f(\mathbf{y}_S | \mathbf{\sigma}), \quad (3.5)
\]

where \( \phi \) denotes the density of a multivariate Gaussian variable \( G_S \) centered at zero and with a covariance matrix of the same structure as \( \Sigma(\mathbf{\sigma}) \), \( \Sigma_S \). \( \Sigma_S \) has possibly a lower dimension than \( \Sigma \) if there is a population cluster from which no unit enters the sample. The differentiation between \( /Sigma_1 \) and \( /Sigma_2 \) represents a cluster in which students are gathered. Students for example in the PISA study, schools all over the United States are a cluster in which students are gathered. Students of these schools share a single random effect. However, not every school is sampled and for those that are not sampled, a prediction of the random effect is impossible. Hence, \( G_S \) is a subvector of \( G \).

When \( G_S \neq G \), the proportion between the terms \( A \) and \( B \) in (3.5) does not reflect the relationship in the finite population likelihood (that is introduced in Equation (4.1)). We can restablish the correct proportion by rescaling the weights \( w_j \), \( i \in U \), such that \( \sum_{i \in S} w_i = |S| =: n \) rather than \( \sum_{i \in S} w_i = N \). In the following, we assume that survey-weights are appropriately rescaled and for ease of notation, we write sometimes \( G \) and \( \mathbf{y} \) instead of their survey sample analogues \( G_S \) and \( \mathbf{y}_S \). As is pointed out in Section 3.2, we must assume \( G \perp S \) in order to assure consistency because the term \( B \) is not survey-weighted.

A common estimation method in the framework of mixed models is the EM algorithm (Laird and Ware 1982). Though, the analytical effort in the integration step is hardly manageable as the conditional distribution of \( G \) given the sample observations \( \mathbf{y} \) in the sample \( s \) is \( q \)-dimensional and possibly nonnormal. We suggest to replace the analytical integration by a Monte-Carlo approximation—a method that does not only account for random effect crossings (like in the formulation in McCulloch 1997; Booth and Hobert 1999) but also allows to respect the survey design \( P_S \). As the conditional distribution of the random effects given the observed data—necessary for the conditional expectation—depends on \( P_S \) and the exponential family \( F \), it is not pre-implemented in statistical software and therefore, again an MC-sampler is necessary in the integration step. One such MC-sampler suggested in Booth and Hobert (1999) is importance sampling where we choose to center the proposal around \( G \)’s current mode with mixture equal to the inverse negative Hessian of the log-likelihood.

We denote by \( \hat{H}_s \) the Hessian of (3.5) with respect to \( \mathbf{y}_S \) given a parameter vector \( \mathbf{\psi} \) and a sample realization \( S = s \). Then, the final estimation algorithm for GLMMs under a variable random effects structure and complex survey design is given in Algorithm 3.1. It is a Monte-Carlo EM (MCEM) algorithm as the E-step of the classical EM algorithm is, like already mentioned, replaced by Monte-Carlo integration. Furthermore, sampling from the required distribution is not generally implementable which is the reason why we employ additionally importance sampling in the MC-integration.

### Algorithm 3.1 MCEM-algorithm for GLMMs with Importance Sampling (wGLMM)

**Require:** \( K, B_1 \in \mathbb{N}, \beta, \sigma > 0, \) \( \text{iter} \gg 0 \)

\[
\text{while } k < K \text{ and } d > \varepsilon \text{ do} \\
\text{Calculate } \hat{y}_k \leftarrow \arg \max_{\mathbf{y} \in \mathbb{R}^q} \hat{L}(\mathbf{y}, \mathbf{y}_S, \mathbf{\psi}, S) \quad \text{Start EM-Loop} \\
\text{Calculate the Hessian } \hat{H}_{s,k} \text{ at } \hat{y}_k \\
\text{for } b = 1, \ldots, B_k \text{ do} \\
\text{Sample } G_b \sim N \left( \hat{y}_{k,b}, \hat{H}^{-1}_{s,k} (\hat{L}(\mathbf{y}, \mathbf{y}_S, \mathbf{\psi}, S)) \right) \\
\text{Sampling from Proposal} \\
\text{Denote by } \mathbf{y}_b \text{ the realization of } G_b \\
\text{for } b = 1, \ldots, B_k \text{ do} \\
\text{Choose } G_b \sim N \left( \hat{y}_{k,b}, \hat{H}^{-1}_{s,k} (\hat{L}(\mathbf{y}_b, \mathbf{y}_S, \mathbf{\psi}, S)) \right) \\
\text{End E-Step} \\
\text{End for} \\
\text{Calculate } w_{MC}^{MC} \leftarrow w_{imp}^{MC} \cdot \exp \left( \frac{1}{2} \text{det}(\hat{H}_{s,k}) \cdot \left( \frac{1}{2} (\mathbf{y}_b - \hat{y}_{k,b})^T \hat{H}_{s,k} (\mathbf{y}_b - \hat{y}_{k,b}) \right) \right) \\
\text{Normalize } w_{MC}^{MC} \leftarrow \frac{w_{imp}^{MC}}{\sum_{b=1}^{B_k} w_{MC}^{MC}} \\
\text{M-Step} \\
\text{Choose } G_{s,k} \leftarrow \arg \max_{\mathbf{y}_s} \sum_{b=1}^{B_k} w_{MC}^{MC} \cdot \hat{L}(\mathbf{y}_b, \mathbf{y}_S, \mathbf{\psi}, S) \\
\text{End M-Step} \\
\text{Ensure: } \mathbf{\psi}_k \leftarrow \hat{y}_k \\
\text{end for} \\
\text{end for}
\]

### 3.2. Theoretical Properties

The EM-algorithm is known to converge to a stationary point of the likelihood distribution (Dempster, Laird, and Rubin 1977) and a closer inspection of (3.5) shows: The log-likelihood is separable between the fixed effects parameters \( \beta \) (and \( \phi \)) in term A and the random effects parameters \( \sigma \) in term B. Furthermore, if no special structure is assumed for the covariance matrix \( \Sigma \) besides positiveness and symmetry, the log-density of the zero-centered normal distribution is concave in \( \mathbf{\sigma} \). Thus the maximizer \( \hat{\mathbf{\sigma}} \) is unique. Concerning \( \beta \), which only appends in term A, we have a unique stationary point for common link functions \( g^{-1} \) (Wedderburn 1976), too. In that case, the EM-algorithm converges consequently to the unique stationary point of the distribution which is thus the maximum.

However, this needs not be necessarily the case when using MC-integration in the E-step. A general overview about convergence properties of the MCEM-algorithm is given in Neath et al. (2013), including the case of GLMMs. There are several notions of convergence in the MCEM-setting and they depend on the number of Monte-Carlo draws in each E-step, on whether they are increasing and on the employed sampler. Generally, the number \( B \) of realized random draws must increase along the E-steps (as mentioned in the Algorithm 3.1), in order to ensure that the MC-error decreases when the estimates get closer to the
stationary point. By the law of large numbers (Owen 2013, chap 9) we find that

$$\sum_{b=1}^{B} w_{b}^{MC} \hat{L}L(y_{b}, \hat{\psi}, y, s) \xrightarrow{a.s.} E_{m}(\hat{L}L)(G)$$

(3.6)

in each E-step, where \( E_{m}(\hat{L}L) \) is the model expectation of the HT-estimator of the log-likelihood:

$$E_{m}(\hat{L}L)(G|Y = y, \hat{\psi}, S = s) = \int L(y_{s}, \hat{\psi}, s) \cdot h(y_{s}|y, s) \lambda(d y_{s}).$$

(3.7)

Neath et al. (2013) summarize conditions under which the solutions of the MCEM converge almost surely to a connected set of stationary points of the likelihood function, and as the likelihood is strictly concave in our set-up, we state that our estimator \( \hat{\psi} \) converges almost surely to the maximizer of the survey-weighted joint likelihood.

