Journal of Data Science 21 (1), 158–172
January 2023

DOI: 10.6339/22-JDS1057
Statistical Data Science

Sampling-based Gaussian Mixture Regression for Big Data

JooChul Lee[1], Elizabeth D. Schifano[1], and HaiYing Wang[1,*]

[1]*Department of Statistics, University of Connecticut, Storrs, CT 06269, USA*

Abstract

This paper proposes a nonuniform subsampling method for finite mixtures of regression models to reduce large data computational tasks. A general estimator based on a subsample is investigated, and its asymptotic normality is established. We assign optimal subsampling probabilities to data points that minimize the asymptotic mean squared errors of the general estimator and linearly transformed estimators. Since the proposed probabilities depend on unknown parameters, an implementable algorithm is developed. We first approximate the optimal subsampling probabilities using a pilot sample. After that, we select a subsample using the approximated subsampling probabilities and compute estimates using the subsample. We evaluate the proposed method in a simulation study and present a real data example using appliance energy data.

Keywords *EM algorithm; massive data; optimal probabilities; supsampling*

1 Introduction

A finite mixture of regression (FMR) model is a statistical tool widely used in various fields such as econometrics, psychology, genetics, marketing, and engineering (e.g., McLachlan and Peel, 2004). FMR models can identify heterogeneous groups in the population and examine the linear relationships between a response and a set of covariates for different groups. For statistical estimation, the maximum likelihood estimator (MLE) is mainly used to estimate unknown parameters. Since the MLE has no closed-form solution, iterative numerical algorithms are often implemented. The expectation-maximization (EM) (Dempster et al., 1977) is a classic approach for conducting maximum likelihood estimation in FMR models.

The EM algorithm, however, can cause an excessive computing burden for fitting FMR models on massive data. Subsampling is a technique to reduce the computational task by using a subsample extracted from the full data. In the linear regression framework, Drineas et al. (2006) and Ma et al. (2014) developed sampling algorithms that use statistical leverage scores of the covariate matrix to specify non-uniform subsampling probabilities. Wang et al. (2019) proposed a deterministic sub-data selection method, which does not involve random sampling. Wang et al. (2018) proposed optimal subsampling strategies in logistic regression. They defined optimal subsampling probabilities by minimizing the asymptotic mean squared error (MSE) of the subsample-based estimator, and extracted sub-data from the full data using approximated optimal subsampling probabilities to obtain estimates. Recently, the optimal subsampling approach based on Wang et al. (2018) has been extended to different model settings, including multinomial logistic regression (e.g., Yao and Wang, 2019), generalized linear models (GLMs) (e.g., Ai et al., 2021b; Lee et al., 2021), quantile regression (e.g., Wang and Ma, 2021; Ai et al.,

[*] Corresponding author. Email: haiying.wang@uconn.edu.

Received May 29, 2022; Accepted July 2, 2022

2021a), quasi-likelihood models (e.g., Yu et al., 2022), and additive hazards models (e.g., Zuo et al., 2021).

In this paper, we develop optimal subsampling strategies for Gaussian FMR models. We first suggest a general estimator based on a subsample and establish its asymptotic normality. To select informative data points, we specify optimal subsampling probabilities by minimizing the asymptotic MSE of the resultant estimators. However, since the specified probabilities depend on unknown parameters, they cannot be calculated directly. To handle this challenge, we propose a two-step algorithm. In the first step, we use a pilot sample to approximate the optimal subsampling probabilities. In the second step, we select a subsample based on the approximated subsampling probabilities and run the EM-Algorithm to obtain estimates using the subsample.

The rest of the paper is organized as follows. In Section 2, we first briefly review Gaussian mixture regression, and then present subsample-based estimation under the Gaussian FMR model. Asymptotic results of the resulting estimators are investigated, and the two-step algorithm to conduct the optimal subsampling strategy is presented. Section 3 assesses the performance of the proposed methods using simulated and real data sets. Section 4 summarizes the paper. Proofs of the theoretical results are collected in Supplementary materials.

2 Mixture Regression Models and Optimal Subsampling Strategy

2.1 Finite Mixture of Gaussian Linear Regressions

In this section, we review a finite mixture of Gaussian linear regressions. Suppose that y is a response and $\mathbf{x}$ is a d dimensional covariate with the first entry being one. The conditional density function of y given $\mathbf{x}$ is

$$f(y|\mathbf{x};\boldsymbol{\theta}) = \sum_{j=1}^{J} p_j f_j(y|\mathbf{x};\boldsymbol{\beta}_j,\sigma_j), \tag{1}$$

where J is a given number of components, p_j's are the component weights satisfying $p_j > 0$ for each j and $\sum_{j=1}^{J} p_j = 1$, $f_j(y|\mathbf{x};\boldsymbol{\beta}_j,\sigma_j)$ is the density of a normal distribution with mean $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_j$ and variance σ_j^2, $\boldsymbol{\beta}_j$ is a $d \times 1$ vector of unknown regression coefficients including an intercept, and $\boldsymbol{\theta} = (\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_J,\sigma_1,\dots,\sigma_J,p_1,\dots,p_{J-1})$. For identifiability of the finite mixture of Gaussian linear regressions, we assume that the density function $f(y|\mathbf{x};\boldsymbol{\theta})$ has common support on $(\mathbf{x}, y)$ and is identifiable in $\boldsymbol{\theta}$ up to the permutation of the components of the mixture. The maximum likelihood estimator (MLE), $\hat{\boldsymbol{\theta}}$, for the unknown parameter $\boldsymbol{\theta}$ is the maximizer of the following log-likelihood,

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{J} p_j f_j(y_i|\mathbf{x}_i;\boldsymbol{\beta}_j,\sigma_j)\right). \tag{2}$$

The EM algorithm is an iterative algorithm that can be used to optimize (2). Define that z_{ij} is equal to one if y_i belongs to the jth component and zero otherwise for $i = 1,\dots,n$, and write $\mathbf{z}_i = (z_{i1},\dots,z_{iJ})$. Then, the EM algorithm can be used to find the MLE by maximizing the complete-data log-likelihood of $\{(\mathbf{x}_i,\mathbf{z}_i,y_i) : i = 1,\dots,n\}$,

$$\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{J} z_{ij} \log\left\{p_j f_j(y_i|\mathbf{x}_i;\boldsymbol{\beta}_j,\sigma_j)\right\}. \tag{3}$$

In the expectation step (E-step), we calculate the conditional expectation of the complete-data log-likelihood given the current parameter estimates and the observed data,

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \sum_{i=1}^{n}\sum_{j=1}^{J} \tau_{ij}^{(s)} \log\{p_j f_j(y_i|\mathbf{x}_i; \boldsymbol{\beta}_j, \sigma_j)\},$$

where

$$\tau_{ij}^{(s)} = \frac{p_j^{(s)} f_j(y_i|\mathbf{x}_i; \boldsymbol{\beta}_j^{(s)}, \sigma_j^{(s)})}{\sum_{k=1}^{J} p_k^{(s)} f_k(y_i|\mathbf{x}_i; \boldsymbol{\beta}_k^{(s)}, \sigma_k^{(s)})}.$$

In the maximization step (M-step), we update the estimates by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$. For $j = 1, \dots, J$, the updated estimates are

$$\begin{aligned}
\hat{p}_j^{(s+1)} &= \sum_{i=1}^{n} \frac{\tau_{ij}^{(s)}}{n},\\
\hat{\boldsymbol{\beta}}_j^{(s+1)} &= \left(\sum_{i=1}^{n} \tau_{ij}^{(s)} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right)^{-1} \sum_{i=1}^{n} \tau_{ij}^{(s)} y_i \mathbf{x}_i,\\
\hat{\sigma}_j^{2(s+1)} &= \sum_{i=1}^{n} \tau_{ij}^{(s)} (y_i - \mathbf{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_j^{(s)})^2 \Big/ \sum_{i=1}^{n} \tau_{ij}^{(s)}.
\end{aligned}$$

The computing time which is required until convergence is $O(\xi J n d^2)$, where ξ is the number of iterations.

