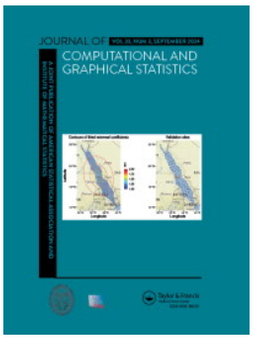

Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ucgs20

Group-Orthogonal Subsampling for Hierarchical Data Based on Linear Mixed Models

Jiaqing Zhu, Lin Wang & Fasheng Sun

To cite this article: Jiaqing Zhu, Lin Wang & Fasheng Sun (2024) Group-Orthogonal Subsampling for Hierarchical Data Based on Linear Mixed Models, Journal of Computational and Graphical Statistics, 33:3, 1037-1046, DOI: 10.1080/10618600.2023.2301093

To link to this article: https://doi.org/10.1080/10618600.2023.2301093

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2024, VOL. 33, NO. 3, 1037–1046
https://doi.org/10.1080/10618600.2023.2301093

Group-Orthogonal Subsampling for Hierarchical Data Based on Linear Mixed Models

Jiaqing Zhu[a], Lin Wang[b], and Fasheng Sun[a]

[a]Department of Statistics, KLAS and School of Mathematics and Statistics, Northeast Normal University, Changchun, China; [b]Department of Statistics, Purdue University, West Lafayette, IN

ABSTRACT

Hierarchical data analysis is crucial in various fields for making discoveries. The linear mixed model is often used for training hierarchical data, but its parameter estimation is computationally expensive, especially with big data. Subsampling techniques have been developed to address this challenge. However, most existing subsampling methods assume homogeneous data and do not consider the possible heterogeneity in hierarchical data. To address this limitation, we develop a new approach called group-orthogonal subsampling (GOSS) for selecting informative subsets of hierarchical data that may exhibit heterogeneity. GOSS selects subdata with balanced data size among groups and combinatorial orthogonality within each group, resulting in subdata that are *D*- and *A*-optimal for building linear mixed models. Estimators of parameters trained on GOSS subdata are consistent and asymptotically normal. GOSS is shown to be numerically appealing via simulations and a real data application. Theoretical proofs, R codes, and supplementary numerical results are accessible online as supplementary materials.

ARTICLE HISTORY
Received April 2023
Accepted December 2023

KEYWORDS
Data reduction; Experimental design; Optimal subsampling; Orthogonal array

1. Introduction

The unprecedented growth of data in modern research poses significant challenges in terms of storage and analysis. First, an individual's computing resources may not have the capacity to store the entire dataset due to its large size. Second, even after the dataset has been loaded into memory, traditional analysis methods may be too slow or even impractical due to the large volume of data (Bates 2014; Gao and Owen 2017).

Subsampling has been widely used to tackle the issue of storage capacity and accelerate data analysis. Several subsampling techniques have been developed to address the challenges of big data, generally aiming to optimize the downstream modeling. For example, for linear regression, Ma and Sun (2015) proposed to use the leverage score to construct nonuniform subsampling probabilities. Using the optimal design theory in experimental design, Wang, Yang, and Stufken (2019) proposed an information-based optimal subdata selection (IBOSS) method based on the *D*-optimality criterion. Inspired by the excellent properties of two-level orthogonal arrays under linear models, Wang et al. (2021) proposed an orthogonal subsampling (OSS) approach and showed that the OSS method typically outperforms existing methods in minimizing the mean squared errors (MSE) of the estimated parameters and maximizing the efficiencies of the selected subdata. Some other subsampling works for linear regression include Li and Meng (2020), Ren and Zhao (2021), Wang (2022), and Yu and Wang (2022), among others. Subsampling methods are also widely studied when other downstream models are considered, for example, the generalized linear model (Ai et al. 2021b), quantile regression (Wang and Ma 2021; Fan, Liu, and Zhu 2021; Ai et al. 2021a; Shao, Song, and Zhou 2022), multiplicative model (Ren, Zhao, and Wang 2023), nonparametric regression (Meng et al. 2020; Sun, Zhong, and Ma 2021; Meng et al. 2022; Zhang et al. 2023), Gaussian process modeling (He and Hung 2022) and the model-free scenario (Mak and Joseph 2018; Shi and Tang 2021). In addition, Meng et al. (2021) proposed the "Lowcon" method to address the presence of model misspecification. Xie, Bai, and Ma (2023) proposed an optimal subsampling method for online streaming data. Yu et al. (2022) considered the optimal subsampling method in a distributed environment. Readers may also refer to Yu, Ai, and Ye (2023) for a comprehensive review of subsampling methodology.

Knowledge discovery in various fields often relies on the analysis of complex data with a hierarchical structure. For example, students could be sampled from within schools, patients from within doctors, medical records from within individuals, or participants in psychological tests from within communities. For more applications, see, for example, Raudenbush (1993), McCulloch and Searle (2004), Bennett and Lanning (2007), Jiang and Nguyen (2007), Gao and Owen (2017), and Gao and Owen (2020). When the covariates of different groups in a dataset come from distinct distributions, they may demonstrate intra-group homogeneity and inter-group heterogeneity. Consequently, selecting a subset of data that has this hierarchical structure requires additional consideration. Existing subsampling methods often assume that the covariates are

CONTACT Fasheng Sun *sunfs359@nenu.edu.cn* Department of Statistics, KLAS and School of Mathematics and Statistics, Northeast Normal University, Changchun, China.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

homogeneous throughout the entire dataset. Using these methods may overlook critical information contained in hierarchical data. Therefore, it is imperative to develop specialized subsampling techniques that can accurately identify and capture the valuable information in such data.

In this article, we investigate the optimal subsampling method for hierarchical data by assuming that the data points come from a linear mixed model, which allows both fixed and random effects and is particularly used to analyze the data with a hierarchical structure, see Jiang and Nguyen (2007) and Gao and Owen (2020). We develop a group-orthogonal subsampling (GOSS) approach to tackle the memory and computational barriers of linear mixed models. GOSS is particularly designed for data with a hierarchical structure and targets two merits of the selected subdata: data size balance among groups and combinatorial orthogonality within each group. First, GOSS achieves data size balance among groups so that all groups contribute equally to the subdata. Second, GOSS selects the subdata from each group that approximate an orthogonal array (OA) to extract informative data points. OAs are universally optimal and have been employed in subdata selection for first-order linear regression (Wang et al. 2021). Our first original contribution lies in extending the theory that establishes the optimality of OAs to the context of the linear mixed model. Consequently, the selected subdata by GOSS is guaranteed to be D- and A-optimal for the generalized least squares (GLS) estimator of a linear mixed model. Numerical results in this article and Appendix demonstrate that GOSS outperforms existing methods in minimizing the MSE of parameter estimators and the prediction error over the full data.

Regarding the computing time, for a large full data size N with R groups of p-dimensional observations and a fixed subdata size n, the computational complexity is $O(Np\log(n/R))$, which is a little faster than $O(Np\log n)$ from OSS and as low as $O(Np)$ from IBOSS. In addition, GOSS is naturally suitable for distributed parallel computing to further accelerate the computation. Theoretical results are provided to show the consistency and asymptotic normality of the GLS estimator obtained on the selected subdata.

The rest of the article is organized as follows. Section 2 introduces the notations of the linear mixed model and the fundamental framework for the GOSS method. Section 3 introduces the OA and derives their theoretical optimality for obtaining the GLS estimator of a linear mixed model. Section 4 proposes the GOSS method and investigates the asymptotic property of the estimator based on the GOSS subdata. Sections 5 and 6 evaluate the GOSS algorithm via simulation studies and a real-world application. Section 7 concludes the article. Technical proofs and R codes are provided in supplementary materials.

2. The Framework

Denote the full data as $\{\mathbf{x}_{ij}, y_{ij}\}_{i=1,\ldots,R}^{j=1,\ldots,C_i}$, which include R groups and C_i observations in the ith group for $i = 1,\ldots,R$, so that the full data size is $N = \sum_{i=1}^{R} C_i$. Here $\mathbf{x}_{ij}$ is a p-vector of covariates for the jth unit in the ith group, the first component of $\mathbf{x}_{ij}$ is 1, and y_{ij} is its response. Consider the following linear mixed model,

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + a_i + e_{ij}, \mathbf{x}_{ij} \in \mathbb{R}^{p\times 1}, i = 1,2,\ldots,R, j = 1,2,\ldots,C_i, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ is a vector of fixed effects, a_i is the iid random effect associated with the ith group, $a_i \sim (0,\sigma_A^2)$, and $e_{ij} \sim (0,\sigma_E^2)$ is the error term independent from a_i. In the model in (1), two observations in the same group are assumed to have constant correlation $\sigma_A^2/(\sigma_A^2+\sigma_E^2)$, and observations from different groups are uncorrelated. More details about the linear mixed models can be found in Jiang and Nguyen (2007).

