Score-Matching Representative Approach for Big Data Analysis with Generalized Linear Models

by

Keren Li
B.Sc. (Nankai University) 2001
M.S. (Louisiana State University) 2004

Thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Mathematics
in the Graduate College of the
University of Illinois at Chicago, 2018

Chicago, Illinois

Defense Committee:
Jie Yang, Chair and Advisor
Samad Hedayat
Min Yang
Jing Wang
Hua Yun Chen, Division of Epidemiology and Biostatistics, School of Public Health

This work is dedicated to my wife and my parents, who have always loved me unconditionally, had me better and more fulfilled than I could have ever imagined.

ACKNOWLEDGMENT

I would like to acknowledge everyone who played a role in my struggle to achieve my dream of becoming a Ph. D.

First of all, I would like to thank my advisor, Prof. Jie Yang, who has provided patient advice and guidance throughout the research process and helped me a lot during my Ph. D. study. He also has given much time, effort and knowledge to aid in the completion of my research papers and this dissertation. Thank you for your patience and friendship.

Secondly, special thanks go to Prof. Min Yang, who supported me in the first summer here and provided me lots of opportunities. I would also like to thank Prof. Jing Wang. She has cultivated an appreciation for teaching. One of my work was inspired in her class. My sincere thank goes to Prof. Samad Hedayat, whose wisdom and experience was greatly appreciated, as well as Prof. Hua Yun Chen, agreed to serve in my committe member in such a short time. Thank all professors taught me for your unwavering support.

Also, thank Prof. Hsin-Hsiung Huang from University of Central Florida, one of my best friends, who offered me lots of helps and suggestions.

And lastly, many thanks go to Xuelong, Shuang, Yue, Hani, and many others, who contributed to our discussion sessions.

Finally, I am grateful to have had this opportunity attending MSCS, UIC. This experience has afforded me the opportunity to achieve great success.

PREFACE

The idea of representative originally came from a joint research project with Prof. Jie Yang, when we tried to develop a subsampling method for generalized linear regression with categorical response for big data. The data we dealt with is the famous flight on-time performance data. For subsampling method, an initial value is always required. Prof. Jie Yang suggested to discretize the only continuous variable, namely distance, in our model into 25 levels by using the interval center as smoothing value. Later we found our subsampling method could not even compare with the initial value no matter how we improved our method. We guessed it was because we discretized distance in too many levels. Then I reduced the pieces of interval down to 20, 10, even 4. Surprisingly, the initial result can still beat the developed method. It indicates that even with a coarse grid, the binning method could still work well in some situations. Thereafter, we tried more comprehensive simulations, where the performance of traditional smoothing choices, interval center, is not that satisfactory. We realized that the key is the choice of representative (smoothing value). We named this method representative method, where later on, we found this terminology has been used for smoothing value on Wikipedia already. The difference is that we focus on the choice of representative while the data binning focuses on how to partition the data. So, the scope of this paper is given partitioning of data, find representatives.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF FIGURES

LIST OF ABBREVIATIONS

MR	Mean representative method
SMR	Score-matching representative method
LM	Linear model
GLM	Generalized linear model
RMSE	Root mean squared error
IBOSS	Information-based subsampling
OLS	Ordinary least square
WLS	Weighted least square
MLE	Maximum likelihood estimator (or estimate)
Δ	Maximum in-block distance
$\tilde{\Delta}$	Maximum distance between block representative and data block

SUMMARY

We propose a fast and efficient strategy, called the representative approach, with linear models and generalized linear models for big data analysis, and in particular for distributed dataset.

With a given partitioning of big dataset, this approach constructs a representative data point for each data block and fits the target model on the representative dataset. In terms of time complexity, it is as fast as the subsampling approaches in the literature. As for efficiency, its accuracy of estimated parameters appears to be better than the divide-and-conquer method. Additionally, the representative approach is especially useful when analyzing massive data distributed stored on different nodes, since the generation of representatives is conditional independent. Overall, we recommend two representative approaches, mean representative (MR) and score-matching representative (SMR), along with theoretical justifications, for big data analysis with generalized linear models.

Comprehensive simulation studies confirm that MR is a good solution for linear models and pre-analysis for GLMs, while SMR outperforms the subsampling and divide-and-conquer methods, even with moderate size of block, for general GLMs. With properly chosen data partition, SMR estimate appears to be even comparable with the full data estimate. Using the Airline on-time performance data as an illustrative real big data example, we show that MR and SMR are as good as the full data estimate when available.

For GLMs with flat inverse link functions and moderate coefficients of the continuous variables, we recommend MR. Otherwise, we recommend SMR solution with MR as an initial step with a finer partition.

CHAPTER 1

INTRODUCTION

In the past decade, big data or massive data has drawn dramatically increasing attention all over the world. It was in the 2009 ASA Data Expo competition when people found out that no statistical software was available to analyze the massive *Airline on-time performance data*. At that time, the airline data file, about 12GB in size, consisted of 123,534,969 records of domestic flights in the United States from October 1987 to April 2008 (Kane et al., 2013 (1)) . Up to February 2017, the airline on-time performance data collected from the Bureau of Transportation Statistics consisted of 353 files with 169,609,446 valid records in total.

The response in the Airline on-time performance data was treated as a binary variable `Late Arrival` with 1 standing for late by 15 minutes or more (Wang et al., 2016 (2)). Generalized linear models (GLMs) have been widely used for modeling binary response, as well as Poisson, Gamma, and Inverse Gaussian responses (McCullagh and Nelder, 1989 (3); Dobson and Barnett, 2008 (4)). In order to fit a GLM with p predictors, a typical algorithm searching for the maximum likelihood estimate (MLE) based on the full data of size N requires $O(\zeta_N Np^2)$ time to run, where ζ_N is the number of iterations required for convergence of the full data MLE algorithm (Wang et al., 2017 (5)).

Staring in 2009, substantial efforts have been made on developing both methodologies and algorithms towards big data analysis (see, for example, Wang et al. (2016) (2), for a good

survey on relevant statistical methods and computing). Divide-and-conquer, also known as divide-and-recombine, split-and-conquer, or split-and-merge, first partitions a big dataset into K blocks, fits the target model block by block, and then aggregates the K fits to form a final one (Wang et al., 2016 (2)). A divide-and-conquer algorithm proposed by Lin and Xi (2011) (6) reaches time complexity of $O(\zeta_{N/K} N p^2)$, where $\zeta_{N/K}$ is the number of iterations required by a GLM MLE algorithm with N/K data points. The accuracy of the estimated parameters based on the divide-and-conquer algorithm relies on the block size N/K, which typically depends on the computer memory. Therefore, as N goes to infinity, K has to increase accordingly. Typically, its accuracy is not as good as the full data estimate.

Another popular strategy for big data analysis is subsampling. For example, leveraging technique has been used to sample a more informative subset of the full data for linear regression problems (Ma and Sun, 2014 (7)). Inspired by D-optimality in optimal design theory, Wang et al. (2018) (8) proposed an information-based subsampling technique, called IBOSS, for big data linear regression problems. Its time complexity is $O(Np)$ while the ordinary least square (OLS) estimate for linear models takes the time complexity of $O(Np^2)$. Motivated by A-optimality, Wang et al. (2017) (5) developed an efficient two-step subsampling algorithm for large sample logistic regression, which is also a special case of generalized linear models. The time complexity of the A-optimal subsampler is also $O(Np)$. Compared with the divide-and-conquer strategy, the subsampling approach requires much less computational cost. Nevertheless, its accuracy relies on the subsample size and is typically not as good as the divide-and-conquer estimate.

In the computer science literature, data binning technique employing a binned version of continuous variables is a commonly used discretization technique for data pre-processing, which bins continuous variables into categorical variables coded by so-called smoothing values (see, for example, Kotsiantis and Kanellopoulos (2006) (9)). It mainly focuses on how to partition data into blocks or bins, while the smoothing values are usually chosen from class labels, boundary points, center, mean, or median of data block, whose performance could not be guaranteed, especially for nonlinear models.

Inspired by data binning but different from it, the representative approach proposed in this dissertation assumes a given data partitioning and concentrates on constructing the best smoothing values, which we call *representatives*, more efficiently according to the regression model. The recommended score-matching representative (SMR) approach runs as fast as subsampling approaches, while estimates model parameters comparable to the divide-and-conquer method. Unlike the data binning technique serving as a data pre-analysis method, the representative approach provides a solution with reasonable accurate level for big data analysis.

Actually, in representative approach, the GLM is fitted on K representatives obtained from the original N data points ($\mathsf{K} \ll \mathsf{N}$). The time complexity is only $\mathsf{O}(\mathsf{Np})$, same as the subsampling approaches. The K representatives are not a subset of the N data points, but summarize the information from each single of the N data points. The accuracy of parameter estimates are comparable with or better than the divide-and-conquer estimate. Moreover, by matching the score function of GLMs, the SMR estimate is comparable with the full data estimate for our comprehensive experiments.

The representative approach provides an ideal solution for the so-called *distributed database* (see, for example, Özsu and Valduriez (2011) (10)), which is dispersed over a network of interconnected computers and is often the case in practice for massive data. By exchanging only the estimated parameters and the representative data points among parallel computing computers, the representative approach can work well even with slow-speed network connection.

The remainder of this dissertation proceeds as follows.

In Chapter 2 we describe necessary preliminary knowledges such as the generalized linear model framework, data partitions including grid partition, feather space oriented, and clustered partition, data oriented, as well as natural partition and distributed database.

In Chapter 3, starting from a toy example of univariate linear regression, we describe the framework of representative approaches. By comparing mean, median and mid-point representatives, we recommend mean representative (MR) for big data linear regression. Also, some theoretical justifications of block-center kind representatives, as well as simulations are given.

In Chapter 4, we develop the score-matching representative (SMR) along with its theoretical justifications. Based on our comprehensive simulation studies comparing SMR with MR, A-optimal subsampling, and Divide-conquer, we recommend SMR for big data generalized linear regressions.

In Chapter 5 we use the Airline on-time performance data as an illustrative real big data analysis example. We show that the MR and SMR estimates are as accurate as the full data estimate when the latter is available.

We conclude in Chapter 6 and discuss the future work.

CHAPTER 2

PRELIMINARY

2.1 Generalized Linear Model and Score Function

Given the original data set $\{(\mathbf{x}_i, y_i), i = 1, 2, \ldots, N\}$ with covariates $\mathbf{x}_i \in \mathbb{R}^d$ and response $y_i \in \mathbb{R}$, we consider a generalized linear model assuming independent response random variable Y_i's and the corresponding predictors $\mathbf{X}_i = (h_1(\mathbf{x}_i), \ldots, h_p(\mathbf{x}_i))^T \in \mathbb{R}^p$. For model-based data analysis with fairly general known functions $h_1(\cdot), \ldots, h_p(\cdot)$, we would rather regard the data set as $\mathscr{D} = \{(\mathbf{X}_i, y_i), i = 1, \ldots, N\}$. For simplicity, we denote $\mathbf{X}_i = (x_{i1}, \ldots, x_{ip})^T$, $i = 1, \ldots, N$. For many applications, $h_1(\mathbf{x}_i) \equiv 1$ corresponds to intercept.

In this dissertation, we only consider independent observations following an exponential family distribution in the natural form with probability density function

$$f(y_i) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\},$$

where θ_i is the natural parameter and ϕ is the dispersion parameter, $b(\cdot)$ and $c(\cdot, \cdot)$ are known functions.

Following McCullagh and Nelder (1989) (3), there exists a link function g and regression parameters $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$, such that

$$\mathbb{E}(Y_i) = \mu_i \text{ and } \eta_i = g(\mu_i) = \mathbf{X}_i^T\boldsymbol{\beta} \tag{2.1}$$

For typical applications, the link function g is one-to-one and differentiable. The connection between parameter θ_i and observer expected value μ_i is given by $\theta_i = h(\mu_i)$. If $\theta_i = \eta_i \ \forall i$ holds, we call the corresponding link function canonical. Note that we do not transform the response y_i, but rather its expected value μ_i.

It is known that

$$\mathbb{E}(Y_i) = \mu_i = b'(\theta_i)$$

$$\mathrm{Var}(Y_i) = \sigma_i^2 = \phi b''(\theta_i)$$

provided $b(\theta)$ is twice differentiable. $V(\theta) := b''(\theta)$ is called the variance function of the GLM.

The log-likelihood function is then

$$\begin{aligned} l(\boldsymbol{\beta}, \phi, \mathbf{y}) &= \sum_{i=1}^{n} \left[\frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right] \\ &= \sum_{i=1}^{n} \left[\frac{y_i h(\mu_i) - b(h(\mu_i))}{\phi} + c(y_i, \phi) \right] \\ &= l(\mu, \phi, \mathbf{y}) \end{aligned}$$

According to McCullagh and Nelder (3) (1989, Section 2.5), the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ is given by

$$\begin{aligned}\hat{\boldsymbol{\beta}} =& \arg\max_{\boldsymbol{\beta}}\{l(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X})\} \\ =& \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left[\frac{y_i\theta(\mathbf{X}_i, \boldsymbol{\beta}) - b(\theta(\mathbf{X}_i, \boldsymbol{\beta}))}{\phi} + c(y_i, \phi)\right]\end{aligned}$$

where $\theta(\mathbf{X}_i, \boldsymbol{\beta}) = \theta_i = h(g^{-1}(\mathbf{X}_i^T\boldsymbol{\beta}))$. MLE is obtained by solving the score equation

$$s(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \boldsymbol{\beta}} = 0$$

where $\mathbf{y} = (y_1, \ldots, y_N)^T$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)^T$.

Suppose g is continuous. The score function can be written as

$$\begin{aligned}s(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}) &= \sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{n} (y_i - \mu_i)\frac{1}{\phi b''(\theta_i)}\frac{1}{g'(\mu_i)}\mathbf{X}_i \\ &= \sum_{i=1}^{n} (y_i - g^{-1}(\eta_i))\frac{1}{\phi b''((b')^{-1}(g^{-1}(\eta_i)))g'(g^{-1}(\eta_i))}\mathbf{X}_i \\ &= \sum_{i=1}^{n} (y_i - G(\eta_i))\nu(\eta_i)\mathbf{X}_i\end{aligned}$$

where

$$\nu(\eta) = \{\phi b''((b')^{-1}(g^{-1}(\eta)))g'(g^{-1}(\eta))\}^{-1} = \phi^{-1}\frac{d\theta}{d\eta},\ G(\eta) = g^{-1}(\eta) \tag{2.2}$$

For canonical link, the effective part of $\nu(\eta) = 1$, since it is a constant. For more ν and G examples of commonly used GLMs, see Table VI.

2.2 Data Partition

The main focus of binning methods is on how to partition data into blocks or bins. Many different partition methods have been proposed in the literature (see Fahad et al. (2014) (11) for a good survey). From our point of view, there are at least two types of partitioning methods.

One type is to partition the feature space $\mathbb{R}^p$ or its subset, with cut points obtained from the summary information of data. We call it *grid partition*. Grid partition methods (Kotsiantis and Kanellopoulos, 2006 (9)) divide the space into rectangular cells at the given cut points on each covariate, such as quantiles (equal-depth) or equal-width points, which is usually satisfactory with a moderate number of predictors. Its time complexity is $O(Np)$. That is, the data is arranged into blocks according to the ranges of the covariates $\mathbf{x}_i$'s or the predictors $\mathbf{X}_i$, for example, a block I_k may be defined as $\{\mathbf{x} = (x_1, \ldots, x_d)^T \in \mathbb{R}^d \mid x_i \in [a_{ki}, b_{ki}), i = 1, \ldots, d\}$.

Another partition type is based on clustering algorithms. We call it *clustered partition*. The clustering methods aim to split the dataset into blocks such that observations in-block are similar and consistent according to specific parameters, e.g. k-means clustering, described in detail by Hartigan (1975) (12). Pakhira (2014) (13) proposed a linear k-means with time complexity $O(Np)$. The hierarchical clustering methods partition data using an agglomerative hierarchical clustering method (Johnson, 1967 (14)). The density-based methods find blocks defined as regions of density (Ester et al., 1996 (15)). The mixture model based clustering

methods classify by a law of multivariate probability distributions, e.g. EM (Ghahramani and Jordan, 1994 (16)).

For a massive dataset, employing partition methods could be very challenging especially with large $\mathtt{p}$ due to the curse of dimensionality. Considering computational cost, some coarse partitioning may be applied, where the binning methods are used only for pre-analysis.

2.3 Natural Partition and Distributed Database

In practice, massive data are often provided in parts or blocks. Each file may contain some unique combinations of covariate values. For example, the Airline on-time performance data up to February 2017 are stored into 353 individual data files, labeled by months. We call such kind of data partition a *natural partition*. In the natural partition, the sizes of blocks are usually large in terms of largest distance in block, binning methods may not reach a good result for regular choices of smoothing values. Therefore, a sub-partition is necessary after a natural partition for analysis purpose.

An example of natural partition is the *distributed database* (Özsu and Valduriez, 2011 (10)), where the data may be stored in different hard disks, multiple computers, even not located in the same physical location but interconnected. The communications between nodes are usually slow and restricted. Therefore, operations on distributed database are expected to be independent or conditional independent. Any re-split and combine operation is also unfriendly to distributed database, while sub-partitioning is allowed, as long as the latter does not require any communication between nodes. In this paper, the representative methods we developed obey the independent or conditional independent rule.

CHAPTER 3

REPRESENTATIVE APPROACHES AND LINEAR MODELS

In this dissertation, we suppose a partitioning is given, that is, if $I = \{1, 2, \ldots, N\}$ is the index set of the whole dataset, then there exists a partitioning $\{I_1, I_2, \ldots, I_K\}$ of I with nonempty, disjoint blocks I_k's, such that, $I = \cup_{k=1}^{K} I_k$.

The kth data block $\mathscr{D}_k = \{(\mathbf{X}_i, y_i), i \in I_k\}$ has block size $n_k = |I_k|$, the number of observations in block, for $k = 1, \ldots, K$. Therefore $\{\mathscr{D}_k\}_{k=1}^{K}$ is a partitioning of original full data $\mathscr{D}$. In the kth block $\mathscr{D}_k$, $\mathbf{X}_k$ and $\mathbf{y}_k$ are the corresponding design matrix and response vector respectively.

3.1 A Toy Example

First, we consider only one explanatory variable, x, and assume that the statistical relationship between x and the response variable y is linear, which can be written as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$i = 1, 2, \ldots, N$. We will assume that y_i's have been standardized so that the intercept is removed from the model

$$y_i = \beta x_i + \epsilon_i, \quad i = 1, 2, \ldots, N, 0 \leq x_i \leq 1.$$

The ordinary least square (OLS) estimator of β is given by

$$\hat{\beta} = (\sum_{i=1}^{n} x_i^2)^{-1} \sum_{i=1}^{n} x_i y_i$$

Our purpose is given partition of data, find a "representative" value for each block to form a set of new data with reduced size, on which the new estimate exactly equals the estimate on the full dataset. We build a new dataset by creating pseudo record $(\tilde{x}_k, \tilde{y}_k)$, to replace all records within kth block, $k = 1, \ldots, K$. The new records along with the corresponding numbers of records in block, form a weighted dataset, $\{(n_k, \tilde{x}_k, \tilde{y}_k), k = 1, \ldots, K\}$. That is, a new dataset $\{(x_i^*, y_i^*) = \sum_{k=1}^{K} (\tilde{x}_k, \tilde{y}_k) \mathbb{1}\{i \in I_k\}, i = 1, \ldots, N\}$.

The weighted least square (WLS) estimator of $\tilde{\beta}$ on the new dataset is then given by

$$\begin{aligned} \tilde{\beta} &= (\sum_{k=1}^{K} \sum_{i \in I_k} \tilde{x}_k^2)^{-1} (\sum_{k=1}^{K} \sum_{i \in I_k} \tilde{x}_k \tilde{y}_k) \\ &= (\sum_{k=1}^{K} n_k \tilde{x}_k^2)^{-1} (\sum_{k=1}^{K} n_k \tilde{x}_k \tilde{y}_k) \end{aligned}$$

There are two criterions for choosing representatives. First, fitted regression parameter on the representatives $\{(n_k, \tilde{x}_k, \tilde{y}_k), k = 1, \ldots, K\}$ holds the equation $\tilde{\beta} = \hat{\beta}$, or $\tilde{\beta}$ is unbiased. Second, the choice of representatives should be independent, i.e., each partition elects their representative only base on the data in the given block, unaffected by other blocks.

