Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Optimal Distributed Subsampling for Maximum Quasi-Likelihood Estimators With Massive Data

Jun Yu, HaiYing Wang, Mingyao Ai & Huiming Zhang

To cite this article: Jun Yu, HaiYing Wang, Mingyao Ai & Huiming Zhang (2022) Optimal Distributed Subsampling for Maximum Quasi-Likelihood Estimators With Massive Data, Journal of the American Statistical Association, 117:537, 265-276, DOI: 10.1080/01621459.2020.1773832

To link to this article: https://doi.org/10.1080/01621459.2020.1773832

View supplementary material

Published online: 07 Jul 2020.

Submit your article to this journal

Article views: 3641

View related articles

View Crossmark data

Citing articles: 70 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uasa20

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 117, NO. 537, 265–276: Theory and Methods
https://doi.org/10.1080/01621459.2020.1773832

Taylor & Francis
Taylor & Francis Group

Check for updates

Optimal Distributed Subsampling for Maximum Quasi-Likelihood Estimators With Massive Data

Jun Yu[a], HaiYing Wang[b], Mingyao Ai[c], and Huiming Zhang[d]

[a]School of Mathematics and Statistics, Beijing Institute of Technology, Beijing, China; [b]Department of Statistics, University of Connecticut, Storrs, CT; [c]LMAM, School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing, China; [d]School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing, China

ABSTRACT

Nonuniform subsampling methods are effective to reduce computational burden and maintain estimation efficiency for massive data. Existing methods mostly focus on subsampling with replacement due to its high computational efficiency. If the data volume is so large that nonuniform subsampling probabilities cannot be calculated all at once, then subsampling with replacement is infeasible to implement. This article solves this problem using Poisson subsampling. We first derive optimal Poisson subsampling probabilities in the context of quasi-likelihood estimation under the A- and L-optimality criteria. For a practically implementable algorithm with approximated optimal subsampling probabilities, we establish the consistency and asymptotic normality of the resultant estimators. To deal with the situation that the full data are stored in different blocks or at multiple locations, we develop a distributed subsampling framework, in which statistics are computed simultaneously on smaller partitions of the full data. Asymptotic properties of the resultant aggregated estimator are investigated. We illustrate and evaluate the proposed strategies through numerical experiments on simulated and real datasets. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received January 2019
Accepted May 2020

KEYWORDS
Big data; Distributed subsampling; Poisson sampling; Quasi-likelihood

1. Introduction

Nowadays, the sizes of collected data are ever increasing, and the incredible sizes of big data bring new challenges for data analysis. Although many traditional statistical methods are still valid with big data, it is often computationally infeasible to perform statistical analysis due to relatively limited computing power. In this scenario, the bottleneck for big data analysis is the limited computing resources, and extracting useful information from massive datasets is a primary goal.

In general, there are two computational barriers for big data analysis: the first is that the dataset is too large to be held in a computer's memory; and the second is that the computation takes too long to obtain the results. Faced with these two challenges, current research on statistical inference for big datasets can be categorized into two basic approaches. One approach utilizes parallel computing platforms by dividing the whole dataset into subsets to compute; the results from subsets are then combined to obtain a final estimator (see Lin and Xi 2011; Duchi, Agarwal, and Wainwright 2012; Li, Lin, and Li 2013; Kleiner et al. 2015; Schifano et al. 2016; Jordan, Lee, and Yang 2019, and the references therein). The other approach uses subsampling to reduce the computational burden by carrying out intended calculations on a subsample drawn from the full data (see Drineas et al. 2011; Dhillon et al. 2013; Ma, Mahoney, and Yu 2015; Quiroz et al. 2019, among others).

A key tactic of subsampling methods is to specify nonuniform sampling probabilities to include more informative data points with higher probabilities. Typical examples are the leverage score-based subsampling (see Drineas et al. 2011; Mahoney 2012; Ma, Mahoney, and Yu 2015) and optimal subsampling method under the A-optimality criterion (Wang, Zhu, and Ma 2018). Wang, Yang, and Stufken (2019) proposed the information based optimal subdata selection for linear models which selects the subsample deterministically without random sampling.

It is worth mentioning that most of the current subsampling strategies focus on linear regression models and logistic regression models. However, many more complicated models are required in mining massive data because a linear regression model or a logistic regression model may not be sufficient to fit a complicated large dataset. For example, the paper citation dataset (*https://www.aminer.cn/citation*) contains text information for over four million research papers. Although we can extract numerical features from these texts, a linear regression or a logistic regression is clearly not adequate to model the number of citations for these papers. As another example, the airline dataset (*http://stat-computing.org/dataexpo/2009/the-data.html*) has more than one hundred million observations, and a primary goal is to model the airline delays, which are right skewed and always positive. A log or power transform

CONTACT Mingyao Ai *myai@math.pku.edu.cn* LMAM, School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing 100871, China.

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

© 2020 American Statistical Association

may help to alleviate the skewness, but a Gamma regression may give better interpretability. More details about these two datasets will be provided in Section 5. To support more statistical models, this article focuses on the quasi-likelihood estimator which only requires assumptions on the moments of the response variable and the form of the distribution is not specified.

Subsampling with replacement according to unequal probabilities requires accessing subsampling probabilities for the full data all at once. This takes a large memory to implement and may reduce the computational efficiency. To overcome this challenge, we propose an algorithm based on Poisson sampling (Särndal, Swensson, and Wretman 1992). Compared with subsampling with replacement, Poisson subsampling also has a high estimation efficiency with nonuniform subsampling probabilities. To utilize parallel computing facilities, a distributed version of the algorithm is also developed which enables us to select subsamples in parallel or in different locations simultaneously. To the best of our knowledge, theoretical and methodological discussions with statistical guarantees on optimal subsampling from massive data are limited for statistical models beyond linear regression models and logistic regression models. This article not only develops optimal subsampling method for quasi-likelihood estimators but also solves storage constraints imposed by large scale datasets.