Let $\mathbf{X} = (\mathbf{X}_1^T,\ldots,\mathbf{X}_R^T)^T \in \mathbb{R}^{N\times p}$ with $\mathbf{X}_i = (\mathbf{x}_{i1},\ldots,\mathbf{x}_{iC_i})^T = (\mathbf{1}_{C_i},\mathbf{Z}_i)$ and $\mathbf{Z}_i = (\mathbf{z}_{i1},\ldots,\mathbf{z}_{iC_i})^T$ and $\mathbf{Y} = (\mathbf{Y}_1^T,\ldots,\mathbf{Y}_R^T)^T \in \mathbb{R}^{N\times 1}$ with $\mathbf{Y}_i = (y_{i1},\ldots,y_{iC_i})^T$, for $i = 1,\ldots,R$. The $\mathbf{Z}_i$ may be distinctly distributed for different i.

We are commonly interested in the estimator of $\boldsymbol{\beta}$, whose GLS estimator based on the full data is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}$$

when σ_A^2 and σ_E^2 are known, where $\mathbf{V} = \mathrm{cov}(\mathbf{Y}) = \sigma_E^2\mathbf{I}_N + \sigma_A^2\mathbf{A}$, and $\mathbf{A} \in \mathbb{R}^{N\times N}$ is a block diagonal matrix with the ith block $\mathbf{1}_{C_i}\mathbf{1}_{C_i}^T$. The estimator $\hat{\boldsymbol{\beta}}$ needs $O(Np^2)$ time complexity to calculate, which is not an easy task when N is big. When σ_A^2 and σ_E^2 are unknown, they are estimated from data, making the process even slower.

Now consider taking a subset of size n from the full data, where n_i points are from the ith group so that $n = \sum_{i=1}^{R} n_i$. Denote the selected subdata as $\{\mathbf{x}_{ij}^*, y_{ij}^*\}_{i=1,\ldots,R}^{j=1,\ldots,n_i}$. Let $\mathbf{X}^* = (\mathbf{X}_1^{*T},\ldots,\mathbf{X}_R^{*T})^T$ with $\mathbf{X}_i^* = (\mathbf{x}_{i1}^*,\ldots,\mathbf{x}_{in_i}^*)^T = (\mathbf{1}_{n_i},\mathbf{Z}_i^*)$ and $\mathbf{Z}_i^* = (\mathbf{z}_{i1}^*,\ldots,\mathbf{z}_{in_i}^*)^T$, $\mathbf{Y}^* = (\mathbf{Y}_1^{*T},\ldots,\mathbf{Y}_R^{*T})^T$ with $\mathbf{Y}_i^* = (y_{i1}^*,\ldots,y_{in_i}^*)^T$. The GLS estimator based on the subdata is given by

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}^{*T}\mathbf{V}^{*-1}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\mathbf{V}^{*-1}\mathbf{Y}^*, \tag{2}$$

where $\mathbf{V}^* = \mathrm{cov}(\mathbf{Y}^*) = \sigma_E^2\mathbf{I}_n + \sigma_A^2\mathbf{A}^*$, and $\mathbf{A}^* \in \mathbb{R}^{n\times n}$ is a block diagonal matrix with the ith block $\mathbf{1}_{n_i}\mathbf{1}_{n_i}^T$. The σ_A^2 and σ_E^2 in (2) may also be replaced by their estimators trained from the subdata. We will see that the accuracy of the estimators for σ_A^2 and σ_E^2 does not depend much on the subsampling strategies. Therefore, we will focus on selecting the subdata that allows the best estimation of $\boldsymbol{\beta}$. From simple algebra,

$$\mathrm{E}(\hat{\boldsymbol{\beta}}^*) = \boldsymbol{\beta} \text{ and } \mathrm{var}(\hat{\boldsymbol{\beta}}^*) = (\mathbf{X}^{*T}\mathbf{V}^{*-1}\mathbf{X}^*)^{-1} = \mathbf{M}^{*-1},$$

where

$$\mathbf{M}^* = \mathbf{X}^{*T}\mathbf{V}^{*-1}\mathbf{X}^* \tag{3}$$

is the information matrix of the subdata. The optimal subdata $\mathbf{X}^*$ maximizes the information $\mathbf{M}^*$ or, in other words, minimizes $\mathrm{var}(\hat{\boldsymbol{\beta}}^*)$ in some manner, which can be obtained by minimizing an optimality function of $\mathbf{M}^{*-1}$. Denote ψ as the optimality function. Finding the optimal subdata is to solve the following optimization problem:

$$\mathbf{X}^{*opt} = \arg\min_{\mathbf{X}^*\subseteq\mathbf{X}} \psi(\mathbf{M}^{*-1})$$
$$s.t.\ \mathbf{X}^* \text{ contains } n \text{ points.} \tag{4}$$

This is akin to the fundamental idea behind optimal experimental design (Kiefer 1959). Popular options for ψ include the determinant and trace, which correspond to the D- and A-optimality, respectively. Both of these two optimal criteria have specific statistical meanings. Specifically, D-optimal design minimizes the volume of the confidence ellipsoid centered at $\hat{\boldsymbol{\beta}}^*$ by maximizing the determinant $|\mathbf{M}^*|$, while A-optimal design minimizes the average variance of the components of $\hat{\boldsymbol{\beta}}^*$ by minimizing the trace $\mathrm{tr}(\mathbf{M}^{*-1})$.

The optimization problem in (4) is not easy to solve. Exhaustive search for solving the problem requires $O(N^n n^2 p)$ operations, which is infeasible for even moderate sizes of $\mathbf{X}$ and $\mathbf{X}^*$. There are many types of algorithms for finding optimal designs and among them, exchange algorithms are among the most popular. For the reasons argued in Wang et al. (2021), these algorithms are cumbersome in solving the subsampling problem in (4). To this end, we will initially derive theoretical results to establish the optimality of using an OA for the problem defined in (4). Following that, we will develop a computationally tractable subsampling approach called GOSS, which selects subdata approximating an OA. Consequently, instead of directly searching for the optimization in (4), GOSS efficiently uses an OA to approximate its solution.

3. Optimality of OA for Linear Mixed Model

An OA of strength 2 on s levels is a matrix with combinatorial orthogonality, that is, entries of the matrix come from a fixed finite set of s levels, arranged in such a way that all ordered pairs of the levels appear equally often in every selection of two columns of the matrix. For a comprehensive introduction to OA, see Hedayat, Sloane, and Stufken (1999). In this article, we consider $s = 2$, and denote the two levels by -1 and 1. Here is an example of 4×3 orthogonal array, where each of the ordered pairs $\{(-1,-1),(-1,1),(1,-1),(1,1)\}$ occurs once:

$$\begin{pmatrix} -1 & -1 & -1 \\ -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}.$$

The combinatorial orthogonality of OA is actually a type of balance that ensures that all columns are considered fairly and rows distributed dissimilarly to cover as much different information as possible. It has been shown that any OA with combinatorial orthogonality is simultaneously D- and A-optimal under a first-order linear model (Dey and Mukerjee 2009). These optimality properties of OA have been used in Wang et al. (2021) for subsampling problems under linear models.

Recall that in (4), for linear mixed model, the D-optimality criterion selects subdata that minimizes the determinant $|\mathbf{M}^{*-1}|$, that is, maximizes $|\mathbf{M}^*|$. Notice that $\mathbf{V}^* = \mathrm{diag}\{\mathbf{V}_i^*\}_{i=1}^R$, with $\mathbf{V}_i^* = \mathrm{cov}(\mathbf{Y}_i^*)$ being the covariance matrix for the ith group, we thus can decompose $\mathbf{M}^*$ in (3) by

$$\mathbf{M}^* = \sum_{i=1}^R \mathbf{X}_i^{*T} \mathbf{V}_i^{*-1} \mathbf{X}_i^* = \sum_{i=1}^R \mathbf{M}_i^*,$$

where $\mathbf{M}_i^* = \mathbf{X}_i^{*T} \mathbf{V}_i^{*-1} \mathbf{X}_i^*$ is the information matrix for the ith group of the subdata. We first study the optimal $\mathbf{X}_i^*$ to maximize $|\mathbf{M}_i^*|$ when the number of points in $\mathbf{X}_i^*$ is given. To facilitate the presentation of the theoretical results below, without loss of generality, we assume that each covariate in $\mathbf{Z}_i$ has been scaled to $[-1, 1]$.