2.2 Estimation and Optimal Subsampling Strategy

2.2.1 Subsample-based Estimation

Since iterative calculations for enormous data cause excessive computational burden, we consider subsample-based estimation to obtain the parameter estimates in this section. We denote the full data as $\mathcal{D}_n = \{(\mathbf{x}_i, y_i) : i = 1, \dots, n\}$. Let $\{\pi_i\}_{i=1}^n$ be the subsampling probabilities assigned to all observations satisfying $\sum_{i=1}^n \pi_i = 1$.

We consider a random subsample of size r selected from the full data $\mathcal{D}_n$ based on the subsampling probabilities. Then, the subsampling estimator $\tilde{\boldsymbol{\theta}}$ can be obtained by maximizing the target function

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^{r} \frac{1}{\pi_i^*} \log\left(\sum_{j=1}^{J} p_j f_j(y_i^*|\mathbf{x}_i^*; \boldsymbol{\beta}_j, \sigma_j)\right), \tag{4}$$

where $\mathbf{x}_i^*$'s, y_i^*'s and π_i^*'s, are covariates, responses, and subsampling probabilities in the subsample, respectively. The EM algorithm can be applied to optimize the target function in (4). The details of the algorithm are presented in Algorithm 1.

Now, we derive the asymptotic distribution of $\tilde{\boldsymbol{\theta}}$. The following assumptions are needed to establish it.

Assumption 1. *For $\boldsymbol{\theta}$ in a neighborhood of true parameter $\boldsymbol{\theta}_t = (\boldsymbol{\beta}_1^t, \dots, \boldsymbol{\beta}_J^t, \sigma_1^t, \dots, \sigma_J^t, p_1^t, \dots, p_{J-1}^t)$, $\mathbf{M} = n^{-1}\partial^2\ell(\boldsymbol{\theta})/(\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}})$ goes to a positive-definite matrix in probability as $n \to \infty$.*

Algorithm 1 EM Algorithm for target function (4).

Estimates can be obtained by maximizing the sampled complete-data target function,

$$\ell_c^*(\boldsymbol{\theta}) = \sum_{i=1}^{r} \frac{1}{\pi_i^*} \sum_{j=1}^{J} z_{ij}^* \log\left\{p_j f_j(y_i^*|x_i^*)\right\},$$

where z_{ij}^* is equal to one if y_i^* belongs to the jth component and zero otherwise.
E-step: Given the current estimate $\boldsymbol{\theta}^{(s)}$,

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \sum_{i=1}^{r} \frac{1}{\pi_i^*} \sum_{j=1}^{J} \tau_{ij}^{*(s)} \log p_j f_j(y_i^*|\mathbf{x}_i^*; \boldsymbol{\beta}_j, \sigma_j),$$

where

$$\tau_{ij}^{*(s)} = \frac{p_j^{(s)} f_j(y_i^*|\mathbf{x}_i^*; \boldsymbol{\beta}_j^{(s)}, \sigma_j^{(s)})}{\sum_{k=1}^{J} p_k^{(s)} f_k(y_i^*|\mathbf{x}_i^*; \boldsymbol{\beta}_k^{(s)}, \sigma_k^{(s)})}.$$

M-step: Updates the estimate $\boldsymbol{\theta}^{(s+1)}$ by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$. For $j = 1, \ldots, J$,

$$\hat{p}_j^{(s+1)} = \left(\sum_{i=1}^{r} \frac{1}{\pi_i^*}\right)^{-1} \sum_{i=1}^{n} \frac{\tau_{ij}^{*(s)}}{\pi_i^*},$$

$$\hat{\boldsymbol{\beta}}_j^{(s+1)} = \left(\sum_{i=1}^{r} \frac{\tau_{ij}^{*(s)} \mathbf{x}_i^* \mathbf{x}_i^{*\mathrm{T}}}{\pi_i^*}\right)^{-1} \sum_{i=1}^{n} \frac{\tau_{ij}^{*(s)} y_i^* \mathbf{x}_i^*}{\pi_i^*},$$

$$\hat{\sigma}_j^{2(s+1)} = \left(\sum_{i=1}^{r} \frac{\tau_{ij}^{*(s)}}{\pi_i^*}\right)^{-1} \sum_{i=1}^{n} \frac{\tau_{ij}^{*(s)} (y_i^* - \mathbf{x}_i^{*\mathrm{T}} \hat{\boldsymbol{\beta}}_j^{(s)})^2}{\pi_i^*}.$$

Repeat until convergence.

Assumption 2. *$n^{-2}\sum_{i=1}^{n} \pi_i^{-1}\|\mathbf{x}_i\|^8 = O_P(1)$ and $n^{-2}\sum_{i=1}^{n} \pi_i^{-1}(y_i - \boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}_i)^8$ is uniformly bounded for $\boldsymbol{\beta}_j$ and j=1,...,J.*

Assumption 3. *The parameter space $\boldsymbol{\Theta}$ is compact. For any three elements θ_j, θ_k, and θ_l of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the third partial derivative $\left|\partial^3 \ell^*(\boldsymbol{\theta}; \mathbf{x}, y)/(\partial\theta_j \partial\theta_k \partial\theta_l)\right|$ is bounded by an integrable function $B(\mathbf{x}, y)$.*

Assumption 4. *Denote $\dot{\ell}_i(\boldsymbol{\theta}) = \partial \log f(y_i|\mathbf{x}_i; \boldsymbol{\theta})/\partial\boldsymbol{\theta}$. There exists some $\delta > 0$ such that $n^{-2-\delta}\pi_i^{-1-\delta}\sum_{i=1}^{n} \|\dot{\ell}_i(\boldsymbol{\theta})\|^{(2+\delta)} = O_P(1)$.*

Assumption 1 is a condition to guarantee that the log-likelihood function is convex for large n, and Assumptions 2 and 3 are conditions on the subsampling probabilities. Assumption 4 is needed for the Lindeberg-Feller Central Limit Theorem.

Theorem 1. *Let $\tilde{\boldsymbol{\theta}}$ be the maximizer of (4). Under the Assumptions 1-4, as $r, n \to \infty$, if $r/n = o(1)$,*

$$\sqrt{r}\mathbf{V}^{-1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_t) \longrightarrow N(\mathbf{0}, \mathbf{I}) \quad \textit{in distribution}, \tag{5}$$

where

$$\mathbf{V} = \mathbf{M}_t^{-1}\mathbf{V}_\pi\mathbf{M}_t^{-1}, \ \mathbf{M}_t = \frac{1}{n}\frac{\partial^2 \ell(\boldsymbol{\theta}_t)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}},$$

and

$$\mathbf{V}_\pi = \sum_{i=1}^{n} \frac{\dot{\ell}_i(\boldsymbol{\theta}_t)\dot{\ell}_i(\boldsymbol{\theta}_t)^{\mathrm{T}}}{n^2\pi_i}.$$

2.2.2 Optimal Subsampling Probability and Two-step Algorithm

In this section, we specify the optimal subsampling probabilities based on the result in Theorem 1. First, we consider minimizing the trace of $\mathbf{V}$, tr$(\mathbf{V})$, which can be viewed as minimizing the asymptotic MSE of $\tilde{\boldsymbol{\theta}}$.