The rest of the article is organized as follows. In Section 2, we introduce the model setup, present the general Poisson subsampling algorithm, and derive theoretical results for the resultant estimator. Section 3 presents optimal subsampling strategies based on the A- and L-optimality criteria for quasi-likelihood estimators. Some practical issues to approximate and implement the optimal subsampling procedures are also considered with theoretical justifications. Section 4 designs a distributed version of the Poisson subsampling algorithm and presents asymptotic properties of the resultant estimators. Section 5 provides numerical results on simulated and real datasets. All proofs are deferred in the supplementary materials.

2. Preliminaries

In this section, we first provide a brief overview of quasi-likelihood estimation and then present the general Poisson subsampling algorithm.

2.1. Models and Assumptions

We adopt the notations for quasi-likelihood estimator discussed in Chen, Hu, and Ying (1999). Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ be a sequence of independent and identically distributed (iid) random variables with each covariate $\boldsymbol{x}_i \in \mathbb{R}^d$ and response $y_i \in \mathbb{R}$. The conditional expectation of the response y_i given $\boldsymbol{x}_i$ is

$$E(y_i|\boldsymbol{x}_i) = \psi({\boldsymbol{\beta}_t}^T \boldsymbol{x}_i), \quad i = 1, 2, \ldots, N, \tag{1}$$

for some true regression parameter vector $\boldsymbol{\beta}_t \in \mathbb{R}^d$, where $\psi(\cdot)$ is a twice continuously differentiable function such that $\dot{\psi}(t) := d\psi(t)/dt > 0$ for all t. The quasi-likelihood estimator $\hat{\boldsymbol{\beta}}_{\text{QLE}}$ is the solution to the following estimation equation:

$$Q(\boldsymbol{\beta}) := \sum_{i=1}^N \{y_i - \psi(\boldsymbol{\beta}^T \boldsymbol{x}_i)\}\boldsymbol{x}_i = \mathbf{0}. \tag{2}$$

The inference procedure based on (2) is very general, and a typical example is the maximum likelihood estimation for generalized linear models (Mccullagh and Nelder 1989). More details can be found in Fahrmeir and Tutz (2001), Chen (2011), and the references therein.

2.2. General Poisson Subsampling Algorithm

Let p_i be the probability to sample the ith data point for $i = 1, \ldots, N$, and let S be a set of subsample observations and the corresponding sampling probabilities. A general Poisson subsampling algorithm is presented in Algorithm 1.

Algorithm 1: General Poisson subsampling algorithm

Initialization $S = \varnothing$;
for $i = 1, \ldots, N$ **do**
  Generate a Bernoulli variable $\delta_i \sim \text{Bernoulli}(p_i)$;
  if $\delta_i = 1$ **then**
    Update $S = S \cup \{(\boldsymbol{x}_i, y_i, p_i)\}$;
Estimation: Solve the following weighted estimation equation to obtain $\tilde{\boldsymbol{\beta}}$ based on the subsample S,

$$Q^*(\boldsymbol{\beta}) = \sum_S \frac{1}{p_i}\{y_i - \psi(\boldsymbol{\beta}^T \boldsymbol{x}_i)\}\boldsymbol{x}_i = \mathbf{0}. \tag{3}$$

An advantage of Poisson subsampling is that the decision of inclusion for each data point $(\boldsymbol{x}_i, y_i)$ is made on the basis of p_i only. We do not need to use all p_i for $i = 1, \ldots, N$ together. In Algorithm 1, p_i can be used one-by-one or block-by-block to generate δ_i while scanning through the full data. Therefore, there is no memory constraint problem for massive data.

The subsample size, say r^*, in Algorithm 1 is random such that $E(r^*) = \sum_{i=1}^N p_i$. We use $r = \sum_{i=1}^N p_i$ to denote the expected subsample size, and further assume $r < N$ throughout this article, which is natural in the big data setting.

To establish our asymptotic results, we need the following assumptions.

Assumption 1. The regression parameter lies in the l_1 ball $\Lambda = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_1 \le B\}$, with $\boldsymbol{\beta}_t$ and $\hat{\boldsymbol{\beta}}_{\text{QLE}}$ being the inner points of Λ, where B is a constant.

Assumption 2. Suppose that

(i) $E(\|\boldsymbol{x}_1\|^9) < \infty$, (ii) $E(|y_1|^6) < \infty$,
(iii) $E\left\{\sup_{\boldsymbol{\beta}\in\Lambda} \psi^6(\boldsymbol{\beta}^T \boldsymbol{x}_1)\right\} < \infty$, (iv) $E\left\{\sup_{\boldsymbol{\beta}\in\Lambda} \dot{\psi}^6(\boldsymbol{\beta}^T \boldsymbol{x}_1)\right\} < \infty$.

Assumption 3. Let $\Sigma_\psi(\boldsymbol{\beta}) = N^{-1}\sum_{i=1}^N \dot{\psi}(\boldsymbol{\beta}^T \boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T$, and further assume that it satisfies $\lim_{N\to\infty} \inf_{\boldsymbol{\beta}\in\Lambda} \lambda_{\min}\{\Sigma_\psi(\boldsymbol{\beta})\} > 0$ with probability approaching one, where $\lambda_{\min}(A)$ means the smallest eigenvalue of matrix A.