The choice is not unique, since there is only one equation while $2K$ variables. By adding some constrains, we have

$$n_k\tilde{x}_k^2 = \sum_{i\in I_k} x_i^2,\ n_k\tilde{x}_k\tilde{y}_k = \sum_{i\in I_k}(x_i y_i) \tag{3.1}$$

for $k = 1,\ldots,K$. So we have

$$\tilde{x}_k = n_k^{-\frac{1}{2}}(\sum_{i\in I_k} x_i^2)^{\frac{1}{2}},\ \tilde{y}_k = n_k^{-\frac{1}{2}}\sum_{i\in I_k}(x_i y_i)(\sum_{i\in I_k} x_i^2)^{\frac{1}{2}}$$

for $k = 1,\ldots,K$. This choice gives exactly the same fitted regression parameter as the full data. Note that the choice of representative to exactly meet full data fit is not unique.

Also, we have an unbiased representative choice

$$\tilde{x}_k = n_k^{-1}\sum_{i\in I_k} x_i = n_k^{-1}\mathbb{1}_{n_k}^T\mathbf{x}_k,\ \tilde{y}_k = n_k^{-1}\sum_{i\in I_k} y_i = n_k^{-1}\mathbb{1}_{n_k}^T\mathbf{y}_k \tag{3.2}$$

where $\mathbf{x}_k = (x_{k_1},\ldots,x_{k_{k_n}})^T$, $\mathbf{y}_{(k)} = (y_{k_1},\ldots,y_{k_{k_n}})^T$, with $I_k = \{k_1,\ldots,k_{k_n}\}$.

3.2 Representative Approaches

Inspired by the data binning and the toy example, we propose the *representative approach* for model-based regression analysis on partitioned dataset. Unlike binning method, main attention of this dissertation is paid on the choice of representative (smoothing value) of predictors (other than covariates) for each block. The procedure is to construct representative data point $(\tilde{\mathbf{X}}_k, \tilde{y}_k)$ for data block $\mathscr{D}_k$, $k = 1,\ldots,K$, and then fit the regression model based on the (weighted)

representative dataset $\tilde{\mathscr{D}} = \{(n_k, \tilde{\mathbf{X}}_k, \tilde{y}_k), k = 1, \ldots, K\}$. The number of unique observations is usually significantly smaller than the original dataset size N. It is only related to the number of blocks K. The procedure may be repeated if the construction of representative data points depends on the fitted model.

Note that unlike subsampling approaches, a representative data point may not belong to the original dataset, it is essentially a pseudo data generated from the data block and the current system information.

We suggest two principles to choose representatives.

First, the choice of representatives should be *conditional independent*, i.e., each block elects its representative only base on data in-block and the current system information, unaffected by the observations of other blocks, to reduce computational costs. Consequently, representative methods will facilitate parallel computing, especially for distributed database, since the generation of representative for each block happens in block, without communicating with other blocks.

Second, the goal of the representative approach is to make the fitted regression parameter $\tilde{\boldsymbol{\beta}}$ based on the weighted representative dataset is considerably close to, if not equal to, the full data estimate $\hat{\boldsymbol{\beta}}$.

Assume a natural partition has been provided for the representative approach. One may perform a sub-partition on each original data block in order to improve the efficiency of the representative approach, such as a linear k-means (Pakhira, 2014 (13)), both with a time complexity $O(Np)$.

There are many representative choices that could possibly work, if the given partitioning is delicate. In practice, we may not be able to reach an "appropriate delicate" partitioning in big data sense or there is a high computational cost. For example, consider applying a grid partition on a set of data with 20 predictor variables. If each dimension is split into 8 intervals, then the theoretical maximum number of blocks is $8^{20} \approx 1.15 \times 10^{18}$, such that the actual number of non-empty blocks may be comparable to the size of full dataset, which fails to reduce the scale of dataset. If the splits is 4 or even smaller, then some representative estimates may be far away from the full data estimate due to the coarse partition. In such a situation, the choice of representatives is very important with respect to the scarce computational resource. Theoretical analysis in Section 3.3 and simulation studies in Section 3.4 show that taking block mean is a good choice of representatives.

3.3 Center Representatives for LMs

Suppose we have p continuous predictors and the data follows a multivariate linear model.

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_d x_{ip} + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

After some data transformation and min-max normalization, β_0 can be removed and $0 \leq x_{ij} \leq 1$, for $i = 1, \ldots, n, j = 1, \ldots, p$, and the linear model can be rewritten as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \text{ where}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}^T, \; X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ & \cdots & \cdots & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^T \\ \mathbf{X}_2^T \\ \cdots \\ \mathbf{X}_n^T \end{bmatrix}, \; \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_p \end{bmatrix}, \; \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdots \\ \epsilon_n \end{bmatrix}$$

$$\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T, \quad i = 1, 2, \ldots, n$$

OLS estimator of β is given by

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y = \left(\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T \right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i y_i$$

Building equations like Equation 3.1 may lead to an insolvable situation with $p+1$ parameters but only $p(p+1)$ equations. Inspired by Equation 3.2, using block center is a naive choice of the representative for LM. It is also popular in the data binning literature. More specifically, given the kth data block $\mathscr{D}_k = \{(\mathbf{X}_i, y_i), i \in I_k\}$ with block size n_k, options for its representative $\tilde{\mathbf{X}}_k$ include

- mid-point of the rectangular block, when a grid partitioning is given;
- component-wise median;
- mean, that is, $\tilde{\mathbf{X}}_k = n_k^{-1} \sum_{i \in I_k} \mathbf{X}_i$.

Then the weighted representative data for the kth block is $(n_k, \tilde{\mathbf{X}}_k, \tilde{y}_k)$ with $\tilde{y}_k = n_k^{-1} \sum_{i \in I_k} y_i$, which will be justified in Theorem 4.1.1. An algorithm of center representative methods is given in Algorithm 1

Algorithm 1: Center Representative Methods

Data: Partitioning of $\mathscr{D}$: $\{\mathscr{D}_k = (\mathbf{X}_k, \mathbf{y}_k)\}_{k=1}^{K}$
Result: Center representative estimator $\boldsymbol{\beta}$ for Linear Model.
for $k = 1, \ldots, K$ **do**
  Calculate $\tilde{\mathbf{X}}_k$ by either mid-point, median, or mean in block;
  Calculate $\tilde{\mathbf{y}}_k = n_k^{-1} \sum_{i \in I_k} y_i$;
  Set n_k the number of records in block;
end
Fit linear regression on the representative dataset $\tilde{\mathscr{D}} = \{(n_k, \tilde{\mathbf{X}}_k^{(t)}, \tilde{y}_k^{(t)})\}_{k=1}^{K}$ to get $\tilde{\boldsymbol{\beta}}$.

A comprehensive simulation study with linear models below shows that the third option using block means, called the *mean representative approach* (MR), is more efficient in terms of mean square error than the mid-point and median options, as well as the IBOSS subsampling approach proposed in Wang et al. (2018) (8).

3.3.1 Mid-point representative approach

If a grid partitioning is given, the mid-point representative does not need any detail information of predictors in block, only the cut points for each predictor to generate the representative predictors. That is, it does not need read blocks for representative predictor. The only information read from each block is the response variables for generating the representative responses.

So, mid-point is a fast solution with very low computational cost, since it is only related to the cut points and response variables. The computing time of generating mid-point representatives is $O(Kp + N)$, since this procedure does not read the covariate information in block but

the cut points and responses of block. The time required by a WLS algorithm on K data points is $O(Kp^2)$, thus required total time complexity is $O(Kp^2 + N)$.

Based on the simulation studies for linear models in Section 3.4, the performance of mid-point is the worst among three center representatives. The estimate is away from the full data estimate when the partition grid is not delicate, especially when there are unbounded or large range covariates. Therefore, mid-point is essentially a pre-analysis for big data, quick but not accurate in most situations. Nevertheless, when the calculation burden is too heavy, mid-point may provide some initial analysis results.

3.3.2 Median representative approach

Median representative provides a better result than mid-point in the sense of accurate, but still is not comparable to mean representative, based on the simulation studies showed in Section 3.4 for the linear model. And median also has bias issue on intercept estimation.

The median is not affected by outliers, which contains more information if the model assumption is true or close. Thus it is robust if there is a model misspecification.

Median representative takes $O(Np)$ to generate representative set, and the WLS estimation also consumes $O(Kp^2)$. Consequently, the total time complexity of either median or mean is $O(Np + Kp^2)$. If $K \ll N$ and $p \ll N$, the overall time complexity is $O(Np)$ for either median or mean representative approaches, while the full data OLS estimation takes $O(Np^2)$ and the IBOSS (Wang et al., 2018 (8)) requires $O(Np)$.

3.3.3 Mean representative approach

The time complexity of MR is the same as median representative for linear models, while the performance of MR is comparable to full data estimate based on the simulation studies in Section 3.4.

The following theorem shows that for linear models, the MR estimate is unbiased and asymptotically efficient as maximum distance in block $\Delta = \max_k \max_{i,j \in I_k} \|\mathbf{X}_i - \mathbf{X}_j\| \to 0$, which are advantages over mid-point and median representative approaches.

Theorem 3.3.1. *Suppose* $\sum_{i=1}^{N} \mathbf{X}_i \mathbf{X}_i^T$ *is positive definite. For linear model* $y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \epsilon_i$, $i = 1, \ldots, N$, *with* ϵ_i *iid* $\sim N(0, \sigma^2)$, *the MR estimator* $\tilde{\boldsymbol{\beta}}$ *has mean* $\boldsymbol{\beta}$ *and covariance* $\mathrm{Cov}(\tilde{\boldsymbol{\beta}}) = \sigma^2 (\sum_{k=1}^{K} n_k \tilde{\mathbf{X}}_k^T \tilde{\mathbf{X}}_k)^{-1}$ *when* Δ *is sufficiently small.*

Additionally, the difference between $\mathrm{Cov}(\tilde{\boldsymbol{\beta}})$ *and* $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ *based on the full data shrinks to zero in terms of largest eigenvalue as* Δ *goes to zero, and* $\left\| \mathrm{Cov}(\tilde{\boldsymbol{\beta}}) - \mathrm{Cov}(\hat{\boldsymbol{\beta}}) \right\|_2 = O(\Delta^2)$, *where the induced matrix norm* $\|A\|_2 = \max_{\|x\|=1} \|Ax\|$ *is square root of the largest eigenvalue of* $A^T A$.

Proof. Denote by $\mathbf{X}_k$, $\mathbf{y}_k$, $\boldsymbol{\epsilon}_k$ the predictor matrix, response vector and error vector respectively of kth block.

Consider

$$\begin{aligned} \left\| \sum_{k=1}^{K} \mathbf{X}_k^T \mathbf{X}_k - \sum_{k=1}^{K} n_k \tilde{\mathbf{X}}_k^T \tilde{\mathbf{X}}_k \right\|_2 &= \left\| \sum_{k=1}^{K} \sum_{i \in I_k} (\mathbf{X}_i - \tilde{\mathbf{X}}_k)(\mathbf{X}_i - \tilde{\mathbf{X}}_k)^T \right\|_2 \\ &\leq \sum_{k=1}^{K} \left\| \sum_{i \in I_k} (\mathbf{X}_i - \tilde{\mathbf{X}}_k)(\mathbf{X}_i - \tilde{\mathbf{X}}_k)^T \right\|_2 \end{aligned}$$

Denote by $\delta_k = \max_{i,j\in I_k} \|\mathbf{X}_i - \mathbf{X}_j\|$, so we can rewrite $\sum_{i\in I_k}(\mathbf{X}_i - \tilde{\mathbf{X}}_k)(\mathbf{X}_i - \tilde{\mathbf{X}}_k)^T = \delta_k^2 \sum_{i\in I_k} a_i a_i^T$, with $\|a_i\| \leq 1$. By the definition of matrix norm,

$$\begin{aligned}\left\|\sum_{i\in I_k} a_i a_i^T\right\|_2 &= \max_{\|x\|=1}\left\|\sum_{i\in I_k} a_i a_i^T x\right\| \\ &\leq \max_{\|x\|=1}\sum_{i\in I_k}\|a_i\|^2\|x\| \\ &\leq n_k\end{aligned}$$

Therefore we have

$$\left\|\sum_{k=1}^{K}\mathbf{X}_k^T\mathbf{X}_k - \sum_{k=1}^{K} n_k\tilde{\mathbf{X}}_k^T\tilde{\mathbf{X}}_k\right\|_2 \leq \sum_{k=1}^{K} n_k\delta_k^2 \leq \Delta^2 N \tag{3.3}$$

Denote by λ_1 and λ_1^* the smallest eigenvalues of $\sum_{k=1}^{K}\mathbf{X}_k^T\mathbf{X}_k$ and $\sum_{k=1}^{K} n_k\tilde{\mathbf{X}}_k^T\tilde{\mathbf{X}}_k$. By the assumption in theorem, $\lambda_1 > 0$. By Equation 3.3, we have $\lambda_1^* > \lambda_1 - \Delta^2 N > \lambda_1/2 > 0$ if $\Delta^2 < \lambda_1/(2N)$. That is, $\sum_{k=1}^{K} n_k\tilde{\mathbf{X}}_k^T\tilde{\mathbf{X}}_k$ is invertible when Δ is sufficiently small.

Therefore, we have the WLS from mean representative dataset

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= (\sum_{k=1}^{K} n_k\tilde{\mathbf{X}}_k\tilde{\mathbf{X}}_k^T)^{-1}\sum_{k=1}^{K} n_k\tilde{\mathbf{X}}_k\tilde{y}_k \\ &=(\sum_{k=1}^{K} n_k^{-1}\mathbf{X}_k^T\mathbb{1}_{n_k}\mathbb{1}_{n_k}^T\mathbf{X}_k)^{-1}\sum_{k=1}^{K} n_k^{-1}\mathbf{X}_k^T\mathbb{1}_{n_k}\mathbb{1}_{n_k}^T\mathbf{y}_k \\ &=\boldsymbol{\beta} + (\sum_{k=1}^{K} n_k^{-1}\mathbf{X}_k^T\mathbb{J}_{n_k}\mathbf{X}_k)^{-1}\sum_{k=1}^{K} n_k^{-1}\mathbf{X}_k^T\mathbb{J}_{n_k}\boldsymbol{\epsilon}_k\end{aligned}$$

It is easy to verify that MR estimator is unbiased,

$$\mathbb{E}(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathbb{E}((\sum_{k=1}^{K} n_k^{-1}\mathbf{X}_k^T\mathbb{J}_{n_k}\mathbf{X}_k)^{-1}\sum_{k=1}^{K} n_k^{-1}\mathbf{X}_k^T\mathbb{J}_{n_k}\epsilon_k) = \boldsymbol{\beta}$$

Also, the covariance matrix of MR estimator is given by

$$\begin{aligned}
&\mathrm{Cov}(\tilde{\boldsymbol{\beta}}) = \sum_{k=1}^{K}\mathrm{Cov}((\sum_{k=1}^{K} n_k^{-1}\mathbf{X}_k^T\mathbb{J}_{n_k}\mathbf{X}_k)^{-1}n_k^{-1}\mathbf{X}_k^T\mathbb{J}_{n_k}\epsilon_k)\\
=&\sigma^2\sum_{k=1}^{K}(\sum_{k=1}^{K} n_k^{-1}\mathbf{X}_k^T\mathbb{J}_{n_k}\mathbf{X}_k)^{-1}n_k^{-2}\mathbf{X}_k^T\mathbb{J}_{n_k}^2\mathbf{X}_k(\sum_{k=1}^{K} n_k^{-1}\mathbf{X}_k^T\mathbb{J}_{n_k}\mathbf{X}_k)^{-1}\\
=&\sigma^2(\sum_{k=1}^{K}\frac{1}{n_k}\mathbf{X}_k^T\mathbb{J}_{n_k}\mathbf{X}_k)^{-1}\\
=&\sigma^2(\sum_{k=1}^{K} n_k\tilde{\mathbf{X}}_k^T\tilde{\mathbf{X}}_k)^{-1}
\end{aligned}$$

and the matrix norm of difference between the Fisher information matrices of OLS and MR is given by

$$\sum_{k=1}^{K}\mathbf{X}_k^T\mathbf{X}_k - \sum_{k=1}^{K} n_k\tilde{\mathbf{X}}_k^T\tilde{\mathbf{X}}_k$$

Consider the induceda matrix norm of difference between covariance matrices of OLS and MR estimators

$$
\begin{aligned}
&\left\|\mathrm{Cov}(\tilde{\boldsymbol{\beta}}) - \mathrm{Cov}(\hat{\boldsymbol{\beta}})\right\|_2 \\
=&\sigma^2\left\|(\sum_{k=1}^{K} n_k \tilde{\mathbf{X}}_k^T \tilde{\mathbf{X}}_k)^{-1} - \sigma^2(\sum_{k=1}^{K} \mathbf{X}_k^T \mathbf{X}_k)^{-1}\right\|_2 \\
=&\sigma^2\left\|(\sum_{k=1}^{K} n_k \tilde{\mathbf{X}}_k^T \tilde{\mathbf{X}}_k)^{-1}(\sum_{k=1}^{K} \mathbf{X}_k^T \mathbf{X}_k - \sum_{k=1}^{K} n_k \tilde{\mathbf{X}}_k^T \tilde{\mathbf{X}}_k)(\sum_{k=1}^{K} \mathbf{X}_k^T \mathbf{X}_k)^{-1}\right\|_2 \\
\leq&\sigma^2\left\|(\sum_{k=1}^{K} n_k \tilde{\mathbf{X}}_k^T \tilde{\mathbf{X}}_k)^{-1}\right\|_2\left\|(\sum_{k=1}^{K} \mathbf{X}_k^T \mathbf{X}_k - \sum_{k=1}^{K} n_k \tilde{\mathbf{X}}_k^T \tilde{\mathbf{X}}_k)\right\|_2\left\|(\sum_{k=1}^{K} \mathbf{X}_k^T \mathbf{X}_k)^{-1}\right\|_2 \\
\leq&\sigma^2 \lambda_1^{-1} \lambda_1^{*-1} \sum_{k=1}^{K} n_k \delta_k^2 \\
\leq&2\sigma^2 \sum_{k=1}^{K} n_k \delta_k^2 \lambda_1^{-2} \\
\leq&2\Delta^2 \sigma^2 N \lambda_1^{-2}
\end{aligned}
$$

Thus when Δ goes to zero, $\mathrm{Cov}(\tilde{\boldsymbol{\beta}})$ converges to $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ in terms of largest eigenvalue. □

3.3.3.1 Practical significance when Δ goes to zero

The difference between Fisher information matrices of OLS estimator $\hat{\beta}$ and $\tilde{\beta}$ can be written as $\sum_{k=1}^{K} n_k \Sigma_k$, with Σ_k the sample covariance matrix for predictors in the kth data block. So, with smaller Δ, the smaller the largest eigenvalue of sample covariance matrix for each data block, and more similarity of covariates between records in data block. From Theorem 3.3.1, with small values of Δ, the difference between MR estimator and OLS is negligible.

$\Delta = 0$ means there is no distinct for covariates in block, the only thing may differ is the response variable due to the randomness. It could happen when all the covariates are categorical or with finite values and the data is naturally partitioned by distinct covariate values. In this case, MR actually replaces the responses in the data block by their average.

From the proof of Theorem 3.3.1, we find that the convergence rate of covariance matrix is $O(\Delta^2)$. More specifically, if the number of blocks K is fixed, the optimal partitioning is to minimize $\sum_{k=1}^{K} n_k \delta_k^2$.

Assume the average density of kth covariate block is f_k, then we have $n_k = c \cdot \delta_k^p f_k$ for some constant c. Thus the goal is equivalent to minimize

$$\sum_{k=1}^{K} n_k^{1+2/p} f_k^{-2/p}$$

subject to $\sum_{k=1}^{K} n_k = N$.