Theorem 2. *The optimal subsampling probabilities that minimize $tr(\mathbf{V})$ are*

$$\pi_i^{\mathbf{V}} = \frac{\left\|\mathbf{M}_t^{-1}\dot{\ell}_i(\boldsymbol{\theta}_t)\right\|}{\sum\limits_{k=1}^{n}\left\|\mathbf{M}_t^{-1}\dot{\ell}_i(\boldsymbol{\theta}_t)\right\|}, \quad i = 1, \ldots, n. \tag{6}$$

The computing time takes $O\{[J(d+2)-1]^2 n\}$ to calculate the optimal subsampling probabilities presented in Theorem 2.

We further consider minimizing the asymptotic MSE of linearly transformed subsample estimators to alleviate the computation time. We assign the optimal subsampling probabilities by minimizing the trace of $\mathbf{V}_\pi$ which is equivalent to minimizing the asymptotic MSE of $\mathbf{M}\tilde{\boldsymbol{\theta}}$. In addition to that, we also focus on the asymptotic MSE of the coefficient estimator $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_J)$. Denote $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$, $\boldsymbol{\theta}_{-\boldsymbol{\beta}} = (\sigma_1, \ldots, \sigma_J, p_1, \ldots, p_{J-1})$,

$$\mathbf{M}_{t,11} = \frac{1}{n}\frac{\partial^2 \ell(\boldsymbol{\theta}_t)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}, \quad \mathbf{M}_{t,12} = \frac{1}{n}\frac{\partial^2 \ell(\boldsymbol{\theta}_t)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}_{-\boldsymbol{\beta}}^{\mathrm{T}}}, \quad \text{and } \mathbf{M}_{t,22} = \frac{1}{n}\frac{\partial^2 \ell(\boldsymbol{\theta}_t)}{\partial\boldsymbol{\theta}_{-\boldsymbol{\beta}}\partial\boldsymbol{\theta}_{-\boldsymbol{\beta}}^{\mathrm{T}}}.$$

From the result of Theorem 1, we can derive

$$\sqrt{r}\mathbf{V}_\beta^{-1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_t) \longrightarrow N(\mathbf{0}, \mathbf{I}) \quad \text{in distribution},$$

where

$$\mathbf{V}_\beta = (\mathbf{M}_{\boldsymbol{\beta}}^{inv}, \mathbf{M}_{\boldsymbol{\theta}_{-\boldsymbol{\beta}}}^{inv})\mathbf{V}_\pi(\mathbf{M}_{\boldsymbol{\beta}}^{inv}, \mathbf{M}_{\boldsymbol{\theta}_{-\boldsymbol{\beta}}}^{inv})^{\mathrm{T}}, \ \mathbf{M}_{\boldsymbol{\beta}}^{inv} = (\mathbf{M}_{t,11} - \mathbf{M}_{t,12}\mathbf{M}_{t,22}^{-1}\mathbf{M}_{t,12}^{\mathrm{T}})^{-1},$$

$$\mathbf{M}_{\boldsymbol{\theta}_{-\boldsymbol{\beta}}}^{inv} = -\mathbf{M}_{\boldsymbol{\beta}}^{inv}\mathbf{M}_{t,12}\mathbf{M}_{t,22}^{-1},$$

and $\boldsymbol{\beta}_t = (\boldsymbol{\beta}_1^t, \ldots, \boldsymbol{\beta}_J^t)$. We minimize the trace of $\mathbf{V}_\beta$ to determine the subsampling probabilities.

Theorem 3. *The optimal subsampling probabilities that minimize $tr(\mathbf{V}_\pi)$ are*

$$\pi_i^{\mathbf{V}_\pi} = \frac{\left\|\dot{\ell}_i(\boldsymbol{\theta}_t)\right\|}{\sum\limits_{k=1}^{n}\left\|\dot{\ell}_k(\boldsymbol{\theta}_t)\right\|}, \quad i = 1, \ldots, n, \tag{7}$$

Algorithm 2 Two-step Algorithm.

1. Draw a pilot sample of size r_0 with the uniform sampling probability to obtain an estimate $\tilde{\boldsymbol{\theta}}_0$. By replacing $\boldsymbol{\theta}$ with $\tilde{\boldsymbol{\theta}}_0$, approximate the optimal subsampling probabilities in (6), (7), or (8).
2. Draw a subsample of size r with replacement based on the approximated optimal subsampling probabilities in the previous step. Combine the pilot sample in the previous step with the subsample in the second step and run Algorithm 1 with the data of size $r_0 + r$ to obtain the estimate $\breve{\boldsymbol{\theta}}$.

and the optimal subsampling probabilities that minimize $tr(\mathbf{V}_\beta)$ *are*

$$\pi_i^{\mathbf{V}_\beta} = \frac{\left\| (\mathbf{M}_{\boldsymbol{\beta}}^{inv}, \mathbf{M}_{\boldsymbol{\theta}_{-\boldsymbol{\beta}}}^{inv}) \dot{\ell}_i(\boldsymbol{\theta}_t) \right\|}{\sum_{k=1}^{n} \left\| (\mathbf{M}_{\boldsymbol{\beta}}^{inv}, \mathbf{M}_{\boldsymbol{\theta}_{-\boldsymbol{\beta}}}^{inv}) \dot{\ell}_k(\boldsymbol{\theta}_t) \right\|}, \quad i = 1, \dots, n. \tag{8}$$

The calculations of the subsampling probabilities $\pi_i^{\mathbf{V}_\pi}$ and $\pi_i^{\mathbf{V}_\beta}$ take $O\{[J(d+2)-1]n\}$ and $O\{[(Jd)^2 + (2J-1)^2]n\}$ time, respectively. Compared to $\pi_i^{\mathbf{V}}$, they require less computing time.

Since the subsampling probabilities in (6), (7), and (8) depend on the unknown parameters, we propose an implementable two-step algorithm. In the first step, we approximate the proposed subsampling probabilities by replacing $\boldsymbol{\theta}$ with an estimate $\tilde{\boldsymbol{\theta}}_0$ obtained from a pilot sample. In the next step, we draw a subsample with replacement based on the approximated subsampling probabilities, and run Algorithm 1 with the subsample. We present the practical algorithm in Algorithm 2.

Remark 1. *Based on the standard error formula proposed by Wang et al. (2018), we suggest to estimate the variance-covariance matrix of the resultant estimator using* $\breve{\mathbf{V}} = \breve{\mathbf{M}}^{-1}\breve{\mathbf{V}}_\pi\breve{\mathbf{M}}^{-1}$ *for statistical inference, where*

$$\breve{\mathbf{M}} = \sum_{i=1}^{r_0+r} \frac{\ddot{\ell}_i^*(\breve{\boldsymbol{\theta}})}{n(r_0+r)\pi_i^*}, \quad \breve{\mathbf{V}}_\pi = \sum_{i=1}^{r_0+r} \frac{\dot{\ell}_i^*(\breve{\boldsymbol{\theta}})\dot{\ell}_i^*(\breve{\boldsymbol{\theta}})^{\mathrm{T}}}{n^2(r_0+r)(\pi_i^*)^2}, \quad \dot{\ell}_i^*(\boldsymbol{\theta}) = \frac{\partial \log f(y_i^*|\mathbf{x}_i^*; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

and $\ddot{\ell}_i^*(\boldsymbol{\theta}) = \partial \dot{\ell}_i^*(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$.

3 Numerical Examples

3.1 Simulation

In this section, we conduct a simulation study to assess the performance of the proposed method. We consider three different models.