Assumption 4. Assume that both $\psi(\boldsymbol{\beta}^T\boldsymbol{x}_i)$ and $\dot{\psi}(\boldsymbol{\beta}^T\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T$ are $m(\boldsymbol{x}_i)$-Lipschitz continuous. To be precise, for all $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \Lambda$, there exist $m_1(\boldsymbol{x}_i)$ and $m_2(\boldsymbol{x}_i)$ such that $\|\psi(\boldsymbol{\beta}_1^T\boldsymbol{x}_i) - \psi(\boldsymbol{\beta}_2^T\boldsymbol{x}_i)\| \le m_1(\boldsymbol{x}_i)\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|$ and $\|\dot{\psi}(\boldsymbol{\beta}_1^T\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T - \dot{\psi}(\boldsymbol{\beta}_2^T\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T\|_s \le m_2(\boldsymbol{x}_i)\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|$, where $\|A\|_s$ denotes the spectral norm of matrix A. Further assume that both $E\{m_1^3(\boldsymbol{x}_i)\}$ and $E\{m_2(\boldsymbol{x}_i)\}$ are finite.

Assumption 5. Assume that $\max_{i=1,\ldots,N}(Np_i)^{-1} = O_P(r^{-1})$.

Assumption 1 is required to guarantee consistency and it is commonly used in the literature such as Newey and McFadden (1994). Assumption 2 imposes some moment assumptions, and similar conditions were also assumed in Chen, Hu, and Ying (1999). Conditions (iii) and (iv) in Assumption 2 are satisfied by many examples of generalized linear models such as linear regressions, logistic regressions, and binomial regressions when the covariate distributions are sub-Gaussian. Assumption 3 is mainly to ensure that the quasi-likelihood estimator is unique, since this condition indicates that the quasi log-likelihood function is convex (cf. Tzavelas 1998; Rao et al. 2007; Chen 2011). Assumption 4 adds restrictions on smoothness. Similar assumptions are common in statistics (see, e.g., van der Vaart 1998, chap. 5). Assumption 5 restricts the weights in the estimation equation (3). It is mainly to protect the estimation equation from being dominated by data points with extremely small subsampling probabilities. This assumption is quite common in classic sampling techniques (see, e.g., Breidt and Opsomer 2000; Berger and Torres 2016). In this article, we allow the subsampling probability p_i to dependent on the observed data, so we use the O_P notation in Assumption 5.

To facilitate the presentation, denote the full data by $\mathcal{F}_N = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$. The following theorems establish consistency to the full data QLE and asymptotically normality of $\tilde{\boldsymbol{\beta}}$ from Algorithm 1.

Theorem 1. If Assumptions 1–5 hold, then as $N \to \infty$ and $r \to \infty$, $\tilde{\boldsymbol{\beta}}$ is consistent to $\hat{\boldsymbol{\beta}}_{\text{QLE}}$ in conditional probability, given $\mathcal{F}_N$ in probability. Moreover, the rate of convergence is $r^{-1/2}$. That is, with probability approaching one, for any $\epsilon > 0$, there exists a finite Δ_ϵ and r_ϵ such that

$$P(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{QLE}}\| \ge r^{-1/2}\Delta_\epsilon|\mathcal{F}_N) < \epsilon \tag{4}$$

for all $r > r_\epsilon$.

Theorem 2. If Assumptions 1–5 hold, then as $N \to \infty$ and $r \to \infty$, conditional on $\mathcal{F}_N$ in probability,

$$V^{-1/2}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{QLE}}) \longrightarrow N(0, I) \tag{5}$$

in distribution, where

$$V = \Sigma_\psi(\hat{\boldsymbol{\beta}}_{\text{QLE}})^{-1}V_c\Sigma_\psi(\hat{\boldsymbol{\beta}}_{\text{QLE}})^{-1} \tag{6}$$

and

$$V_c = \frac{1}{N^2}\sum_{i=1}^N \frac{\{y_i - \psi(\hat{\boldsymbol{\beta}}_{\text{QLE}}^T\boldsymbol{x}_i)\}^2\boldsymbol{x}_i\boldsymbol{x}_i^T}{p_i} - \frac{1}{N^2}\sum_{i=1}^N \{y_i - \psi(\hat{\boldsymbol{\beta}}_{\text{QLE}}^T\boldsymbol{x}_i)\}^2\boldsymbol{x}_i\boldsymbol{x}_i^T. \tag{7}$$

Remark 1. When $r/N \to 0$, the second term on the right-hand-side of (7) can be ignored. In this case, the result is the same as that for sampling with replacement in logistic regression (see Wang, Zhu, and Ma 2018). However, when $r/N \to c \in (0, 1]$, Poisson subsampling will lead to a smaller variance.

3. Optimal Poisson Subsampling

In this section, we derive optimal subsampling probabilities to better approximate $\hat{\boldsymbol{\beta}}_{\text{QLE}}$.

3.1. Optimal Subsampling Strategies

The result in Theorem 2 can be used to find optimal subsampling probabilities that minimize the asymptotic mean squared error (MSE) of $\tilde{\boldsymbol{\beta}}$ in approximating $\hat{\boldsymbol{\beta}}_{\text{QLE}}$. This is equivalent to minimizing $\text{tr}(V)$, which corresponds to the A-optimality in the language of optimal design (see Pukelsheim 2006).

Theorem 3. For ease of presentation, define

$$\hbar_i^{\text{MV}} = |y_i - \psi(\hat{\boldsymbol{\beta}}_{\text{QLE}}^T\boldsymbol{x}_i)|\|\Sigma_\psi(\hat{\boldsymbol{\beta}}_{\text{QLE}})^{-1}\boldsymbol{x}_i\|, \quad i = 1, \ldots, N, \tag{8}$$

and let $\hbar_{(1)}^{\text{MV}} \le \hbar_{(2)}^{\text{MV}} \le \cdots \le \hbar_{(N)}^{\text{MV}}$ denote the order statistics of $\{\hbar_i^{\text{MV}}\}_{i=1}^N$. For convenience, denote $\hbar_{(N+1)}^{\text{MV}} = +\infty$, and assume that $\hbar_{(N-r)}^{\text{MV}} > 0$. The asymptotic MSE of $\tilde{\boldsymbol{\beta}}$, $\text{tr}(V)$, attains its minimum, if p_i's in Algorithm 1 are chosen to be