We write \( E_{m}(\hat{L}L) \) instead of \( E_{m}(\hat{L}L) \) in order to underline that the conditional density \( h(\cdot|y, s) \neq h(\cdot|y, U) \): As the conditional density given a sample \( s \) differs from that given the population \( U \), the estimator is not the expected value of \( E_{m}(\hat{L}L) = \int L \cdot h(\cdot|y, U) \) but a separate estimator \( E_{m}(\hat{L}L) = \int L \cdot h(\cdot|y, s) \) given \( s \). We find for the conditional density \( h \) (which is not Gaussian if \( Y \) is not normal)

$$h(y_{s}|y, s) = \frac{\exp(L(y_{s}, y, \psi, s))}{\int \exp(L(y_{s}, y, \psi, s)) P_{S}(S) d y} \propto \exp(L(y_{s}, y, \psi, s)).$$

(3.8)

When the design meets certain requirements that are discussed below, especially when \( S \perp G \), we claim that \( L(y_{s}, y, \psi, s) \xrightarrow{p} L(y, y, \psi, U) \). Therefore, also the exponential thereof converges (with weight rescaling, multiplied by a constant) in probability. We combine both findings to analyze the overall convergence.

Assume we have that the following asymptotic assumptions hold for the superpopulation and the survey design:

1. The finite population \( U \) is an element of a sequence of growing finite populations \( U_{1} \subset U_{2} \subset U_{k} \subset \ldots \) where \( |U_{k}| = N_{k} \to \infty \).
2. The number of domains for the random effects grows as \( k \to \infty \), that is, taking our setting, \( z_{j} \in \mathbb{R}^{q} \) (\( j \in U_{k} \)), we observe that \( q \to \infty \) as \( k \to \infty \).
3. \( \lim_{k \to \infty} E_{k}[S] \in [0, 1] \).
4. \( P_{S}(i | S) > 0 \) for all \( i \in U_{k} \) (\( k \in \mathbb{N} \)).
5. \( \lim_{k \to \infty} E_{k}[S] \max_{i \in U_{k}} |\tau_{i} - \mu_{i}| < \infty \).
6. \( \| P_{S}(S; Y, X, Z, G) - P_{S}(S; Y, X, Z) \| \to 0 \) for every sample with \( P_{S}(S = s) > 0 \).
7. The expected dimension of \( G \) goes to infinity, that is, the number of domains that are expected to enter with at least one unit the sample grows unboundedly as \( k \to \infty \).

These assumptions exclude a theoretical setting where for example the random effects represent clusters in the population and the design and only the number of secondary sampling units per cluster increases.

Conditions 3, 4, and 5 are the Conditions H1, H2, and H4 from Chauvet (2014). Together with the fact that the summands in the log-likelihood (3.5) take the role of the ‘variable of interest’ in Chauvet (2014), and the existence of the summands second moments (H3 in Chauvet (2014)), we can assure that the HT estimator for term A in (3.5) is consistent. Condition 2 and 6 assure the consistency of the ML variance estimator for a Gaussian covariance matrix. Condition 6 and 7 assure that the design is asymptotically independent of \( G \) which is important to assure that term \( B \) is a good—rescaled—estimator for \( \log(\phi(\gamma | \sigma) \) which in turn is only consistent when \( q \to \infty \).

Under these assumptions, we find that

$$P_{\lim_{n,N \to \infty}} \left( \lim_{B \to \infty} \sum_{b=1}^{B} w_{b}^{MC} \hat{L}L(y_{b}, \hat{\psi}, y, S) \right)$$

(3.9)

implying that we have only convergence in probability when the number of MC-samples for the integration increases within the estimation process as stated in McCulloch (1997); Booth and Hobert (1999) and in Algorithm 3.1.

The population log-likelihood (4.1), which we discuss in the next section, where we derive the algorithm, is continuously differentiable with respect to \( \psi \). The same applies to the expectation given the observed data. So the implicit function

$$\hat{\psi} = \arg \max_{\psi} E_{m}(\hat{L}L)(G | y, \psi, S)$$

(3.10)

is a model-design consistent estimator, too, that according to Dempster, Laird, and Rubin (1977) will converge to a stationary point of the log-likelihood. As the log-likelihood, however, has a unique maximum for common link functions (Wedderburn 1976), this means that the estimator \( \hat{\psi} \) is a consistent estimator, meaning that

$$\lim_{N, k \to \infty} P(\| \psi - \hat{\psi} \| > \varepsilon) = 0 \quad \forall \varepsilon > 0.$$  

(3.11)

In a certain sense, our result is a development of what Rao, Verret, and Hidiroglou (2013) call a “marginal model”—as we only require one survey weight per observation—and the classical Pseudo-ML approach of Asparouhov and Muthen (2006) and Rabe-Hesketh and Skrondal (2006). In these works, the classical Pseudo-ML approach gets tractable because the log-likelihood \( f(\hat{L}L)(\psi) \) is assumed to be nested due to the random effects structure, where the random effects of different stages are independent of each other. However, the easier analytically tractability is at the expense of a strict random effects structure. As we circumvent the need for analytical tractability by the application of MC-integration, we have less restrictions on the random effect structure: The survey weight that we employ is the inverse of the product of conditional probabilities on different
stages, hence it is the marginal survey weight as the inverse of the unit’s marginal inclusion probability. Though, the less restrictive random effects structure comes with a price: We must assume $S \perp G$, which is not necessary in Rabe-Hesketh and Skrondal (2006). Furthermore, as one reviewer pointed out, there is some information loss if the random effects structure is indeed nested: The marginal survey weight does not differentiate at which level of a multistage cluster design which inclusion probability is involved. This may lead to a loss in efficiency in the estimation. On the other hand, the integral approximation methods in Rabe-Hesketh and Skrondal (2006) become more inaccurate the higher the integration dimension becomes. In this term, MC-integration helps to maintain efficiency. The impact of survey weights per sampling stage on the efficiency may depend on the survey design, though. We refer to the appendix where we study the impact of sampling stage-based survey weights in a simulation study.

4. Theoretical Derivation of the Estimation Algorithm

4.1. Point Estimation

As motivated in the previous section, Algorithm 3.1 relies on an approximation and estimation of the finite population log-likelihood

$$\mathcal{L}(y, \psi, \gamma, U) = \sum_{i \in U} \log f(y_i | \gamma, \beta) + \log \phi(\gamma | \sigma),$$

where $U$ denotes the index set $\{1, \ldots, N\}$ and $Y_j = y_j$ is the $j$th realization of the DGP (3.1) to (3.3) under $G = \gamma$. In our working example, $y_j$ would be the outcome “student $j$ achieved one of the two highest reading proficiency levels” ($y_j = 1$) or “did not achieve” ($y_j = 0$) for all 15-year-old students in the United States (the ensemble of these students thus is $U$). $G$ then denotes a school-specific effect on individual reading skills. The students actually drawn into the PISA study are then elements of the random variable realization $S = s$. An adequate estimator if students were sampled from all schools in the US would be the Horvitz-Thompson estimator (Horvitz and Thompson 1952) like in (3.5) with $B = \log \phi(\gamma | \sigma)$. However, as some schools do not enter the PISA study, only a subvector $\gamma_S$ of $\gamma$ could enter the sample log-likelihood. This yields a random part $B_S$ in (3.5) instead of $B$ of the shape

$$B_S = -0.5 \log \det \Sigma_S - 0.5 \log (2\pi) - 0.5 \gamma_S^T \Sigma_S^{-1} \gamma_S$$

that has the same structure as $B$ but is of lower dimension. Hence, in order to give the random effects $G$ adequate ‘importance’ (comparing to term $A$) in the subsequent optimization process, we require a rescaling of the survey weights $w_j, j \in U$. As $E \left( \sum_{j \in U} \delta_S(j) w_j \right) = N$, we suggest to multiply the survey weights by the factor $n/N$ where $n$ is the expected sample size: The likelihood part $B_S$ has the correct size dimension for a simple random sample of size $n$ and as the term cannot be marginally weighted, the term $A$ must be adjusted.