Model 1. We examine a two-component Gaussian mixture regression model,

$$p_1 f_1(y_i|\mathbf{x}_i; \boldsymbol{\beta}_1, \sigma_1) + p_2 f_2(y_i|\mathbf{x}_i; \boldsymbol{\beta}_2, \sigma_2),$$

where $f_j(y|\mathbf{x}_i; \boldsymbol{\beta}_j, \sigma_j)$ is the density of a normal distribution with mean $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j$ and variance σ_j^2, and $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j}, \beta_{2j}, \beta_{3j})$ is a 4-dimensional vector with the intercept β_{0j} for $j = 1, 2$.

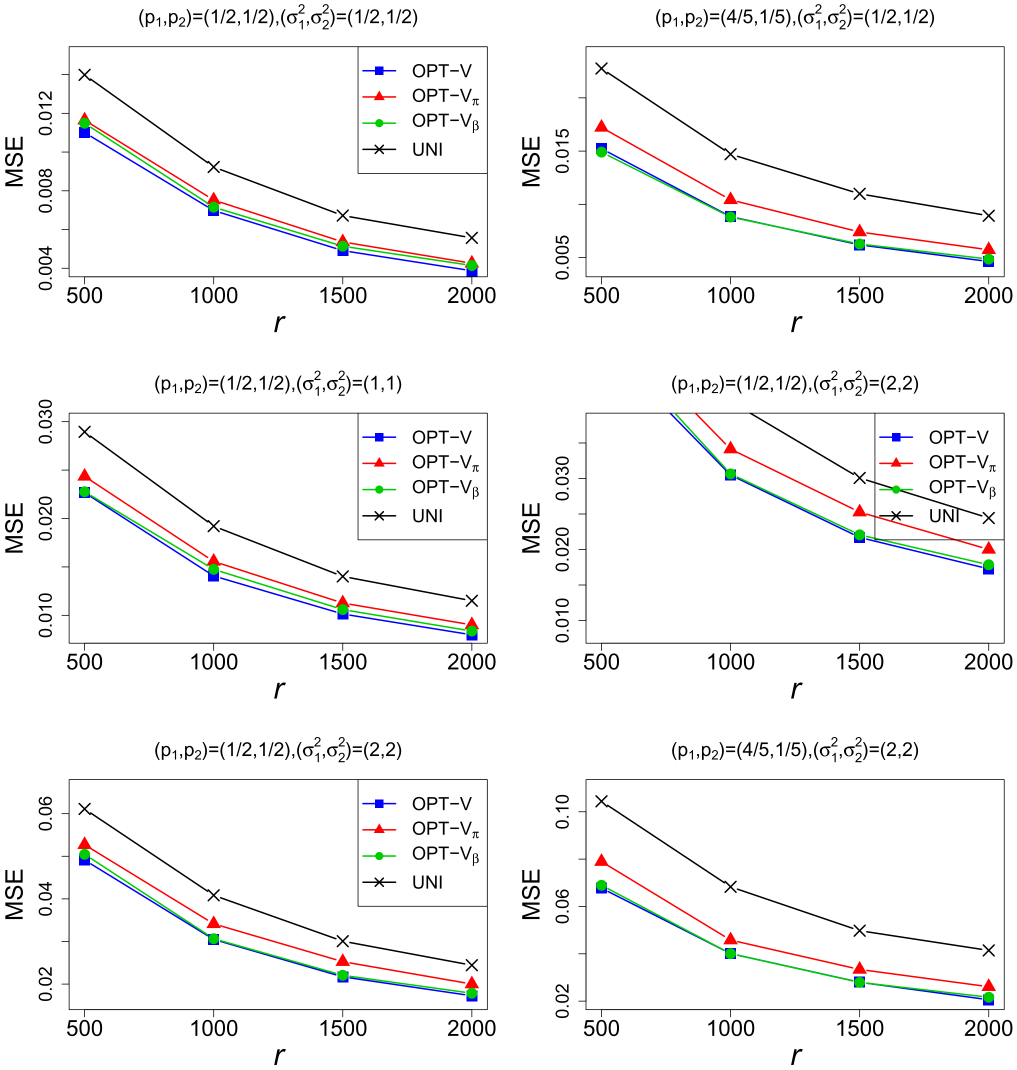

Figure 1: MSEs of Model 1 for varied (σ_1^2, σ_2^2), (p_1, p_2), and r. OPT-$\mathbf{V}$, OPT-$\mathbf{V}_\pi$, and OPT-$\mathbf{V}_\beta$ use $\pi_i^{\mathbf{V}}$, $\pi_i^{\mathbf{V}_\pi}$, and $\pi_i^{\mathbf{V}_\beta}$, respectively. UNI uses uniform subsampling probabilities.

Covariate $\mathbf{x}_i$ follows a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{\Sigma}$, $N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}_{lm} = 0.5^{I(l \neq m)}$ for $l, m = 1, \dots, 3$ and $I()$ is the indicator function. The true values of coefficients are $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \{(1, 1, 1, 1), (4, 4, 4, 4)\}$. We set $n = 10^5$, $(\sigma_1^2, \sigma_2^2) \in \{(1/2, 1/2), (1, 1), (2, 2)\}$, and $(p_1, p_2) \in \{(1/2, 1/2), (4/5, 1/5)\}$.

Model 2. Model setup is the same for Model 1, except for true values of coefficients, $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \{(1, 1, 1, 1), (-4, -4, -4, -4)\}$.

Model 3. We examine a three-component Gaussian mixture regression model,

$$p_1 f_1(y_i|\mathbf{x}_i; \boldsymbol{\beta}_1, \sigma_1) + p_2 f_2(y_i|\mathbf{x}_i; \boldsymbol{\beta}_2, \sigma_2) + p_3 f_3(y_i|\mathbf{x}_i; \boldsymbol{\beta}_3, \sigma_3),$$

where $f_3(y_i|\mathbf{x}_i; \boldsymbol{\beta}_3, \sigma_3)$ is the density of a normal distribution with mean $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_3$ and variance σ_3^2 and $\boldsymbol{\beta}_3$ is a 4-dimensional vector with an intercept. We consider the same distribution of the covariates