$$p_i^{\text{MV}} = r\frac{\hbar_i^{\text{MV}} \wedge M}{\sum_{j=1}^N(\hbar_j^{\text{MV}} \wedge M)}, \tag{9}$$

where $a \wedge b = \min(a, b)$,

$$M = \frac{1}{r-k}\sum_{i=1}^{N-k}\hbar_{(i)}^{\text{MV}}, \tag{10}$$

and

$$k = \min\left\{s \,\middle|\, 0 \le s \le r, \quad (r-s)\hbar_{(N-s)}^{\text{MV}} < \sum_{i=1}^{N-s}\hbar_{(i)}^{\text{MV}}\right\}, \tag{11}$$

that is, k satisfies

$$(r-k+1)\hbar_{(N-k+1)}^{\text{MV}} \ge \sum_{i=1}^{N-k+1}\hbar_{(i)}^{\text{MV}} \quad \text{and} \tag{12}$$

$$(r-k)\hbar_{(N-k)}^{\text{MV}} < \sum_{i=1}^{N-k}\hbar_{(i)}^{\text{MV}}.$$

Remark 2. In (9), if $r\hbar_{(N)}^{\text{MV}}/(\sum_{j=1}^N\hbar_j^{\text{MV}}) < 1$, then $\hbar_{(N)}^{\text{MV}} < M = r^{-1}\sum_{j=1}^N\hbar_j^{\text{MV}}$ and the optimal subsampling probabilities reduce to $p_i^{\text{MV}} = r\hbar_i^{\text{MV}}/(\sum_{j=1}^N\hbar_j^{\text{MV}})$. In this case, all p_i^{MV}'s are smaller than one and the inclusion of any data point in the subsample is random. If $r\hbar_i^{\text{MV}}/(\sum_{j=1}^N\hbar_j^{\text{MV}}) \ge 1$ for some i, then some p_i^{MV}'s will be equal to one. For this scenario, k is the number of p_i^{MV}'s that are one and M is the threshold that satisfies

$$\max_{i=1,\ldots,N}\frac{r(\hbar_i^{\text{MV}} \wedge M)}{\sum_{j=1}^N(\hbar_j^{\text{MV}} \wedge M)} = 1. \tag{13}$$

From (10) and (12), we see that

$$\hbar^{\mathrm{MV}}_{(N-k)} < M \le \hbar^{\mathrm{MV}}_{(N-k+1)}. \tag{14}$$

Remark 3. To determine the value of k, we need to find and sort at most r largest values of $\hbar^{\mathrm{MV}}_i$'s. Thus, the required time to find the value of k is $O(N + r\log r)$ using partition based partial selection algorithm. The simulation results reveal that when $r/N \to c > 0$, it also works well if we select M as some quantile of $\{\hbar_i\}_{i=1}^N$.

As observed in (8), the optimal subsampling probability $\boldsymbol{p}^{\mathrm{MV}} = \{p^{\mathrm{MV}}_i\}_{i=1}^N$ depends on data through both the covariates and the responses directly. For the covariates, the terms $\|\Sigma_\psi(\hat{\boldsymbol{\beta}}_{\mathrm{QLE}})^{-1}\boldsymbol{x}_i\|$ describe the structure information of the covariates and they are similar to statistical leverage scores. The direct effect of the responses on the optimal subsampling probabilities is through $|y_i - \psi(\hat{\boldsymbol{\beta}}^T_{\mathrm{QLE}}\boldsymbol{x}_i)|$. Intuitively, including data points with lager values of $|y_i - \psi(\hat{\boldsymbol{\beta}}^T_{\mathrm{QLE}}\boldsymbol{x}_i)|$ will improve the robustness of the subsample estimator.

The optimal subsampling strategy derived in the previous section requires the calculation of $\|\Sigma_\psi(\hat{\boldsymbol{\beta}}_{\mathrm{QLE}})^{-1}\boldsymbol{x}_i\|$ for $i = 1, 2, \ldots, N$, which takes $O(Nd^2)$ time even if $\Sigma_\psi(\hat{\boldsymbol{\beta}}_{\mathrm{QLE}})$ is available. To further reduce the computation time, Wang, Zhu, and Ma (2018) proposed to minimize $\mathrm{tr}(V_c)$. This criterion essentially is the linear optimality (L-optimality) criterion in optimal experimental design (see Pukelsheim 2006), which is to improve the quality of the estimator for some linear combinations of unknown parameters.

The following theorem gives the optimal subsampling probabilities that minimize $\mathrm{tr}(V_c)$.

Theorem 4. Let

$$\hbar^{\mathrm{MVc}}_i = |y_i - \psi(\hat{\boldsymbol{\beta}}^T_{\mathrm{QLE}}\boldsymbol{x}_i)|\|\boldsymbol{x}_i\|, \quad i = 1, \ldots, N, \tag{15}$$

and let $\hbar^{\mathrm{MVc}}_{(1)} \le \hbar^{\mathrm{MVc}}_{(2)} \le \cdots \le \hbar^{\mathrm{MVc}}_{(N)}$ denote the order statistics of $\{\hbar^{\mathrm{MVc}}_i\}_{i=1}^N$. For convenience, denote $\hbar^{\mathrm{MVc}}_{(N+1)} = +\infty$ and assume that $\hbar^{\mathrm{MVc}}_{(N-r)} > 0$. The trace of V_c defined in (7) attains its minimum if p_i's in Algorithm 1 are selected as

$$p^{\mathrm{MVc}}_i = r\frac{\hbar^{\mathrm{MVc}}_i \wedge M}{\sum_{j=1}^N \hbar^{\mathrm{MVc}}_j \wedge M}, \tag{16}$$