So, if the random effects realization were observed, maximization of the HT-estimate with $S = s$ would represent a sort of pseudo joint likelihood estimate for the superpopulation parameter vector $\psi$. This idea is summarized in Figure 1 and reflects the considerations given in Isaki and Fuller (1982) about (simple) regression modeling in complex samples. In fact, the sampling approach in superpopulation models was studied in several influential works such as Chambless and Boyle (1985), Yi, Rao, and Li (2016), Lumley and Scott (2017) and Kott (2018).

However, random effects are unobserved heterogeneity. That is, estimates cannot be gathered from (3.5). A common method for ML-estimation under partially unobserved data is the EM-algorithm (Dempster, Laird, and Rubin 1977). This algorithm consists of two iterating steps. In the E-step, $k$ the expected likelihood of the observed data is recalculated given the current maximizer $\psi_k$. And in the M-step, the expected likelihood given

![Figure 1](image_url)
the observed data is maximized. The EM-algorithm was first proposed for GLMMs in Laird and Ware (1982). As the integration becomes difficult to manage analytically in a nonnormal setting that we have described here, the E-step can be replaced by Monte-Carlo integration. In the case of GLMMs, this was suggested in McCulloch (1997) and Booth and Hobert (1999). Chen, Zhang, and Davidian (2002) use a MCEM-approach for GLMM estimation, too, but they relax the model on the normality assumption of \( G \) whereas we relax the model with respect to the survey design. The novelty of our approach is that the EM-algorithm is not applied to the population log-likelihood (4.1) but on its HT-estimator. Zipunnikov and Booth (2006) introduced weights in their MCEM-algorithm, too, but their focus lies rather on a correction for overdispersion and heteroscedasticity.

The different motivation for weighting in our setting—namely the design \( P_\text{S} \)—has also implications on the conditional density against which we integrate for the E-step: As pointed out in the previous section, we integrate with respect to a realization of \( h(\cdot|y, S) \) instead of \( h(\cdot|y, U) \) (see Equation (3.7)), which would return an estimator \( \hat{E}_m(\hat{L}\hat{C}) \) that would be preferable due to removing noise from the MC-integration steps.

On the other hand, conditioning on \( S \) allows to respect the (potentially complex) sampling design in both steps of the algorithm—the maximization and the expectation—in contrast to the pairwise likelihood in Yi, Rao, and Li (2016) and the pseudo-likelihood approach of Rabe-Hesketh and Skrondal (2006). There, the likelihood \( \hat{L}\hat{C} \) is survey-weighted but for the conditional density it is assumed that \( h(\cdot|y, S) = h(\cdot|y, U) \). In consequence, the expectation of \( A \) becomes scale-variant but is congruent between sampling stages and random effects. Furthermore, it can be argued that the additional noise through the MC-integration is counter-weighted by the fact that Rabe-Hesketh, Skrondal, and Pickles (2002) integrate a linear approximation to the GLMM and face in addition numerical integration errors through adaptive quadrature, which is avoided here. The linear approximation in Rabe-Hesketh, Skrondal, and Pickles (2002) however, yields an additional error that does not depend on the numerical integration method and can yield an additional bias (Yi, Rao, and Li 2016).

### 4.2. Standard Errors

For the MCEM-algorithm on GLMMs, the literature (e.g., Tan, Tian, and Fang 2007) suggests to use as variance estimators of the fixed effects estimator \( \hat{\beta} \) the inverse of the observed Fisher information introduced in Louis (1982). In the case of MC-integration, the Fisher information is replaced by a MC-estimate thereof (Booth and Hobert 1999). The observed information matrix corresponds to the negative expected Hessian of the joint likelihood minus the expected variance of the gradient of (4.1) in a classical model-based estimation. In our case, the derivation is rather with respect to the HT-estimator of (4.1), as we take into account the survey design \( P_\text{S} \). That is, the expectation (with respect to \( G \)) of the first and second derivatives of (3.5) is required. Originally, the observed Fisher information given in Louis (1982) was intended for the traditional EM-algorithm. Oakes (1999) derives the observed information matrix of the EM-algorithm differently but arrives at the same result at the stationary point.

Replacing the integral by an MC-integral is—following the MCEM training procedure—easy: After the convergence, the simulated random variables \( y_1, \ldots, y_\beta \) and the corresponding importance weights are available and allow once again a MC-integration to approximate the integral. That is, we suggest to use the estimated information matrix

\[
I_\beta = -\hat{E}_m\left(\frac{\partial^2 \hat{L}\hat{C}}{\partial \beta^2}\right) - \hat{E}_m\left(\nabla_{\beta} \hat{L}\hat{C} \cdot \nabla_{\beta} \hat{L}\hat{C}\right),
\]

(4.3)

where \( \frac{\partial^2 \hat{L}\hat{C}}{\partial \beta^2} \) denotes the Hessian matrix and \( \nabla_{\beta} \hat{L}\hat{C} \) the gradient of the joint likelihood with respect to the parameter vector \( \beta \) at position \( \hat{\beta} \). From the final E-step, one can keep the simulated random effects and importance weights in order to define Monte-Carlo-based estimators

\[
\hat{E}_m\left(\frac{\partial^2 \hat{L}\hat{C}}{\partial \beta^2}\right) = \sum_{b=1}^{B} w_b^{MC} \cdot \frac{\partial^2 \hat{L}\hat{C}(G_b, y, \hat{\psi}, S)}{\partial \beta^2}
\]

(4.4)

and

\[
\hat{E}_m\left(\nabla_{\beta} \hat{L}\hat{C} \cdot \nabla_{\beta} \hat{L}\hat{C}\right) = \sum_{b=1}^{B} w_b^{MC} \nabla_{\beta} \hat{L}\hat{C}(G_b, y, \hat{\psi}, S) \cdot \nabla_{\beta} \hat{L}\hat{C}(G_b, y, \hat{\psi}, S).
\]

(4.5)

These estimators have been suggested in Oakes (1999) and Tan, Tian, and Fang (2007) for the case without survey weights. Booth and Hobert (1999) used the inverse of the information matrix, too, but not as a final variance estimator. Instead, Booth and Hobert (1999) used the inverse observed Fisher matrix to determine stopping rules for the number of MC random generations in the Monte-Carlo E-step.

Chen, Zhang, and Davidian (2002), on the other hand, have found numerical problems with the MC-version of the information matrix (the MC-estimated Fisher information (4.3) is not necessarily positive definite). In consequence, we suggest to use the closest positive semidefinite symmetric matrix when the estimate of \( \hat{I}_\beta \) is not positive semidefinite (Higham 1988).

Having an estimator of the information matrix, its inverse \( \hat{I}_\beta^{-1} \) is a lower bound for the variance-covariance matrix and the square roots of the diagonal elements are thus lower bounds of the standard errors of the fixed effects estimator \( \hat{\beta} \) (Rao 1992).

### 5. Numerical Aspects of the Implementation

In a given survey realization \( S = s \), sampling from the distribution \( h(\cdot|y, s) \) is necessary for the MC-integration in the E-step. However, as we have already mentioned, \( h(\cdot|y, s) \) is high-dimensional and not pre-implemented in statistical software packages because it depends on \( S = s \) under \( P_\text{S} \) and \( Y = y \). It is thus very specific for a dataset. Furthermore, if \( Y \) is not Gaussian, the distribution is not normal, neither. Consequently, Monte-Carlo sampling from a good proposal distribution is required.