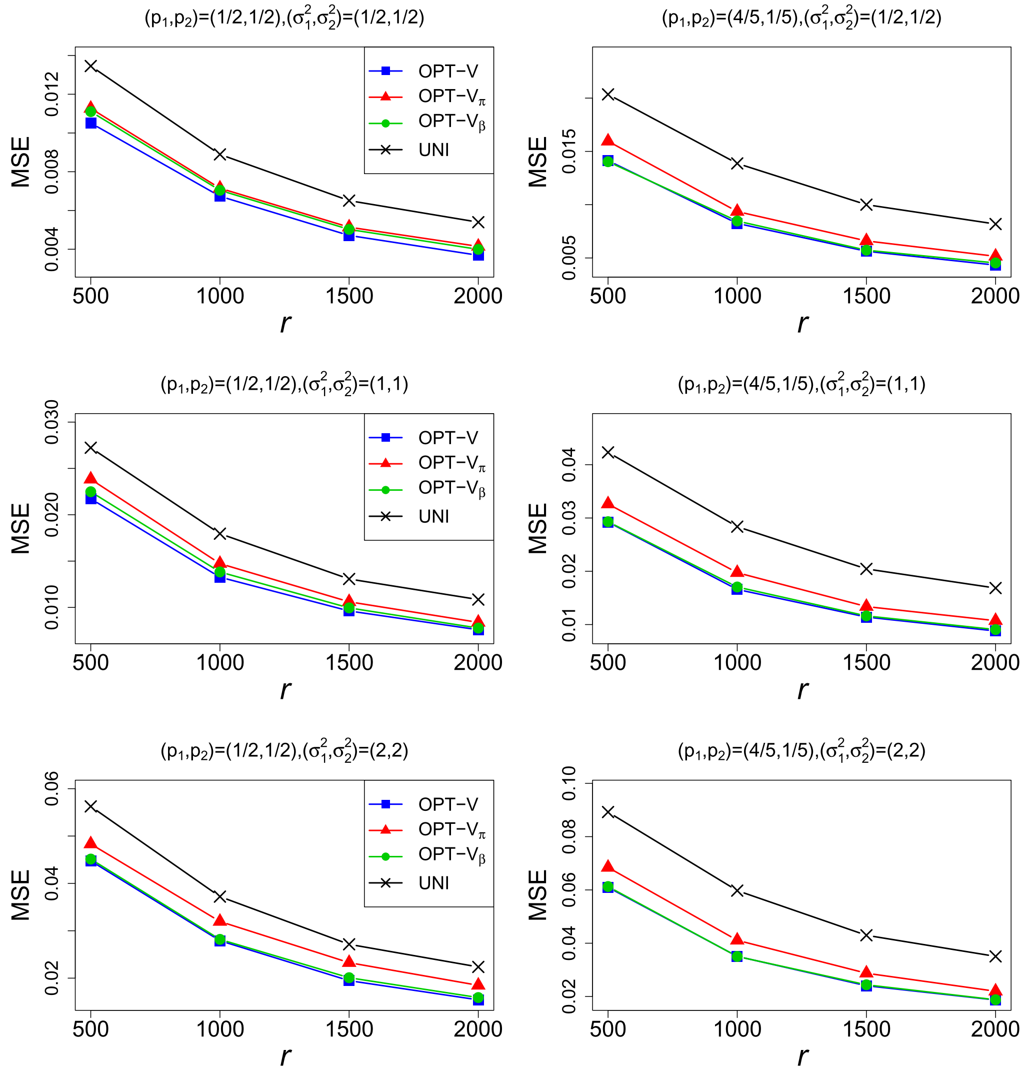

Figure 2: MSEs of Model 2 for varied (σ_1^2, σ_2^2), (p_1, p_2), and r. OPT-$\mathbf{V}$, OPT-$\mathbf{V}_\pi$, and OPT-$\mathbf{V}_\beta$ use $\pi_i^{\mathbf{V}}$, $\pi_i^{\mathbf{V}_\pi}$, and $\pi_i^{\mathbf{V}_\beta}$, respectively. UNI uses uniform subsampling probabilities.

as in Model 1. We set $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3) = \{(1, 1, 1, 1), (-4, -4, -4, -4), (4, 4, 4, 4)\}$, $(\sigma_1^2, \sigma_2^2, \sigma_3^2) \in \{(1/2, 1/2, 1/2), (1, 1, 1), (2, 2, 2)\}$, and $(p_1, p_2, p_3) \in \{(1/3, 1/3, 1/3), (1/2, 1/4, 1/4)\}$.

We calculate empirical MSEs based on $\sum_{k=1}^{K} \|\tilde{\boldsymbol{\theta}}_{(k)} - \boldsymbol{\theta}\|^2/K$ where K is the number of replications, and $\tilde{\boldsymbol{\theta}}_{(k)}$ is the estimate of $\boldsymbol{\theta}$ provided from the k-th subsample. For comparison, we consider the proposed subsampling probabilities, $\pi_i^{\mathbf{V}}$ (OPT-$\mathbf{V}$), $\pi_i^{\mathbf{V}_\pi}$ (OPT-$\mathbf{V}_\pi$), $\pi_i^{\mathbf{V}_\beta}$ (OPT-$\mathbf{V}_\beta$), and uniform subsampling probabilities (UNI). For UNI, subsamples of size $r_0 + r$ are used. We set $K = 1000$, $r_0 = 500$, and $r = 500, 1000, 1500, 2000$. For the initial values, k-means clustering is first conducted to form J groups. Then, we implement ordinary least square regressions for each group to obtain the initial values, $\boldsymbol{\beta}_j^{(0)}$ and $\sigma_j^{(0)}$ for $j = 1, 2, \ldots, J$. We set $J = 2$ for Model 1 and 2, and $J = 3$ for Model 3.

Figures 1, 2, and 3 present the simulation results for MSE. We observe that the proposed methods give better performance than UNI, with the smaller MSEs in all cases. OPT-$\mathbf{V}$ and

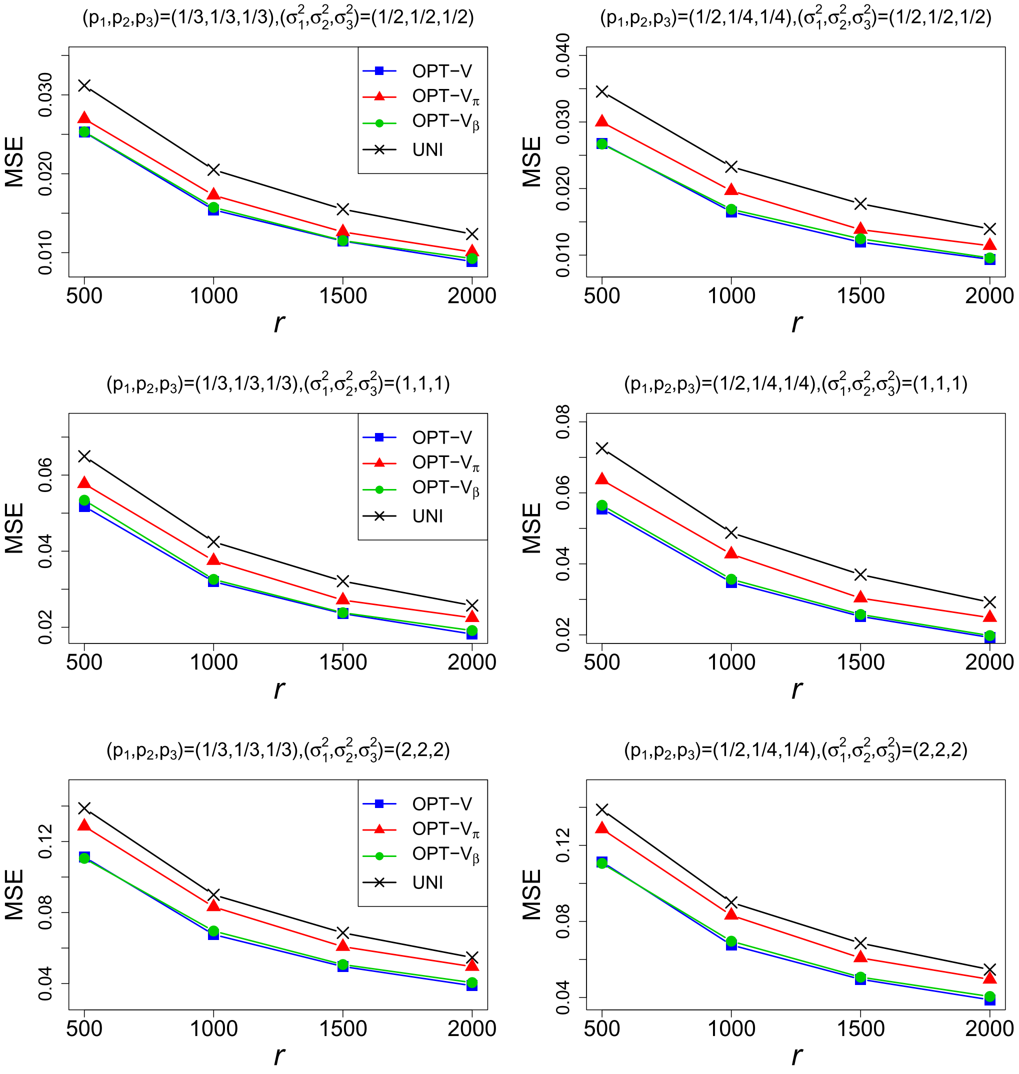

Figure 3: MSEs of Model 3 for varied $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$, (p_1, p_2, p_3), and r. OPT-$\mathbf{V}$, OPT-$\mathbf{V}_\pi$, and OPT-$\mathbf{V}_\beta$ use $\pi_i^{\mathbf{V}}$, $\pi_i^{\mathbf{V}_\pi}$, and $\pi_i^{\mathbf{V}_\beta}$, respectively. UNI uses uniform subsampling probabilities.