where

$$M = (r-k)^{-1}\sum_{i=1}^{N-k}\hbar^{\mathrm{MVc}}_{(i)}, \tag{17}$$

and

$$k = \min\left\{s \,\middle|\, 0 \le s \le r, \quad (r-s)\hbar^{\mathrm{MVc}}_{(N-s)} < \sum_{i=1}^{N-s}\hbar^{\mathrm{MVc}}_{(i)}\right\}, \tag{18}$$

that is, k satisfies

$$(r-k+1)\hbar^{\mathrm{MVv}}_{(N-k+1)} \ge \sum_{i=1}^{N-k+1}\hbar^{\mathrm{MV}}_{(i)} \quad \text{and}$$

$$(r-k)\hbar^{\mathrm{MVc}}_{(N-k)} < \sum_{i=1}^{N-k}\hbar^{\mathrm{MVc}}_{(i)}.$$

The structural results for $\boldsymbol{p}^{\mathrm{MVc}} = \{p^{\mathrm{MVc}}_i\}_{i=1}^N$ and $\boldsymbol{p}^{\mathrm{MV}}$ are similar. The difference is in the covariate effect: $\boldsymbol{p}^{\mathrm{MV}}$ uses $\|\Sigma_\psi(\hat{\boldsymbol{\beta}}_{\mathrm{QLE}})^{-1}\boldsymbol{x}_i\|$ while $\boldsymbol{p}^{\mathrm{MVc}}$ uses $\|\boldsymbol{x}_i\|$. The computational benefits is obvious, only $O(Nd)$ time is required to compute $\boldsymbol{p}^{\mathrm{MVc}}$ while $O(Nd^2)$ is needed for $\boldsymbol{p}^{\mathrm{MV}}$.

3.2. Practical Implementation

For ease of presentation, we use a unified notation p^{os}_i to denote the optimal subsampling probabilities p^{MV}_i or p^{MVc}_i derived in Theorems 3 or 4, respectively. To be precise,

$$p^{\mathrm{os}}_i = r\frac{\hbar^{\mathrm{os}}_i \wedge M}{\sum_{j=1}^N(\hbar^{\mathrm{os}}_j \wedge M)} = r\frac{\hbar^{\mathrm{os}}_i \wedge M}{N\Psi}, \quad i = 1, \ldots, N, \tag{19}$$

where $M = (r-k)^{-1}\sum_{i=1}^{N-k}\hbar^{\mathrm{os}}_{(i)}$, $\Psi = N^{-1}\sum_{j=1}^N(\hbar^{\mathrm{os}}_j \wedge M)$, and $\hbar^{os}_i$ is either $\hbar^{\mathrm{MV}}_i$ or $\hbar^{\mathrm{MVc}}_i$.

To practically implement the optimal subsampling probabilities, we need to replace the unknown $\hat{\boldsymbol{\beta}}_{\mathrm{QLE}}$ by a pilot estimator, say $\tilde{\boldsymbol{\beta}}_0$, which can be obtained by taking a uniform subsample. Some other sampling distributions can also be used to obtain the pilot estimator as long as they satisfy Assumption 5 and are computationally feasible to implement. Furthermore, to take advantage of Poisson subsampling and determine the inclusion of each data point separately, we use the pilot sample to approximate M and Ψ.

In the setting of subsampling for computational efficiency, it is typical that $r \ll N$ and the number of cases that $\hbar^{\mathrm{os}}_i > M$ is small. Thus, taking $M = \infty$ will not significantly affect the optimal subsampling probabilities. In facts, if $r\hbar^{\mathrm{os}}_{(N)}/(\sum_{j=1}^N \hbar^{\mathrm{os}}_j) \le 1$, then taking $M = \infty$ does not affect the optimal subsampling probabilities at all. Simulation results in Section 5 show that taking $M = \infty$ does not reduce the estimation efficiency as long as r/N is small.

Let $\tilde{S}_{r_0}$ be the set of the pilot subsample and

$$\hat{\Psi} = \frac{1}{|\tilde{S}_{r_0}|}\sum_{\tilde{S}_{r_0}}|y_i - \psi(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_i)|h(\boldsymbol{x}_i), \tag{20}$$

where $|\tilde{S}_{r_0}|$ is the size of $\tilde{S}_{r_0}$, and $h(\boldsymbol{x}) = \|\boldsymbol{x}\|$ for MVc or $h(\boldsymbol{x}) = \|\Sigma_\psi(\tilde{\boldsymbol{\beta}}_0)^{-1}\boldsymbol{x}\|$ for MV with $\Sigma_\psi(\tilde{\boldsymbol{\beta}}_0)$ calculated as $\Sigma_\psi(\tilde{\boldsymbol{\beta}}_0) = |\tilde{S}_{r_0}|^{-1}\sum_{\tilde{S}_{r_0}}\dot{\psi}(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_i^*)\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}$. Let $\tilde{p}^{\mathrm{os}}_i$ be the approximated subsampling probabilities with $\hat{\boldsymbol{\beta}}_{\mathrm{QLE}}$, M, and Ψ in (19) replaced by the pilot estimator $\tilde{\boldsymbol{\beta}}_0$, $M = \infty$, and $\hat{\Psi}$. The weighted estimator with $\tilde{p}^{\mathrm{os}}_i$ inserted in (3) may be sensitive to data points with $y_i - \dot{\psi}(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_i) \approx 0$ if they are included in the subsample. To make the estimator more stable and robust, we adopt the idea of shrinkage-based subsampling method proposed in Ma, Mahoney, and Yu (2015). To be specific, we use the following subsampling probabilities

$$\tilde{p}^{\mathrm{sos}}_i = (1-\varrho)\frac{r|y_i - \psi(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_i)|h(\boldsymbol{x}_i)}{N\hat{\Psi}} + \varrho r N^{-1}, \quad i = 1, \ldots, N, \tag{21}$$

where $\varrho \in (0, 1)$.