McCulloch (1997) suggested a Metropolis algorithm, which can take much time to scan the sample space of \( G \) and yields
correlated random generations of $G$. Booth and Hobert (1999) suggested rejection sampling or importance sampling. Rejection sampling, however, may become inefficient in the case of low acceptance rates (i.e., when the proposal is not chosen wisely). The normal distribution centered at 0 with covariance matrix $\Sigma$ would be an option for a proposal, for both rejection and importance sampling. It has the advantage that only the conditional density of $Y|G_S$ must be evaluated. However, if the variance components of $\Sigma$ are moderate-to-large, and the modes of $G|Y$ have therefore a moderate distance to zero, this proposal is inefficient. It leads to high rejection rates and often almost zero-valued importance weights. Therefore, for the importance sampling, Booth and Hobert (1999) suggest a $t$-distribution centered at the current mode of $G_S$ and with the inverse negative Hessian as covariance matrix.

Zipunnikov and Booth (2006) propose a quasi MC-method named randomized spherical radial integration, arguing that a transformation of the random effects in combination with random orthogonal matrices allows to reduce the MC-draws and consequently the computational burden. However, if there is no predefined structure on the covariance matrix $\Sigma$ (except for symmetry), and there are crossed random effects, the dimension of the required random orthogonal matrix is $q \times q$, which is cumbersome to generate. In our working example, there are 148 high schools sampled into the survey and thus, the dimension of the required random orthogonal matrix would be $148 \times 148$.

For the named inconveniences of the other algorithms, we suggest to use importance sampling in each E-step with proposal distribution

$$G_s \sim N(\hat{\gamma}_s, -\hat{H}_s^{-1}(\hat{\Sigma}_s))$$

(5.1)

in Algorithm 3.1. In order to improve sampling from the tails, the covariance matrix can be scaled by a factor slightly greater than 1. This proposal distribution (5.1) was applied in Pinheiro and Bates (1995) and is relatively easy to compute.

If $Z^T Z$ is not block diagonal with $Z$ only containing zeros and ones, and/or the number of random effects is large ($q \gg 0$) calculus of $H$ and its inversion is cumbersome, too. Furthermore, realizations of multivariate normals are usually gathered by multiplying an iid standard normal vector by the square root of $-\hat{H}_s^{-1}$. This is a computational burden that must be repeated in each Monte-Carlo E-step when the proposal (5.1) is used, because the Hessian is updated in each step. A solution is the modified BFGS-algorithm described in Powell (1987) that returns the mode $\hat{\psi}_s$ in each MCEM-step together with a $q \times q$ matrix $C$ such that $CC^T = -\hat{H}_s^{-1}$. We can then use in each training step the matrix $C$ resulting from the optimization to generate random vectors with the respective mode and covariance.

Note that the mode $\hat{\psi}_s$ given $\hat{\psi}$ is a unique maximizer of the log-likelihood (under the link functions given in Wedderburn (1976)): The GLM-part $A$ of the likelihood (3.5) is strictly concave in the linear predictor $\eta$ and the multivariate normal part (here denoted by $B$) is strictly concave in $\psi$, too. Thus, $A+B$ is strictly concave and thus there is a unique maximum.

Another point of the numerical implementation is that the MC-integration in the E-step turns Algorithm 3.1 into a stochastic optimization procedure. Though we have already discussed the convergence of the MCEM algorithm (Neath et al. 2013), the application might yield some instabilities. Especially in the binary data case, we found like McCulloch (1997) that one gets quickly into a neighborhood of the ML-estimators but then values might keep oscillating and even can leave the neighborhood. We therefore keep track of the expected log-likelihood and store the values that have returned the highest estimated expected log-likelihood.

6. Applications

6.1. Application to Real World Data

We estimate a logit mixed model on our working example, the U.S. PISA study data from 2000. Indeed, this working example allows us a direct comparison to the results presented in Rabe-Hesketh and Skrondal (2006). In order to soften the impact of MC-error in our algorithm, we estimate the model 500 times and present here the boxplots of these estimates. The exact model is as follows: The endogenous variable is an indicator on whether a student achieved one of the top two reading proficiency levels. The explanatories are

- female: Dummy on the student’s gender
- ise: International socio-economic index
- high school: Dummy on whether the highest education level of either parent is high school
- college: Dummy on whether the highest education level of either parent is high school
- one for: Dummy on whether one parent is foreign born
- both for: Dummy on whether both parents are foreign born

And for the schools, a random intercept is assumed whose variance is denoted “revar.” The results of the point estimates are summarized in Figure 2. The results of Rabe-Hesketh and Skrondal (2006) are added in the plot. Most point estimates agree.

![Figure 2](image-url)

Figure 2. Weighted logit mixed model parameter estimates (boxplots) on the reading proficiency application versus the results obtained by Rabe-Hesketh and Skrondal (2006).
with those of Rabe-Hesketh and Skrondal (2006). Nonetheless, our estimation procedure returns a higher impact of the parents’ educational achievement and a lower effect of one parent being foreign born. The estimate of the intercept is lower with our proposed algorithm, however, as the intercept has rarely a causal interpretation, we consider our deviations from Rabe-Hesketh and Skrondal’s results to be minor. In addition, consider that in average, the proposed estimation Algorithm 3.1 returns a less biased intercept estimator than the pseudo-loglikelihood and Skrondal’s results to be minor. In addition, consider that interpretation, we consider our deviations from Rabe-Hesketh proposed algorithm, however, as the intercept has rarely a causal foreign born. The estimate of the intercept is lower with our educational achievement and a lower effect of one parent being

\[ Y_i \sim \text{Ber} \left( \frac{\exp \eta_i^k}{1 + \exp \eta_i^k} \right) = \text{Ber} \left( \mu_i^k \right). \]  

(6.3)

We have thus a random intercept logistic regression model. The finite population is equally split into the \( d = 1, \ldots, 500 \) domains. That is, depending on the size \(|U|\), every domain \( d \) contains \(|U_d| = N^{(1)} \in \{5, 10, 20, 50, 100\} \) units. The explanatory for unit \( i \in U, x_{1,i} \) and \( x_{2,i} \) are once realized (per finite population of size 500 \( \cdot N^{(1)} \)) and generated from the Bernoulli distribution with probability 0.5. In fact, we have for a unit \( i \) belonging to domain \( d, i \in U_d, \)

\[ X_{1,i} = X_{1,d} \sim \text{Ber}(0.5) \]  

(6.4)

\[ X_{2,i} = X_{2,d} \sim \text{Ber}(0.5) \]  

(6.5)

Sample Randomization. Rabe-Hesketh and Skrondal (2006) oversample on the first stage of their sampling design (a two-stage cluster design) domains with a large absolute random intercept, i.e. the probability \( \pi_d^{(1)} \) that a cluster \( d \) enters the sample depends on the realization \( G_d^k = \gamma_d^k \) and corresponds to

\[ \pi_d^{(1)} = \begin{cases} 0.25, & \text{if } |\gamma_d^k| > 1 \\ 0.75, & \text{else} \end{cases} \]  

(6.6)

Consequently, the number of sampled domains is random, too. A similar design of the inclusion probability is implemented on the second stage given the realization \( Y_i^k = \gamma_i^k \): 

\[ \pi_{id}^{(2)} = \begin{cases} 0.25, & \text{if } |\gamma_i^k - \mu_i^k| > 1 \\ 0.75, & \text{else} \end{cases} \]  

(6.7)

The design is thus nonignorable at both stages and if it was not accounted for, the random effects’ variance would be systematically underestimated.