OPT-$\mathbf{V}_\beta$ show smaller MSEs than OPT-$\mathbf{V}_\pi$ since they minimize the asymptotic MSEs of the unknown parameter estimator $\tilde{\boldsymbol{\theta}}$ and the coefficient estimator $\tilde{\boldsymbol{\beta}}$, respectively. The MSEs of the proposed methods decrease when the subsample size r increases, which agrees with the asymptotic result of the resultant estimator.

We investigate the proposed standard error using the diagonal elements of the variance-covariance matrix in Remark 1. Using the formula, we estimate $\mathrm{tr}(\mathbf{V})$, i.e, the MSE of $\tilde{\boldsymbol{\theta}}$. Also, we calculate empirical variances for each of the estimators and then compute the sum of all these variances (EmpVar), and compare it with the average estimated MSE (AveMSE). Table 1 presents the results for AveMSE and EmpVar for Model 1. AveMSE provides similar results to EmpVar for all the cases.

We also examine components with different variances using data from Model 1. The variances $(\sigma_1^2, \sigma_2^2) \in \{(1/2, 1), (1, 2)\}$ are considered. Other settings are the same as those of Model 1.

Table 1: Average estimated MSE (AveMSE) and empirical variance (EmpVar) of Model 1 for varied (σ_1^2, σ_2^2), (p_1, p_2), and r.

	$(p_1, p_2) =$	(1/2, 1/2)				(4/5, 1/5)			
	$r =$	500	1000	1500	2000	500	1000	1500	2000
$(\sigma_1^2, \sigma_2^2) = (1/2, 1/2)$									
OPT-**V**	AveMSE	0.011	0.007	0.005	0.004	0.015	0.009	0.006	0.005
	EmpVar	0.011	0.007	0.005	0.004	0.015	0.009	0.006	0.005
OPT-$\mathbf{V}_\pi$	AveMSE	0.012	0.007	0.005	0.004	0.017	0.010	0.007	0.006
	EmpVar	0.012	0.007	0.005	0.004	0.017	0.010	0.007	0.006
OPT-$\mathbf{V}_\beta$	AveMSE	0.012	0.007	0.005	0.004	0.016	0.009	0.006	0.005
	EmpVar	0.011	0.007	0.005	0.004	0.015	0.009	0.006	0.005
$(\sigma_1^2, \sigma_2^2) = (1, 1)$									
OPT-**V**	AveMSE	0.023	0.014	0.010	0.008	0.033	0.019	0.013	0.010
	EmpVar	0.023	0.014	0.010	0.008	0.032	0.019	0.013	0.010
OPT-$\mathbf{V}_\pi$	AveMSE	0.025	0.016	0.011	0.009	0.037	0.022	0.015	0.012
	EmpVar	0.024	0.016	0.011	0.009	0.036	0.022	0.015	0.012
OPT-$\mathbf{V}_\beta$	AveMSE	0.024	0.015	0.011	0.008	0.033	0.019	0.013	0.010
	EmpVar	0.023	0.015	0.011	0.008	0.032	0.018	0.013	0.010
$(\sigma_1^2, \sigma_2^2) = (2, 2)$									
OPT-**V**	AveMSE	0.050	0.030	0.022	0.017	0.072	0.041	0.028	0.021
	EmpVar	0.049	0.030	0.022	0.017	0.068	0.040	0.028	0.021
OPT-$\mathbf{V}_\pi$	AveMSE	0.054	0.035	0.025	0.020	0.080	0.048	0.033	0.026
	EmpVar	0.053	0.034	0.025	0.020	0.079	0.046	0.033	0.026
OPT-$\mathbf{V}_\beta$	AveMSE	0.051	0.031	0.023	0.018	0.073	0.042	0.029	0.022
	EmpVar	0.051	0.031	0.022	0.018	0.069	0.040	0.028	0.022

We can see that the MSEs for the proposed methods are smaller than UNI in Figure 4, and the average estimated MSEs are quite close to the empirical MSEs in Table 2.

To evaluate computational efficiency, we record the average CPU time for Model 1 under a MacBook Pro with 2.5 GHz Intel Core i7 processor and 16 GB memory. The simulation is conducted using the R programming language. We also provide the computing time for the full data. As shown in Table 3, the sampling strategies based on the optimal subsampling probabilities can save the computing time compared to the full data approach (Full). As expected, the computing times for OPT-$\mathbf{V}_\pi$ and OPT-$\mathbf{V}_\beta$ were less than for OPT-**V**. Since UNI does not need additional computation for calculating the subsampling probabilities, it is the fastest algorithm.

We conduct additional simulations using data from Model 1 to check the computational advantage on larger datasets. We increase the full data size $n = 5 \times 10^5, 10^6, 5 \times 10^6, 10^7$ and consider a 10-dimensional vector of the regression coefficients for each component. Table 4 shows the results when $r_0 = 500$, $r = 2000$, and $K = 100$. We observe that the proposed subsampling algorithms save more computation time as the full data size increases. OPT-**V**, OPT-$\mathbf{V}_\beta$, and OPT-$\mathbf{V}_\pi$ are about 7.7, 8.7, 21.8 times faster than Full at $n = 5 \times 10^5$, and about 11.2, 12.1, 32.2 times faster than Full at $n = 10^7$.

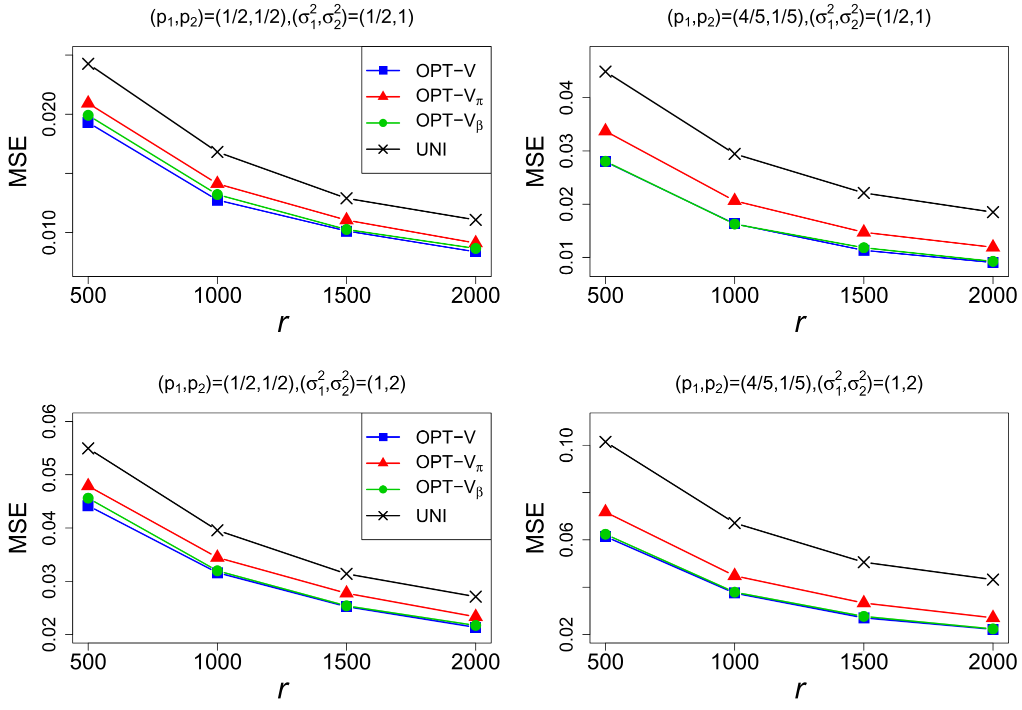

Figure 4: MSEs of Model 1 for varied (p_1, p_2), and r when $(\sigma_1^2, \sigma_2^2) \in \{(1/2, 1), (1, 2)\}$. OPT-$\mathbf{V}$, OPT-$\mathbf{V}_\pi$, and OPT-$\mathbf{V}_\beta$ use $\pi_i^{\mathbf{V}}$, $\pi_i^{\mathbf{V}_\pi}$, and $\pi_i^{\mathbf{V}_\beta}$, respectively. UNI uses uniform subsampling probabilities.