Lemma 1. For $i = 1, 2, \ldots, R$, let n_i be the number of points in $\mathbf{X}_i^*$ and $\gamma_i = \sigma_E^2/(\sigma_E^2 + n_i\sigma_A^2)$, then

$$|\mathbf{M}_i^*| \leqslant \gamma_i \left(\frac{n_i}{\sigma_E^2}\right)^p,$$

with equality if and only if $\mathbf{Z}_i^*$ forms a two-level OA with n_i runs.

Lemma 1 shows that given the number of points in $\mathbf{Z}_i^*$, it should form an OA to maximize $|\mathbf{M}_i^*|$. To find the subdata that maximizes $|\mathbf{M}^*|$, we are concerned about two questions. First, following Lemma 1, does aggregating the OA subdata in each group maximize the overall information $|\mathbf{M}^*|$? Second, what are the optimal settings for n_i, $i = 1, \ldots, R$? The following theorem, guiding our later algorithm, answers the two questions.

Theorem 1. For a subdata set $\mathbf{X}^*$ with n points, $\mathbf{M}^*$ in (3) satisfies that

$$|\mathbf{M}^*| \leqslant \frac{n^{p-1}}{\sigma_E^{2p}} \left[\sum_{i=1}^R \gamma_i n_i\right] \leqslant \frac{Rn^p}{\sigma_E^{2(p-1)}(R\sigma_E^2 + n\sigma_A^2)}, \quad (5)$$

where n_i is the number of points of the ith group in $\mathbf{X}_i^*$ and $\gamma_i = \sigma_E^2/(\sigma_E^2 + n_i\sigma_A^2)$. In addition, (i) the first equality in (5) holds when each $\mathbf{Z}_i^*$ forms a two-level OA, and further, (ii) the second equality holds if and only if the runsize of each OA selected from each group is equal, that is, $n_1 = \cdots = n_R$.

By Theorem 1, the D-optimal subdata should have a group orthogonality, that is, equal-sized groups with each group forming an OA. The following result shows that such group-orthogonal subdata is also A-optimal.

Theorem 2. For a subdata set $\mathbf{X}^*$ with n points, $\mathbf{M}^*$ in (3) satisfies that

$$\mathrm{tr}(\mathbf{M}^{*-1}) \geqslant \sigma_E^2 \left(\frac{1}{\sum_{i=1}^R \gamma_i n_i} + \frac{p-1}{n}\right) \quad (6)$$

$$\geqslant \frac{1}{n}\left(p\sigma_E^2 + \frac{n}{R}\sigma_A^2\right), \quad (7)$$

where (i) the equality in (6) holds when each $\mathbf{Z}_i^*$ forms a two-level OA, and (ii) the equality in (7) holds if and only if the runsize of each OA selected from each group is equal, that is, $n_1 = \cdots = n_R$.

Theorems 1 and 2 suggest selecting the group-orthogonal subdata for fitting linear mixed models. It is also worth noting that the optimal subdata is independent of σ_A^2 and σ_E^2. That is, we do not need to estimate σ_A^2 and σ_E^2 before subsampling, which further simplifies our calculation. To this end, we propose the GOSS algorithm, which is specifically designed for hierarchical data and holds for any σ_A^2 and σ_E^2.

4. Group-Orthogonal Subsampling

In this section, we propose the GOSS method. By the discussion in Section 3, the optimal subdata should have the same group size and form an OA in each group. Recall that Wang et al. (2021) introduced the OSS algorithm to select subdata that best approximates an OA. Hence, GOSS can employ OSS to select the subdata from each group. Specifically, we sequentially select data points from the ith group to minimize the discrepancy function:

$$L\left(\mathbf{Z}_i^*\right) = \sum_{1 \leqslant j < j' \leqslant n_i} \left[(p-1) - \left\|\mathbf{z}_{ij}^*\right\|^2 /2 - \left\|\mathbf{z}_{ij'}^*\right\|^2 /2 + \delta\left(\mathbf{z}_{ij}^*, \mathbf{z}_{ij'}^*\right) \right]^2, \tag{8}$$

where

$$\delta\left(\mathbf{z}_{ij}^*, \mathbf{z}_{ij'}^*\right) = \sum_{k=2}^{p} \delta_1\left(x_{ijk}^*, x_{ij'k}^*\right),$$

and $\delta_1(x, y)$ is 1 if both x and y have the same sign and 0 otherwise. The function $L\left(\mathbf{Z}_i^*\right)$ measures the distance between $\mathbf{Z}_i^*$ and an OA. Therefore, the subdata for the ith group obtained by minimizing (8) can well approximate an OA. The details of the OSS approach can be found in Section C of the Appendix.

Other than the orthogonality within each group, GOSS needs to make sure that the group size of the selected subdata are balanced. Therefore, for the desired subdata size n, we choose $m = n/R$ points from each group. After we have subdata from all groups, we aggregate all the subdata and obtain the GLS estimator for a linear mixed model. Algorithm 1 outlines the proposed GOSS algorithm.

Algorithm 1 GOSS algorithm

Input: Full data $\mathbf{Z} = \left(\mathbf{Z}_1^T, \ldots, \mathbf{Z}_R^T\right)^T$, $\mathbf{Y} = \left(\mathbf{Y}_1^T, \ldots, \mathbf{Y}_R^T\right)^T$, subdata size n

Output: The subdata-based GLS estimator of $\breve{\boldsymbol{\beta}}^*$

for $i = 1$ to R **do**

Let $m = n/R$. Use the OSS method to minimize the discrepancy function in (8) and select a subdata of size m from group i, denoted as $\left\{\mathbf{Z}_i^*, \mathbf{Y}_i^*\right\}$

end for

Aggregate the R subdata sets as $\mathbf{Z}^* = \left(\mathbf{Z}_1^{*T}, \ldots, \mathbf{Z}_R^{*T}\right)^T$ and $\mathbf{Y}^* = \left(\mathbf{Y}_1^{*T}, \ldots, \mathbf{Y}_R^{*T}\right)^T$. Let $\hat{\sigma}_A^2$ and $\hat{\sigma}_E^2$ be consistent estimators of σ_A^2 and σ_E^2 based on the selected data $\mathbf{X}^* = (\mathbf{1}_n, \mathbf{Z}^*)$ and $\mathbf{Y}^*$. Estimate the coefficient $\boldsymbol{\beta}$ using

$$\breve{\boldsymbol{\beta}}^* = (\mathbf{X}^{*T}\hat{\mathbf{V}}^{*-1}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\hat{\mathbf{V}}^{*-1}\mathbf{Y}^*, \tag{9}$$

where $\hat{\mathbf{V}}^* = \hat{\sigma}_E^2\mathbf{I}_n + \hat{\sigma}_A^2\mathbf{A}^*$ and $\mathbf{A}^*$ is a block diagonal matrix with R blocks of $\mathbf{1}_m\mathbf{1}_m^T$.

Remark 1. The restriction of Algorithm 1 that $m = n/R$ is an integer is mostly for convenience. In the case that $m = n/R$ is not an integer, we may use a combination of $\lfloor m \rfloor$ and $\lceil m \rceil$ to keep the subdata size as n.

Remark 2. We use the method of moments approach proposed by Gao and Owen (2017) (refer to Section D in the Appendix) to estimate σ_A^2 and σ_E^2 in our numerical results in Sections 5 and 6. From Theorem 1 of Gao and Owen (2020), the moment method estimators based on GOSS subdata are consistent with variances

$$\text{var}(\hat{\sigma}_A^2) = O(R^{-1}) \text{ and } \text{var}(\hat{\sigma}_E^2) = O(m^{-1}).$$

Remark 3. The computation in Algorithm 1 is mostly involved in OSS in each group, so the time complexity of Algorithm 1 is $O(Np \ln m)$ (Wang et al. 2021). In addition, Algorithm 1 is naturally suited for distributed and parallel computing. We can simultaneously process each group of the full data, which will dramatically accelerate the subsampling process.

Compared to OSS, GOSS offers two main novel advantages. First, GOSS suggests that subsampling should be groupwise for hierarchical data, and the group size of the subdata should be the same. This is to ensure that the contribution of groups in the subdata are balanced. OSS, by contrast, directly subsamples the full data, resulting in unbalanced contributions from groups. Second, compared to OSS, which only ensures the combinatorial orthogonality of the entire subdata, GOSS further ensures the combinatorial orthogonality of the subdata in each group. This groupwise orthogonality adds an additional layer of value to the subdata. As detailed in the theory presented in Section 3, it will significantly benefit the fitting of a linear mixed model.