**Results.** In the following, we present boxplots in order to give an idea about the distribution of our proposed model parameter estimators. In Figure 3, we added the average result of the point estimator that Rabe-Hesketh and Skrondal (2006) obtained in their simulation study. We find that our method offers an advantage over the pseudo-loglikelihood estimation under weight scaling especially for the fixed effects. The bias reduces with increasing cluster sizes. However, we underestimate the standard deviation of the random effects. While the MCEM outperforms the pseudo-loglikelihood approach under weight scaling for \( N_j^{(1)} = 10 \) and has approximately the same bias for \( N_j^{(1)} = 20 \), it is worse for very small and very large clusters. However, note that these results agree with the application to the real world data: The average (and median) number of students within high schools in the dataset is about 14, which is close to the expected sample size within clusters for \( N_j^{(1)} = 20 \). There, we found that the estimated variance of the school random effect using MCEM is slightly larger than that found by Rabe-Hesketh and Skrondal (2006). The same result that occurs here for \( N_j^{(1)} = 10, 20 \).

### 6.2. Simulation Studies on GLMMs under Complex Design

#### 6.2.1. Replication of Rabe-Hesketh and Skrondal (2006)

**Data-Generating Process.** The DGP in Rabe-Hesketh and Skrondal (2006) can be summarized as follows: \( k = 1, \ldots, 1000 \) Monte-Carlo finite populations were generated according to

\[ n_i^k = 1 + x_{1,i} + x_{2,i} + G_d^k, \quad i = 1, \ldots, 500 \cdot N^{(1)}, \]  

\( N^{(1)} \in \{5, 10, 20, 50, 100\} \)  

(6.1)

\[ G_d^k \sim N(0,1), \quad d = 1, \ldots, 500 \]  

(6.2)

\[ Y_i^k \sim \text{Ber} \left( \frac{\exp \eta_i^k}{1 + \exp \eta_i^k} \right) = \text{Ber} \left( \mu_i^k \right). \]  

(6.3)

We arrange the parameters as

\[ \eta_i^k = 4 - 2x_{1,i} - x_{2,i} + G_d^k + (1, x_{1,i}) \cdot G_{d2}^k, \quad i = 1, \ldots, N \]  

(6.8)

\[ G_{d1}^k \sim N(0,2^2), \quad d_1 = 1, \ldots, 20 \]  

(6.9)

\[ G_{d2}^k \sim N \left( 0, \begin{pmatrix} 0.7 & 0.5 \\ 0.5 & 1.3 \end{pmatrix} \right), \quad d_2 = 1, \ldots, 30 \]  

(6.10)

\[ Y_i^k \sim \text{FGLM}(\eta_i^k). \]  

(6.11)

We arrange the parameters as

\[ \psi^T = (\beta_0, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2, \rho, \sigma_Z^2) = (4, -2, -1, 4, 0.7, 0.5, 1.3). \]  

\( \psi \) is chosen such that the linear predictors return sensible conditional expectations \( \mu_i^k = g(\eta_i^k) \) under both, the identity link \( g = \text{id} \) and the logit link \( g = \text{logit}^{-1} \). The distribution \( F \) is once Gaussian and once Bernoulli. That is, we run simulations for a linear mixed model and a logit mixed model with crossed random effects structures: Domains (clusters) for the first random effect are indexed by \( d_1 \) and domains for the second random effect are indexed by \( d_2 \) and are not nested within domains \( d_1 = 1, \ldots, 20 \). Crossing of both domain types \( d_1 \) and \( d_2 \) does not follow a pattern, the cross-classifications \( U_{d1} \cap U_{d2} \) was once assigned randomly and kept constant over all simulations (that is, \( z \) does not change over the simulations). All domains of type 1 have the same number of units \( N_{d1} = 200 \). The affiliation to domains of type 2 is also uniform, that means that \( N_{d2} \) is either 133 or 134. The expanators are also constant along the simulation runs and were generated once according to \( X_1 \sim N(2,1) \) and \( X_2 \sim \text{Exp}(1) \).
**Ignorable Design.** The simulation is designed to show the strengths and weaknesses of the proposed wGLMM approach in the case of non-ignorable survey designs. Furthermore, we show that the loss of efficiency under ignorable survey designs due to weighting is acceptable. For this reason, we first implement an ignorable sampling design under the DGP: Given the explainatories, the random sample $S$ is independent of the endogenous variable, $(Y|X,Z) \perp S$. From the finite population realization $k$, $k = 1, \ldots, 2000$, a stratified random sample is drawn. The strata are the domains $d_1 = 1, \ldots, 20$ and the sample size in each domain is equal to $n_{d_1} = 10$. The domains $d_2$, in contrast, are unplanned as they randomly cross domains $d_1$. This leads to random sample sizes for domains $d_2$ and possibly to a sample size $n_{d_2} = 0$. In this set-up, alternative algorithms that were reviewed in the introduction would have troubles to take into account either random effects $G_{d_1}$ (e.g. the algorithm introduced in Rabe-Hesketh and Skrondal (2006)) or survey-weights (e.g., the R-package lme4, Bates (2011)). Because of the unplanned sampling in domain type 2, it might happen that the observed realizations $y_S \neq y$ and thus the outlined rescaling of the survey weights is necessary.

The units within the strata are sampled with unequal probabilities. Hereby, the first design ($\pi ps$) is ignorable. The inclusion probability for unit $i$ in stratum $d_1$ is proportional to $x_{2,i}$

$$\pi_i = \frac{x_{2,i}}{\sum_{j \in U_{d_1}} x_{2,j}}.$$  

As the elements $x_2$ are observed in the sample, survey-weighting should not have an effect under this sampling design. Probability proportional to size sampling designs are described by Lohr.
In this set-up, the weights are ignorable in expectation (Kott 2018). Therefore, this design is some sort of control for the estimation quality when weighting is not necessary. Using the weights nonetheless should lead to some loss in efficiency but should not impact the estimator’s expected outcome.

**Simulation Results under Ignorable Design.** The benchmark for our results are the results that the R-package lme4 (Bates 2011) returns - without the inclusion of survey weights. The results are presented in Tables 1 and 2.

We find that for the fixed effects and a Gaussian-dependent variable, the MCEM algorithm yields unbiased estimators and there is only a slight loss in efficiency in comparison to ignoring the survey weights in lme4. In addition, the results for the random effects variance components are reasonable, too. There is a slight bias concerning the variance components of the effects crossing the survey design. This bias, however, is about the same range as the one that is returned by lme4.

For the binary dependent variable, on the other hand, the integral approximation in lme4 leads partially to tremendous errors such that any efficiency loss that might result from the (unnecessary) inclusion of survey weights in our MCEM algorithm is clearly overweighted by that error. Furthermore, we remark that some simulation runs of the lme4 functions stopped and would have required further fine-tuning of the estimation hyperparameters whereas our algorithm managed to finish these particular simulation runs when lme4 failed.

Nonetheless, especially for the crossed random effects’ parameters a bias for the MCEM algorithm is observable.

In the second step, we report results for the discussed standard error estimator. As discussed in previous sections, the Fisher information of the EM algorithm (Louis 1982) was suggested by several authors. However, to the best of our knowledge, we are the first ones to present its performance in a MC simulation study. We benchmark the standard error estimators with the MC-estimated standard error:

$$ SE^{MC}(\hat{\beta}_i) = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \text{err}_2^2(\hat{\beta}_i^{(k)})} \quad (6.12) $$

where

$$ \text{err}_2^2(\hat{\beta}_i^{(k)}) = (\hat{\beta}_i^{(k)} - \beta_i)^2, \quad i = 0, 1, 2. \quad (6.13) $$

The results are summarized in Tables 3 and 4; negative variance estimates (negative diagonal elements of $\hat{I}_\beta^{-1}$) were disregarded. In cases where $\hat{I}_\beta^{-1}$ was not positive definite, the closest positive definite matrix was taken (‘nearPD’ in the table).