3.2 Real Data Example

In this section, we apply the proposed methods to appliance energy data.[1] This dataset includes appliances energy consumption and humidity collected with an internet-connected energy monitoring system and a ZigBee wireless sensor network, respectively (Candanedo et al., 2017). For the response, appliances energy consumption is used, and three humidities in different areas are considered as covariates: kitchen area (H-Kit), living room area (H-Liv), and laundry area (H-Lau). We use the log-transformed response and covariates. The full data size is $n = 19,735$. Figure 5 presents the distribution of the log-transformed energy consumption showing two peaks (around 4.2 and 5.8), and the relationships between energy consumption and humidities monitored in different areas.

Table 5 provides the parameter estimates from the full data and the average of parameter estimates for subsampling methods based on 1000 subsamples with $r_0 = 500$ and $r = 1500$. We observe that the estimates from the optimal subsampling probabilities are close to those from the full data. For the average CPU time, OPT-$\mathbf{V}$, OPT-$\mathbf{V}_\pi$, and OPT-$\mathbf{V}_\beta$ methods took 0.065, 0.060, and 0.064 seconds to calculate the estimates, respectively. It took 0.343 seconds to obtain estimates from the full data.

To compare OPT-$\mathbf{V}$, OPT-$\mathbf{V}_\pi$, and OPT-$\mathbf{V}_\beta$ with UNI, we set $r_0 = 500$, and $r = 500, 1000,$

[1] The data is available at the UCI Machine Learning repository https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction

Table 2: Average estimated MSE (AveMSE) and empirical variance (EmpVar) of Model 1 for varied (p_1, p_2), and r when $(\sigma_1^2, \sigma_2^2) \in \{(1/2, 1), (1, 2)\}$.

	$(p_1, p_2) =$	$(1/2, 1/2)$				$(4/5, 1/5)$			
	$r =$	500	1000	1500	2000	500	1000	1500	2000
$(\sigma_1^2, \sigma_2^2) = (1/2, 1)$									
OPT-**V**	AveMSE	0.0175	0.0107	0.0077	0.0060	0.0253	0.0141	0.0096	0.0071
	EmpVar	0.0166	0.0102	0.0075	0.0057	0.0258	0.0144	0.0096	0.0072
OPT-$\mathbf{V}_\pi$	AveMSE	0.0186	0.0117	0.0085	0.0067	0.0298	0.0177	0.0124	0.0096
	EmpVar	0.0182	0.0114	0.0083	0.0065	0.0315	0.0187	0.0129	0.0099
OPT-$\mathbf{V}_\beta$	AveMSE	0.0180	0.0112	0.0081	0.0063	0.0257	0.0145	0.0098	0.0074
	EmpVar	0.0171	0.0106	0.0076	0.0061	0.0261	0.0144	0.0100	0.0074
$(\sigma_1^2, \sigma_2^2) = (1, 2)$									
OPT-**V**	AveMSE	0.0375	0.0229	0.0165	0.0128	0.0533	0.0298	0.0201	0.0151
	EmpVar	0.0348	0.0220	0.0157	0.0119	0.0544	0.0309	0.0208	0.0156
OPT-$\mathbf{V}_\pi$	AveMSE	0.0403	0.0254	0.0185	0.0147	0.0610	0.0359	0.0250	0.0193
	EmpVar	0.0382	0.0247	0.0179	0.0140	0.0641	0.0377	0.0265	0.0204
OPT-$\mathbf{V}_\beta$	AveMSE	0.0381	0.0237	0.0171	0.0133	0.0541	0.0304	0.0206	0.0155
	EmpVar	0.0356	0.0225	0.0159	0.0122	0.0552	0.0312	0.0212	0.0159

Table 3: Average of computing time (in seconds) using data from Model 1 for different r at fixed $r_0 = 500$, $(\sigma_1^2, \sigma_2^2) = (2, 2)$, and $(p_1, p_2) = (1/2, 1/2)$. The computing time (in seconds) calculated from the full data is provided.

	r			
	500	1000	1500	2000
OPT-**V**	0.0778	0.0859	0.0881	0.0943
OPT-$\mathbf{V}_\pi$	0.0560	0.0610	0.0630	0.0655
OPT-$\mathbf{V}_\beta$	0.0704	0.0764	0.0780	0.0846
UNI	0.0156	0.0209	0.0241	0.0284
Full data CPU seconds: 0.6903				

1500, 2000. We calculate the MSEs based on $\sum_{k=1}^{K} \|\tilde{\boldsymbol{\theta}}^{(k)} - \hat{\boldsymbol{\theta}}\|^2/K$ where K is the number of replications, $\tilde{\boldsymbol{\theta}}^{(k)}$ is the estimate of $\boldsymbol{\theta}$ provided from the k-th subsample, and $\hat{\boldsymbol{\theta}}$ is the estimate calculated from the full data. Figure 6 shows the results for MSE based on 1000 subsamples of the size $r_0 + r$ from the full data. OPT-**V**, OPT-$\mathbf{V}_\pi$, and OPT-$\mathbf{V}_\beta$ provide smaller MSE than UNI.

4 Conclusion

In the big data era, statistical modeling with a large amount of data causes computational burden. In this article, we proposed an optimal subsampling method under mixtures of lin-

Table 4: Average of computing time (in seconds) using data from Model 1 with varied full data size n at fixed $r = 2000$, $r_0 = 500$, $(\sigma_1^2, \sigma_2^2) = (2, 2)$, and $(p_1, p_2) = (1/2, 1/2)$. The number of coefficients for each component is 10, and repetition is 100. The computing time (in seconds) for the full data is provided.

	n			
	5×10^5	10^6	5×10^6	10^7
OPT-$\mathbf{V}$	0.6785	1.2601	6.1786	12.1675
OPT-$\mathbf{V}_\pi$	0.2410	0.4383	2.0605	4.2224
OPT-$\mathbf{V}_\beta$	0.6051	1.1485	5.5159	11.2271
UNI	0.0401	0.0409	0.0403	0.0413
Full	5.2446	10.8181	66.0913	135.9436