Next, we discuss the asymptotic behavior of the slope estimator. Let $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_{-1}^T)^T$, where β_1 is the intercept and $\boldsymbol{\beta}_{-1}$ the slope parameter. In practice, we are typically more interested in the estimation of $\boldsymbol{\beta}_{-1}$. Write the $\breve{\boldsymbol{\beta}}^*$ in (9) as $\breve{\boldsymbol{\beta}}^* = (\breve{\beta}_1^*, \breve{\boldsymbol{\beta}}_{-1}^{*T})^T$. We next study the asymptotic normality of $\breve{\boldsymbol{\beta}}_{-1}^*$ as an estimator of $\boldsymbol{\beta}_{-1}$. Write the subdata design matrix $\mathbf{Z}_i^*$ from each group as

$$\mathbf{Z}_i^* = \mathbf{L}_i^* + \mathbf{D}_i^*,$$

where $\mathbf{L}_i^*$ is a two-level OA, and $\mathbf{D}_i^*$ is the difference between $\mathbf{Z}_i^*$ and $\mathbf{L}_i^*$. Let $\mathbf{D}^* = (\mathbf{D}_1^{*T}, \ldots, \mathbf{D}_R^{*T})^T$ and $||\mathbf{D}^*||_\infty$ be the entrywise max norm, that is, the maximum absolute value of the entries in $\mathbf{D}^*$. We have the following theorem.

Theorem 3. For a fixed number of groups R, suppose that the maximum norm of $\mathbf{D}^*$ is $||\mathbf{D}^*||_\infty = o(1)$ as $n = Rm \to \infty$, $\mathrm{E}|e_{ij}^3| < \infty$, and $\hat{\sigma}_A^2$ and $\hat{\sigma}_E^2$ are consistent estimators of σ_A^2 and σ_E^2, respectively. For the estimator of the slope parameter in (9), $\breve{\boldsymbol{\beta}}_{-1}^*$, we have

$$\sqrt{n}\left(\breve{\boldsymbol{\beta}}_{-1}^* - \boldsymbol{\beta}_{-1}\right) \xrightarrow{d} N(\mathbf{0}, \sigma_E^2\mathbf{I}_{p-1}), \quad as\ n \to \infty,$$

where "$\xrightarrow{d}$" denotes convergence in distribution.

Theorem 3 indicates that the slope estimator based on a GOSS subdata is asymptotically normal with a covariance matrix $\sigma_E^2\mathbf{I}_{p-1}$ and an average variance σ_E^2, which is the smallest possible average variance for an estimator of $\boldsymbol{\beta}_{-1}$. Because the subdata size n is typically finite, the smaller asymptotic variance guarantees that the estimator based on a GOSS subdata is more accurate than other subdata.

5. Simulation Studies

In this section, we evaluate the performance of GOSS with simulation studies. Let the number of groups $R = 20$. The first

10 groups have the same data size, and the last 10 groups have the same data size, that is, $C_1 = \cdots = C_{10}$ and $C_{11} = \cdots = C_{20}$. Four cases are considered to generate the design matrix $\mathbf{Z} = (\mathbf{z}_{ij,k})$ of the full data for $j = 1, \ldots, C_i$, $i = 1, \ldots, 20$, and $k = 2, \ldots, p$. Cases 1 and 2 consider homogeneous data, where data in all groups are from an identical distribution. Cases 3 and 4 consider heterogeneous data with different group means, simulating heterogeneity among the groups. Specifically, we consider the following settings:

Case 1. The covariates $\mathbf{z}_{ij}$'s are independent and follow a multivariate uniform distribution: $\mathbf{z}_{ij,k} \sim U[-1, 1]$, $k = 2, \ldots, p$.

Case 2. The covariates $\mathbf{z}_{ij}$'s follow a multivariate normal distribution: $\mathbf{z}_{ij} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, with

$$\boldsymbol{\Sigma} = \left(0.5^{I(k \neq k')}\right), k, k' = 2, \ldots, p.$$

Case 3. The covariates $\mathbf{z}_{ij}$'s follow a uniform distribution: $\mathbf{z}_{ij,k} \sim U[\theta_{i1}, \theta_{i2}]$, where $U[\theta_{i1}, \theta_{i2}]$ is a shift of $U[-1, 1]$ such that the centers of groups vary within $\{-0.5, -0.45, \ldots, 0.45\}$. Thus, we set $\theta_{i1} = -1 + (i - 11)/20$ and $\theta_{i2} = 1 + (i - 11)/20$.

Case 4. The covariates $\mathbf{z}_{ij}$'s follow a multivariate normal distribution: $\mathbf{z}_{ij} \sim N(\mu_i \mathbf{1}, \boldsymbol{\Sigma})$, with μ_i varying within $\{-2, -1.8, \ldots, 1.8\}$.

The response data are generated from the linear mixed model (1) with the true value of $\boldsymbol{\beta}$ being a 51×1 vector of unity which includes an intercept and 50 slope parameters, so $p = 51$. The error term is generated from $e_{ij} \sim N(0, 9)$. We consider two settings of the random effect, namely, $a_i \sim N(0, 0.5)$ and $a_i \sim t(3)$, to illustrate the impact of the distribution and variance of the random effect. Here $a_i \sim N(0, 0.5)$ simulates smaller random effects and thus lower correlations between responses within groups, while $a_i \sim t(3)$ simulates larger random effects and higher correlations within groups.

5.1. Comparison of Performance

The simulation is repeated for $B = 200$ times. We compare the following different subsampling methods: UNIF (simple random subsampling with uniform weights), LEV (leveraging subsampling), IBOSS, OSS, GUNIF (Group-UNIF), GLEV (Group-LEV), GIBOSS (Group-IBOSS), and GOSS. The

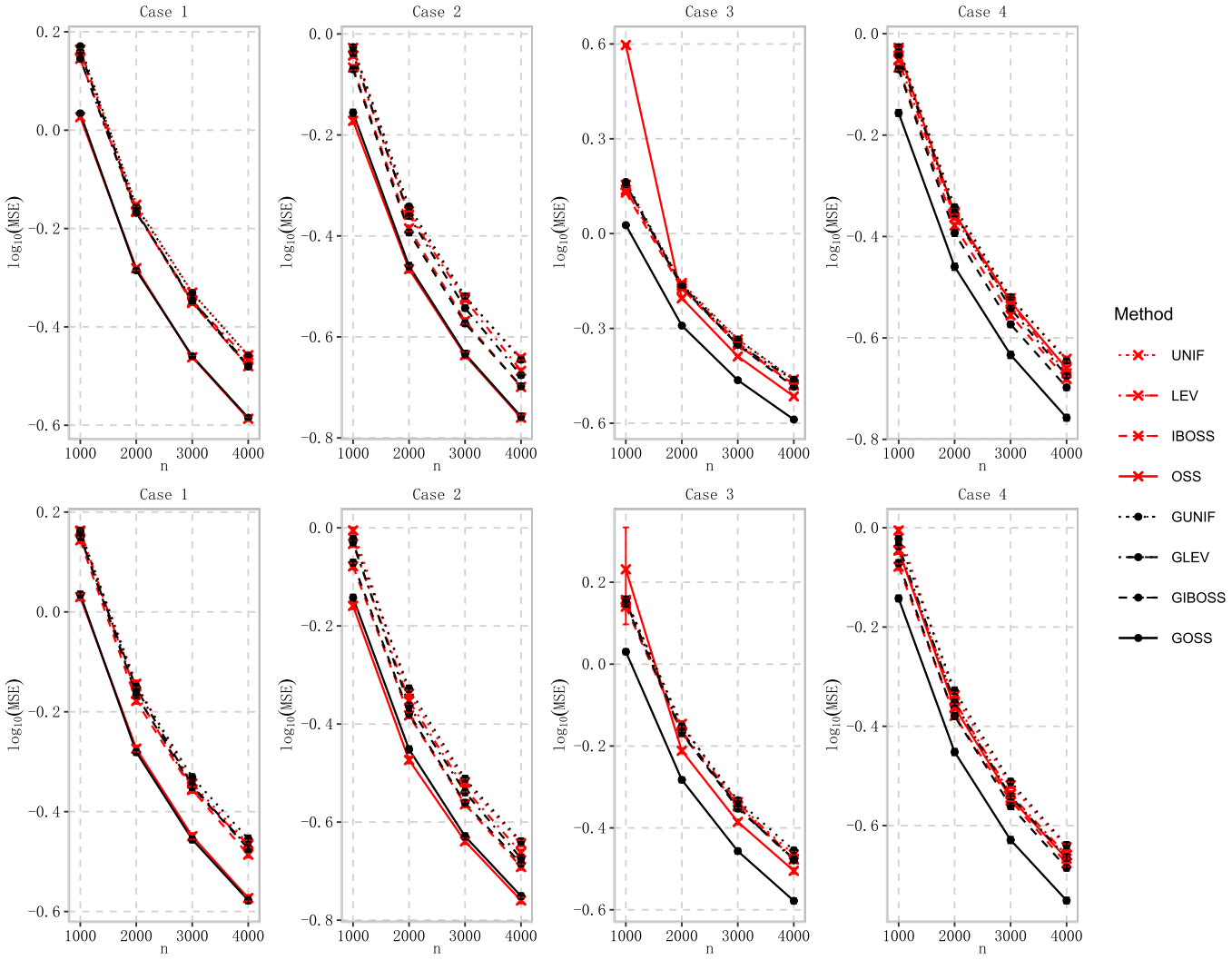

Figure 1. The $\log_{10}$(MSE) of the estimated slope parameters for different subdata sizes n. The upper panels are for $a_i \sim N(0, 0.5)$ and the lower panels for $a_i \sim t(3)$. The full data size is $N = 1.5 \times 10^5$. The bars represent standard errors obtained from 200 replicates. Some bars are very narrow, so they seem to be invisible.