Let $w_k = n_k/N$. Then $\sum_{k=1}^{K} w_k = 1$. The goal is to minimize

$$\sum_{k=1}^{K} w_k^{1+2/p} f_k^{-2/p} \tag{3.4}$$

subject to $\sum_{k=1}^{K} w_k = 1$. Rewrite Equation 3.4, we have

$$\sum_{k=1}^{K-1} w_k^{1+2/p} f_k^{-2/p} + (1 - \sum_{k=1}^{K-1} w_k)^{1+2/p} f_K^{-2/p}$$

For $k = 1, \ldots, K-1$,

$$\frac{\partial}{\partial w_k}[\sum_{k=1}^{K-1} w_k^{1+2/p} f_k^{-2/p} + (1 - \sum_{k=1}^{K-1} w_k)^{1+2/p} f_K^{-2/p}] = 0$$

implies $w_k/w_K = f_k/f_K$. That is, $w_k \sim f_k$. Therefore, there exists some constant c', such that

$$\delta_k = c \cdot n_k^{1/p} f_k^{-1/p} = c'$$

Consequently, the optimal partitioning should keep all blocks with approximately even size, and minimize the largest size. Therefore, k-means partition is the optimal choice for MR in linear regressions.

In case the grids of partition are coarse, the MR variance could be large and away from the variance of OLS.

3.3.3.2 Homogeneous linear representative family

In general, for any choice of linear combinations of records in block, the representative estimator is unbiased. That is, given collection of vectors $\{\alpha_k \in \mathbb{R}^p\}_{k=1}^K$, take the representatives as the linear combinations of records in each block, $(\tilde{\mathbf{X}}_k, \tilde{y}_k) = (\mathbf{X}_k^T \alpha_k, \mathbf{y}_k^T \alpha_k)$, then the estimator

$$\tilde{\beta} = (\sum_{k=1}^{K} n_k \mathbf{X}_k^T \alpha_k \alpha_k^T \mathbf{X}_k)^{-1} \sum_{k=1}^{K} n_k \mathbf{X}_k^T \alpha_k \alpha_k^T \mathbf{y}_k$$

is unbiased. For convenience and regularization, we require all components of α_k to be nonnegative and $\alpha_k^T \alpha_k = n_k^{-1}$. The covariance matrix of $\tilde{\boldsymbol{\beta}}$ is given by

$$\sigma^2 (\sum_{k=1}^{K} n_k \mathbf{X}_k^T \alpha_k \alpha_k^T \mathbf{X}_k)^{-1}$$

Such family of representative choices is called *homogeneous linear representative family*. MR belongs to this family.

3.4 Simulation Study with LMs

For an illustration purpose, we first run simulation studies based for a linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_7 x_{i7} + \epsilon_i \tag{3.5}$$

where $i = 1, \ldots, N$ and ϵ_i's are iid $\sim N(0, \sigma^2)$. Note that linear regression models are actually special cases of the generalized linear models with normal distributed responses and identity link (see Table VI).

Following Wang et al. (2017) (5), we assume $\beta_0 = 0$, $\beta_1 = \cdots = \beta_7 = 0.5$. We also set $\sigma^2 = 1$ for simulating responses. For simulating the predictors $\mathbf{x}_i = (x_{i1}, \ldots, x_{i7})^T$, we consider 6 unbounded distributions in Wang et al. (2017) (5), as well as a bounded distribution as follows:

1. **mzNormal**: $N_7(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ having diagonal 1 and off-diagnonal 0.5;
2. **nzNormal**: $N_7(\mathbf{1.5}, \boldsymbol{\Sigma})$, a case with imbalanced responses;
3. **ueNormal**: $N_7(\mathbf{0}, \boldsymbol{\Sigma}_u)$ with $\boldsymbol{\Sigma}_u$ having diagonals $\{1^2, \ldots, 7^2\}$ and off-diagnal 0.5;

4. **mixNormal**: $0.5N_7(\mathbf{1}, \mathbf{\Sigma}) + 0.5N_7(-1, \mathbf{\Sigma})$, a case with bimodal $\mathbf{x}_i$;
5. $\mathbf{T_3}$: Multivariate t with 3 degrees of freedom $\mathbf{t}_3(\mathbf{0}, \mathbf{\Sigma})/10$, a case with heavy tails;
6. **EXP**: $\exp(\lambda = 2)$, a case with a heavier tail on the right;
7. **BETA**: $\text{Beta}(\alpha = 0.5, \beta = 0.5)$, a bounded case with "U" shaped distribution.

The main-effects predictors in Equation 3.5 are for illustration purpose. The representative approach can actually work with general predictors included, such as interactions of covariates.

For illustration purpose, we choose a moderate sample size $N = 10^6$ in this simulation study. Since a natural partition for the simulated data is not available, we use two data-driven partitions:

- An equal-depth partition with $m = 4$ splits for each predictor, that is, use the three sample quartiles (25%, 50%, and 75% quantiles) as cut points for each predictor and partition the whole data into up to $4^7 = 16384$ blocks;
- A k-means partition with $K = 1000$.

A great advantage of grid partition is the significant reduction in complexity, especially when the original data provided in files, where a natural grid performed already. Quantiles are more preferred than equal division points since the latter sometimes may lead to imbalance blocks, where some of them may contain very small numbers of observations and some contain large proportions of observations, especially in the case of high skewed data where too many empty blocks are produced such that it is not feasible for the model fitting. The k-means partition emphasizes more on similarity in-block, but dissimilarity between blocks.

TABLE I: Average (std) of RMSEs (10^{-3}) of 20 simulations for linear model with $N = 10^6$

Simulation setup	Full data	Equal-depth ($m = 4$)			k-means ($K = 1000$)		IBOSS
		Mid	Med	MR	Med	MR	
mzNormal	1.3(0.1)	239.3(0.7)	28.8(0.0)	**1.5**(0.1)	3.3(0.1)	**1.6**(0.1)	6.8(0.4)
nzNormal	1.3(0.1)	239.3(0.7)	28.8(0.0)	**1.5**(0.1)	3.3(0.1)	**1.6**(0.1)	6.8(0.4)
ueNormal	.45(.04)	251.7(0.8)	43.6(0.1)	**.47**(.04)	13.2(0.2)	**.44**(.04)	2.3(0.2)
mixNormal	1.3(0.1)	201.3(0.7)	16.4(0.0)	**1.4**(0.1)	2.0(0.1)	**1.4**(0.1)	7.5(0.4)
T_3	6.6(0.5)	483.6(1.0)	107.2(0.5)	**9.0**(0.6)	11.2(0.5)	**8.4**(0.4)	12.0(0.9)
EXP	1.9(0.1)	369.3(1.0)	76.8(0.2)	**2.1**(0.1)	30.1(0.3)	**2.0**(0.1)	6.0(0.4)
BETA	3.2(0.2)	27.9(0.3)	12.8(0.3)	**3.3**(0.2)	33.5(0.3)	**3.3**(0.2)	18.2(1.2)

Table I shows the root mean squared errors (RMSEs) $(\sum_{i=1}^{7}(\tilde{\beta}_i - \beta_i)^2/7)^{1/2}$ between the estimated parameter values $\tilde{\boldsymbol{\beta}}$'s based different methods and the true value $\boldsymbol{\beta}$ across different simulation settings each with 20 independent simulations. Average of RMSEs and standard deviation of average are provided in Table I and all future tables related to RMSE.

In terms of RMSE, Table I clearly shows that even with moderate size of block, the MR outperforms both mid-point (Mid) and median (Med) representative approaches, as well as IBOSS proposed by Wang et al. (2018) (8) with subsample size $20,000$, which is larger than the number of non-empty blocks. Compared with the true parameter value, MR is comparable even with the estimates based on the full data. As for **ueNormal**, average of MR is slightly smaller than full data fit. But by paired t-test, this discrepancy is not significant. If the data and partitioning are given, then the result of MR is deterministic. Thus box-plots may provide more information than standard deviations.

TABLE II: Average intercepts (10^{-5}) estimation of 20 simulations for linear model with $N = 10^6$

Simulation setup	full data	median	MR
mzNormal	-6.0	-24.3	-6.0
nzNormal	-75.6	-30154.6	-61.8
ueNormal	-6.0	11.6	-5.9
mixNormal	0.3	-11.9	0.4
T_3	22.5	21.4	22.5
EXP	-2.5	-10381.7	6.5
BETA	-75.9	4366.3	-98.9

Note that the RMSE of MR based on equal-depth partition obtained from average $11488 \sim 16384$ non-empty blocks are comparable with the RMSE from k-means with $K = 1000$. It indicates representative approaches based on clustered partition are more efficient, which is confirmed by our theoretical justifications in Theorem 4.1.1.

3.4.1 Naive labeling algorithm for grid partition

Suppose each predictor X_j is split into m_j intervals, with cut points $\{s_1^{(j)}, s_2^{(j)}, \ldots, s_{m_j-1}^{(j)}\}$. We purpose an algorithm to label observations such that observations in the same block share the same label. Given a piece of observation $x = (x_1, x_2, \ldots, x_p)$, its label id is given by

$$id = \sum_{j=1}^{p} \left([\prod_{i=j+1}^{p} m_i][\sum_{k=1}^{m_i-1} \mathbb{1}\{x_j > s_k^{(j)}\}] \right)$$

where we make convention that $\prod_{i=p+1}^{p} m_i := 1$.

On the other hand, given id, we can recover the location of each predictor of that point. Define $c_1 = id\%/\%(\prod_{i=2}^{p} m_i)$ and $r_1 = id \mod (\prod_{i=2}^{p} m_i)$ to be the integer quotient and the

reminder respectively, where %/% is the integer division operator and mod is the modulus operator. Then define $c_j = r_{j-1}\%/\%(\prod_{i=j+1}^{p} m_i)$ and $r_j = r_{j-1} \mod (\prod_{i=j+1}^{p} m_i)$. That is, x_j locates in the c_jth interval.

3.4.2 CPU time of MR

For all experiments in this dissertation, we use the R programming language (R version 3.4.4). For the IBOSS method and the A-optimal subsampling method, we use the packages provided by Haiying Wang on his website. As for center representatives, "data.table" package is used to achieve the calculating by group. Function "data.Cluster()" is used to perform k-means partition. All computations are carried out on a single thread of a server running Ubuntu 16.04.4 with Intel Xeon CPU E5-2695 v4 @ 2.10GHz and 377GB memory.

The CPU time for MR and IBOSS is shown in Table III for linear models with $N = 10^6$ and $p = 7$, using both equal-depth with $m = 4$ and k-means with $K = 1000$ partition. When we increase the number of parameters p to 10, in **mzNormal**, the number of non-empty blocks is around 375000, one-third of sample size. If $p = 14$, the number of non-empty blocks is almost equal to the sample size $N = 10^6$. Equal-depth partitioning is not a good option for large p, since the number of blocks increases exponentially as p increases and approaches N quickly. Therefore, only clustering methods will be considered when p is not small.

For illustration purpose, we also use equal-depth and k-means partition on the simulation studies showing the relation of the CPU time against the number of parameters p and sample size N respectively. The computational time to apply such a k-means is high though. According to the time complexity analysis in Section 3.3, the computational time of MR should roughly

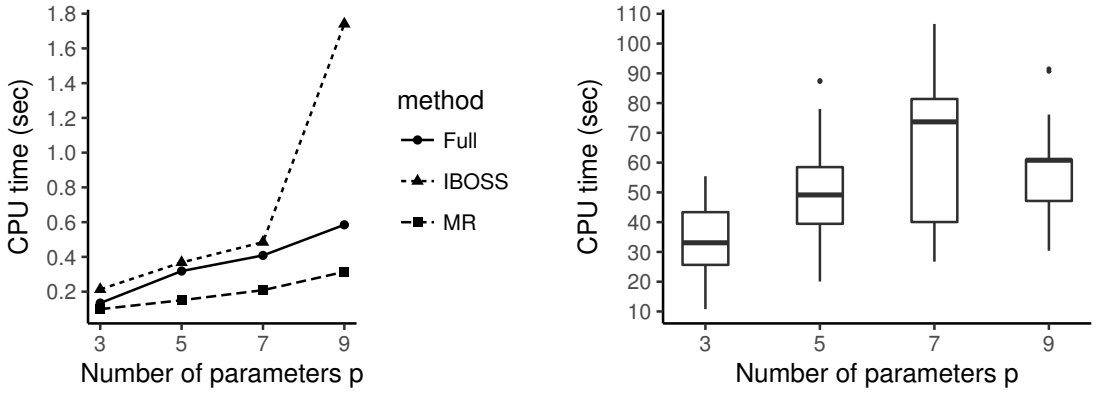
TABLE III: Average CPU time (sec) of MR and IBOSS over 100 simulations for linear model with $N = 10^6$, $p = 7$

Simulation setup	Full	Equal-depth Partition	Equal-depth MR	k-means Partition	k-means MR	IBOSS
mzNormal	0.463	2.551	0.498	70.669	0.188	2.018
nzNormal	0.507	2.404	0.492	71.130	0.191	1.916
ueNormal	0.373	2.545	0.485	74.665	0.214	1.973
mixNormal	0.388	2.528	0.417	65.670	0.189	1.865
T_3	0.451	2.384	0.550	96.920	0.211	1.793
EXP	0.473	2.357	0.549	113.779	0.243	1.826
BETA	0.401	2.405	0.521	134.842	0.228	1.789

proportional to the number of parameters p and sample size N. Our simulation studies in Figure 1 and Figure 2 confirm our conclusion. With a prior partitioning provided, MR is comparable to IBOSS method in terms of computational time. But also, if there is no partitioning provided, and a k-means partitioning is required, then a very high computational cost has to be paid additionally.

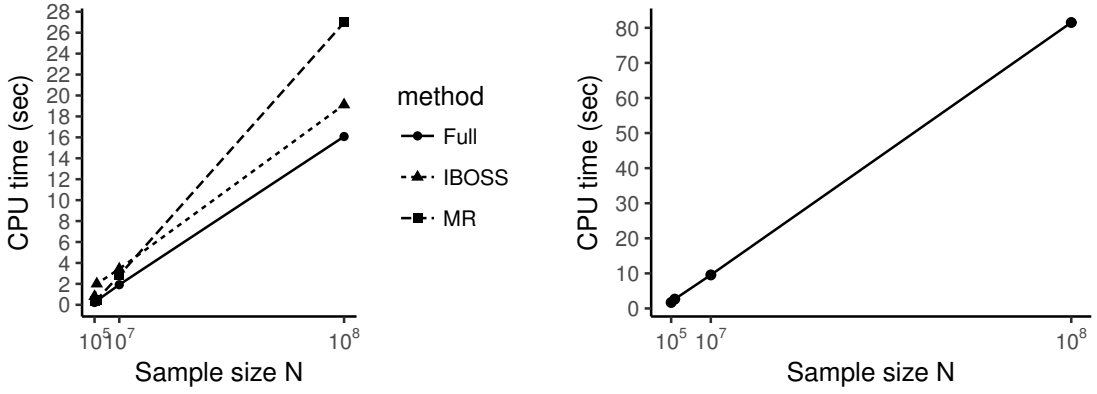
Equal depth partition can be performed quickly for small p, and linear to N. As stated in Theorem 3.3.1, k-means partition is the best partitioning when the number of partitions is fixed, but it is slow currently using "data.Cluster()" shown in Figure 1 (b) and Figure 3. The latter shows that computational time of k-means using "data.Cluster()" is high for large N.

How to obtain a more efficient partition is very important to representative approaches, but out of the scope of this dissertation work.



(a) MR with k-means provided (b) k-means partitioning

Figure 1: Average CPU time of full data fit, MR and IBOSS over 20 simulations against $\mathtt{p}$ for linear model with $\mathtt{N} = 10^6$, **mzNormal**, under k-means ($\mathtt{K} = 1000$)



(a) MR with equal-depth provided (b) Equal-depth partitioning

Figure 2: Average CPU time of full data fit, MR and IBOSS over 20 simulations against $\mathtt{N}$ for linear model with $\mathtt{p} = 7$, **mzNormal**, under equal-depth ($\mathtt{m} = 4$)

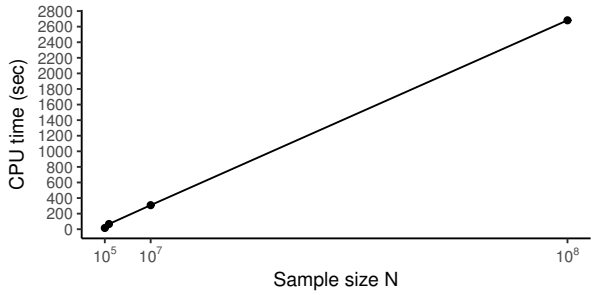


Figure 3: Average CPU time of k-means partition ($\mathsf{K} = 1000$) over 20 simulations against N for linear model with $\mathsf{p} = 7$, **mzNormal**

CHAPTER 4

REPRESENTATIVES FOR GENERALIZED LINEAR MODELS

4.1 Center Representatives for GLMs

4.1.1 Theoretical justification for center representatives for GLMs

The simulation studies in Section 3.4 imply that the maximum distance within data blocks, denoted as $\Delta = \max_k \max_{i,j\in I_k} \|\mathbf{X}_i - \mathbf{X}_j\|$, may play an important role in extracting data information more efficiently. Given the data set $\mathscr{D} = \{(\mathbf{X}_i, y_i), i = 1, \ldots, N\}$. Recall that in the generalized linear model of Section 2.1, $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate (MLE) based on the full data set $\mathscr{D} = \{(\mathbf{X}_i, y_i), i = 1, \ldots, N\}$ and $\tilde{\boldsymbol{\beta}}$ is the MLE based on the representative data set $\tilde{\mathscr{D}} = \{(n_k, \tilde{\mathbf{X}}_k, \tilde{y}_k), k = 1, \ldots, K\}$. We denote $\tilde{\Delta} = \max_k \max_{i\in I_k} \|\mathbf{X}_i - \tilde{\mathbf{X}}_k\|$.

Theorem 4.1.1. *Suppose the log-likelihood function* $l(\boldsymbol{\beta})$ *is strictly concave on a compact set* $B \in \mathbb{R}^p$ *and the maximum can be achieved in the interior of* B. *If* $\tilde{y}_k = n_k^{-1} \sum_{i\in I_k} y_i$, *then* $\left\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right\| \to 0$ *as* $\tilde{\Delta} \to 0$*; additionally,* $\left\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right\| = O(\tilde{\Delta}^{\frac{1}{2}})$.

Lemma 4.1.1 (Theorem 2.1 of Kanniappan and Sastry (1983) (17))**.** *Let* $f : X \to Y$ *and* $f_n : X \to Y$ *be continuous, convex maps. Suppose* $f_n \to f$ *uniformly. Then the sequence of* $\arg\inf_x f_n(x)$ *converges to* $\arg\inf_x f(x)$.

Proof of Theorem 4.1.1:

The log-likelihood contributed by the kth block essentially is

$$l_k(\boldsymbol{\beta}) = \sum_{i \in I_k} [y_i \theta(\mathbf{X}_i^T \boldsymbol{\beta}) - b(\theta(\mathbf{X}_i^T \boldsymbol{\beta}))]$$

Consider the log-likelihood of representative from kth block,

$$\begin{aligned} \tilde{l}_k(\boldsymbol{\beta}) =& n_k[\tilde{y}_k \theta(\tilde{\mathbf{X}}_k^T \boldsymbol{\beta}) - b(\theta(\tilde{\mathbf{X}}_k^T \boldsymbol{\beta}))] \\ =& (n_k \tilde{y}_k - \sum_{i \in I_k} y_i)\theta(\tilde{\eta}_k) + \sum_{i \in I_k} [y_i \theta(\tilde{\mathbf{X}}_k^T \boldsymbol{\beta}) - b(\theta(\tilde{\mathbf{X}}_k^T \boldsymbol{\beta}))] \end{aligned} \tag{4.1}$$

Since $\tilde{y}_k = n_k^{-1} \sum_{i \in I_k} y_i$, the first term in Equation 4.1 equals 0.