It becomes clear that the inverse of the Fisher information is only a lower bound of the variance and hence, the square root thereof is only a lower bound of the standard errors. In contrast to lme4, the standard error estimator is biased downwards. This is even the case under the Gaussian DGP with ignorable design, i.e. in the most favorable setting.

**Table 1.** LMM under ignorable design.

| Coef | Method   | True | $Q_{0.05}$ | $Q_{0.25}$ | Mean | Median | $Q_{0.75}$ | $Q_{0.95}$ |
|------|----------|------|------------|------------|------|--------|------------|------------|
| $\beta_0$ | wGLMM    | 4.0  | 2.73       | 3.47       | 3.99 | 3.98   | 4.52       | 5.27       |
|       | lme4     | 4.0  | 2.85       | 3.52       | 3.97 | 3.99   | 4.44       | 5.10       |
| $\beta_1$ | wGLMM    | −2.0 | −2.51      | −2.22      | −2.00| −2.01  | −1.79      | −1.46      |
|       | lme4     | −2.0 | −2.47      | −2.19      | −2.00| −1.99  | −1.81      | −1.54      |
| $\beta_2$ | wGLMM    | −1.0 | −1.29      | −1.12      | −1.00| −1.01  | −0.90      | −0.71      |
|       | lme4     | −1.0 | −1.20      | −1.08      | −1.00| −1.00  | −0.91      | −0.78      |
| $\sigma_1^2$ | wGLMM   | 4.0  | 1.62       | 2.66       | 3.81 | 3.62   | 4.85       | 6.66       |
|      | lme4     | 4.0  | 1.86       | 2.76       | 3.86 | 3.68   | 4.77       | 6.54       |
| $\sigma_2^2$ | wGLMM   | 0.7  | 0.05       | 0.22       | 0.50 | 0.38   | 0.63       | 1.35       |
|      | lme4     | 0.7  | 0.00       | 0.27       | 1.07 | 0.75   | 1.55       | 3.13       |
| $\rho$  | wGLMM    | 0.5  | 0.12       | 0.37       | 0.56 | 0.56   | 0.75       | 1.02       |
|      | lme4     | 0.5  | −0.73      | 0.00       | 0.35 | 0.44   | 0.75       | 1.11       |
| $\sigma_3^2$ | wGLMM   | 1.3  | 0.55       | 0.90       | 1.22 | 1.15   | 1.47       | 2.04       |
|      | lme4     | 1.3  | 0.51       | 0.89       | 1.30 | 1.24   | 1.64       | 2.27       |

**Table 2.** GLMM under ignorable design.

| Coef | Method   | True | $Q_{0.05}$ | $Q_{0.25}$ | Mean | Median | $Q_{0.75}$ | $Q_{0.95}$ |
|------|----------|------|------------|------------|------|--------|------------|------------|
| $\beta_0$ | wGLMM    | 4.0  | 1.32       | 2.39       | 3.18 | 3.19   | 3.93       | 5.02       |
|       | lme4     | 4.0  | 2.40       | 3.44       | 6.57 | 4.21   | 5.43       | 19.79      |
| $\beta_1$ | wGLMM    | −2.0 | −2.43      | −1.92      | −1.58| −1.58  | −1.22      | −0.71      |
|       | lme4     | −2.0 | −17.26     | −2.70      | −3.82| −2.13  | −1.72      | −1.24      |
| $\beta_2$ | wGLMM    | −1.0 | −1.40      | −1.00      | −0.81| −0.80  | −0.59      | −0.32      |
|       | lme4     | −1.0 | −6.32      | −1.36      | −1.74| −1.04  | −0.86      | −0.62      |
| $\sigma_1^2$ | wGLMM   | 4.0  | 0.00       | 0.22       | 1.66 | 1.36   | 2.62       | 4.71       |
|      | lme4     | 4.0  | 1.02       | 2.27       | 184.92| 3.87   | 6.81       | 369.63     |
| $\sigma_2^2$ | wGLMM   | 0.7  | 0.00       | 0.06       | 0.21 | 0.17   | 0.28       | 0.56       |
|      | lme4     | 0.7  | 0.00       | 0.15       | 159.54| 0.81   | 2.98       | 48.66      |
| $\rho$  | wGLMM    | 0.5  | −7.57      | −0.52      | −54.67| 0.20   | 0.69       | 1.84       |
|      | lme4     | 0.5  | −7.57      | −0.52      | −54.67| 0.20   | 0.69       | 1.84       |
| $\sigma_3^2$ | wGLMM   | 1.3  | 0.00       | 0.29       | 0.58 | 0.57   | 0.84       | 1.26       |
|      | lme4     | 1.3  | 0.10       | 0.70       | 102.22| 1.45   | 2.94       | 175.72     |
For the logit regression, the standard error estimator of the inverse Fisher information behaves similar like in the linear regression—it is a lower bound. However, in comparison with inverse Fisher information behaves similar like in the linear model assumptions such as the ignorability of the design do not hold. For this scenario, we account for when the weights are included. A similar non-ignorable survey design was studied in the simulation of Pfeffermann et al. (1998): In this article, the authors constructed the inclusion probability as a function of the random effects.

**Simulation Results under Nonignorable Design.** As discussed above, survey-weighting becomes essential under a non-ignorable survey design. This becomes especially clear when looking on the intercept in Tables 5 and 6.

In the linear case, except for the estimates on the fixed effect intercept, however, the results are similar to those under the ignorable design due to the construction of the sample informativity. The relatively large bias of the intercept (about 22%) is especially critical when the regression model is used for prediction, for example in small area estimation.

The biases of the variance components under the MCEM algorithm are only marginally greater than those under ignorable survey design. Furthermore, they are (overall) relatively smaller than those that result under ignorance of the survey design.

For nonlinear models such as the logit regression, on the other hand, things change dramatically as can be seen in Table 6. The combination of integration error and ignored sample informativity leads to severely biased estimators for both fixed effects and variance components under `lme4`.

Note that under this case, any inference stemming from standard error estimators must be erroneous. Standard error estimators are reported in Tables 7 and 8.

It is observable that the lower bound standard error estimators for the MCEM algorithm behave similarly like under an ignorable survey design, which makes us confident that the bias only stems from the lower bound property. Then, an improved estimator similar to the square roots of the diagonals of the inverse Fisher information could yield approximately correct standard error estimators. On the other hand, for the Gaussian-dependent variable, the standard error estimator of

---

**Table 3. SE Estimator in the LMM under ignorable design.**

|        | SE\(_{MC}\) | \(Q_{0.05}\) | \(Q_{0.25}\) | Mean | Median | \(Q_{0.75}\) | \(Q_{0.95}\) |
|--------|-------------|-------------|-------------|------|--------|-------------|-------------|
| Louis (1982), \(\beta_0\) | 0.10 | 0.25 | 0.31 | 0.38 | 0.34 | 0.38 | 0.62 |
| Equation (4.3) | \(\beta_1\) | 0.33 | 0.11 | 0.12 | 0.16 | 0.14 | 0.17 | 0.27 |
| \(\beta_2\) | 0.17 | 0.11 | 0.13 | 0.13 | 0.13 | 0.14 | 0.17 |
| nearPD Louis (1982) | \(\beta_0\) | 0.78 | 0.16 | 0.30 | 0.37 | 0.34 | 0.41 | 0.60 |
| \(\beta_1\) | 0.33 | 0.10 | 0.12 | 0.16 | 0.14 | 0.18 | 0.28 |
| \(\beta_2\) | 0.17 | 0.11 | 0.12 | 0.13 | 0.13 | 0.14 | 0.17 |
| lme4 | \(\beta_0\) | 0.69 | 0.57 | 0.63 | 0.68 | 0.67 | 0.72 | 0.80 |
| \(\beta_1\) | 0.28 | 0.23 | 0.26 | 0.28 | 0.28 | 0.30 | 0.34 |
| \(\beta_2\) | 0.13 | 0.11 | 0.12 | 0.13 | 0.13 | 0.13 | 0.14 |