Note that when $\tilde{\boldsymbol{\beta}}_0$ and $\hat{\Psi}$ are calculated from the pilot subsample, $\tilde{p}^{\mathrm{sos}}_i$ depends on the ith observation $(\boldsymbol{x}_i, y_i)$ only.

Thus, each $\tilde{p}_i^{\text{sos}}$ can be calculated when scanning the data from hard drive line-by-line or block-by-block; there is no need to calculate $\tilde{p}_i^{\text{sos}}$'s all at once. Therefore, there is no need to load the full data into memory to calculate all $\tilde{p}_i^{\text{sos}}$'s and this is very computationally beneficially in terms of memory usage.

In (21), $\tilde{\boldsymbol{p}}^{\text{sos}} = \{\tilde{p}_i^{\text{sos}}\}_{i=1}^N$ is a convex combination of $\tilde{\boldsymbol{p}}^{\text{os}} = \{\tilde{p}_i^{\text{os}}\}_{i=1}^N$ and the uniform subsampling probability, and it shares the strengths of both. When ϱ is larger, the corresponding estimator will be more stable since the estimation equation will not be inflated by data points with extremely small values of $\tilde{p}_i^{\text{os}}$. The rankings of $\tilde{p}_i^{\text{sos}}$ and $\tilde{p}_i^{\text{os}}$ are the same, so the estimator still enjoys the benefits of the optimal subsampling strategy. The shrinkage term not only increases small subsampling probabilities, but also shrinks large subsampling probabilities and thus protects the effects of potential outliers to some extent.

Since we approximate Ψ and take $M = \infty$, some $\tilde{p}_i^{\text{sos}}$ may be larger than one. Thus, we need to use inverses of $\tilde{p}_i^{\text{sos}} \wedge 1$'s as weights in the subsample QLE estimator. For transparent presentation, we summarize the practical procedure with approximated quantities in Algorithm 2.

Algorithm 2: Practical algorithm

Pilot subsampling: Run Algorithm 1 with average subsample size r_0 and $\boldsymbol{p}^{\text{UNIF}} = \{p_i := r_0/N\}_{i=1}^N$ to take a subsample set $\tilde{S}_{r_0}$, and use it to obtain an estimate $\tilde{\boldsymbol{\beta}}_0$ and $\hat{\Psi}$ as in (20).

Initialization: $S_0 = \tilde{S}_{r_0}$;

for $i = 1, \ldots, N$ **do**

  Generate $\delta_i \sim \text{Bernoulli}(1, p_i)$ with $p_i = \tilde{p}_i^{\text{sos}} \wedge 1$, where $\tilde{p}_i^{\text{sos}}$ is defined in (21);

  if $\delta_i = 1$ **then**

    Update $S_i = S_{i-1} \cup \{(y_i, \boldsymbol{x}_i, p_i)\}$

  else

    Set $S_i = S_{i-1}$

Estimation: Solve the following weighted estimating equation to obtain the estimate $\breve{\boldsymbol{\beta}}$ based on the subsample set S_N.

$$Q^*(\boldsymbol{\beta}) = \sum_{S_N} \frac{1}{p_i}[y_i - \psi(\boldsymbol{\beta}^T \boldsymbol{x}_i)]\boldsymbol{x}_i = 0.$$

For estimators obtained from Algorithm 2, we derive asymptotic properties as follows.

Theorem 5. Under Assumptions 1–4, if $r_0 r^{-1/2} \to 0$, then for the estimator $\breve{\boldsymbol{\beta}}$ obtained from Algorithm 2, as $r \to \infty$ and $N \to \infty$, with probability approaching one, for any $\epsilon > 0$, there exist finite Δ_ϵ and r_ϵ such that

$$P(\|\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{QLE}}\| \geq r^{-1/2}\Delta_\epsilon | \mathcal{F}_N) < \epsilon$$

for all $r > r_\epsilon$.

Theorem 6. If Assumptions 1–4 hold and $r_0 r^{-1/2} \to 0$, then as $r_0 \to \infty$, $r \to \infty$ and $N \to \infty$, conditionally on $\mathcal{F}_N$ in probability,

$$V^{-1/2}(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{QLE}}) \to N(0, I) \text{ in distribution,}$$

where $V = \Sigma_\psi(\hat{\boldsymbol{\beta}}_{\text{QLE}})^{-1} V_c \Sigma_\psi(\hat{\boldsymbol{\beta}}_{\text{QLE}})^{-1}$ and

$$V_c = \frac{1}{N^2}\sum_{i=1}^N \frac{\{1 - (p_i^{\text{sos}} \wedge 1)\}\{y_i - \psi(\hat{\boldsymbol{\beta}}_{\text{QLE}}^T \boldsymbol{x}_i)\}^2 \boldsymbol{x}_i \boldsymbol{x}_i^T}{p_i^{\text{sos}} \wedge 1},$$

with

$$p_i^{\text{sos}} := (1-\varrho)\frac{r|y_i - \psi(\hat{\boldsymbol{\beta}}_{\text{QLE}}^T \boldsymbol{x}_i)|\|\Sigma_\psi(\hat{\boldsymbol{\beta}}_{\text{QLE}})^{-1}\boldsymbol{x}_i\|}{\sum_{j=1}^N |y_j - \psi(\hat{\boldsymbol{\beta}}_{\text{QLE}}^T \boldsymbol{x}_j)|\|\Sigma_\psi(\hat{\boldsymbol{\beta}}_{\text{QLE}})^{-1}\boldsymbol{x}_j\|} + \varrho\frac{r}{N},$$

for MV criterion and

$$p_i^{\text{sos}} := (1-\varrho)\frac{r|y_i - \psi(\hat{\boldsymbol{\beta}}_{\text{QLE}}^T \boldsymbol{x}_i)|\|\boldsymbol{x}_i\|}{\sum_{j=1}^N |y_j - \psi(\hat{\boldsymbol{\beta}}_{\text{QLE}}^T \boldsymbol{x}_j)|\|\boldsymbol{x}_j\|} + \varrho\frac{r}{N},$$

for MVc criterion.