Consider the first order Taylor expansion of $\tilde{l}_k$ about $\tilde{\mathbf{X}}_k$ at $\mathbf{X}_i$, and by the Cauchy-Schwarz Inequality, we have

$$\begin{aligned} \left|\tilde{l}_k(\boldsymbol{\beta}) - l_k(\boldsymbol{\beta})\right| &= \sum_{i \in I_k} \{[y_i \theta'(\mathbf{X}_i^T \boldsymbol{\beta}) - b'(\theta(\mathbf{X}_i^T \boldsymbol{\beta}))\theta'(\mathbf{X}_i^T \boldsymbol{\beta})](\tilde{\mathbf{X}}_k - \mathbf{X}_i)^T \boldsymbol{\beta} + o(\left\|\tilde{\mathbf{X}}_k - \mathbf{X}_i\right\|)\} \\ &= \sum_{i \in I_k} \{[(y_i - G(\mathbf{X}_i^T \boldsymbol{\beta}))v(\mathbf{X}_i^T \boldsymbol{\beta})](\tilde{\mathbf{X}}_k - \mathbf{X}_i)^T \boldsymbol{\beta} + o(\left\|\tilde{\mathbf{X}}_k - \mathbf{X}_i\right\|)\} \\ &\leq (\sum_{i \in I_k} \left\|\tilde{\mathbf{X}}_k - \mathbf{X}_i\right\|^2 \sum_{i \in I_k} [(y_i - G(\mathbf{X}_i^T \boldsymbol{\beta}))v(\mathbf{X}_i^T \boldsymbol{\beta})\|\boldsymbol{\beta}\|]^2)^{\frac{1}{2}} + \sum_{i \in I_k} o(\left\|\tilde{\mathbf{X}}_k - \mathbf{X}_i\right\|) \\ &\leq n_k \tilde{\Delta} \|\boldsymbol{\beta}\| (n_k^{-1} \sum_{i \in I_k} [(y_i - G(\mathbf{X}_i^T \boldsymbol{\beta}))v(\mathbf{X}_i^T \boldsymbol{\beta})]^2)^{\frac{1}{2}} + \sum_{i \in I_k} o(\tilde{\Delta}) \end{aligned}$$

Denote $F_k = (n_k^{-1} \sum_{i \in I_k} [(y_i - G(\mathbf{X}_i^T \boldsymbol{\beta}))v(\mathbf{X}_i^T \boldsymbol{\beta})]^2)^{\frac{1}{2}}$.

Therefore for sufficient small $\tilde{\Delta}$, we have

$$\frac{1}{N}\left|\tilde{l}(\boldsymbol{\beta}) - l(\boldsymbol{\beta})\right| \leq \frac{1}{N}\sum_{k=1}^{K} n_k \tilde{\Delta}\|\boldsymbol{\beta}\| F_k + \frac{1}{N}\sum_{i=1}^{N} o(\tilde{\Delta}) \tag{4.2}$$

$$\leq \tilde{\Delta}\|\boldsymbol{\beta}\| \max_k F_k + o(\tilde{\Delta}) \tag{4.3}$$

$$\leq M\tilde{\Delta} \tag{4.4}$$

for some $M > 0$. That is, $\frac{1}{N}\tilde{l}(\boldsymbol{\beta})$ converges to $\frac{1}{N}l(\boldsymbol{\beta})$ uniformly as $\tilde{\Delta}$ goes to 0 for $\boldsymbol{\beta}$ on a compact set.

By Lemma 4.1.1, the MLE of $\tilde{l}$ converges to the MLE of l as $\tilde{\Delta} \to 0$.

The strict concavity of $l(\boldsymbol{\beta})$ implies the existence of unique $\hat{\boldsymbol{\beta}} \in B$, such that $\hat{\boldsymbol{\beta}} = \arg\max_{\beta} l(\boldsymbol{\beta})$. Let $\tilde{\boldsymbol{\beta}}$ maximizes $\tilde{l}(\boldsymbol{\beta})$. Then we have that

$$\frac{1}{N}\left|\tilde{l}(\tilde{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}})\right| \leq M\tilde{\Delta}$$

Therefore

$$0 \leq \frac{1}{N}(l(\hat{\boldsymbol{\beta}}) - l(\tilde{\boldsymbol{\beta}})) \leq 2M\tilde{\Delta}$$

Consider the second Taylor expansion of $l(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}$

$$l(\boldsymbol{\beta}) = l(\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\Big|_{\beta=\hat{\beta}} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + o(\left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|^2)$$

since $\frac{\partial l}{\partial \beta}(\hat{\beta}) = 0$. The smallest eigenvalue of $\frac{\partial^2 l}{\partial \beta \partial \beta^T}\Big|_{\beta=\hat{\beta}}$, λ_1, is positive, i.e., $\lambda_1 > 0$, since $l(\beta)$ is strictly concave.

Therefore,

$$\left|\tilde{l}(\beta) - l(\beta)\right| > \frac{\lambda_1}{2}\left\|\beta - \hat{\beta}\right\|^2 \tag{4.5}$$

for small enough $\left\|\beta - \hat{\beta}\right\|$. It can be verified that

$$\left\|\tilde{\beta} - \hat{\beta}\right\| \leq \sqrt{\frac{4MN}{\lambda_1}}\tilde{\Delta}^{\frac{1}{2}}$$

Actually, if $\left\|\tilde{\beta} - \hat{\beta}\right\|^2 > 4MN/\lambda_1\tilde{\Delta}$, then by Equation 4.5,

$$l(\hat{\beta}) - l(\tilde{\beta}) > 2MN\tilde{\Delta}$$

or

$$l(\tilde{\beta}) < l(\hat{\beta}) - 2MN\tilde{\Delta} \tag{4.6}$$

From Equation 4.4 and Equation 4.6, we know that

$$\tilde{l}(\tilde{\beta}) < l(\hat{\beta}) - MN\tilde{\Delta} \leq \tilde{l}(\hat{\beta})$$

That leads to a contradiction.

That is,

$$\left\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right\| = O(\tilde{\Delta}^{\frac{1}{2}})$$

□

Remark. *Denote the partition of the predictor space* $\mathbb{R}^p$ *or its subset by* $\{B_1, \ldots, B_K\}$. *Assume the predictors* $\mathbf{X}_1, \ldots, \mathbf{X}_N \in \mathbb{R}^p$ *are iid* $\sim F$ *with a finite expectation, and for* $k = 1, \ldots, K$, $[(y - G(\mathbf{X}^T\boldsymbol{\beta}))v(\mathbf{X}^T\boldsymbol{\beta})]^2$ *on* B_k *has finite expectation. That is,* $\max_k F_k$ *is bounded in probability, for* $k = 1, \ldots, K$. *Therefore, if the distribution of data is fixed, increasing of* N *will not affect the convergency and convergent rate.*

Remark. *For MR approach,* $\tilde{\mathbf{X}}_k = n_k^{-1} \sum_{i \in I_k} \mathbf{X}_i$. *It can be verified that* $\tilde{\Delta} \leq \Delta$ *instantly. That is, for MR approach, if* $\Delta \to 0$, *then* $\tilde{\boldsymbol{\beta}} \to \hat{\boldsymbol{\beta}}$.

For median representative approach, $\tilde{\Delta}$ *is also controlled by* Δ. *Consider in a block we have data* $\mathbf{X}_i = (x_{i1}, \ldots, x_{ip})^T$, $i = 1, \ldots, n$ *and the median representative is given by* $\tilde{\mathbf{X}} = (\tilde{x}_1, \ldots, \tilde{x}_p)^T$. *Let* $x_{(1)j}, \ldots, x_{(n)j}$ *be ordered statistic of* $x_{1j}, \ldots, x_{nj}$, $j = 1, \ldots, p$.

$$\begin{aligned}
\max_i \left\|\mathbf{X}_i - \tilde{\mathbf{X}}\right\| &= \sum_{j=1}^{p} (x_{ij} - \tilde{x}_j)^2 \\
&\leq \sum_{j=1}^{p} (x_{(1)j} - x_{(n)j})^2 \\
&\leq p \max_{ij} \|\mathbf{X}_i - \mathbf{X}_j\|
\end{aligned}$$

Therefore, if $\Delta \to 0$, *then* $\tilde{\boldsymbol{\beta}} \to \hat{\boldsymbol{\beta}}$ *for median representative.*

As for mid-point representative, we should redefine Δ as the maximum grid size to enjoy the results of Theorem 4.1.1

When all the covariates are categorical or have finite discrete values, one may partition the data according to distinct $\mathbf{X}_i$'s. In this case, $\Delta = 0$.

Corollary 4.1.1. *If $\Delta = 0$, then the mid-point, median, and mean representative approaches are the same and all satisfy $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$.*

Proof. In each block, all the predictor variables are the same, i.e., $\mathbf{X}_i \equiv \bar{\mathbf{X}}_k$ for $i \in I_k$,$k = 1, \ldots, K$. Consider the log-likelihood of MR from kth data block,

$$\begin{aligned}\tilde{l}_k =& n_k[\bar{y}_k \theta(\bar{\mathbf{X}}_k^T \boldsymbol{\beta}) - b(\theta(\bar{\mathbf{X}}_k^T \boldsymbol{\beta}))] \\ =& \sum_{i \in I_k} [y_i \theta(\mathbf{X}_i^T \boldsymbol{\beta}) - b(\theta(\mathbf{X}_i^T \boldsymbol{\beta}))] \\ =& l_k\end{aligned}$$

Therefore, MR method exactly meets the full data fit. Similarly, mid-point and median representatives meet the full data estimate as well. □

4.1.2 Simulation studies with logistic model

The MR approach works very well for linear models and has been validated for GLMs when Δ is sufficiently small. Nevertheless, simulation studies for moderate Δ with general GLMs in this section show that MR approach is not that satisfactory.

Logistic regression model is one of the most widely used generalized linear models. It connects an expected Bernoulli response or binomial response with linear predictors by the logit link $g(\mu) = \log[\mu/(1-\mu)]$. Wang et al. (2017) (5) proposed a subsampling strategy inspired by A-optimality.

Table IV shows results from a comprehensive simulation study with logistic regression models. We use the same setup in Section 3.4 for simulating the predictors, as well as the true parameter values. For comparison purpose, we include (Wang et al., 2017 (5))'s A-optimal subsampling estimates with subsample size $20,000$, which is bigger than the average number of representatives ($11488 \sim 16384$). Also, a new representative approach proposed in Section 4.2, SMR, is included.

Based on Table IV, for logistic regression models, mid-point representative does not work so well as others, its RMSEs to true value are above 0.2 for unbounded cases. MR performs very good in the bounded case **BETA**, while RMSEs of MR are below or around 0.03 for almost all the cases except for **ueNormal**, where a high proportion of the linear predictor η_i's are extremely large. Compared with MR, median representative approach still has a bias issue in estimating intercept β_0, although its RMSE for estimating $\boldsymbol{\beta}$ seems a slightly better than MR for the four normal distributions. Both median and mean representative approaches perform better than A-optimal subsampler for most simulation settings except for the non-equal-variance case **ueNormal**.

Table IV shows that SMR with MR estimates as its initial values performs uniformly the best, even comparable with the estimates based on the full data. Based on equal-depth partition

TABLE IV: Average (std) of RMSEs (10^{-3}) of 20 simulations for *logistic* model with $N = 10^6$

Simulation setup	Full data	Equal-depth ($m = 4$)				k-means ($K = 1000$)			A-opt
		Mid	Med	MR	SMR	Med	MR	SMR	
mzNormal	3.6	266.8	8.8	20.3	**4.0**	15.7	17.7	**4.0**	28.0
	(0.2)	(0.6)	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)	(0.3)	(0.2)
nzNormal	7.3	206.4	12.5	20.5	**9.7**	13.0	15.4	**8.6**	435.5
	(0.4)	(0.9)	(0.7)	(0.5)	(0.6)	(0.3)	(0.3)	(0.4)	(0.1)
ueNormal	1.8	344.0	143.8	170.2	**4.5**	205.1	208.5	**4.1**	13.5
	(0.2)	(0.6)	(0.2)	(0.1)	(0.2)	(0.4)	(0.4)	(0.4)	(0.1)
mixNormal	5.1	220.5	10.9	19.9	**5.4**	17.9	17.7	**6.2**	31.6
	(0.4)	(0.6)	(0.3)	(0.3)	(0.4)	(0.3)	(0.3)	(0.4)	(0.2)
T_3	18.3	484.5	81.0	31.3	**21.8**	22.9	21.7	**21.4**	140.5
	(1.0)	(0.9)	(0.7)	(0.6)	(0.8)	(1.6)	(1.5)	(1.5)	(1.0)
EXP	6.5	374.8	49.8	23.3	**13.8**	14.4	12.9	**9.0**	269.6
	(0.5)	(1.0)	(0.7)	(0.5)	(0.5)	(0.6)	(0.5)	(0.5)	(0.4)
BETA	7.2	26.4	16.5	7.8	**7.7**	36.3	7.3	**6.9**	245.1
	(0.3)	(0.6)	(0.6)	(0.4)	(0.4)	(0.5)	(0.3)	(0.3)	(0.7)

with $m = 4$ (up to 16,384 representatives), the RMSE of MR is 0.1702 on average for **ueNormal** simulation setup, while SMR pulls the RMSE back to 0.0045. With a better partition, such as k-means, SMR can achieve a similar accuracy level with only 1000 representatives.

As a conclusion, when the predictors are bounded or the proportion of extremely large linear predictors is low, MR is a fast and low-cost (computationally cheaper) solution for big data analysis using generalized linear models. It is better than mid-point, median, or A-optimal approaches. Nevertheless, MR may not be satisfactory if higher accuracy level is desired. In that case, MR is used as a pre-analysis of SMR for generalized linear models, while the latter has a huge improvement over different distributions and different partition methods, and benefits from the large sample size.

TABLE V: Average intercept (10^{-5}) estimate of 20 simulations for logistic model with $N = 10^6$ based on equal-depth with $m = 4$

Simulation setup	Full data	Med	MR
mzNormal	-92.6	-99.7	-89.3
nzNormal	8.9	-27441.2	555.1
ueNormal	72.8	42.6	25.8
mixNormal	-52.1	-61.8	-50.0
T_3	-6.6	-8.5	-7.4
EXP	142.0	-5444.0	3985.7
BETA	311.5	4800.3	348.8

4.2 Score-Matching Representative Approach for GLMs

Section 4.1.2 shows that for moderate Δ with general GLMs, MR approach is not so satisfactory when there are extreme values in either predictors or coefficients. In this section, we propose a much more efficient representative approach, called *score-matching representative* (SMR) approach, for GLMs. Its asymptotic efficiency is even better than divide-and-conquer approach, with reduced time complexity due to the representative strategy.

Recall that preliminary results in Section 2.1, the MLE $\hat{\boldsymbol{\beta}}$ solves the score equation $s(\boldsymbol{\beta}) = 0$. It is usually solved by the Fisher scoring method, which iteratively updates the score function with the current estimate of $\boldsymbol{\beta}$.

Inspired by the Fisher scoring method, given some initial values of the estimated parameters, our score-matching representative approach builds data representatives by matching the values of the score function block by block, then applies the Fisher Scoring method on the representative dataset and gets estimated parameter values. We may use the current estimated

parameter value as initial value for the next iteration, and repeat this procedure for a few times till a certain accuracy level is achieved. According to our comprehensive simulation studies (see Section 4.1.2), three iterations are satisfactory for typically applications.

4.2.1 Score-matching representative approach

Let $s_k(\boldsymbol{\beta})$ denote the score function contributed by kth data block $\mathscr{D}_k = \{(\mathbf{X}_i, y_i), i \in I_k\}$, and $\tilde{s}_k(\boldsymbol{\beta})$ denote the score function based on the weighted representative data of kth block $(n_k, \tilde{\mathbf{X}}_k, \tilde{y}_k)$.

Suppose the estimated parameter is $\tilde{\boldsymbol{\beta}}^{(t)}$ at the tth iteration. For the $(t+1)$th iteration, our strategy is to find the representative $(\tilde{\mathbf{X}}_k, \tilde{y}_k)$ carrying the same score as the kth data block at $\tilde{\boldsymbol{\beta}}^{(t)}$, that is,

$$\sum_{i \in I_k} \nu(\mathbf{X}_i^T \tilde{\boldsymbol{\beta}}^{(t)})(y_i - G(\mathbf{X}_i^T \tilde{\boldsymbol{\beta}}^{(t)}))\mathbf{X}_i = n_k \, \nu(\tilde{\mathbf{X}}_k^T \tilde{\boldsymbol{\beta}}^{(t)})(\tilde{y}_k - G(\tilde{\mathbf{X}}_k^T \tilde{\boldsymbol{\beta}}^{(t)}))\tilde{\mathbf{X}}_k \tag{4.7}$$

which implies $s_k(\tilde{\boldsymbol{\beta}}^{(t)}) = \tilde{s}_k(\tilde{\boldsymbol{\beta}}^{(t)})$. Multiplying by $\tilde{\boldsymbol{\beta}}^{(t)}$ both sides of Equation 4.7, we get

$$\sum_{i \in I_k} \nu(\eta_i)(y_i - G(\eta_i))\eta_i = n_k \, \nu(\tilde{\eta}_k)(\tilde{y}_k - G(\tilde{\eta}_k))\tilde{\eta}_k \tag{4.8}$$

where $\eta_i = \mathbf{X}_i^T \boldsymbol{\beta}^{(t)}$ and $\tilde{\eta}_k = \tilde{\mathbf{X}}_k^T \boldsymbol{\beta}^{(t)}$. The weight $\nu(\eta_i)\eta_i$ of y_i in Equation 4.8 suggests that we take $\tilde{y}_k$ as a weighted average of y_i's for the SMR approach, that is,

$$\tilde{y}_k = [\sum_{i \in I_k} \nu(\eta_i)\eta_i]^{-1} \sum_{i \in I_k} \nu(\eta_i)\eta_i y_i \tag{4.9}$$

Note that $\tilde{y}_k \to n_k^{-1} \sum_{i \in I_k} y_i$ as Δ goes to 0. That is, $\tilde{y}_k$ in Equation 4.9 is a natural generalization of the mean representative.

Since $\tilde{y}_k$ in Equation 4.9 does not rely on $\tilde{\eta}_k$, we can further obtain $\tilde{\eta}_k$ by solving Equation 4.8.

Theorem 4.2.1. *There exists an* $\tilde{\eta}_k \in [\min_{i \in I_k} \eta_i,\ \max_{i \in I_k} \eta_i]$ *that solves Equation 4.8.*

Proof. Define $S(\eta; \tilde{y}_k) := \nu(\eta)(\tilde{y}_k - G(\eta))\eta$, the Equation 4.8 is equivalent to

$$n_k^{-1} \sum_{i \in I_k} S(\eta_i; \tilde{y}_k) = S(\tilde{\eta}_k; \tilde{y}_k) \tag{4.10}$$

by definition of $\tilde{y}_k$ in Equation 4.9. This is a 1-dimensional nonlinear equation about $\tilde{\eta}_k$. There exists at least one solution between $\min_{i \in I_k}\{\eta_i\}$ and $\max_{i \in I_k}\{\eta_i\}$, since $S(\cdot\,; \tilde{y}_k)$ is continuous. □

Remark. *The existence of this representative choice is guaranteed by the existence of the solution to Equation 4.8 about* $\tilde{\eta}_k$*. The solution of* $\tilde{\eta}_k$ *may not be unique since* S *usually is not monotone. In that case, we choose the* $\tilde{\eta}_k$ *with* $\tilde{\mathbf{X}}_k$ *closest to* $\bar{\mathbf{X}}_k$ *to keep* $\tilde{\Delta}$ *as small as possible, thus consistent with the mean representative.*

Remark. *If the original predictors contain the intercept term 1, the corresponding representative predictor may not be exactly 1.*

After plugging $\tilde{\eta}_k$ into Equation 4.7, we get the representative $\tilde{\mathbf{X}}_k$ for SMR:

$$\tilde{\mathbf{X}}_k = [n_k \nu(\tilde{\eta}_k)(\tilde{y}_k - G(\tilde{\eta}_k))]^{-1} \sum_{i \in I_k} \nu(\eta_i)(y_i - G(\eta_i))\mathbf{X}_i \tag{4.11}$$

Remark. *The solution of* $\tilde{\mathbf{X}}_k$ *is extremely sensitive to the accuracy of* $\tilde{\eta}_k$*, so is* $\tilde{s}_k(\tilde{\boldsymbol{\beta}}^{(t)})$*, when* $\nu(\tilde{\eta}_k)(\tilde{y}_k - G(\tilde{\eta}_k))$ *is close to 0. Since our criterion is* $s_k(\tilde{\boldsymbol{\beta}}^{(t)}) = \tilde{s}_k(\tilde{\boldsymbol{\beta}}^{(t)})$*, we should take a strict tolerance on the solution to Equation 4.8. Some adjustment is considered if* $\tilde{\mathbf{X}}_k$ *is away from* $\bar{\mathbf{X}}_k$ *even with a high accuracy requirement taken already.*

The representative $\tilde{\mathscr{D}}_k = (n_k, \tilde{\mathbf{X}}_k, \tilde{y}_k)$ now carries the same scoring value as the data block, $\tilde{s}_k(\beta) = s_k(\beta)$, $k = 1, \ldots, K$ for the $(t+1)$th iteration. The size of representative dataset only relies on the number of blocks K, significant smaller than the original sample size N. We apply the Fisher Scoring algorithm on the representative dataset $\tilde{\mathscr{D}} = \{(n_k, \tilde{\mathbf{X}}_k, \tilde{y}_k), k = 1, \ldots, K\}$ and get $\tilde{\boldsymbol{\beta}}^{(t+1)}$.