---

**Table 4. SE Estimator in the GLMM under ignorable design.**

|        | SE\(_{MC}\) | \(Q_{0.05}\) | \(Q_{0.25}\) | Mean | Median | \(Q_{0.75}\) | \(Q_{0.95}\) |
|--------|-------------|-------------|-------------|------|--------|-------------|-------------|
| Louis (1982), \(\beta_0\) | 1.40 | 0.43 | 0.59 | 0.74 | 0.72 | 0.84 | 1.11 |
| Equation (4.3) | \(\beta_1\) | 0.67 | 0.17 | 0.25 | 0.33 | 0.31 | 0.37 | 0.51 |
| \(\beta_2\) | 0.38 | 0.17 | 0.23 | 0.27 | 0.27 | 0.30 | 0.91 |
| nearPD Louis (1982) | \(\beta_0\) | 1.40 | 0.28 | 0.57 | 0.72 | 0.71 | 0.83 | 1.10 |
| \(\beta_1\) | 0.67 | 0.17 | 0.25 | 0.33 | 0.31 | 0.37 | 0.51 |
| \(\beta_2\) | 0.38 | 0.18 | 0.23 | 0.27 | 0.27 | 0.30 | 0.91 |
| lme4 | \(\beta_0\) | 10.60 | 0.00 | 0.97 | 1.96 | 1.20 | 1.53 | 3.17 |
| \(\beta_1\) | 7.14 | 0.00 | 0.45 | 0.96 | 0.58 | 0.79 | 1.73 |
| \(\beta_2\) | 2.93 | 0.00 | 0.24 | 0.47 | 0.29 | 0.37 | 0.70 |

---

Nonignorable Design. The second design (ups), in contrast, is non-ignorable. This design is selected to underpin the importance of survey weights when model assumptions such as the ignorability of the design do not hold. For this scenario, we define the inclusion probability of unit \(i\) in stratum \(d_1\) in three steps (and due to the model dependency, for each finite population realization \(k\)):

\[
e_i^k := y_i^k - \mu_i^k
\]

\[
\bar{e}_i^k = \begin{cases} 0.1, & \text{if } e_i^k \leq \tau_{d_1,0.25} \\ 0.2, & \text{if } \tau_{d_1,0.25} < e_i^k \leq \tau_{d_1,0.5} \\ 0.4, & \text{if } e_i^k > \tau_{d_1,0.5} \end{cases}
\]

\[
\bar{\pi}_i^k = \frac{\sum_{j \in U_{d_1}} \bar{e}_i^k}{\sum_{j \in U_{d_1}} \bar{\pi}_i^k} \cdot 200,
\]

where \(\tau_{d_1,p}\) is the \(p\)-quantile of the residuals \(e_i^k\) in domain \(d_1\) in the \(k\)th finite population. Consequently, the inclusion probability is a function of an unobserved model error and thus non-ignorable, \(Y \mid X, Z, \pi \sim \mathcal{S}\) (Pfeffermann 1993). Under this set-up, the expectation of the error in the sample does not equal zero, not even so when conditioned on the explanatory variables (Kott 2018). This scenario corresponds to the assumption violation of Kott (2018) that makes survey weights nonignorable. As the regression concept seeks to minimize the random error and observations with a highly positive error are oversampled, we expect the regression intercept to be biased upwards. However, the information on the oversampling of units with positive error realizations is contained in the survey weights and thus be
the intercept of $\beta_1$ would lead to wrong inference, too, due to the bias. For the binary case, the large biases in all three fixed effects overweight any variance component that is caught by the standard error estimator of $\beta_1$. Hence, there is a severe underestimation.

7. Discussion

In this article, we proposed to combine the Monte-Carlo EM-algorithm, that allows for flexible mixed models (McCulloch 1997; Booth and Hobert 1999; Chen, Zhang, and Davidian
with the inclusion of survey weights (e.g. Chambless and Boyle 1985). To the best of our knowledge, survey weighting in GLMMs has until now come from another direction (e.g., Peffermann et al. 1998; Rabe-Hesketh and Skrondal 2006; Rao, Verret, and Hidiroglo 2013; Yi, Rao, and Li 2016). The previous approaches for survey weighting in mixed models make it difficult to include a single, unit-specific survey weight and crossed random effect patterns. We have circumvented this problem by the use of the MCEM. In addition, we find that our proposed estimator is design consistent for the finite population parameters. These are in turn model-consistent for the parameters in the data-generating process (DGP). Overall, the suggested MCEM yields thus a model-design consistent estimator at the sample level. Therefore, our approach is a powerful tool in both econometric research, where the focus lies on the DGP, and survey statistics, where a model under the finite population can assist the estimation process.

For example, small area estimation often uses a multi-level framework in order to increase the efficiency of mean estimators in small domains Battese, Harter, and Fuller (1988). However, those estimators are only model-ubnased under the correct model and ignorable sampling. The survey weights thus could protect the point estimators when the design is non-ignorable. In combination with Box–Cox transformations (Box and Cox 1964) to improve error distribution assumptions as suggested in Li and Lahiri (2007), Gurka et al. (2006) and Rojas-Perilla et al. (2017), a powerful and flexible tool might result. Another, yet to be studied application, would be to replace the survey weights in the algorithm with nonresponse propensities in order to correct for a missing-at-random nonresponse pattern. Furthermore, our approach could be of use in the context of imputation: Instead response propensity scores, a (mixed) survey weighted regression imputation becomes possible.

Besides the theoretical background, we have discussed the computational problems that occur for flexible covariance matrices and the resulting high dimensionality of the random vector G. With the application to U.S. PISA 2000 data, we demonstrate the applicability in real data situations. A final simulation study revealed that the fixed effects estimators are not harmed by survey weighting when the design is ignorable and profit when this is not the case. When the domains, for which a random effect pattern is assumed, reflect sampling stages, we get the same evidence for the variance component estimation. However, further investigation is necessary for how to stabilize the variance component estimation for unplanned domains, especially when the endogenous variable is binary. We have furthermore contrasted our estimation algorithm with an established estimation procedure and find it competitive.

Finally, we have evaluated the inverse observed Fisher information for standard error estimation. To the best of our knowledge, this method was proposed in the literature but not evaluated in a simulation study before. Though the tendency of the standard error estimator is stable across scenarios, we find that further research must be done—the Fisher information can lead at most to lower bounds for the standard errors.

To conclude, we introduced unit-specific weights into the generalized linear mixed model framework. This is important as most surveys only provide a single survey weight that must thus account for the complete sampling design. The proposed methodology is, therefore, an important tool to reconcile classical statistical modeling with the reality of empirical research where only samples generated by complex surveys are available.

Acknowledgments

This research is supported by the Research Institute for Official and Survey Statistics (RIFOSS) at Trier University. Furthermore, the authors would like to thank Ralf Münnich and three anonymous reviewers for valuable comments that helped to improve the article.

Source Code

The source code is available on https://github.com/jburgard/wglmm.