4. Distributed Poisson Subsampling

In this section, we discuss the distributed optimal Poisson subsampling procedure. For large datasets, it is common to analyze them on multiple machines. This motivates us to develop divide-and-conquer subsampling procedures that take advantages of parallel and distributed computational architectures. Although Poisson subsampling can be easily implemented in parallel, pooling the subsample sets from multiple machines together may still result in a subsample set that exceeds the memory limit of a single machine. In addition, transferring data may be time consuming and subject to security issues. Thus, this method can only be used when the subsample size on each machine is not that big. We propose to aggregate estimators derived in different machines to approximate the full data quasi-likelihood estimator. Here we assume that the entire dataset of size N are stored in K different machines, and let $\mathcal{F}_{Nj}$ ($j = 1, \ldots, K$) denote the data stored in the jth machine. For simplicity, assume that the number of observations in different machines are all equal to n, and denote the observations in $\mathcal{F}_{Nj}$ as $\{(y_{ji}, \boldsymbol{x}_{ji})\}_{i=1}^n$. We present the distributed optimal Poisson subsampling procedure in Algorithm 3.

Remark 4. The first step in Algorithm 3 can be implemented by sampling the data machine-by-machine and pooling all the subsamples together. Since r_0 is usually small in our setting, the time of communication can be ignored.

The results of consistency and asymptotic normality are presented in the following theorems.

Theorem 7. Under Assumptions 1–4, if the estimator $\tilde{\boldsymbol{\beta}}_0$ based on the first step sample exists, $r_0(Kr)^{-1/2} \to 0$ and the partition number K satisfies $K = O(r^\eta)$ for some η in $[0, 1/3]$, then conditional on $\mathcal{F}_N$, for the estimator $\tilde{\boldsymbol{\beta}}_{Kr}$ obtained from Algorithm 3, as $r \to \infty$ and $n \to \infty$, with probability approaching one, for any $\epsilon > 0$, there exist finite Δ_ϵ and r_ϵ such that

$$P(\|\tilde{\boldsymbol{\beta}}_{Kr} - \hat{\boldsymbol{\beta}}_{\text{QLE}}\| \geq (Kr)^{-1/2}\Delta_\epsilon | \mathcal{F}_N) < \epsilon$$

for all $r > r_\epsilon$.

Algorithm 3: Distributed optimal Poisson subsampling

Step 1: Obtain the pilot estimator

for $i = 1, \ldots, N$ **do**

Generate $\delta_i \sim \text{Bernoulli}(1, p_i)$ with $p_i = r_0/N$;

if $\delta_i = 1$ **then**

Add (x_i, y_i, p_i) to the subsample set S_{r_0}

For the obtained subsample S_{r_0}, calculate the pilot estimator $\tilde{\boldsymbol{\beta}}_0$, $\hat{\Psi}$, and $\dot{Q}_0$ according to (3), (20) and (23), respectively.

Step 2: Subsampling and compression

foreach $\mathcal{F}_{Nj}, j = 1, \ldots, K$ **do**

Initialization: $S_{j0} = \varnothing$;

for $i = 1, \ldots, n$ **do**

Calculate the corresponding subsampling probabilities $\tilde{p}_{ji}^{\text{sos}}$ according to (21);

Generate $\delta_{ji} \sim \text{Bernoulli}(1, p_{ji})$ with $p_{ji} = \tilde{p}_{ji}^{\text{sos}} \wedge 1$;

if $\delta_{ji} = 1$ **then**

Update $S_{ji} = S_{ji-1} \cup \{(y_{ji}, \boldsymbol{x}_{ji}, p_{ji})\}$.

else

Set $S_{ji} = S_{ji-1}$.

Obtain $\tilde{\boldsymbol{\beta}}_j$ by solving

$$Q_j^*(\boldsymbol{\beta}) = \frac{1}{n} \sum_{S_{jn}} \frac{1}{p_{ji}} \{y_{ji} - \psi(\boldsymbol{\beta}^T \boldsymbol{x}_{ji})\} \boldsymbol{x}_{ji} = \mathbf{0}, \tag{22}$$

and calculate

$$\dot{Q}_j^*(\tilde{\boldsymbol{\beta}}_j) = -\frac{1}{n} \sum_{S_{jn}} \frac{1}{p_{ji}} \dot{\psi}(\tilde{\boldsymbol{\beta}}_j^T \boldsymbol{x}_{ji}) \boldsymbol{x}_{ji} \boldsymbol{x}_{ji}^T. \tag{23}$$

Step 3: Combination

Combine the K estimators and the pilot estimator by calculating

$$\tilde{\boldsymbol{\beta}}_{Kr} = \left\{ \sum_{j=0}^{K} \dot{Q}_j^*(\tilde{\boldsymbol{\beta}}_j) \right\}^{-1} \sum_{j=0}^{K} \dot{Q}_j^*(\tilde{\boldsymbol{\beta}}_j) \tilde{\boldsymbol{\beta}}_j. \tag{24}$$