We may repeat this procedure for a few times, say 3 times, to achieve the desired accuracy level (see Figure 4, Figure 5, Figure 6 in Section 4.3). See Algorithm 2 for the complete procedure of SMR.

4.2.1.1 Time complexity of SMR

Since each time we are dealing with K representative data points and its accuracy depends on Δ instead of N, the computational cost has been significantly reduced as well. For general GLM, to calculate all η_i, the time is $O(Np)$. For kth data block, to calculate $\tilde{y}_k$ through Equation 4.9, the time is $O(n_k)$. The time to list Equation 4.8 is $O(n_k)$ and ζ_r iterations is required to solve this 1-dimensional nonlinear equation. Also, $\tilde{\mathbf{X}}_k$ in Equation 4.11 requires $O(n_k p)$. Along with the time to proceed GLM MLE on K representative points, $O(\zeta_K K p^2)$, 3-iteration SMR requires $O(Np + N\zeta_r + \zeta_K K p^2)$. If $\zeta_r, \zeta_K, K, p \ll N$, the time complexity of

Algorithm 2: Score-Matching Representative Method

Data: Partitioning of $\mathscr{D}$: $\{\mathscr{D}_k = (\mathbf{X}_k, \mathbf{y}_k)\}_{k=1}^K$, threshold δ
Result: SMR estimator $\tilde{\boldsymbol{\beta}}$ for Generalized Linear Model.

Get the mean representative set $\tilde{\mathscr{D}}^{(0)} = \{(n_k, \tilde{\mathbf{X}}_k^{(0)}, \tilde{y}_k^{(0)})\}_{k=1}^K$;
Initialize allocation $\tilde{\boldsymbol{\beta}}^{(0)}$ by apply Fisher Scoring method on $\tilde{\mathscr{D}}^{(0)}$;
Set $t = 0$;
while $\left\|\tilde{\boldsymbol{\beta}}^{(t)} - \tilde{\boldsymbol{\beta}}^{(t-1)}\right\| > \delta$ **do**
    Set $t := t + 1$;
    for $k = 1, \ldots, K$ **do**
        Calculate $\eta_i^{(t)} := \mathbf{X}_i^T \beta^{(t-1)}$ for $i \in I_k$;
        Calculate $\tilde{y}_k^{(t)}$ by Equation 4.9;
        Solve the nonlinear equation Equation 4.8 for $\tilde{\eta}_k^{(t)}$;
        Calculate $\tilde{\mathbf{X}}_k^{(t)}$ by Equation 4.11;
    end
    Update the regression parameter by applying Fisher Scoring method on the representative dataset $\tilde{\mathscr{D}}^{(t)} = \{(n_k, \tilde{\mathbf{X}}_k^{(t)}, \tilde{y}_k^{(t)})\}_{k=1}^K$ to get $\tilde{\boldsymbol{\beta}}^{(t)}$;
end
$\tilde{\boldsymbol{\beta}} := \tilde{\boldsymbol{\beta}}^{(t)}$

SMR is essentially $O(Np)$. Similarly, the time complexity of MR is $O(Np+\zeta_K Kp^2)$, thus $O(Np)$ as well when N is relatively large.

For linear model, Equation 4.8 is a quadratic equation with explicit form of solutions, therefore, the time complexity of SMR is $O(Np + Kp^2)$.

4.2.1.2 Observed scoring updating

Besides of apply the regular Fisher Scoring algorithm on the representative dataset, another updating of $\tilde{\beta}^{(t)}$ is given by

$$\tilde{\beta}^{(t+1)} = \tilde{\beta}^{(t)} + [\sum_{k=1}^{K} \frac{1}{n_k} \tilde{s}_k(\tilde{\beta}^{(t)}) \tilde{s}_k(\tilde{\beta}^{(t)})^T]^{-1} \tilde{s}(\tilde{\beta}^{(t)}) \tag{4.12}$$

But it turns out a poor performance for coarse partitioning.

4.2.2 Commonly used GLMs

4.2.2.1 Canonical links

For canonical link, $\nu(\eta) = 1/\phi$, since $\phi\eta = \theta$. Then

$$\begin{aligned}
\tilde{y}_k &= [\sum_{i \in I_k} \eta_i]^{-1} \sum_{i \in I_k} y_i \eta_i \\
S(\eta; \tilde{y}_k) &= \tilde{y}_k \eta - g^{-1}(\eta)\eta \\
\tilde{\mathbf{X}}_k &= [n_k \tilde{y}_k - n_k g^{-1}(\tilde{\eta}_k)]^{-1} \sum_{i \in I_k} [(y_i - g^{-1}(\eta_i))\mathbf{X}_i] \\
&= [n_k(\tilde{y}_k - \tilde{\mu}_k)]^{-1} \sum_{i \in I_k} [(y_i - \mu_i)\mathbf{X}_i]
\end{aligned}$$

where $\eta_i = \mathbf{X}_i^T \beta$, $\tilde{\mu}_k = g^{-1}(\tilde{\eta}_k)$, $\mu_i = g^{-1}(\eta_i)$.

Linear model:

Linear model is a special case of GLMs with normally distributed response and identity link. The density of Y_i has the exponential family form

$$f(y_i, \mu_i, \sigma) = \exp\left\{ \frac{y_i\mu_i - \frac{\mu_i^2}{2}}{\sigma^2} - \frac{1}{2}\left[\ln\left(2\pi\sigma^2\right) + \frac{y_i^2}{\sigma^2}\right]\right\},$$

This implies for $\theta_i = \mu_i$, $\phi = \sigma^2$, $b(\theta_i) = \theta_i^2/2$, and also the identity link function, i.e., $g(\mu_i) = \mu_i$, $g^{-1}(\eta_i) = \eta_i$, thus $G(\eta) = \eta$

For SMR, we have explicit solutions to Equation 4.7:

$$\begin{aligned}
\tilde{y}_k &= [\sum_{i\in I_k} \eta_i]^{-1} \sum_{i\in I_k} y_i\eta_i \\
\tilde{\eta}_k &= \frac{\tilde{y}_k \pm D^{1/2}}{2} \\
\tilde{\mathbf{X}}_k &= [n_k\tilde{y}_k - n_k\tilde{\eta}_k]^{-1} \sum_{i\in I_k} [(y_i - \eta_i)\mathbf{X}_i]
\end{aligned}$$

The solution of presentative linear predictor always exits since the discriminant is always non-negative,

$$\begin{aligned}
D &= \tilde{y}_k^2 - 4n_k^{-1} \sum_{i\in I_k} (\tilde{y}_k\eta_i - \eta_i^2) \\
&= n_k^{-1} \sum_{i\in I_k} (\tilde{y}_k - 2\eta_i)^2 \geq 0
\end{aligned}$$

This solution does not belong to homogeneous linear representative family, since it is an iterative representative choice.

Binary response with logit link:

$g(\mu) = \ln \frac{\mu}{1-\mu}$. So we have , $G(\eta) = g^{-1}(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)}$

$$\tilde{y}_k = [\sum_{i \in I_k} \eta_i]^{-1} \sum_{i \in I_k} y_i \eta_i$$

$$S(\eta; \tilde{y}_k) = \tilde{y}_k \eta - \frac{\exp(\eta)}{1 + \exp(\eta)} \eta$$

$$\tilde{\mathbf{X}}_k = [n_k \tilde{y}_k - n_k \frac{\exp(\tilde{\eta}_k)}{1 + \exp(\tilde{\eta}_k)}]^{-1} \sum_{i \in I_k} [(y_i - \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}) \mathbf{X}_i]$$

Poisson response with log link:

The likelihood function of Poisson has the exponential family form

$$f(y_i; \lambda_i) = \exp\{y_i \log \lambda_i - \lambda_i - \log y_i!\} = \exp\{y_i \theta_i - \exp \theta_i - \log y_i!\}$$

where $g(\mu) = \ln \mu$. So we have , $G(\eta) = g^{-1}(\eta) = \exp(\eta)$

$$\tilde{y}_k = [\sum_{i \in I_k} \eta_i]^{-1} \sum_{i \in I_k} y_i \eta_i$$

$$S(\eta; \tilde{y}_k) = \tilde{y}_k \eta - \exp(\eta)\eta$$

$$\tilde{\mathbf{X}}_k = [n_k \tilde{y}_k - n_k \exp(\tilde{\eta}_k)]^{-1} \sum_{i \in I_k} [(y_i - \exp(\eta_i)\mathbf{X}_i]$$

4.2.2.2 Non-canonical links

Binary response with probit link:

$g(\mu) = \Phi^{-1}(\mu)$ and $\theta = \log \Phi(\eta) - \log \Phi(-\eta)$. Thus,

$$\nu(\eta) = \frac{\phi(\eta)}{\Phi(\eta)\Phi(-\eta)}$$

$$G(\eta) = \Phi(\eta)$$

Binary response with complementary Log-log link:

We have $g(\mu) = \ln(\ln(1 - \mu))$ and $\theta = \log(\exp\{\exp(\eta)\} - 1)$. Thus,

$$\nu(\eta) = \frac{d\theta}{d\eta} = \frac{\exp(\eta)}{1 - \exp\{-\exp(\eta)\}}$$

$$G(\eta) = g^{-1}(\eta) = 1 - \exp\{-\exp(\eta)\}$$

The representative choices should be

$$\tilde{y}_k = [\sum_{i\in I_k} \frac{\exp(\eta_i)}{1-\exp\{-\exp(\eta_i)\}}\eta_i]^{-1} \sum_{i\in I_k} y_i\eta_i$$

$$S(\eta;\tilde{y}_k) = (\tilde{y}_k - (1-\exp\{-\exp(\eta)\}))\frac{\exp(\eta)}{1-\exp\{-\exp(\eta)\}}\eta$$

$$\tilde{\mathbf{X}}_k = [n_k\frac{\exp(\tilde{\eta})}{1-\exp\{-\exp(\tilde{\eta})\}}(\tilde{y}_k - (1-\exp\{-\exp(\tilde{\eta}_k)\}))]^{-1}\cdot$$
$$\sum_{i\in I_k} \frac{\exp(\eta_i)}{1-\exp\{-\exp(\eta_i)\}}[(y_i - (1-\exp\{-\exp(\eta_i)\})\mathbf{X}_i]$$

Binary response with log-log link:

We have $g(\mu) = \ln(-\ln(\mu))$ and $\theta = -\log(\exp\{\exp(\eta)\} - 1)$. Thus,

$$\nu(\eta) = \frac{d\theta}{d\eta} = \frac{\exp(\eta)}{\exp\{-\exp(\eta)\}-1}$$

$$G(\eta) = g^{-1}(\eta) = \exp\{-\exp(\eta)\}$$

Binary response with cauchit link:

$g(\mu) = \tan(\pi(\mu - 1/2))$ and $\theta = \log(\pi/2 + \arctan(\eta))$–$\log(\pi/2 - \arctan(\eta)) = \log \operatorname{arccot}(\eta)$–$\log \operatorname{arccot}(-\eta)$. Thus,

$$\nu(\eta) = \frac{\pi}{(1+\eta^2)(\pi/2+\arctan(\eta))(\pi/2-\arctan(\eta))}$$

$$G(\eta) = \frac{1}{\pi}\arctan(\eta) + \frac{1}{2}$$

Table VI

TABLE VI: Examples of $\nu(\eta)$ and $G(\eta)$

Distribution of Y	link function g	effective $\nu(\eta)$	$G(\eta)$
Normal(μ)	*identity*	1	η
Bernoulli(μ)	*logit*	1	$\exp(\eta)\{1+\exp(\eta)\}^{-1}$
	progit	$\phi(\eta)\{\Phi(\eta)\Phi(-\eta)\}^{-1}$	$\Phi(\eta)$
	cloglog	$\exp(\eta)\{1-\exp[-\exp(\eta)]\}^{-1}$	$1-\exp\{-\exp(\eta)\}$
	loglog	$\exp(\eta)\{\exp\{-\exp(\eta)\}-1\}^{-1}$	$\exp\{-\exp(\eta)\}$
	cauchit	$\pi\{(1+\eta^2)(\pi^2/4-\arctan^2(\eta))\}^{-1}$	$\arctan(\eta)/\pi+1/2$
Poisson(μ)	*log*	1	$\exp(\eta)$
Gamma	*reciprocal*	1	$1/\eta$
Inverse Gaussian	*inverse squared*	1	$1/\sqrt{\eta}$

For more GLMs, see Table VI.

4.2.3 Justification of SMR

First of all, if the current estimate of regression parameter is exactly the full data estimate $\hat{\boldsymbol{\beta}}$, then the representative sets have the score functions with value 0 at the current estimation. The new estimate $\tilde{\boldsymbol{\beta}}$ is equal to the initial value. So, the full data fit $\hat{\boldsymbol{\beta}}$ is a stationary point of the SMR. Similarly like MR, when $\tilde{\Delta}$ goes to 0, SMR estimator also converges to $\hat{\boldsymbol{\beta}}$.

Theorem 4.2.2. *Suppose the regularity condition in Theorem 4.1.1 holds. Then with Equation 4.9 and Equation 4.11, SMR estimator $\tilde{\boldsymbol{\beta}}^{(t)}$ converges to $\hat{\boldsymbol{\beta}}$ as $\tilde{\Delta}$ goes to zero for any t.*

Proof. When $\tilde{\Delta}$ goes to 0, the discrepancy between η_i in the kth block and $\tilde{\eta}_k$ in Equation 4.8 also goes to 0. Therefore $\tilde{y}_k$ in Equation 4.9 converges to block mean. By Theorem 4.1.1, $\tilde{\boldsymbol{\beta}}^{(t)}$ converges to $\hat{\boldsymbol{\beta}}$ as $\tilde{\Delta}$ goes to 0 for any t. □

Next, we provide a theorem for score-matching methods which include SMR with Equation 4.9 and Equation 4.11 as a special case:

Theorem 4.2.3. *Consider a more general iterative representative approach with estimated parameter* $\tilde{\boldsymbol{\beta}}^{(t)}$ *at its* t*th iteration. Suppose for the* $(t+1)$ *iteration, for each* $k = 1, \ldots, K$*, the obtained representative* $(\tilde{\mathbf{X}}_k, \tilde{y}_k)$ *satisfies the following three conditions:*

(1) *The representative matches the score function at* $\tilde{\boldsymbol{\beta}}^{(t)}$*, that is, Equation 4.7 is true;*

(2) *The representative response* $\tilde{y}_k \in [\min_{i \in I_k} y_i,\ \max_{i \in I_k} y_i]$*;*

(3) $\tilde{\Delta} = O(\Delta)$ *as* Δ *goes to* 0*, that is, there exists an* $M > 0$ *which does not depend on* Δ*, such that,* $\tilde{\Delta} \leq M\Delta$ *for any data partition with small enough* Δ*.*

Then the estimated parameter $\tilde{\boldsymbol{\beta}}^{(t+1)}$ *based on the representative data satisfies*

$$\|\tilde{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}\| \leq \rho(\Delta)\ \|\tilde{\boldsymbol{\beta}}^{(t)} - \hat{\boldsymbol{\beta}}\| + O(\Delta) \tag{4.13}$$

where $\rho(\Delta) = O(\Delta) < 1$ *for small enough* Δ*. Therefore,* $\tilde{\boldsymbol{\beta}}^{(t)} \to \hat{\boldsymbol{\beta}}$ *as* $t \to \infty$ *and* $\Delta \to 0$*.*

Lemma 4.2.1. *If* A *is invertible and sufficiently large compared to* B *in terms of eigenvalues, we have*

$$\begin{aligned}(A+\lambda B)^{-1} &= (\nu(I+\lambda A^{-1}B))^{-1} \\ &= (I+\lambda A^{-1}B)^{-1}A^{-1} \\ &= (I-\lambda A^{-1}B+\lambda^2 A^{-1}BA^{-1}B-\dots)A^{-1} \\ &= A^{-1}-\lambda A^{-1}BA^{-1}+\lambda^2 A^{-1}BA^{-1}BA^{-1}-\dots\end{aligned}$$

Taking $\lambda = 1$*, we have* $(A+B)^{-1} = A^{-1} - A^{-1}BA^{-1} + o(\left\|A^{-1}B\right\|)$

Proof of Theorem Equation 4.13:

Suppose the score function for the full data $s(\boldsymbol{\beta})$ has zero root $\widehat{\boldsymbol{\beta}}$, i.e., $s(\widehat{\boldsymbol{\beta}}) = 0$. The score function of representative data $s^{(t)}(\boldsymbol{\beta}) := s(\boldsymbol{\beta}; \tilde{\mathbf{y}}^{(t)}, \tilde{\mathbf{X}}^{(t)})$ based on the current estimation of regression parameter $\tilde{\boldsymbol{\beta}}^{(t)}$, which has zero root $\tilde{\boldsymbol{\beta}}^{(t+1)}$, satisfies $s^{(t)}(\tilde{\boldsymbol{\beta}}^{(t)}) = s(\tilde{\boldsymbol{\beta}}^{(t)})$.

Consider the first order of Taylor approximation,

$$\begin{aligned}s(\widehat{\boldsymbol{\beta}}) &= s(\tilde{\boldsymbol{\beta}}^{(t)}) + \frac{\partial s}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}^{(t)}}(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^{(t)}) + o(\left\|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^{(t)}\right\|) \\ s^{(t)}(\tilde{\boldsymbol{\beta}}^{(t+1)}) &= s^{(t)}(\tilde{\boldsymbol{\beta}}^{(t)}) + \frac{\partial s^{(t)}}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}^{(t)}}(\tilde{\boldsymbol{\beta}}^{(t+1)} - \tilde{\boldsymbol{\beta}}^{(t)}) + o(\left\|\tilde{\boldsymbol{\beta}}^{(t+1)} - \tilde{\boldsymbol{\beta}}^{(t)}\right\|)\end{aligned}$$

which implies

$$
\begin{aligned}
&\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^{(t+1)} \\
=&(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^{(t)}) - (\tilde{\boldsymbol{\beta}}^{(t+1)} - \tilde{\boldsymbol{\beta}}^{(t)}) \\
=&(I - \tilde{H}(\tilde{\boldsymbol{\beta}}^{(t)})^{-1} H(\tilde{\boldsymbol{\beta}}^{(t)}))(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^{(t)}) + \tilde{H}(\tilde{\boldsymbol{\beta}}^{(t)})^{-1}(o(\left\|\tilde{\boldsymbol{\beta}}^{(t+1)} - \tilde{\boldsymbol{\beta}}^{(t)}\right\|) + o(\left\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^{(t)}\right\|))
\end{aligned}
$$

where

$$
\begin{aligned}
H(\tilde{\boldsymbol{\beta}}^{(t)}) &= \frac{\partial s}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}^{(t)}} = \sum_{k=1}^{K} \sum_{i \in I_k} [(y_i - G(\eta_i))\nu'(\eta_i) - G'(\eta_i)\nu(\eta_i)]\mathbf{X}_i \mathbf{X}_i^T \\
\tilde{H}(\tilde{\boldsymbol{\beta}}^{(t)}) &= \frac{\partial s^{(t)}}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}^{(t)}} = \sum_{k=1}^{K} n_k [(\tilde{y}_k^{(t)} - G(\tilde{\eta}_k^{(t)}))\nu'(\tilde{\eta}_k^{(t)}) - G'(\tilde{\eta}_k^{(t)})\nu(\tilde{\eta}_k^{(t)})]\tilde{\mathbf{X}}_k^{(t)} \tilde{\mathbf{X}}_k^{(t)^T}
\end{aligned}
$$

with $\eta_i^{(t)} = \mathbf{X}_i^T \tilde{\boldsymbol{\beta}}^{(t)}$ and $\tilde{\eta}_k^{(t)} = \tilde{\mathbf{X}}_k^{(t)^T} \tilde{\boldsymbol{\beta}}^{(t)}$.