ORCID

Jan Pablo Burgard © http://orcid.org/0000-0002-5771-6179

References

Asparouhov, T., and Muthen, B. (2006), “Multilevel Modeling of Complex Survey Data,” in Proceedings of the Joint Statistical Meeting in Seattle, pp. 2718–2726. [1040] Bates, D. (2011), “Computational Methods for Mixed Models,” Vignette for lme4. [1045,1046] Battese, G. E., Harter, R. M., and Fuller, W. A. (1988), “An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data,” Journal of the American Statistical Association, 83, 28–36. [1049] Booth, J. G., and Hobert, J. P. (1999), “Maximizing Generalized Linear Mixed Model Likelihoods With an Automated Monte Carlo EM Algorithm,” Journal of the Royal Statistical Society, Series B, 61, 265–285. doi: 10.1111/1467-9868.00176. [1039,1040,1042,1043,1048] Box, G. E. P., and Cox, D. R. (1964), “An Analysis of Transformations,” Journal of the Royal Statistical Society, Series B, pages 211–252. [1049] Chambless, L. E., and Boyle, K. E. (1985), “Maximum Likelihood Methods for Complex Sample Data: Logistic Regression and Discrete Proportional Hazards Models,” Communications in Statistics–Theory and Methods, 14, 1377–1392. [1041,1049] Chauvet, G. (2014), “A Note on the Consistency of the Narain-Horvitz-Thompson Estimator,” arXiv:1412.2887. [1040] Chen, J., Zhang, D., and Davidian, M. (2002), “A Monte Carlo EM Algorithm for Generalized Linear Mixed Models With Flexible Random Effects Distribution,” Biostatistics, 3, 347–360. [1042,1049] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), “Maximum Likelihood From Incomplete Data Via the EM Algorithm,” Journal of the Royal Statistical Society, Series B, 1–38. [1039,1040,1041] Gurka, M. J., Edwards, L. J., Muller, K. E., and Kupper, L. L. (2006), “Extending the Box–Cox Transformation to the Linear Mixed Model,” Journal of the Royal Statistical Society, Series A, 169, 273–288. [1049] Higham, N. J. (1988), “Computing a Nearest Symmetric Positive Semidefinite Matrix,” Linear Algebra and Its Applications, 103, 103–118. [1042] Horvitz, D. G., and Thompson, D. J. (1952), “A Generalization of Sampling Without Replacement From a Finite Universe,” Journal of the American Statistical Association, 47, 663–685. [1039,1041] Isaki, C. T., and Fuller, W. A. (1982), “Survey Design Under the Regression Superpopulation Model,” Journal of the American Statistical Association, 77, 89–96. [1041] Kott, P. S. (2018), “A Design-Sensitive Approach to Fitting Regression Models With Complex Survey Data,” Statistics Surveys, 12, 1–17. [1037,1041,1046,1047] Laird, N. M., and Ware, J. H. (1982), “Random-Effects Models for Longitudinal Data,” Biometrics, 963–974. [1039,1042] Leckie, G. (2009), “The Complexity of School and Neighbourhood Effects and Movements of Pupils on School Differences in Models of Educational Achievement,” Journal of the Royal Statistical Society, Series A, 172, 537–554. [1037]
Li, Y., and Lahiri, P. (2007), “Robust Model-Based and Model-Assisted Predictors of the Finite Population Total,” *Journal of the American Statistical Association*, 102, 664–673. doi: 10.1198/016214507000000158. [1049]

Lohr, S. L. (2010), *Sampling: Design and Analysis*, Brooks Cole, Boston, Mass. [u.a.], 2. ed., internat. ed. edition. Available at: [http://digitale-objekte.hbz-nrw.de/storage/2012/12/05/file_228/4825326.pdf](http://digitale-objekte.hbz-nrw.de/storage/2012/12/05/file_228/4825326.pdf). [1046]

Louis, T. A. (1982), “Finding the Observed Information Matrix When Using the EM Algorithm,” *Journal of the Royal Statistical Society*, Series B, 226–233. [1042,1046,1047,1048]

Lumley, T., and Scott, A. (2017), “Fitting Regression Models to Survey Data,” *Statistical Science*, 32, 265–278. [1037,1041]

McCulloch, C. E. (1997), “Maximum Likelihood Algorithms for Generalized Linear Mixed Models,” *Journal of the American Statistical Association*, 92, 162–170. doi: 10.1080/01621459.1997.10473613. [1039,1040,1042,1043,1048]

Neath, R. C. et al. (2013), “On Convergence Properties of the Monte Carlo EM Algorithm,” in *Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton*, Institute of Mathematical Statistics, pp. 43–62. [1039,1040,1043]

Oakes, D. (1999), “Direct Calculation of the Information Matrix Via the EM,” *Journal of the Royal Statistical Society*, 61, Series B, 479–482. [1042]

Oguz-Alper, M., and Berger, Y. G. (2016), “Empirical Likelihood inference for Regression Parameters When Modelling Hierarchical Complex Survey Data,” in *Proceedings of the Survey Research Methods Section of the American Statistical Association*, Joint Statistical Meetings, Chicago. [1037]

Owen, A. B. (2013), “Monte Carlo Theory, Methods and Examples.” Available at: [http://statweb.stanford.edu/~owen/mc/](http://statweb.stanford.edu/~owen/mc/). [1040]

Pfeffermann, D. (1993), “The Role of Sampling Weights When Modeling Survey Data,” *International Statistical Review*, 517–537. [1037,1047]

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998), “Weighting for Unequal Selection Probabilities in Multilevel Models,” *Journal of the Royal Statistical Society*, Series B, 60, 23–40. [1037,1047,1049]

Pinheiro, J. C., and Bates, D. M. (1995), “Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model,” *Journal of Computational and Graphical Statistics*, 4, 12–35. [1043]

Powell, M. J. D. (1987), “Updating Conjugate Directions by the bfgs Formula,” *Mathematical Programming*, 38, 29. [1043]

Rabe-Hesketh, S., and Skrondal, A. (2006), “Multilevel Modelling of Complex Survey Data,” *Journal of the Royal Statistical Society*, Series A, 169, 805–827. [1037,1038,1040,1041,1042,1043,1044,1045,1049]

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2002), “Reliable Estimation of Generalized Linear Mixed Models Using Adaptive Quadrature,” *The Stata Journal*, 2, 1–21. [1042]

Radhakrishna Rao, C. “Information and the Accuracy Attainable in the Estimation of Statistical Parameters,” in *Breakthroughs in Statistics*, pp. 235–247. Springer. [1042]

Rao, F. V., and Hidiroglou, M. A. (2013), “A Weighted Composite Likelihood Approach to Inference for Two-Level Models From Survey Data,” *Survey Methodology*, 39, 263–282. [1037,1040,1049]

Rojas-Perilla, N., Pannier, S., Schmid, T., and Iázidis, N. (2017), “Data-Driven Transformations in Small Area Estimation,” Technical report, Discussion Paper, School of Business & Economics: Economics. [1049]

Stiratelli, R., Laird, N., and Ware, J. H. (1994), “Random-Effects Models for Serial Observations With Binary Response,” *Biometrics*, 40, 961–971. [1038]

Tan, M., Tian, G-L., and Fang, H-B. (2007), “An Efficient MCEM Algorithm for Fitting Generalized Linear Mixed Models for Correlated Binary Data,” *Journal of Statistical Computation and Simulation*, 77, 929–943. [1042]

RWM Wedderburn. (1976), “On the Existence and Uniqueness of the Maximum Likelihood Estimates for Certain Generalized Linear Models,” *Biometrika*, 63, 27–32. [1038,1039,1040,1043]

Yi, G. Y., Rao, J. N. K., and Li, H. (2016), “A Weighted Composite Likelihood Approach for Analysis of Survey Data Under Two-Level Models,” *Statistica Sinica*, 569–587. [1037,1041,1042,1049]

Zipunnikov, V. V., and Booth, J. G. (2006), “Monte Carlo EM for Generalized Linear Mixed Models Using Randomized Spherical Radial Integration,” Technical report. [1042,1043,1049]