GUNIF, GLEV, and GIBOSS methods select the same number of data from each group using the UNIF, LEV, and IBOSS methods, respectively. We compare these three methods with the GOSS algorithm to demonstrate that the optimality of GOSS is not merely attributed to the balance of subdata sizes among groups, but also to the orthogonality of the subdata within each group. For each subsampling method, we consider the empirical MSE of the slope parameters:

$$\text{MSE} = B^{-1}\sum_{b=1}^{B}||\breve{\boldsymbol{\beta}}_{-1}^{*(b)} - \boldsymbol{\beta}_{-1}||^2, \tag{10}$$

where $\breve{\boldsymbol{\beta}}_{-1}^{*(b)}$ is the GLS estimator of $\boldsymbol{\beta}_{-1}$ based on subdata in the bth repetition.

We first consider the setting of $C_1 = \cdots = C_{10} = 5\times 10^3$ and $C_{11} = \cdots = C_{20} = 2C_1$, resulting in a fixed full data size of $N = 1.5\times 10^5$. Since σ_A^2 and σ_E^2 are unknown in practice, we estimate them based on subdata using the moment method proposed by Gao and Owen (2017) and plug them into the estimator $\breve{\boldsymbol{\beta}}_{-1}^{*(b)}$. Figure S4 in Appendix shows the $\log_{10}$(MSE) of $\hat{\sigma}_A^2$ and $\hat{\sigma}_E^2$ with respect to subdata sizes $n = 10^3, 2\times 10^3, 3\times 10^3$, and 4×10^3 when $a_i \sim N(0, 0.5)$. We observe that all the subdata tend to provide reliable estimates for σ_A^2 and σ_E^2, except for OSS in Case 3 when the subdata size is small ($n = 10^3$).

With $\hat{\sigma}_A^2$ and $\hat{\sigma}_E^2$, Figure 1 plots the $\log_{10}$(MSE) of the plug-in estimator $\breve{\boldsymbol{\beta}}_{-1}^{*(b)}$ with respect to n. For Cases 1 and 2, grouped methods perform similarly to their counterparts because groups are identically distributed, and GOSS and OSS outperform other methods due to the orthogonality of the subdata. For Cases 3 and 4, however, the performance of GOSS dominates all other methods for every subdata size n, although all methods decrease at the same rate. It should be noted that GUNIF, GLEV, and GIBOSS do not outperform their counterparts, indicating that the advantages of the GOSS method go beyond the balancing of group sizes, and within-group orthogonality is crucial in determining its superiority. Moreover, the fact that the GOSS method outperforms other methods in both the upper and lower panels of Figure 1 demonstrates that GOSS is powerful regardless of the size of random effects.

We also consider the performance of GOSS for different full data sizes and show the result in Figure 2. We consider $C_1 = \cdots = C_{10} \in \{10^3, 5\times 10^3, 2.5\times 10^4, 1.25\times 10^5\}$ and $C_{11} = \cdots = C_{20} = 2C_1$, which results in the full data size $N \in \{3\times 10^4, 1.5\times 10^5, 7.5\times 10^5, 3.75\times 10^6\}$. The subdata size

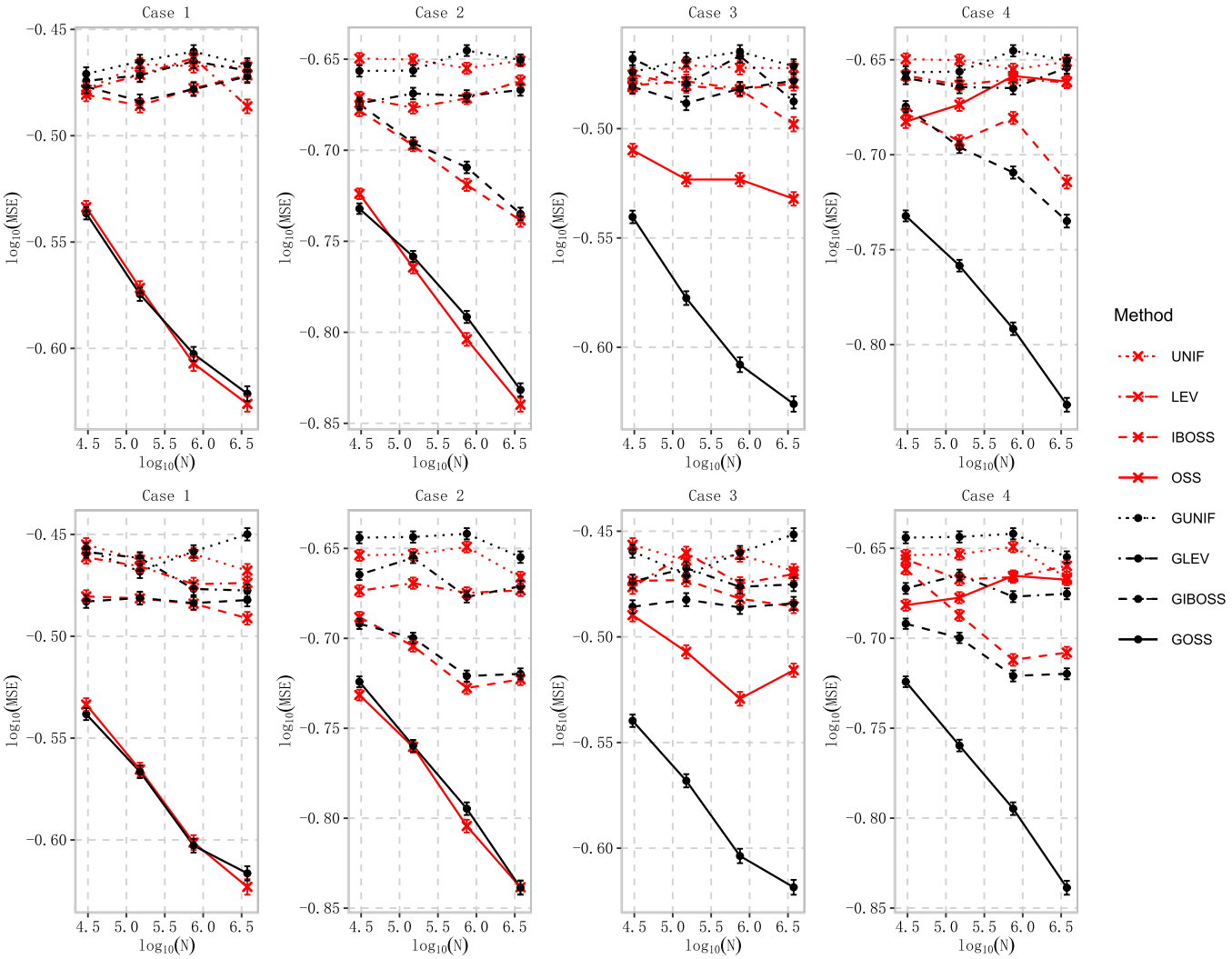

Figure 2. The $\log_{10}$(MSE) of the estimated slope parameters for different full data sizes N. The subdata size is fixed at $n = 4\times 10^3$. The upper panels are for $a_i \sim N(0, 0.5)$, and the lower panels for $a_i \sim t(3)$. The bars represent standard errors obtained from 200 replicates.