Since the condition (2) and (3) hold, we have

$$
\begin{aligned}
\left\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^{(t)}\right\| &= O(\Delta) \\
\left\|\tilde{\boldsymbol{\beta}}^{(t+1)} - \tilde{\boldsymbol{\beta}}^{(t)}\right\| &= O(\Delta)
\end{aligned}
$$

and

$$
\begin{aligned}
\tilde{y}_k^{(t)} &= \bar{y}_k^{(t)} + O(\Delta) \\
\mathbf{X}_i &= \tilde{\mathbf{X}}_k^{(t)} + O(\Delta)
\end{aligned}
$$

for $i \in I_k$. Therefore, we have

$$\tilde{H}(\tilde{\boldsymbol{\beta}}^{(t)}) = H(\tilde{\boldsymbol{\beta}}^{(t)}) + O(\Delta)$$

in terms of matrix norm. By Lemma 4.2.1, we have

$$I - \tilde{H}(\tilde{\boldsymbol{\beta}}^{(t)})^{-1} H(\tilde{\boldsymbol{\beta}}^{(t)}) = I - (H(\tilde{\boldsymbol{\beta}}^{(t)}) + O(\Delta))^{-1} H(\tilde{\boldsymbol{\beta}}^{(t)}) = O(\Delta)$$

When Δ is sufficiently small, the largest eigenvalue $\rho(\Delta) = O(\Delta)$ of $I - \tilde{H}(\tilde{\boldsymbol{\beta}}^{(t)})^{-1} H(\tilde{\boldsymbol{\beta}}^{(t)})$ is strictly lesser than 1. Therefore

$$\left\| \tilde{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}} \right\| \leq \rho(\Delta) \left\| \tilde{\boldsymbol{\beta}}^{(t)} - \hat{\boldsymbol{\beta}} \right\| + O(\Delta) \tag{4.14}$$

while $\tilde{\boldsymbol{\beta}}^{(t)}$ is close to $\hat{\boldsymbol{\beta}}$. This guarantees that $\tilde{\boldsymbol{\beta}}^{(t)} \to \hat{\boldsymbol{\beta}}$ if $t \to \infty$ when the maximum size of block $\Delta \to 0$. □

Remark. *We call $\rho(\Delta)$ in Equation 4.13 the globe rate of convergence, which depends on the size of Δ. Its specific form can be found in the proof of Theorem 4.2.3. Based on our experience, even for moderate size of Δ, $\rho(\Delta)$ can be significantly smaller than 1 and the first few iterations can improve the accuracy level significantly. Nevertheless, the final discrepancy away from the full data estimate still depends on Δ.*

For the proposed SMR approach in Section 4.2.1, the MLE estimator $\hat{\boldsymbol{\beta}}$ based on the full data is a stationary point of the SMR iteration. That is, if the estimated parameter $\tilde{\boldsymbol{\beta}}^{(t)} = \hat{\boldsymbol{\beta}}$ at the tth iteration, then after one SMR iteration, the estimated parameter $\tilde{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}$.

For SMR, the condition (1) of Theorem 4.2.3 holds instantly.

It should be noted that condition (2) of Theorem 4.2.3 does not hold automatically. For example, consider a binomial model. Suppose there is a block with 2 observations: $y_1 = 1$, $y_2 = 0$, and $\eta_1 = 0.5$, $\eta_2 = -1$. SMR representative could be $\tilde{y} = -1$; or if $y_1 = 0$, $y_2 = 1$ and $\eta_1 = 0.5$, $\eta_2 = -1$, then $\tilde{y} = 2$. The invalid responses is due to mixed signs of η_i's. To avoid invalid representative responses, we split the block into two pieces by the signs of η_i's and generate two representatives, one for positive η_i's and the other for negative η_i's. By this way, condition (2) of Theorem 4.2.3 holds for sure since the weights $\nu(\eta_i)\eta_i$ in each part have the same sign.

As for condition (3), simulation studies show that it is almost always the case for SMR approach. Occasionally (approximately 3 out of 16,000), $\tilde{\mathbf{X}}_k$ could be out of the convex hull of its block due to small $\nu(\tilde{\eta}_k)(\tilde{y}_k - G(\tilde{\eta}_k))$ value. For such cases, we replace it with the MR representative. The difference caused for score function is negligible. By this way, the condition (3) of Theorem 4.2.3 holds as well.

Overall, Theorem 4.2.3 explains why SMR approach works so well.

Corollary 4.2.1. *When $\Delta = 0$, MR and SMR generate the same set of representatives. Therefore, both SMR and MR estimates are equal to the full data estimate for GLMs. A special*

case is when all covariates are categorical and the dataset is naturally partitioned by distinct covariate values.

Proof. Since $\Delta = 0$, then in the kth block, there is no diversion for covariates, $\mathbf{X}_i = \bar{\mathbf{X}}_k$, so is $\eta_i = \bar{\eta}_k$, $i \in I_k$. Therefore, $\tilde{y}_k = \bar{y}_k$ by Equation 4.9. Equation 4.7 indicates that $S(\bar{\eta}_k; \tilde{y}_k) = S(\tilde{\eta}_k; \tilde{y}_k)$, where $\tilde{\eta}_k = \bar{\eta}_k$ is a solution. By taking the solution with $\tilde{\mathbf{X}}_k$ closest to $\bar{\mathbf{X}}_k$, we actually have $\tilde{\mathbf{X}}_k = \bar{\mathbf{X}}_k$.

So, SMR meets MR and the full data fit. □

When most covariates are categorical except for a few continuous variables, for example, the flight on-time performance analysis in Section 5, both MR and SMR may work very well.

4.2.4 Asymptotic properties of MR and SMR for big data

In order to study the asymptotic properties of MR and SMR as N goes to ∞, we assume the predictors $\mathbf{X}_1, \ldots, \mathbf{X}_N \in \mathbb{R}^p$ are iid $\sim F$ with a finite expectation, and the partition $\{B_1, \ldots, B_K\}$ of the predictor space $\mathbb{R}^p$ is fixed. To avoid trivial cases, we assume $p_k = F(B_k) > 0$ for each $k = 1, \ldots, K$. Then the index block $I_k = \{i \in \{1, \ldots, N\} \mid \mathbf{X}_i \in B_k\}$ with size n_k. By the strong law of large numbers, as $N \to \infty$, $n_k/N \to p_k > 0$ almost surely. Since we are considering asymptotic properties here, we consider the discrepancy from the true parameter value $\boldsymbol{\beta}$ instead of the estimator $\hat{\boldsymbol{\beta}}$ based on the full data.

For MR approach, as $N \to \infty$,

$$\tilde{\mathbf{X}}_k \to p_k^{-1} \int_{B_k} \mathbf{x}\, F(d\mathbf{x}), \quad \tilde{y}_k \to p_k^{-1} \int_{B_k} G(\boldsymbol{\beta}^T \mathbf{x})\, F(d\mathbf{x}) \tag{4.15}$$

almost surely. If the link function g or $G = g^{-1}$ is linear, then $g(\tilde{y}_k) - \tilde{\mathbf{X}}_k^\intercal \boldsymbol{\beta} \to 0$ and thus MR estimate $\tilde{\boldsymbol{\beta}} \to \boldsymbol{\beta}$. Nevertheless, in general g is nonlinear, the accuracy of MR estimate mainly depends on the size of blocks Δ, not the sample size N. In other words, fixing the data partition, the accuracy of MR estimate is limited by Equation 4.15 thus will not benefit from increasing a large enough sample size for most GLMs.

Different from MR, by matching the score function of the full data, SMR approach can still improve its estimate when the sample size gets bigger even with a fixed data partition.

Actually, for a general GLM, $\mathbb{E}(Y_i) = G(\eta_i)$ and $Y_i - G(\eta_i) \overset{ind}{\sim} (0, \sigma_i^2)$, where $\sigma_i^2 = \mathrm{Var}(Y_i) = h(\eta_i) > 0$. For either a bounded block B_k or a bounded $h(\cdot)$, $\max_{i \in I_k} \sigma_i^2$ is also bounded. By strong law of large numbers and Taylor expansion, as $N \to \infty$ and $n_k \to \infty$, the left hand side of Equation 4.8 after divided by n_k is

$$
\begin{aligned}
\text{LHS} =& n_k^{-1} \sum_{i \in I_k} \nu(\eta_i)\eta_i(y_i - G(\eta_i)) \\
=& n_k^{-1} \sum_{i \in I_k} \nu(\eta_i)\eta_i \left[y_i - G(\mathbf{X}_i^\intercal \boldsymbol{\beta}) - G'(\mathbf{X}_i^\intercal \boldsymbol{\beta}) \mathbf{X}_i^\intercal (\tilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}) + O(\|\tilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}\|^2) \right] \\
\overset{a.s.}{\to} & n_k^{-1} \sum_{i \in I_k} \nu(\eta_i)\eta_i \left[-G'(\mathbf{X}_i^\intercal \boldsymbol{\beta}) \mathbf{X}_i^\intercal (\tilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}) + O(\|\tilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}\|^2) \right] \qquad (4.16) \\
=& -\nu(\tilde{\eta}_k)\tilde{\eta}_k G'(\tilde{\mathbf{X}}_k^\intercal \boldsymbol{\beta}) \tilde{\mathbf{X}}_k^\intercal (\tilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}) + O(\Delta \|\tilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}\|) + O(\|\tilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}\|^2) \qquad (4.17)
\end{aligned}
$$

Equation 4.16 shows that when N increases, the leading discrepancy of LHS caused by response y_i's vanishes. Even if the maximum block size Δ is fixed, when $\tilde{\boldsymbol{\beta}}^{(t)}$ is close to $\boldsymbol{\beta}$, the LHS of Equation 4.8 is small, and so is its right hand side. For blocks with $\tilde{\eta}_k$ away from 0, it indicates

$\tilde{y}_k - G(\tilde{\eta}_k)$ and thus $\tilde{y}_k - G(\tilde{\mathbf{X}}_k^T\boldsymbol{\beta})$ is small. That is, when $N \to \infty$, the SMR representatives $\{(\tilde{\mathbf{X}}_k, \tilde{y}_k), k = 1, \ldots, K\}$ stay close to the true curve, $\mu = E(Y) = G(\mathbf{X}^T\boldsymbol{\beta})$, which leads to a faster convergent rate of SMR estimate towards $\boldsymbol{\beta}$ than MR's.

Equation 4.17 implies that if $G'(\tilde{\mathbf{X}}_k^T\boldsymbol{\beta})$ is relatively large, it may slow down the convergence of SMR estimate. For example, under a Poisson regression model with log link , $G(\eta) = e^{\eta}$. If the initial estimate of the regression parameter is not so close, SMR may have difficulty in converging to the full data estimate. In such kind of situations, one may need a finer partition or smaller Δ to obtain a good initial estimate. For models with fairly flat G functions, such as models with logit link, G' is small for most blocks. Even if the initial estimate for SMR is not so close, we can still accurate estimate after a few iterations.

4.3 More Simulation Studies

In practice, we only need to run a few iterations for SMR to reach the accuracy level comparable with the full data estimate. Our simulation studies in this section show that 3-iteration SMR is comparable with the divide-and-conquer approach (Lin and Xi, 2011 (6)), also known as divide and recombine, split and conquer, or split and merger in the literature (Wang et al., 2016 (2)). In the rest of this paper, we call the 3-iteration SMR simply SMR.

4.3.1 SMR vs MR for linear model

Following the simulation setups in Section 3.4, we simulate 20 datasets of size $N = 1 \times 10^6$ for each of the 7 distributions. MR and SMR are used to obtain the parameter estimate $\tilde{\boldsymbol{\beta}}$. Different from Table I, we show in Table VII, the average RMSE between $\tilde{\boldsymbol{\beta}}$ and the full data estimate $\hat{\boldsymbol{\beta}}$. The improvement of RMSE from MR to SMR is not much based on equal-depth

TABLE VII: Average (std) of RMSEs (10^{-3}) from $\hat{\boldsymbol{\beta}}$ of 20 simulations for linear model with $N = 10^6$

Simulation	Equal-depth ($m = 4$)		k-means ($K = 1000$)	
setup	MR	SMR	MR	SMR
mzNormal	0.710(0.037)	**0.705**(0.037)	0.805(0.047)	**0.029**(0.003)
nzNormal	0.710(0.037)	**0.704**(0.036)	0.805(0.047)	**0.029**(0.005)
ueNormal	0.194(0.024)	**0.193**(0.024)	0.194(0.021)	**0.067**(0.010)
mixNormal	0.816(0.060)	**0.806**(0.060)	0.731(0.063)	**0.004**(0.001)
T_3	6.432(0.324)	**6.351**(0.327)	5.653(0.416)	**0.506**(0.055)
EXP	1.020(0.068)	**0.990**(0.066)	0.750(0.039)	**0.018**(0.003)
BETA	0.686(0.041)	**0.672**(0.041)	0.873(0.062)	**0.010**(0.002)

partition, while the improvements based on k-means partition are truly significant (see also Figure 4).

This simulation study confirms the conclusion in Theorem 4.2.3. That is, when Δ is smaller, $\rho(\Delta)$ is closer to 0 and the improvement from MR to SMR is much more significant.

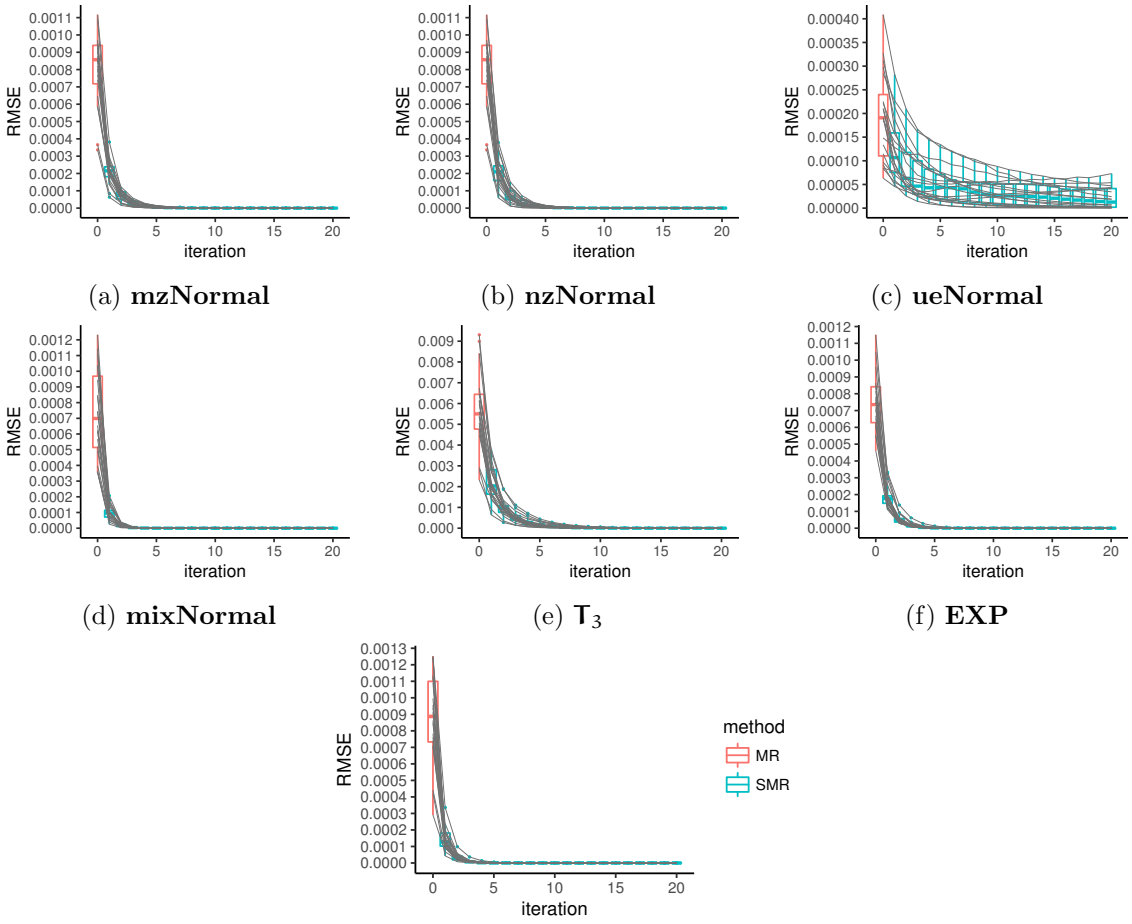
4.3.2 SMR vs divide-and-conquer for logistic models

Logistic regression model is also a special case of GLMs with Bernoulli or binomial response and canonical link, logit, and

$$G(\eta) = g^{-1}(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

$$S(\eta; \tilde{y}_k) = \tilde{y}_k \eta - \frac{\exp(\eta)}{1 + \exp(\eta)} \eta$$

Following the simulation setup in Section 4.1.2, we simulate 20 datasets of size $N = 1 \times 10^6$ for logistic regression models. Divide-and-conquer (DC) proposed by Lin and Xi (2011) (6) and



(a) **mzNormal** (b) **nzNormal** (c) **ueNormal**

(d) **mixNormal** (e) $\mathbf{T}_3$ (f) **EXP**

(g) **BETA**

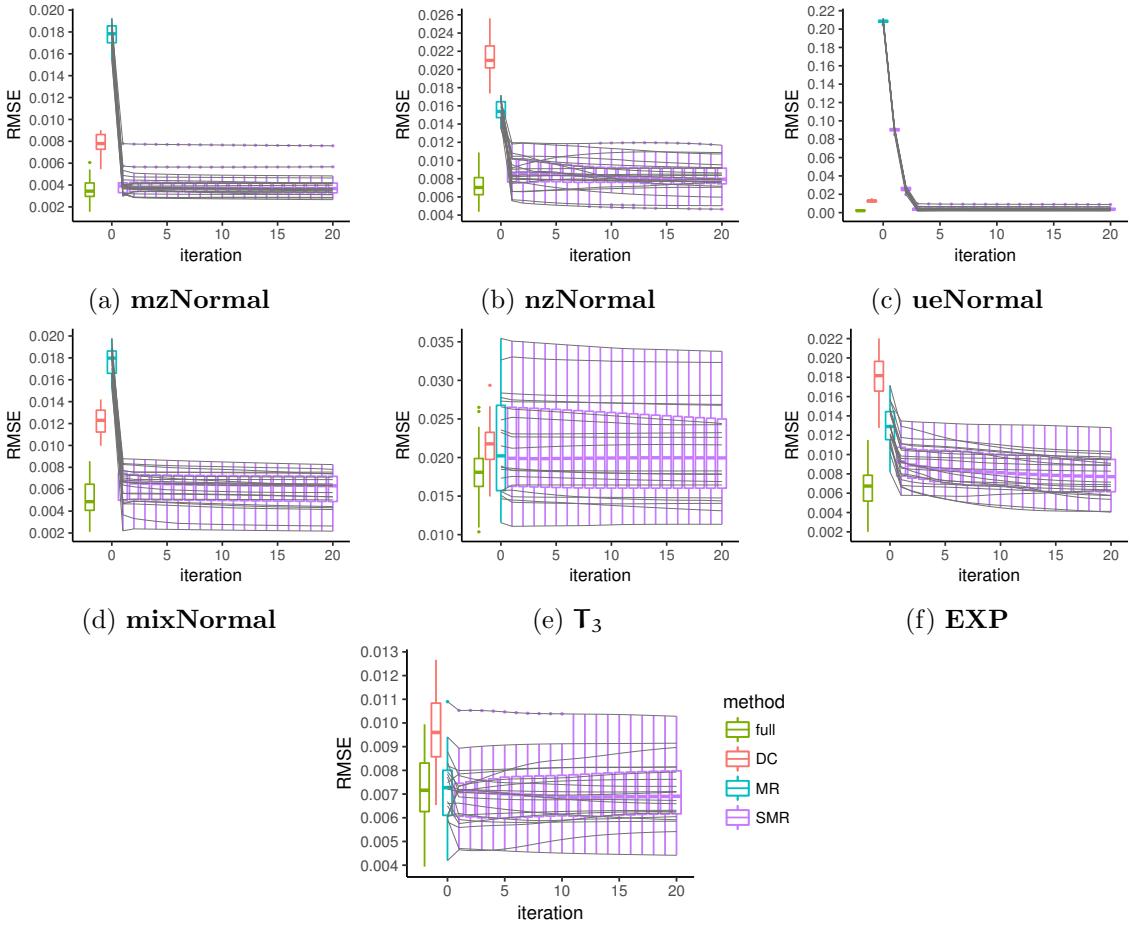
Figure 4: Box-plots of iterative SMR vs MR : RMSE from full data $\hat{\boldsymbol{\beta}}$ for linear model, $\mathsf{N} = 10^6$ with based on k-means with $\mathsf{K} = 1000$. The x-axis is MR and the iterations of SMR, from 1 to 20. Grey lines connect iterations in each simulation.