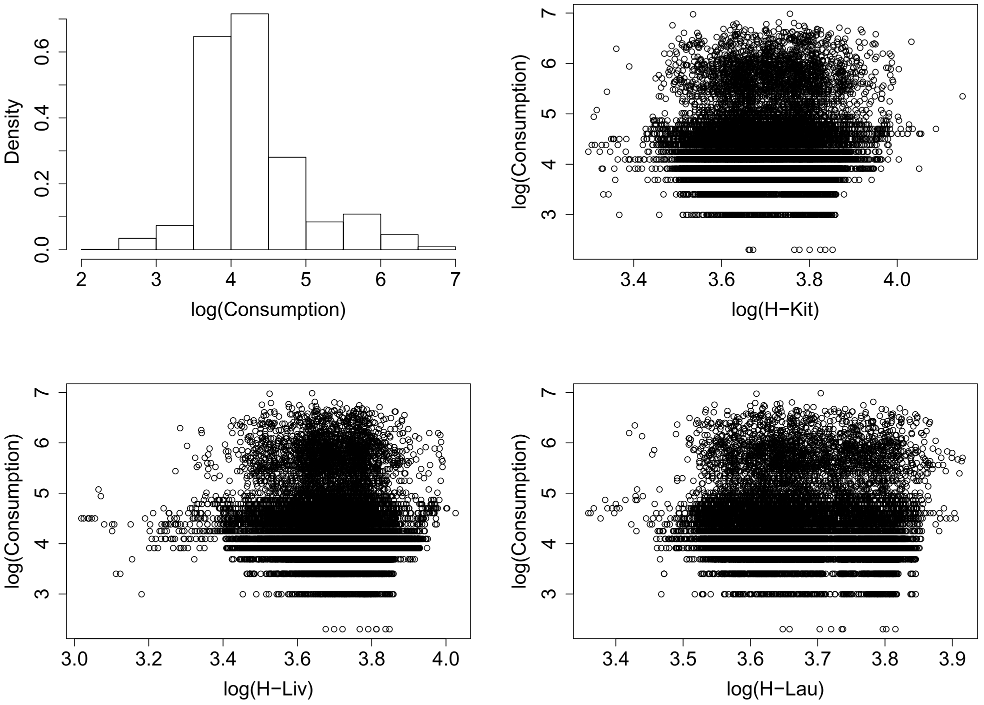

Figure 5: Histogram of log-transformed appliances energy consumption (top-left) and scatter plots between appliances energy consumption and humidity at different areas (top-right, bottom-left, bottom-right).

ear regression models for computational efficiency. We derived the asymptotic results of the subsample-based estimator and proposed the optimal subsampling probabilities. Since the subsampling probabilities cannot be directly calculated, a practical algorithm was also proposed. We first approximated the subsampling probabilities using a pilot sample and then drew a subsample with the approximated probabilities to obtain parameter estimates.

There are important questions for future research. In this paper, we assume that the number of components J is given. We often need to select the number of components in real applications.

Table 5: Average estimates of the proposed methods and UNI for the appliances energy data. 1000 subsamples of $r_0 = 500$, and $r = 1500$ are used. The estimates from the full data is also provided (Full).

	Full		OPT-**V**		OPT-$\mathbf{V}_\pi$		OPT-$\mathbf{V}_\beta$		UNI	
	Com.1	Com.2	Com.1	Com.2	Com.1	Com.2	Com.1	Com.2	Com.1	Com.2
Intercept	7.381	6.788	7.391	6.737	7.381	6.718	7.382	6.705	7.374	6.755
H-Kit	3.657	1.623	3.654	1.637	3.662	1.622	3.658	1.636	3.649	1.660
H-Liv	−1.703	−0.908	−1.700	−0.916	−1.707	−0.896	−1.704	−0.910	−1.699	−0.932
H-Lau	−2.851	−1.005	−2.854	−0.997	−2.853	−0.997	−2.851	−0.994	−2.846	−1.011
Variance	0.150	0.161	0.149	0.160	0.149	0.159	0.150	0.159	0.149	0.158
Mix proportion	0.892	0.108	0.892	0.108	0.892	0.108	0.892	0.108	0.892	0.108

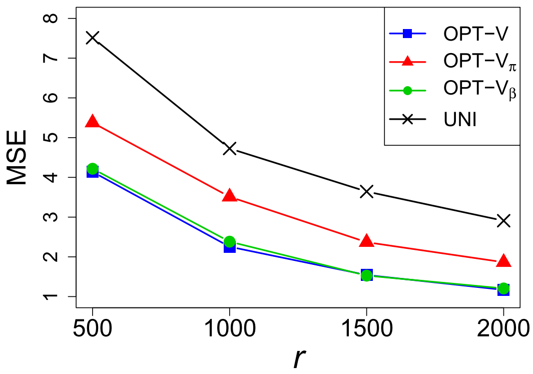

Figure 6: MSEs obtained from 1000 subsamples of the appliances energy data for varied r. OPT-**V**, OPT-$\mathbf{V}_\pi$, and OPT-$\mathbf{V}_\beta$ use $\pi_i^{\mathbf{V}}$, $\pi_i^{\mathbf{V}_\pi}$, and $\pi_i^{\mathbf{V}_\beta}$, respectively. UNI uses uniform subsampling probabilities.

One possible approach is to consider model-selection criteria using the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Lumley and Scott (2015) proposed the design-based AIC and BIC for survey data under a sampling design. Based on the criteria, a modified AIC and BIC can be developed for choosing the number of components in terms of the subsampling framework. Also, we consider the weighted objective function assigning higher weights to less informative data points. For more efficient estimation, it would be interesting to develop unweighted subsample-based estimators by avoiding the inverse probability weighting in a target function (e.g Wang, 2019; Wang and Kim, 2020). These are areas of future research.

Supplementary Material

- Software: R codes used for the proposed methods and algorithms are available on GitHub https://github.com/pedigree07/OPTMixture.
- Supplementary document: The supplementary document provides the proofs of the theorems.

Funding

HaiYing Wang's research was partially supported by the US NSF grant CCF-2105571.

References

Ai M, Wang F, Yu J, Zhang H (2021a). Optimal subsampling for large-scale quantile regression. *Journal of Complexity*, 62: 101512.

Ai M, Yu J, Zhang H, Wang H (2021b). Optimal subsampling algorithms for big data regressions. *Statistica Sinica*, 31: 749–772.

Candanedo LM, Feldheim V, Deramaix D (2017). Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings*, 140: 81–97.

Dempster AP, Laird NM, Rubin DB (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22.

Drineas P, Mahoney MW, Muthukrishnan S (2006). Sampling algorithms for l2 regression and applications. In: *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, 1127–1136.

Lee J, Schifano ED, Wang H (2021). Fast optimal subsampling probability approximation for generalized linear models. *Econometrics and Statistics*, doi: https://doi.org/10.1016/j.ecosta.2021.02.007.

Lumley T, Scott A (2015). Aic and bic for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 3(1): 1–18.

Ma P, Mahoney M, Yu B (2014). A statistical perspective on algorithmic leveraging. In: *International Conference on Machine Learning*, 91–99. PMLR.

McLachlan G, Peel D (2004). *Finite Mixture Models Wiley Series in Probability and Statistics.* Wiley.

Wang H (2019). More efficient estimation for logistic regression with optimal subsamples. *Journal of Machine Learning Research*, 20(132): 1–59.

Wang H, Kim JK (2020). Maximum sampled conditional likelihood for informative subsampling. arXiv preprint: https://arxiv.org/abs/2011.05988.

Wang H, Ma Y (2021). Optimal subsampling for quantile regression in big data. *Biometrika*, 108(1): 99–112.

Wang H, Yang M, Stufken J (2019). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, 114(525): 393–405.

Wang H, Zhu R, Ma P (2018). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522): 829–844.

Yao Y, Wang H (2019). Optimal subsampling for softmax regression. *Statistical Papers*, 60(2): 585–599.

Yu J, Wang H, Ai M, Zhang H (2022). Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association*, 117(537): 265–276.

Zuo L, Zhang H, Wang H, Liu L (2021). Sampling-based estimation for massive survival data with additive hazards model. *Statistics in Medicine*, 40(2): 441–450.