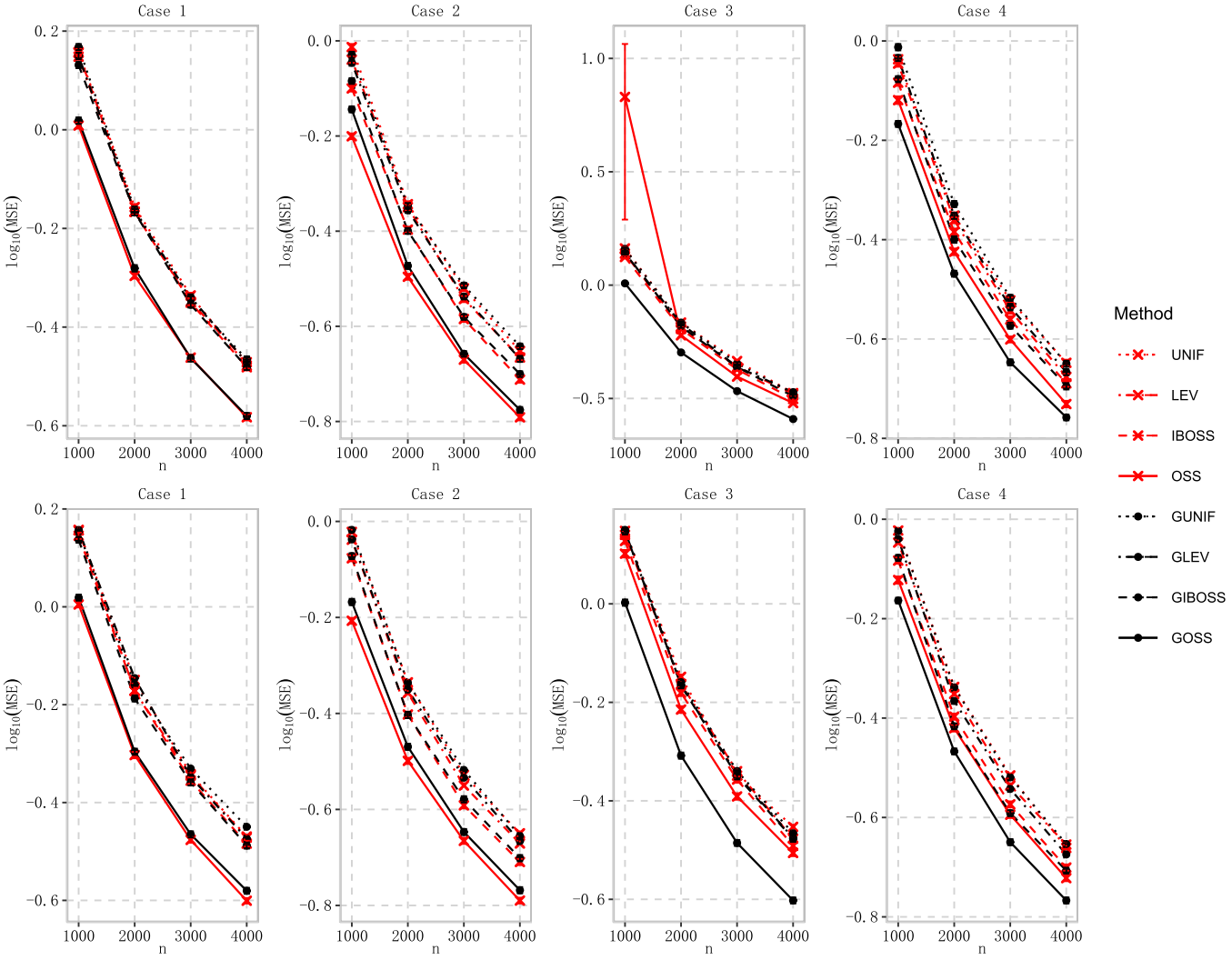

Figure 3. The $\log_{10}$(MSE) of the estimated slope parameters for different subdata sizes n. The upper panels are for $a_i \sim N(0, 0.5)$ and the lower panels for $a_i \sim t(3)$. The full data size is $N = 5.5 \times 10^5$. The bars represent standard errors obtained from 200 replicates. Some bars are very narrow and may be invisible.

is fixed at $n = 4 \times 10^3$. As evidenced by Figure 2, for Cases 1 and 2, grouped methods perform similarly to their counterparts, and both GOSS and OSS exhibit outstanding performance and fast decreasing MSEs as N increases, meaning that they can both extract more information from the full data as the size of the full data increases. For Case 3, OSS fails to extract more information as N increases because of the heterogeneity of the full data, but GOSS keeps its fast decreasing trend and outperforms all other methods significantly. For Case 4, the GOSS method retains its remarkable superiority, even though the IBOSS and GIBOSS also exhibit a slow decreasing trend.

We further examine the performance of GOSS when there is an extreme imbalance among group sizes in full data. To this end, we change the setting of C_i to $C_1 = \cdots = C_{10} = 5\times10^3$ and $C_{11} = \cdots = C_{20} = 10C_1 = 5\times10^4$. Figure S5 in Appendix plots $\log_{10}$(MSE) for $\hat{\sigma}_A^2$ and $\hat{\sigma}_E^2$ with respect to the subdata size n, and Figure 3 shows the $\log_{10}$(MSE) for $\breve{\boldsymbol{\beta}}_{-1}^{*(b)}$ versus n. The GOSS still outperforms all other methods for Cases 3 and 4 because of its balance among groups and within-group orthogonality, which still provides more information even though the group sizes of the full data are extremely unbalanced.

To see the performance of GOSS when the full data size grows and is extremely imbalanced, we further consider $C_1 = \cdots = C_{10} \in \{10^3, 5 \times 10^3, 2.5 \times 10^4\}$ and $C_{11} = \cdots = C_{20} = 10C_1$, with the full data size $N \in \{1.1 \times 10^5, 5.5 \times 10^5, 2.75 \times 10^6\}$. The subdata size is again fixed at $n = 4 \times 10^3$. According to Figure 4, all subsampling methods behave similarly as in Figure 2. One point to note is that for Case 2, the grouped methods appear to be slightly inferior to their counterparts, mainly because of the homogeneous and overlapping information in all groups of the full data. In this case, drawing the same amount of information from each group can result in missing more important information in bigger groups. For Cases 3 and 4, the superiority of GOSS is attributed to the balance of heterogeneous groups, which contain information from different aspects. The balance among these groups enables more accurate modeling and parameter estimation, resulting in a fast downward trend and improved performance.

We have also conducted simulations to evaluate the performance of subsampling methods in estimating the intercept and predicting the response over the full data. Possible model misspecification has also been considered. Due to page limitations, the results are deferred to Section B of the Appendix.