TABLE VIII: Average (std) of RMSEs (10^{-3}) of 20 simulations for logistic models with $N = 10^6$ (MR, SMR: k-means with $K = 1000$; Divide-and-Conquer (DC): 1000 blocks)

Simulation setup	RMSE from true $\boldsymbol{\beta}$				RMSE from full data $\hat{\boldsymbol{\beta}}$		
	Full	MR	SMR	DC	MR	SMR	DC
mzNormal	3.6	17.7	**4.0**	7.7	17.5	**1.7**	6.9
	(0.2)	(0.2)	(0.3)	(0.2)	(0.1)	(0.1)	(0.1)
nzNormal	7.3	15.4	**8.6**	21.3	13.9	**4.9**	20.2
	(0.4)	(0.3)	(0.4)	(0.4)	(0.2)	(0.3)	(0.2)
ueNormal	2.3	208.5	**4.1**	12.8	209.0	**3.5**	13.2
	(0.2)	(0.4)	(0.4)	(0.3)	(0.5)	(0.4)	(0.1)
mixNormal	5.1	17.7	**6.2**	12.2	17.0	**3.2**	11.2
	(0.4)	(0.3)	(0.4)	(0.3)	(0.1)	(0.2)	(0.1)
T_3	18.3	21.7	**21.4**	21.6	11.9	**11.2**	11.9
	(1.0)	(1.5)	(1.5)	(0.9)	(0.9)	(0.9)	(0.1)
EXP	6.5	12.9	**9.0**	18.1	10.5	**5.3**	16.7
	(0.5)	(0.5)	(0.5)	(0.5)	(0.2)	(0.2)	(0.1)
BETA	7.2	7.3	**6.9**	9.6	2.9	**2.2**	5.9
	(0.3)	(0.3)	(0.3)	(0.4)	(0.1)	(0.1)	(0.1)

our SMR are applied for estimating parameters. Table VIII shows that based on a k-means partitioning with $k = 1000$, SMR outperforms the divide-and-conquer method with 1000 blocks for all simulation settings. For illustration purpose, boxplots of comprehensive simulations are listed in Figure 5, as well as Figure 6, which shows that SMR outperforms the divide-and-conquer method with 1000 blocks for all simulations. It also reveals that 3 iterations of SMR are pretty enough to reach a accuracy level comparable to the full data fit.

SMR can be applied to multiple computers (known as *nodes*), and exchanges only the representative data points and estimated parameter values. It can perform well even with limited network connections.

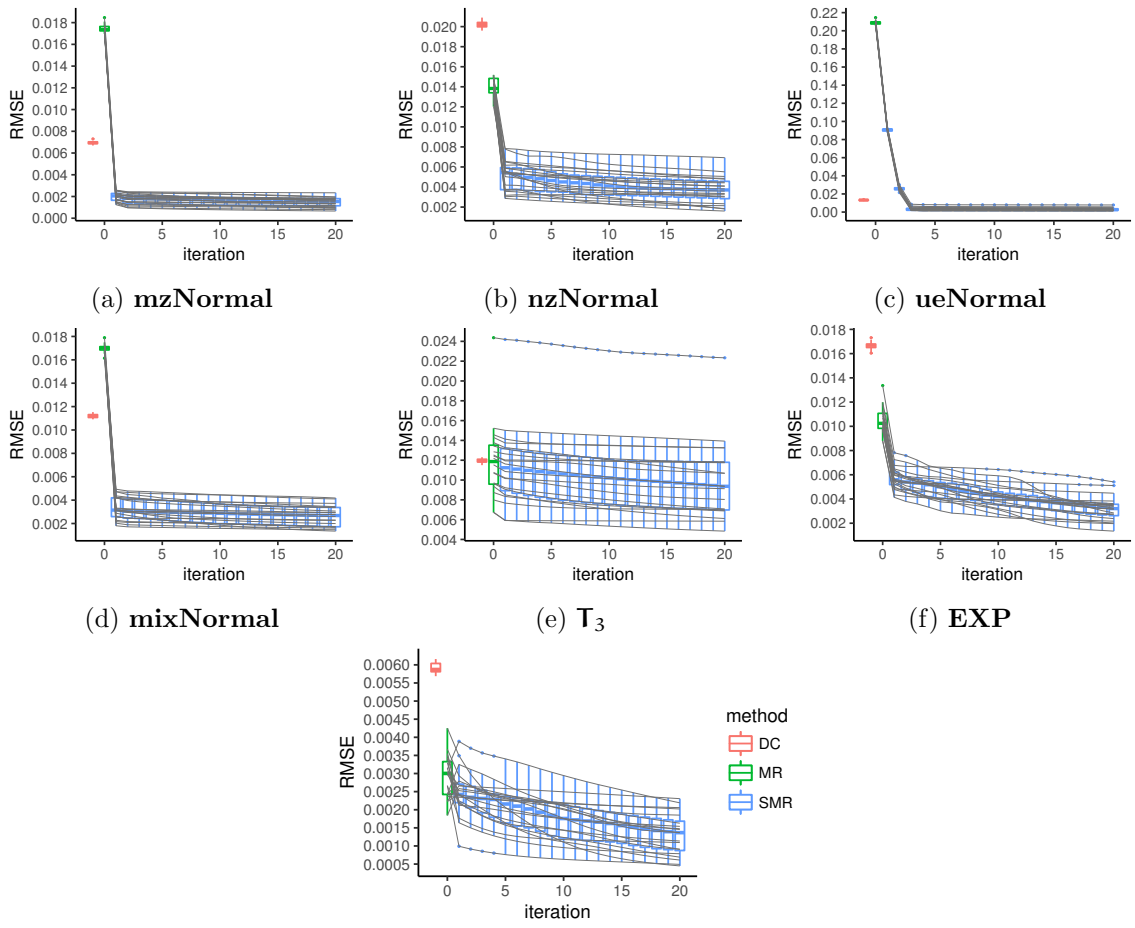


(a) **mzNormal** (b) **nzNormal** (c) **ueNormal**

(d) **mixNormal** (e) $\mathbf{T}_3$ (f) **EXP**

(g) **BETA**

Figure 5: Box-plots of iterative SMR vs full, MR, and divide-conquer: RMSE from true $\boldsymbol{\beta}$ for logistic model with $N = 10^6$ based on k-means with $K = 1000$. The x-axis is full, Divide-and-conquer, MR, and the iterations of SMR, from 1 to 20. Grey lines connect iterations in each simulation.



(a) **mzNormal** (b) **nzNormal** (c) **ueNormal**

(d) **mixNormal** (e) $\mathbf{T}_3$ (f) **EXP**

(g) **BETA**

Figure 6: Box-plots of iterative SMR vs MR, and Divide-and-conquer: RMSE from full data $\hat{\boldsymbol{\beta}}$ for logistic model with $\mathsf{N} = 10^6$ based on k-means with $\mathsf{K} = 1000$. The x-axis is Divide-and-conquer, MR, and the iterations of SMR, from 1 to 20. Grey lines connect iterations in each simulation.

Divide-and-conquer methods typically operate on random partitions, that is, each data block for divide-and-conquer consists of data points from many nodes. Usually there are only limited values for some fields in a single node, on which regression on a single node may not have a feasible solution. Therefore, a heavy communication between nodes are required, which violates the independence computing rule. Thus the computing of divide-and-conquer methods heavily depend on the speed and capacity of network connection.

4.3.3 Some properties of SMR

Both MR and SMR benefit from increasing in sample size and number of splitting blocks with respect to RMSE from true value. In this section, we reveal that SMR can take extra advantage in such situation with respect to RMSE from full data estimate through experiments over different sample sizes and numbers of blocks.

4.3.3.1 Performances of MR, SMR over different sample sizes

We conclude in Section 4.2.4 that SMR can benefit more than MR as sample size increases. In this section, we show that the advantage of SMR is over divide-and-conquer method as well.

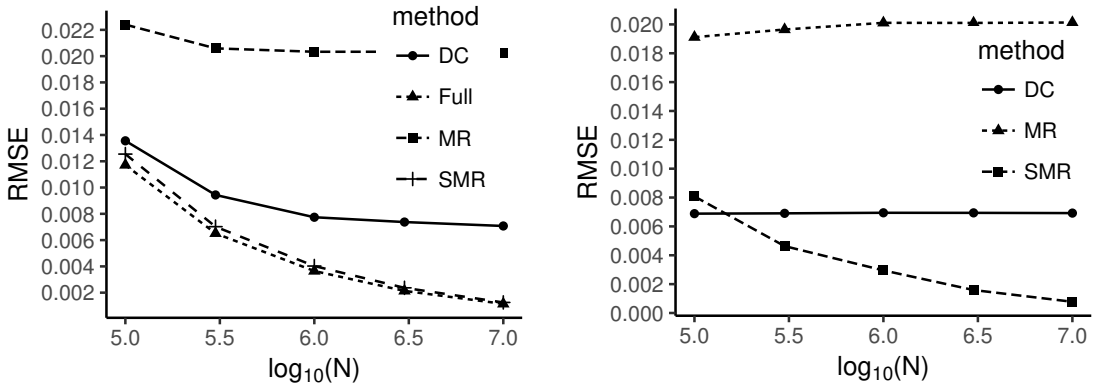
In this simulation study, we use the first simulation setup *mzNormal* for illustration purpose. For MR and SMR, we use an equal-depth partition with $\mathsf{m} = 4$, whose effectiveness is comparable with a k-means partition with $\mathsf{K} = 1000$ according to Section 4.1.2. The partition is fixed as N gets bigger. For divide-and-conquer method, we fix its block size 10^3 for illustration purpose, which is the same as the block size in Section 4.3.2. As N gets bigger, the number of blocks for divide-and-conquer method increases proportionally, which is often the case in practice.

TABLE IX: Average (std) of RMSEs (10^{-3}) of 20 simulations for logistic models with different N (MR, SMR: equal-depth with $m = 4$; Divide-and-conquer (DC): block size 1000), **mzNormal**

N	RMSE from true $\boldsymbol{\beta}$				RMSE from full $\hat{\boldsymbol{\beta}}$		
	Full	MR	SMR	DC	MR	SMR	DC
1×10^5	11.71	22.40	**12.55**	13.56	19.13	**4.04**	6.88
	(0.62)	(0.58)	(0.72)	(0.52)	(0.21)	(0.26)	(0.06)
3×10^5	6.51	20.59	**7.02**	9.44	19.64	**2.32**	6.91
	(0.40)	(0.26)	(0.40)	(0.32)	(0.10)	(0.20)	(0.05)
10×10^5	3.65	20.33	**4.03**	7.74	20.10	**1.48**	6.94
	(0.23)	(0.22)	(0.24)	(0.23)	(0.06)	(0.08)	(0.03)
30×10^5	2.13	20.33	**2.37**	7.37	20.11	**0.79**	6.94
	(0.15)	(0.09)	(0.17)	(0.10)	(0.02)	(0.06)	(0.02)
100×10^5	1.15	20.22	**1.26**	7.08	20.13	**0.39**	6.92
	(0.08)	(0.06)	(0.08)	(0.06)	(0.02)	(0.02)	(0.01)

As shown in Table IX and Figure 7, as N increases, SMR estimate gets closer to full data estimate quickly, and accuracy is improved faster than MR and divide-conquer, which seem not get closer to the full data estimate. Thus, SMR is more efficient with the same sample size, additionally, time complexity of SMR is lower than divide-conquer. The different performances of MR and SMR confirm our conclusion in Section 4.2.4.

The reason MR gets a little away from it is the unbounded covariates, which leads to the ranges of predictors increase along N. The flat trend of the divide-and-conquer method in Figure 7 (b) is mainly due to the increased relative bias as the number of blocks increases (see Figure 1 in Lin and Xi (2011) (6)). As $N \to \infty$, its estimate cannot catch up with full data fit.



(a) RMSE from true $\boldsymbol{\beta}$ (b) RMSE from full data $\hat{\boldsymbol{\beta}}$

Figure 7: RMSE vs $\log_{10}(N)$ of MR, SMR, and divide-conquer for logistic model

4.3.3.2 Performances of MR and SMR with finer partition

According to Theorems 4.1.1 and 4.2.3, the estimate $\tilde{\boldsymbol{\beta}}$ obtained by MR or SMR converges to the full data estimate $\hat{\boldsymbol{\beta}}$ as $\Delta \to 0$. That is, with finer partition, $\tilde{\boldsymbol{\beta}}$ gets closer to $\hat{\boldsymbol{\beta}}$, but not necessarily the true parameter $\boldsymbol{\beta}$ for given dataset, which confirms the conclusion of Theorem 4.2.3. Similar to MR, the maximum size of blocks contributes more than the number of blocks to the convergence rate of SMR. Our simulation studies summarized in Table X and Table XI confirm our conclusions. In addition, Even if the partitioning is coarse, SMR can still work, while finer partitioning improves the performance of course. That is, SMR is more robust to partitions than MR, where the latter is more sensitive to the size of blocks.

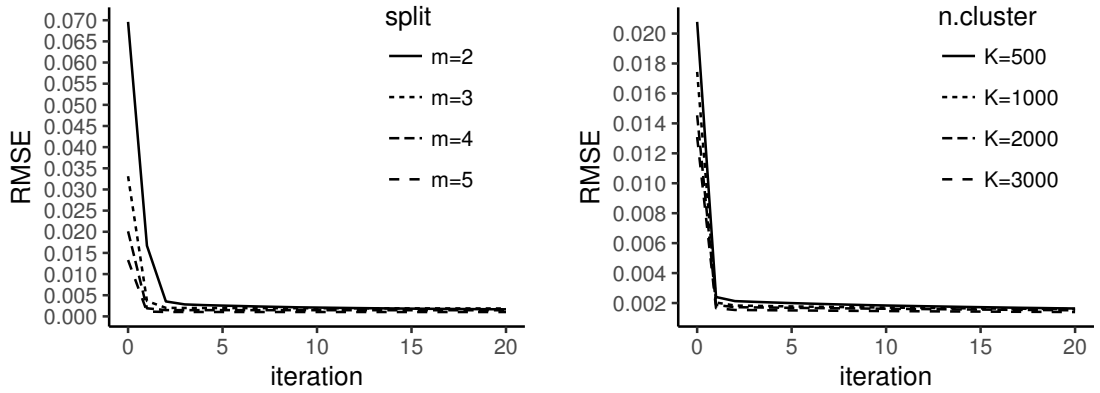
For comparison purpose, we also list the RMSEs of divide-and-conquer (DC) methods for the same number of blocks in Table XI. Note that DC typically uses random partition and prefers as fewer blocks as possible.

TABLE X: Average (std) of RMSEs (10^{-3}) of 20 simulations for logistic model, $N = 10^6$, **mzNormal**, equal-depth partition with different m

m	RMSE from true $\boldsymbol{\beta}$			RMSE from full $\hat{\boldsymbol{\beta}}$	
	Full	MR	SMR	MR	SMR
2	3.65(0.23)	69.59(0.19)	4.51(0.26)	69.61(0.12)	2.83(0.18)
3	3.65(0.23)	33.21(0.22)	3.96(0.32)	33.13(0.08)	1.98(0.13)
4	3.65(0.23)	20.33(0.22)	4.03(0.24)	20.10(0.06)	1.48(0.08)
5	3.65(0.23)	13.67(0.24)	3.70(0.22)	13.32(0.04)	1.04(0.05)

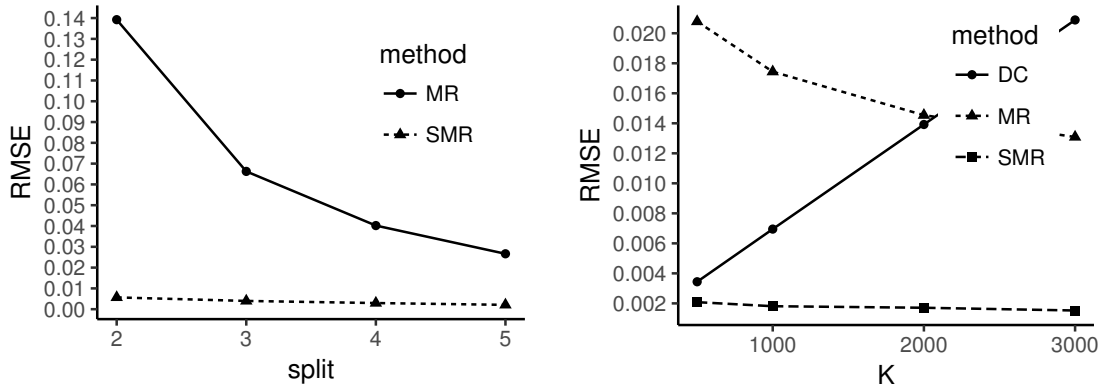
TABLE XI: Average (std) of RMSEs (10^{-3}) of 20 simulations for logistic model, $N = 10^6$, **mzNormal**, k-means partition for MR and SMR with different K, random partition for DC with K blocks

K	RMSE from true $\boldsymbol{\beta}$				RMSE from full $\hat{\boldsymbol{\beta}}$		
	Full	MR	SMR	DC	MR	SMR	DC
500	3.65	21.00	4.33	4.92	20.78	2.08	3.44
	(0.23)	(0.23)	(0.27)	(0.24)	(0.09)	(0.14)	(0.02)
1000	3.65	17.68	4.04	7.73	17.44	1.81	6.95
	(0.23)	(0.25)	(0.25)	(0.25)	(0.08)	(0.14)	(0.03)
2000	3.65	14.84	3.89	14.27	14.54	1.71	13.92
	(0.23)	(0.21)	(0.26)	(0.23)	(0.06)	(0.11)	(0.04)
3000	3.65	13.55	4.26	21.09	13.09	1.53	20.89
	(0.23)	(0.20)	(0.23)	(0.23)	(0.09)	(0.10)	(0.03)



(a) Based on equal-depth with different $\mathfrak{m}$ (b) Based on k-means with different K

Figure 8: Average RMSE from true $\boldsymbol{\beta}$ of 20 simulations for logistic model with $\mathsf{N} = 10^6$, **mzNormal**, based on different partition size



(a) Based on equal-depth with different $\mathfrak{m}$ (b) Based on k-means with different K

Figure 9: Average RMSE from full $\widehat{\boldsymbol{\beta}}$ of 20 simulations vs different partition setups for logistic model with $\mathsf{N} = 10^6$, **mzNormal**

4.3.4 Other GLMs

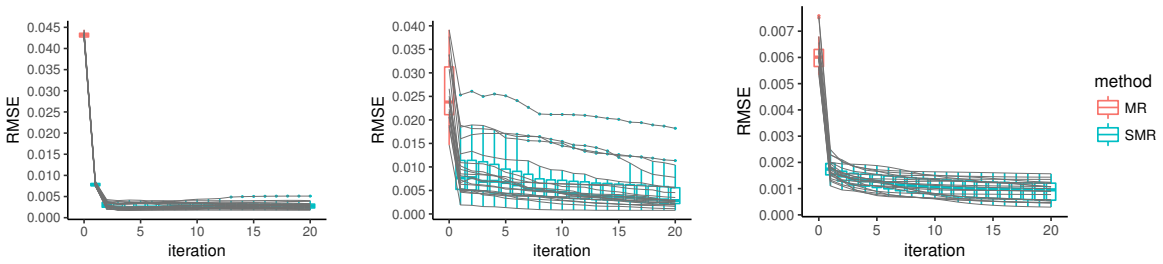
Commonly used GLMs include binary responses with logit, probit, cloglog, loglog, cauchit links, Poisson responses with log link, Gamma responses with reciprocal link, Inverse Gaussian responses with inverse squared link, etc. We provide detailed formulas for $\nu(\eta)$ and $G(\eta)$ in Table VI.

TABLE XII: Average (std) of RMSEs (10^{-3}) for three models, $N = 10^6$, **mzNormal**, k-means ($K = 1000$)

	Binary with cloglog			Poisson with log			Logistic with interactions		
Average	Full	MR	SMR	Full	MR	SMR	Full	MR	SMR
From true	2.62	43.26	3.83	0.22	25.61	9.14	3.49	7.33	3.78
	(0.17)	(0.18)	(0.21)	(0.01)	(1.48)	(1.34)	(0.22)	(0.35)	(0.29)
From full	0	43.21	2.71	0	25.61	9.14	0	6.13	1.39
	-	(0.12)	(0.15)	-	(1.48)	(1.34)	-	(0.13)	(0.07)

In Table XII, we show average RMSE of 20 simulations based on k-means partition with $K = 1000$ for the following three models:

- Binary response with complementary Log-log (cloglog) link



(a) Binary with cloglog (b) Poisson with log (c) Logistic with interactions

Figure 10: Boxplots of RMSE from full $\hat{\boldsymbol{\beta}}$ of 20 simulations for three models with $N = 10^6$, **mzNormal**, based on k-means with $K = 1000$

Cloglog link function $g(\mu) = \ln(\ln(1-\mu))$. Thus, $G(\eta) = 1 - \exp\{-\exp(\eta)\}$ is relatively flat, therefore even the performance of MR is not that close, SMR still converges fast.

- Poisson response with log link

  The canonical link function of Poisson is $g(\mu) = \ln \mu$. So $G(\eta) = \exp(\eta)$ increases exponentially. With a low accuracy initial value MR, the convergence of SMR is slower down, which confirms our conclusion in Section 4.2.4. Also, the variance of both MR and SMR are high. Thus, a good initial value for Poisson is crucial important.

- GLM with interactions

  Either MR or SMR can be applied on interactions directly since they face to the predictor variables other than covariates. $\mathbf{x} = (x_1, x_2, x_3)^T$ follows **mzNormal**, and predictors are $(h_1(\mathbf{x}), \ldots, h_7(\mathbf{x})) = (x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1x_2x_3)$ for a logistic model, both MR and SMR works well.

TABLE XIII: Average CPU time (sec) of MR, SMR, A-optimal, Divide-and-conquer over 20 simulations for logistic model with $N = 10^6$, $p = 7$, under k-means ($K = 1000$)

Simulation setup	full data	MR	SMR	A-opt	DC
mzNormal	21.949	0.459	51.722	2.237	18.242
nzNormal	27.732	0.528	54.821	2.340	19.057
ueNormal	28.987	0.548	58.150	2.329	18.900
mixNormal	27.015	0.442	51.071	2.296	18.765
$\mathbf{T}_3$	15.306	0.308	36.118	2.344	17.667
EXP	22.984	0.338	34.704	2.368	18.079
BETA	18.666	0.324	33.435	2.310	17.725

4.3.5 CPU time of SMR

For illustration purpose, we use k-means partition method with $K = 1000$ in this section. The CPU time for 1-step SMR and A-optimal subsampling with subsample size 20000 and Divide-and-conquer with 1000 random blocks are shown in Table XIII for logistic models with $N = 10^6$ and $p = 7$. The CPU time of SMR is still higher than full data fit and Divide-and-conquer method, since the number of parameters p is of moderate size.

According to the time complexity analysis in Section 4.2.1.1, the computational time of SMR should also roughly proportional to the number of parameters p and sample size N like MR, if the number of blocks K is small. Our simulation studies in Figure 11 and Figure 12 confirm our conclusion, which showing the relation of the CPU time of SMR against the number of parameters p and sample size N. The computational time to apply such a k-means partition is still high though.

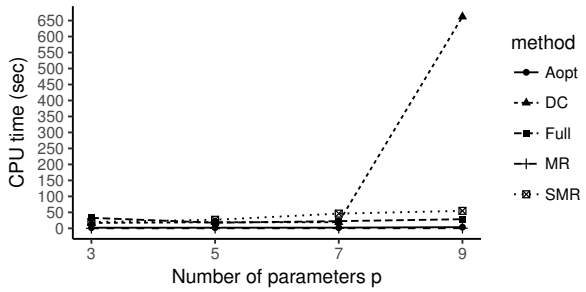


Figure 11: Average CPU time of SMR over 20 simulations against p for logistic model with $N = 10^6$, **mzNormal**, under k-means ($K = 1000$)

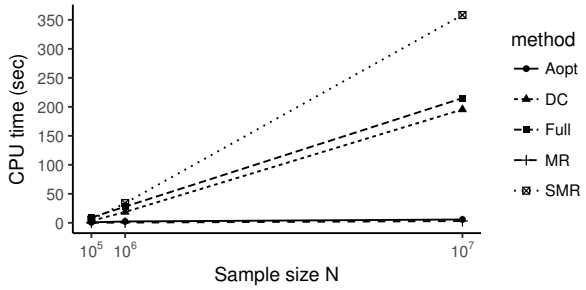


Figure 12: Average CPU time of SMR over 20 simulations against N for logistic model with $p = 7$, **mzNormal**, under k-means ($K = 1000$)

CHAPTER 5

AIRLINE ON-TIME PERFORMANCE DATA

The Airline on-time performance data for US domestic flights of arrival time from October 1987 to February 2017 were collected from Bureau of Transportation Statistics as a real example for big data analysis.

5.1 Descriptive Analysis of Airline Data

The dataset consists of 353 cvs files with total original records number $N = 173,106,219$ and the fields shown in Table XIV.

First we discretize DISTANCE by taking value of interval center for every 200 miles. Then classify data into bins with weighted response pair (Ontime, Delay) by distinct values of YEAR, SEASON, DAYOFWEEK, DEPTIMEBLK, and DISTANCE to do descriptive analysis. There is 1 bin without departure information, 172 bins with no delay or travel information. Thus we remove the following bins in Table XV.

The total valid records number after cleaning is $N = 169,609,446$. Parts of descriptive analysis are given in Table XVI, Table XVII, Table XVIII.

5.2 SMR and MR on Flight Data with Oracle Responses

For illustration purpose, we consider three categorical covariates. Quarter (season, $1 \sim 4$) is used instead of MONTH with 7 levels for simplification purpose. DayOfWeek (day of week, $1 \sim 7$) is still considered since it is a significant covariate. Following convention of O'hare

TABLE XIV: Description of fields in original data

Field Name	Description
YEAR	Year, from 1987 to 2017
QUARTER	Quarter (1-4)
MONTH	Month
DAY_OF_MONTH	Day of Month
DAY_OF_WEEK	Day of Week
FL_DATE	Flight Date (yyyymmdd)
ORIGIN_AIRPORT_ID	Origin Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport. Use this field for airport analysis across a range of years because an airport can change its airport code and airport codes can be reused.
DEST_AIRPORT_ID	Destination Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport. Use this field for airport analysis across a range of years because an airport can change its airport code and airport codes can be reused.
CRS_DEP_TIME	CRS Departure Time (local time: hhmm)
DEP_DELAY	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.
DEP_DELAY_GROUP	Departure Delay intervals, every (15 minutes from < -15 to > 180)
DEP_TIME_BLK	CRS Departure Time Block, Hourly Intervals
CRS_ARR_TIME	CRS Arrival Time (local time: hhmm)
ARR_DELAY	Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.
ARR_DELAY_GROUP	Arrival Delay intervals, every (15-minutes from < -15 to > 180)
ARR_TIME_BLK	CRS Arrival Time Block, Hourly Intervals
CANCELLED	Cancelled Flight Indicator (1=Yes)
CANCELLATION_CODE	Specifies The Reason For Cancellation
DIVERTED	Diverted Flight Indicator (1=Yes)
CRS_ELAPSED_TIME	CRS Elapsed Time of Flight, in Minutes
DISTANCE	Distance between airports (miles)
DISTANCE_GROUP	Distance Intervals, every 250 Miles, for Flight Segment

TABLE XV: Removed records

n of records	Year	SEASON	DAYOFWEEK	DEPTIMEBLK	DISTANCE	Ontime	Delay
26	1987	4	1	0700-0759	0	0	0
706	1987	4	3	1500-1559	0	0	0
1899	1987	4	7	1900-1959	0	0	0
1948	1987	4	7	2200-2259	0	0	0
46311	1989	4	2	0600-0659	0	1	0
46429	1989	4	2	1400-1459	0	0	0
594920	2012	3	4		1	0	1
190	1987	4	1	1700-1759	23	0	0
487	1987	4	2	1800-1859	22	0	0
...	...	...	...	...	...	...	...

TABLE XVI: Ontime-delay ratio over SEASON

SEASON	Ontime ratio	Delay ratio	Number of total records
1	0.788	0.212	41579713
2	0.804	0.196	42361276
3	0.808	0.192	42914516
4	0.801	0.199	42753941

TABLE XVII: Ontime-delay over DAYOFWEEK

DAYOFWEEK	Ontime ratio	Delay ratio	Number of total records
1	0.803	0.197	24988255
2	0.814	0.186	24686764
3	0.802	0.198	24810248
4	0.778	0.222	24914325
5	0.772	0.228	24973697
6	0.832	0.168	21550008
7	0.806	0.194	23686149

TABLE XVIII: Ontime-delay over DEPTIMEBLK

DEPTIMEBLK	Ontime ratio	Delay ratio	Number of total records
0001-0559	0.860	0.140	2476334
0600-0659	0.902	0.098	10358193
0700-0759	0.877	0.123	11921054
0800-0859	0.852	0.148	12060297
0900-0959	0.839	0.161	10924704
1000-1059	0.835	0.165	10270352
1100-1159	0.824	0.176	10768536
1200-1259	0.813	0.187	10832755
1300-1359	0.797	0.203	11141752
1400-1459	0.785	0.215	10036841
1500-1559	0.767	0.233	10432324
1600-1659	0.757	0.243	10344962
1700-1759	0.739	0.261	11628797
1800-1859	0.733	0.267	10426461
1900-1959	0.730	0.270	9374510
2000-2059	0.732	0.268	7724958
2100-2159	0.749	0.251	5207005
2200-2259	0.784	0.216	2592115
2300-2359	0.803	0.197	1087496

TABLE XIX: Description of fields in Airline on-time performance data

Field Name	Description
ArrDel15	binary response variable: arrival delay indicator, 15 minutes or more (1=Yes)
QUARTER	season, "1": January 1-March 31,"2": April 1-June 30, "3": July 1-September 30, "4": October 1-December 31
DayOfWeek	day of week, "1": Monday, "2": Tuesday, "3": Wednesday, "4": Thursday, "5": Friday, "6": Saturday, "7": Sunday
DepTimeBlk	CRS departure time block, "1": 12:00 AM - 05:59 AM, "2": 06:00 AM - 11:59 AM, "3": 12:00 PM - 05:59 PM, "4": 06:00 PM - 11:59 PM
DISTANCE	distance between airports, in miles

airport, we divide the departure time into 4 periods : "12:00 AM - 05:59 AM", "06:00 AM - 11:59 AM" , "12:00 PM - 05:59 PM" , "06:00 PM - 11:59 PM". Therefore DepTimeBlk (departure time block) takes $1 \sim 4$. The only continuous covariate, DISTANCE (distance of flight, $8 \sim 4983$ miles) is taken into consideration, as well as the binary response variable, Delay (arrival delay indicator, 1=YES). For details of variables, see Table XIX.

In order to evaluate the performances of MR and SMR even when full data estimate is not available, we generate the oracle coefficient values by fitting logistic model on the data files from March 2012 to February 2017, and then simulate pseudo responses 10 times through the logistic model with the oracle parameter values. That is, we know the true parameter values $\boldsymbol{\beta}$ in this case, shown in Table XX.

In order to mimic the distributed enviorment, we do not combine files or operate across multiple files. For the sub-partition, the only continuous variable DISTANCE is splitted into 8

TABLE XX: Oracle coefficients of predictors

Predictor	$\boldsymbol{\beta}$	$\boldsymbol{\beta}'$
Intercept	-2.3168	-2.3168
QUARTER2	-0.0074	-0.0074
QUARTER3	0.0024	0.0024
QUARTER4	-0.0952	-0.0952
DAY_OF_WEEK2	-0.1200	-0.1200
DAY_OF_WEEK3	-0.1079	-0.1079
DAY_OF_WEEK4	0.0632	0.0632
DAY_OF_WEEK5	0.0369	0.0369
DAY_OF_WEEK6	-0.2321	-0.2321
DAY_OF_WEEK7	-0.1041	-0.1041
DEP_TIME_BLK2	0.4678	0.4678
DEP_TIME_BLK3	1.0978	1.0978
DEP_TIME_BLK4	1.3058	1.3058
DISTANCE	5.87e-5	5.87e-4

intervals by its quantiles in each individual file. In order to show how the accuracy of parameter estimate is improved with more and more data, we run 4 experiments with the first 12 months, 60 months, 240 months and 353 months (that is, the whole dataset), respectively. In each experiment, we obtain full data estimate (not available for 240 months and 353 months due to big data size), as well as our MR and SMR estimates, which are listed in Table XXI. The dash mark is made for full data fits of 240 months and 353 months because our computer cannot handle such large size dataset directly.

From Table XXI, we can see that the three estimates are about the same. The main reason is that there is only one continuous predictor **DISTANCE**, whose coefficient is 0.0000587. Even multiplied by the largest value of DISTANCE, 4983, the contribution of DISTANCE is

TABLE XXI: Average (std) of RMSEs (10^{-3}) from oracle $\boldsymbol{\beta}$ for airline on-time performance data

Number of months	Full	MR	SMR
12 months	6.094(0.946)	6.087(0.955)	6.082(0.954)
60 months	2.695(0.282)	2.686(0.280)	2.690(0.282)
240 months	-	1.304(0.304)	1.303(0.302)
353 months	-	0.914(0.137)	0.910(0.137)

only 0.03, which is too small compared with the intercept -2.3168. In other words, this is roughly a case with all categorical predictors. According to Corollaries 4.1.1 and 4.2.1, both MR and SMR will match the full data estimate.

Table XXI also shows that as sample size gets bigger, both MR and SMR estimates are getting better, which is especially important when a full data estimate could not be obtained.

In order to show when SMR is better than MR, we enlarge the coefficient of DISTANCE by 10 times to get a new oracle $\boldsymbol{\beta}'$ (see Table XX). Then the maximum contribution of predictor DISTANCE becomes 0.3, which is expected to play a more important role in predicting Delay. We list the corresponding results in Table XXII. As the sample size increases, the improvement of SMR estimate over MR's becomes more in terms of average but still not that significant if we look at the standard deviation term, since the contribution of continuous term to linear predictor is still relative small. In that case, MR and SMR estimates are comparable but not the same as the full data estimate.

TABLE XXII: Average (std) of RMSEs (10^{-3}) from oracle $\boldsymbol{\beta}'$ with coefficient of DISTANCE enlarged by 10 times for airline on-time performance data

Number of months	Full	MR	SMR
12 months	6.530(1.167)	7.460(1.107)	7.167(1.156)
60 months	2.656(0.451)	3.319(0.540)	3.276(0.492)
240 months	-	2.181(0.368)	1.791(0.357)
353 months	-	1.946(0.283)	1.493(0.251)

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

When all predictors of the GLMs are categorical or discrete, the best solution would be partitioning the data according to distinct predictor values if applicable. In this case, $\Delta = 0$ and MR estimate exactly matches the full data estimate.

For GLMs with flat $G(\eta)$ (that is, $G'(\eta)$ is bounded by some moderate number), such as logit, probit, cloglog, loglog, and cauchit links for binomial models, one may check the coefficients of the continuous variables fitted by MR. That is, estimate the contribution of continuous variables to linear predictor by multiplying their coefficients by first quantiles and third quantiles respectively, then compare the products to the intercept and the coefficients of binary dummy covariates of categorical variables. If all linear predictors contributed by continuous variables are relatively small comparing to the intercept or linear predictors contributed by categorical ones, then MR estimate might be good enough. Otherwise, we recommend SMR solution.

For GLMs with unbounded or too large $G'(\eta)$, such as Poisson model, Gamma model, we recommend SMR over MR with a finer partition.

6.2 Dynamic Distributed Computing Framework

A method designed for dynamic distributed dataset is now desired. It should satisfy a scenario as the follows: the central processor mode and local nodes with most data restored

on, are physically distributed; The communications between them are expensive, and may not be synchronized; The data on local nodes may be updated along time. A dynamic distributed computing framework is described as follows.

1. Central processor node sends the initial information to the distributed local nodes;
2. Local nodes calculate independently based on the current local data and the system information, then send the calculating results to the central processor node;
3. The central processor node makes analysis on the data sent by local nodes;
4. Repeat $1 \sim 3$ until some thresholds are triggered.

6.3 Future Works

Firstly, current SMR is good for distributed dataset, it can be improved to fit a dynamic distributed system. The algorithm will be the same, but the properties of SMR in dynamic distributed computing are worth to explore. Working under a dynamic distributed computing framework is one of our most important future works.

The choice of representatives is not unique for GLMs. We may not only consider representatives based on the score functions, but also ones based on the log-likelihoods themselves directly, or ones based on information matrices.

Even for the model described in Theorem 4.2.3, our proposed SMR does not perfectly satisfy all the conditions. An improved SMR under this framework could be a next step.

Also, MR is a quick solution but essentially a pre-analysis for GLMs. The statistical inference for MR is a completion of this representative approach, as well as median representative

approach. Furthermore, could we have other representatives with computing speed comparable to MR, but containing more information?

Data partition, or more specifically, Δ, is critical for both MR and SMR. How to obtain a more efficient partition is very important to representative approaches.

All above topics are under model based framework, nonparametric structure will be another story.

CITED LITERATURE

1. Kane, M., Emerson, J., and Weston, S.: Scalable strategies for computing with massive data. Journal of Statistical Software, 55:1–19, 2013.

2. Wang, C., Chen, M., Schifano, E., Wu, J., and Yan, J.: Statistical methods and computing for big data. Statistics and Its Interface, 9:399–414, 2016.

3. McCullagh, P. and Nelder, J.: Generalized Linear Models. Chapman and Hall/CRC, 2 edition, 1989.

4. Dobson, A. and Barnett, A.: An Introduction to Generalized Linear Models. Chapman & Hall/CRC, 3 edition, 2008.

5. Wang, H., Zhu, R., and Ma, P.: Optimal subsampling for large sample logistic regression. Journal of the American Statistical Association, 2017. to appear, available via `https://arxiv.org/abs/1702.01166`.

6. Lin, N. and Xi, R.: Aggregated estimating equation estimation. Statistics and Its Interface, 4:73–83, 2011.

7. Ma, P. and Sun, X.: Leveraging for big data regression. WIREs Computational Statistics, 7:70–76, 2014.

8. Wang, H., Yang, M., and Stufken, J.: Information-based optimal subdata selection for big data linear regression. Journal of the American Statistical Association, 2018. to appear, available via `https://haiying-wang.uconn.edu/wp-content/uploads/sites/2127/2017/04/IBOSS_Linear.pdf`.

9. Kotsiantis, S. and Kanellopoulos, D.: Discretization techniques: A recent survey. GESTS International Transactions on Computer Science and Engineering, 32(1):47–58, 2006.

10. Özsu, M. and Valduriez, P.: Principles of Distributed Database Systems. Springer Science & Business Media, 3 edition, 2011.

11. Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A., Foufou, S., and Bouras, A.: A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE transactions on emerging topics in computing, 2.3:267–279, 2014.

12. Hartigan, J.: Clustering Algorithms. Wiley, 1975.

13. Pakhira, M.: A linear time-complexity k-means algorithm using cluster shifting. 2014 Sixth International Conference on Computational Intelligence and Communication Networks, pages 1047–1051, 2014.

14. Johnson, S.: Hierarchical clustering schemes. Psychometrika, 32(3):241–254, 1967.

15. Ester, M., Kriegel, H. P., Sander, J., and Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd, 34:226–231, 2014.

16. Ghahramani, Z. and Jordan, M. I.: Supervised learning from incomplete data via an em approach. Advances in neural information processing systems, pages 120–127, 1994.

17. Kanniappan, P. and Sastry, S. M.: Uniform convergence of convex optimization problems. Journal of mathematical analysis and applications, 96.1:1–12, 1983.

VITA

NAME: Keren Li

Education

B.Sc., Nankai University, Tianjin, China, 2001

M.S., Louisiana State University, Baton Rouge, US, 2004

Recent Teaching and Working Experience

Chongqing University of Science and Technology,Chongqing, China, 2010-2014

Visiting Scholar, *University of Illinois at Chicago*, IL, 2014-2015

Teaching/Research Assistant, *University of Illinois at Chicago,* Chicago, IL, Fall 2015 – Summer 2018

Publications

D-optimal Sampling method for Big Data with Multinomial Logistic Models. With Jie Yang. (Ready to submit)

D-optimal Designs for Generalized Linear Models with Mixed Factors. With Jie Yang and Abhyuday Mandal. (Ready to submit)

A New Nonparametric Estimation of Risk-Neutral Density and its Application in Variance Swaps. With Liyuan Jiang, Shuang Zhou and Jie Yang. (Ready to submit)