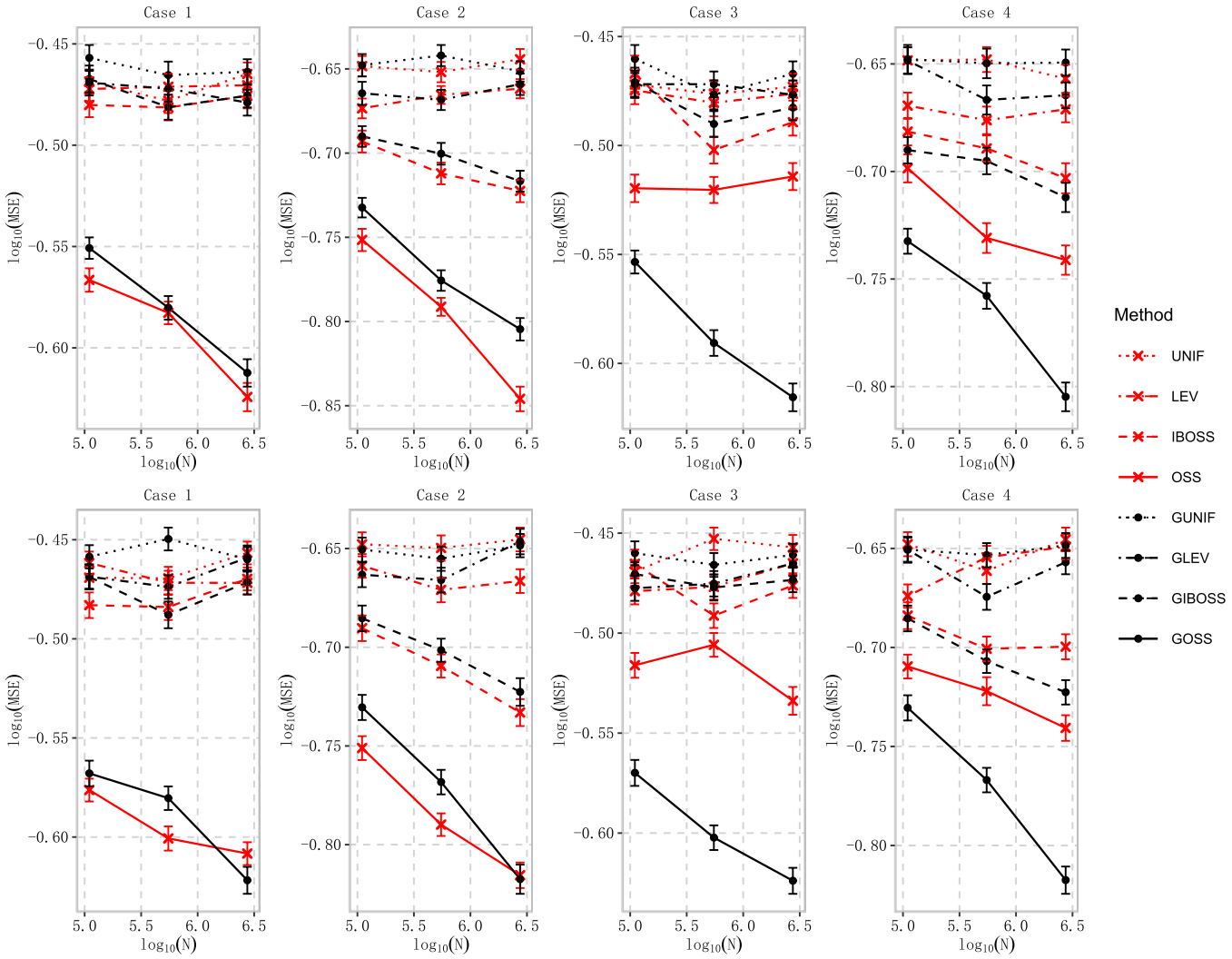

Figure 4. The $\log_{10}$(MSE) of the estimated slope parameters for different full data sizes N, when there is an extreme imbalance in the data sizes among groups. The subdata size is fixed at $n = 4 \times 10^3$. The upper panels are for $a_i \sim N(0, 0.5)$, and the lower panels for $a_i \sim t(3)$. The bars represent standard errors obtained from 200 replicates.

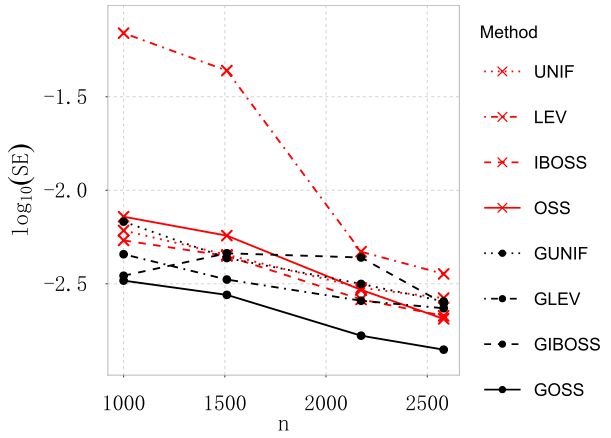

Figure 5. The $\log_{10}$(SE) of $\breve{\boldsymbol{\beta}}^*_{-1}$ with different subdata sizes for the accelerometer dataset.

5.2. Computing Time

Table 1 reports the computation times (including the selection of subdata and the computation of estimators of $\boldsymbol{\beta}$, in seconds) under the setting of $C_1 = \cdots = C_{10} = 5 \times 10^3$, $C_{11} = \cdots = C_{20} = 2C_1$, $p = 6, 51$, and 101, and $n = 10^3$. Covariates are generated as in Case 3 and the random effect $a_i \sim N(0, 0.5)$. The times shown in Table 1 are the mean wall-clock times of 200 repetitions, with each wall-clock time measured using the function **Sys.time**() in **R**. All computations are carried out on a laptop running Windows 10 21H2 with a 3.00GHz Intel Core i7 processor and 16GB memory. As indicated in Table 1, the grouped methods are more time-efficient than the ungrouped method. UNIF and GUNIF require the least computation time as expected. The GOSS is faster than LEV, OSS, and IBOSS and is comparable to GLEV and GIBOSS. Table 2 reports the computation times for different full data sizes N with a fixed dimension $p = 51$ and a fixed subdata size $n = 1000$. The GOSS is faster than LEV, OSS, and GIBOSS and is comparable to IBOSS and GLEV for all full data sizes.

6. Real Data Analysis-Accelerometer Dataset

We analyze the accelerometer dataset to evaluate the performance of the GOSS approach. The data records the vibration of the cooler fan with weights on its blades, which allows us to infer

Table 1. The wall-clock times (in seconds) of subsampling methods with $n = 10^3$.

Method	UNIF	LEV	IBOSS	OSS	GUNIF	GLEV	GIBOSS	GOSS
$p = 6$	0.2240	0.2297	0.2001	0.2602	0.0883	0.1431	0.1313	0.1373
$p = 51$	0.6006	1.2980	1.5579	1.8271	0.3936	0.8973	0.9745	0.8799
$p = 101$	0.9349	3.6877	2.9458	3.6636	0.7431	1.7489	1.8859	1.7723

Table 2. The wall-clock times (in seconds) of subsampling methods with $p = 51$.

Method	UNIF	LEV	IBOSS	OSS	GUNIF	GLEV	GIBOSS	GOSS
$N = 3 \times 10^4$	0.1347	0.1925	0.2984	0.5159	0.0981	0.1868	0.1867	0.1837
$N = 7.5 \times 10^5$	1.2927	2.7587	1.8938	2.8484	0.6611	2.0032	2.7679	1.9937
$N = 3.75 \times 10^6$	6.3441	14.3277	8.8961	11.3674	3.0972	9.3464	17.8353	9.3434

Table 3. The real data wall-clock times (in seconds) of subsampling methods with $n = 1000$.

Method	UNIF	LEV	IBOSS	OSS	GUNIF	GLEV	GIBOSS	GOSS	Full
Time	0.1400	0.2133	0.4142	0.5555	0.1491	0.2844	0.3319	0.2997	426

when the motor failed. To generate different vibration scenarios, the experimenters set 17 different cooler fan speeds ranging from 20% to 100% of the maximum fan speed at 5% intervals. Vibrations were measured by accelerometers at a frequency of 20 milliseconds, with vibration measurements taking 1 min at each speed and generating 3000 recordings at each frequency. Thus, a total of $N = 153{,}000$ vibration records were collected. Further details about the data can be found at Scalabrini Sampaio et al. (2019). At each speed, the accelerometer measures 9000 observations of vibration on x, y, and z axes. We grouped the data according to the 17 different cooler fan speeds. Thus, the number of groups is $R = 17$. For each speed, the vibration on the z axis varies with the vibration on the x and y axes. We take the x and y axes as independent variables and the z axis as the response variable to assess the impact of x and y axes vibrations on the z axis. We thus consider the model

$$z_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 y_{ij} + a_i + e_{ij},\ i = 1, \ldots, 17, j = 1, \ldots, 9000, \tag{11}$$

where a_i denotes the random effect of the cooler fan speed, and e_{ij} is the random error of the response at the same speed.

We consider subdata sizes $n = 1000, 1510, 2173$, and 2581 and assess subsampling methods by examining the difference between the estimator derived from subdata and the estimator obtained from the full data. That is, we consider the squared error (SE)

$$SE = ||\check{\boldsymbol{\beta}}^*_{-1} - \hat{\boldsymbol{\beta}}_{-1}||^2,$$

where $\hat{\boldsymbol{\beta}}_{-1}$ is the GLS estimator of the slope parameter $\boldsymbol{\beta}_{-1} = (\beta_1, \beta_2)^T$ based on the full data, and $\check{\boldsymbol{\beta}}^*_{-1}$ is the estimator from subdata. For the methods UNIF, LEV, GUNIF, and GLEV, we repeat them 200 times due to their randomness and calculate the average SE. OSS, IBOSS, GOSS, and GIBOSS are deterministic methods and are executed only once. Figure 5 plots the SE for different subsampling methods. It is clear that GOSS outperforms all other methods for all subdata sizes in terms of minimizing the SE. Further, the SE for GOSS decreases fast as the subdata size increases, which suggests that GOSS allows a better estimation of the impact of x and y axes vibration on the vibration of the z axis.

Table 3 shows the wall-clock times (average over 200 repetitions) of different subsampling methods for the accelerometer data with $n = 1000$. The comparison in Table 3 is similar with that in Table 1, which shows that GOSS is faster than OSS and IBOSS and is comparable to LEV, GLEV, and GIBOSS.

7. Concluding Remarks

In this article, we present a novel subsampling method called GOSS, which is designed for selecting subdata from large datasets with a hierarchical structure. GOSS achieves data size balance among groups and combinatorial orthogonality within each group, ensuring that the selected subdata is D- and A-optimal for the GLS estimator of a linear mixed model. Extensive simulations and a real-world application demonstrate that GOSS outperforms existing methods in minimizing the MSE of the estimator for the slope parameter, especially in cases where data groups are heterogeneous. Theoretical results establish that the estimator obtained from the GOSS subdata has the minimum variance among all possible subdata, as evidenced by its asymptotic distribution. Additionally, GOSS is faster than competing methods, making it a highly efficient option for accelerating the analysis of big data using a linear mixed model.

Particular aspects associated with this research require more extensive and thorough studies. First, GOSS is developed for scenarios where the full dataset has a fixed number of groups, with the sample size in each group tending toward infinity. However, in real-world applications, we may encounter situations where the number of groups tends toward infinity, while the sample size of each group remains limited. Subsampling methods that can handle this scenario require further study. Second, we have only considered a constant within-group variance for convenience, but it is also common to have varying within-group variances, and addressing this issue is of pressing concern for future research. Third, the data within each group may be sparse or incomplete due to missing values. Investigating suitable subsampling methods to handle sparse and incomplete data is another valuable avenue for exploration.

Supplementary Materials

Title: Proofs of theoretical results and related materials

Appendix: provides proofs of the theoretical results in the article, additional numerical results, the OSS algorithm, and an estimation method for σ_A^2 and σ_E^2.

Code and Data File: provides R code and data to replicate our results and apply the method to other dataset. The R code and data are described in a readme file.

Acknowledgments

The authors would like to thank the Editor, the Associate Editor, and two anonymous reviewers for providing helpful comments on earlier drafts of the manuscript.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This work was supported by the NSFC Grant (Nos. 12371259, 11971098), the Fundamental Research Funds for the Central Universities(2412023YQ003) and the National Key Research and Development Program of China (Nos. 2020YFA0714102, 2022YFA1003701).

ORCID

Jiaqing Zhu http://orcid.org/0000-0001-9008-9093
Lin Wang http://orcid.org/0000-0003-0888-6232
Fasheng Sun http://orcid.org/0000-0003-2410-4018

References

Ai, M., Wang, F., Yu, J., and Zhang, H. (2021a), "Optimal Subsampling for Large-Scale Quantile Regression," *Journal of Complexity*, 62, 101512. [1037]

Ai, M., Yu, J., Zhang, H., and Wang, H. (2021b), "Optimal Subsampling Algorithms for Big Data Regressions," *Statistica Sinica*, 31, 749–772. [1037]

Bates, D. (2014), "Computational Methods for Mixed Models," in *LME4: Mixed-Effects Modeling with R*, pp. 99–118. [1037]

Bennett, J., and Lanning, S. (2007), "The Netflix Prize," in *Proceedings of KDD Cup and Workshop* (Vol. 2007), p. 35, New York, NY, USA. [1037]

Dey, A., and Mukerjee, R. (2009), *Fractional Factorial Plans*, New York: Wiley. [1039]

Fan, Y., Liu, Y., and Zhu, L. (2021), "Optimal Subsampling for Linear Quantile Regression Models," *Canadian Journal of Statistics*, 49, 1039–1057. [1037]

Gao, K., and Owen, A. (2017), "Efficient Moment Calculations for Variance Components in Large Unbalanced Crossed Random Effects Models," *Electronic Journal of Statistics*, 11, 1235–1296. [1037,1040,1042]

Gao, K., and Owen, A. (2020), "Estimation and Inference for Very Large Linear Mixed Effects Models," *Statistica Sinica*, 30, 1741–1771. [1037,1038,1040]

He, L., and Hung, Y. (2022), "Gaussian Process Prediction Using Design-based Subsampling," *Statistica Sinica*, 32, 1165–1186. [1037]

Hedayat, A. S., Sloane, N. J. A., and Stufken, J. (1999), *Orthogonal Arrays: Theory and Applications*, New York: Springer. [1039]

Jiang, J., and Nguyen, T. (2007), *Linear and Generalized Linear Mixed Models and their Applications* (Vol. 1), New York: Springer. [1037,1038]

Kiefer, J. C. (1959), "Optimum Experimental Designs," *Journal of the Royal Statistical Society*, Series B, 21, 272–319. [1039]

Li, T., and Meng, C. (2020), "Modern Subsampling Methods for Large-Scale Least Squares Regression," *International Journal of Cyber-Physical Systems (IJCPS)*, 2, 1–28. [1037]

Ma, P., and Sun, X. (2015), "Leveraging for Big Data Regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, 7, 70–76. [1037]

Mak, S., and Joseph, R. V. (2018), "Support Points," *The Annals of Statistics*, 46, 2562–2592. [1037]

McCulloch, C., and Searle, S. (2004), *Generalized, Linear, and Mixed Models*, New York: Wiley. [1037]

Meng, C., Xie, R., Mandal, A., Zhang, X., Zhong, W., and Ma, P. (2021), "Lowcon: A Design-based Subsampling Approach in a Misspecified Linear Model," *Journal of Computational and Graphical Statistics*, 30, 694–708. [1037]

Meng, C., Yu, J., Chen, Y., Zhong, W., and Ma, P. (2022), "Smoothing Splines Approximation Using Hilbert Curve Basis Selection," *Journal of Computational and Graphical Statistics*, 31, 802–812. [1037]

Meng, C., Zhang, X., Zhang, J., Zhong, W., and Ma, P. (2020), "More Efficient Approximation of Smoothing Splines via Space-Filling basis Selection," *Biometrika*, 107, 723–735. [1037]

Raudenbush, S. (1993), "A Crossed Random Effects Model for Unbalanced Data with Applications in Cross-Sectional and Longitudinal Research," *Journal of Educational Statistics*, 18, 321–349. [1037]

Ren, M., and Zhao, S. (2021), "Subdata Selection based on Orthogonal Array for Big Data," *Communications in Statistics-Theory and Methods*, 52, 5483–5501. [1037]

Ren, M., Zhao, S., and Wang, M. (2023), "Optimal Subsampling for Least Absolute Relative Error Estimators with Massive Data," *Journal of Complexity*, 74, 101694. [1037]

Scalabrini Sampaio, G., Vallim Filho, A. R. d. A., Santos da Silva, L., and Augusto da Silva, L. (2019), "Prediction of Motor Failure Time Using an Artificial Neural Network," *Sensors*, 19, 4342. [1045]

Shao, L., Song, S., and Zhou, Y. (2022), "Optimal Subsampling for Large-Sample Quantile Regression with Massive Data," *Canadian Journal of Statistics*, 51, 420–443. [1037]

Shi, C., and Tang, B. (2021), "Model-Robust Subdata Selection for Big Data," *Journal of Statistical Theory and Practice*, 15, 1–17. [1037]

Sun, X., Zhong, W., and Ma, P. (2021), "An Asymptotic and Empirical Smoothing Parameters Selection Method for Smoothing Spline Anova Models in Large Samples," *Biometrika*, 108, 149–166. [1037]

Wang, H., and Ma, Y. (2021), "Optimal Subsampling for Quantile Regression in Big Data," *Biometrika*, 108, 99–112. [1037]

Wang, H., Yang, M., and Stufken, J. (2019), "Information-Based Optimal Subdata Selection for Big Data Linear Regression," *Journal of the American Statistical Association*, 114, 393–405. [1037]

Wang, L. (2022), "Balanced Subsampling for Big Data with Categorical Covariates," arXiv preprint arXiv:2212.12595. [1037]

Wang, L., Elmstedt, J., Wong, W. K., and Xu, H. (2021), "Orthogonal Subsampling for Big Data Linear Regression," *Annals of Applied Statistics*, 15, 1273–1290. [1037,1038,1039,1040]

Xie, R., Bai, S., and Ma, P. (2023), "Optimal Sampling Designs for Multi-Dimensional Streaming Time Series with Application to Power Grid Sensor Data," *arXiv preprint arXiv:2303.08242. Annals of Applied Statistics (online)*. [1037]

Yu, J., Ai, M., and Ye, Z. (2023), "A Review on Design Inspired Subsampling for Big Data," *Statistical Papers*, 1–44. [1037]

Yu, J., and Wang, H. (2022), "Subdata Selection Algorithm for Linear Model Discrimination," *Statistical Papers*, 63, 1883–1906. [1037]

Yu, J., Wang, H., Ai, M., and Zhang, H. (2022), "Optimal Distributed Subsampling for Maximum Quasi-Likelihood Estimators with Massive Data," *Journal of the American Statistical Association*, 117, 265–276. [1037]

Zhang, Y., Wang, L., Zhang, X., and Wang, H. (2023), "Independence-Encouraging Subsampling for Nonparametric Additive Models," arXiv preprint arXiv:2302.13441. [1037]