Theorem 8. Under Assumptions 1–4, if $r_0(Kr)^{-1/2} \to 0$ and the partition number K satisfies $K = O(r^\eta)$ for some η in $[0, 1/3]$, then for the estimator $\tilde{\boldsymbol{\beta}}_{Kr}$ obtained from Algorithm 3, conditionally on $\mathcal{F}_N$ in probability, as $n \to \infty$, $r \to \infty$ and $r_0 \to \infty$,

$$V_{\text{opt}}^{-1/2}(\tilde{\boldsymbol{\beta}}_{Kr} - \hat{\boldsymbol{\beta}}_{\text{QLE}}) \to N(0, I) \quad \text{in distribution,}$$

where $V_{\text{opt}} = \Sigma_\psi(\hat{\boldsymbol{\beta}}_{\text{QLE}})^{-1} V_{c,\text{opt}} \Sigma_\psi(\hat{\boldsymbol{\beta}}_{\text{QLE}})^{-1}$,

$$V_{c,\text{opt}} = \frac{1}{KN^2} \sum_{i=1}^{N} \frac{\{1 - (p_i^{\text{sos}} \wedge 1)\} \{y_i - \psi(\hat{\boldsymbol{\beta}}_{\text{QLE}}^T \boldsymbol{x}_i)\}^2 \boldsymbol{x}_i \boldsymbol{x}_i^T}{p_i^{\text{sos}} \wedge 1},$$

and p_i^{sos} is defined in Theorem 6.

For statistical inference, we propose to estimate the asymptotic variance-covariance matrix of $\tilde{\boldsymbol{\beta}}_{Kr}$ using

$$\tilde{V} = \left\{ \frac{1}{N} \sum_{j=0}^{K} \dot{Q}_j^*(\tilde{\boldsymbol{\beta}}_j) \right\}^{-1} \tilde{V}_c \left\{ \frac{1}{N} \sum_{j=0}^{K} \dot{Q}_j^*(\tilde{\boldsymbol{\beta}}_j) \right\}^{-1}, \tag{25}$$

where

$$\tilde{V}_c = \frac{1}{N^2} \left\{ \sum_{S_{r_0}} \frac{\{y_{0i}^* - \psi(\tilde{\boldsymbol{\beta}}_0^T \boldsymbol{x}_{0i}^*)\}^2 \boldsymbol{x}_{0i}^* \boldsymbol{x}_{0i}^{*T}}{(r_0/N)^2} (1 - r_0/N) + \sum_{j=1}^{K} \sum_{S_{jn}} \frac{\{y_{ji}^* - \psi(\tilde{\boldsymbol{\beta}}_j^T \boldsymbol{x}_{ji}^*)\}^2 \boldsymbol{x}_{ji}^* \boldsymbol{x}_{ji}^{*T}}{(\tilde{p}_{ji}^{\text{sos}*})^2} (1 - \tilde{p}_{ji}^{\text{sos}*}) \right\}.$$

This formula enables us to know how well $\tilde{\boldsymbol{\beta}}_{Kr}$ approximates $\hat{\boldsymbol{\beta}}_{\text{QLE}}$. When $Kr = o(N)$, we can also draw inference on the true parameter $\boldsymbol{\beta}_t$, since uncertainty of $\hat{\boldsymbol{\beta}}_{\text{QLE}}$ can be ignored under this assumption. It is worth mentioning that if we want to calculate (25), we also need to have $\sum_{S_{jn}} (1 - \tilde{p}_{ji}^{\text{sos}}) \{y_{ji} - \psi(\tilde{\boldsymbol{\beta}}_i^T \boldsymbol{x}_{ji})\}^2 \boldsymbol{x}_{ji} \boldsymbol{x}_{ji}^T / (\tilde{p}_{ji}^{\text{sos}})^2$ calculated on each machine.

Since the pilot estimator $\tilde{\boldsymbol{\beta}}_0$ has to be calculated anyway, our method is valuable even for the case $K = 1$ because this avoids iterative calculation on the Step 1 sample twice.

5. Numerical Studies

In this section, we present examples of numerical experiments using the methods developed in Sections 3 and 4. Computations are performed using R (R Core Team 2018). The performance of a sampling strategy is evaluated by the empirical MSE of the resultant estimator:

$$\text{MSE} = \frac{1}{T} \sum_{t=1}^{T} \|\boldsymbol{\beta}_{\boldsymbol{p}}^{(t)} - \hat{\boldsymbol{\beta}}_{\text{QLE}}\|^2,$$

where $\boldsymbol{\beta}_{\boldsymbol{p}}^{(t)}$ is the estimate from the tth subsample with subsampling probability $\boldsymbol{p}$ and $\hat{\boldsymbol{\beta}}_{\text{QLE}}$ is the quasi-likelihood estimator calculated from the whole dataset. We set $T = 1000$ throughout this section.

5.1. Simulation Studies

We take Poisson regression as an example to evaluate the finite sample performance of the proposed methods throughout this section. We also considered logistic regression and Gamma regression models, the results were similar and thus were omitted. Full data of size $N = 500{,}000$ are generated from a Poisson regression model such that given the covariate $\boldsymbol{x}$, the response y follows a Poisson distribution with mean $E(y|\boldsymbol{x}) = \exp(\boldsymbol{\beta}^T \boldsymbol{x})$. Here we set the true value of $\boldsymbol{\beta}$ as a 7×1 vector of 0.5. We consider the following four scenarios to generate the covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{i7})^T$.

Case 1. The seven covariates are iid from the standard uniform distribution, namely, $x_{ij} \overset{\text{iid}}{\sim} U(0, 1)$ for $j = 1, \ldots, 7$.

Case 2. The second covariate is $x_{i2} = x_{i1} + \varepsilon_i$ with $x_{i1} \sim U(0, 1)$, $\varepsilon_i \overset{\text{iid}}{\sim} U(0, 1)$, and other covariates are $x_{ij} \overset{\text{iid}}{\sim} U(0, 1)$ for $j = 1, 3, \ldots, 7$. In this scenario, the first two covariates are correlated (≈ 0.5).

Case 3. This scenario is the same as Case 2 except that $\varepsilon_i \overset{\text{iid}}{\sim} U([0, 0.1])$. For this case, the correlation between the first two covariates is close to 0.8.

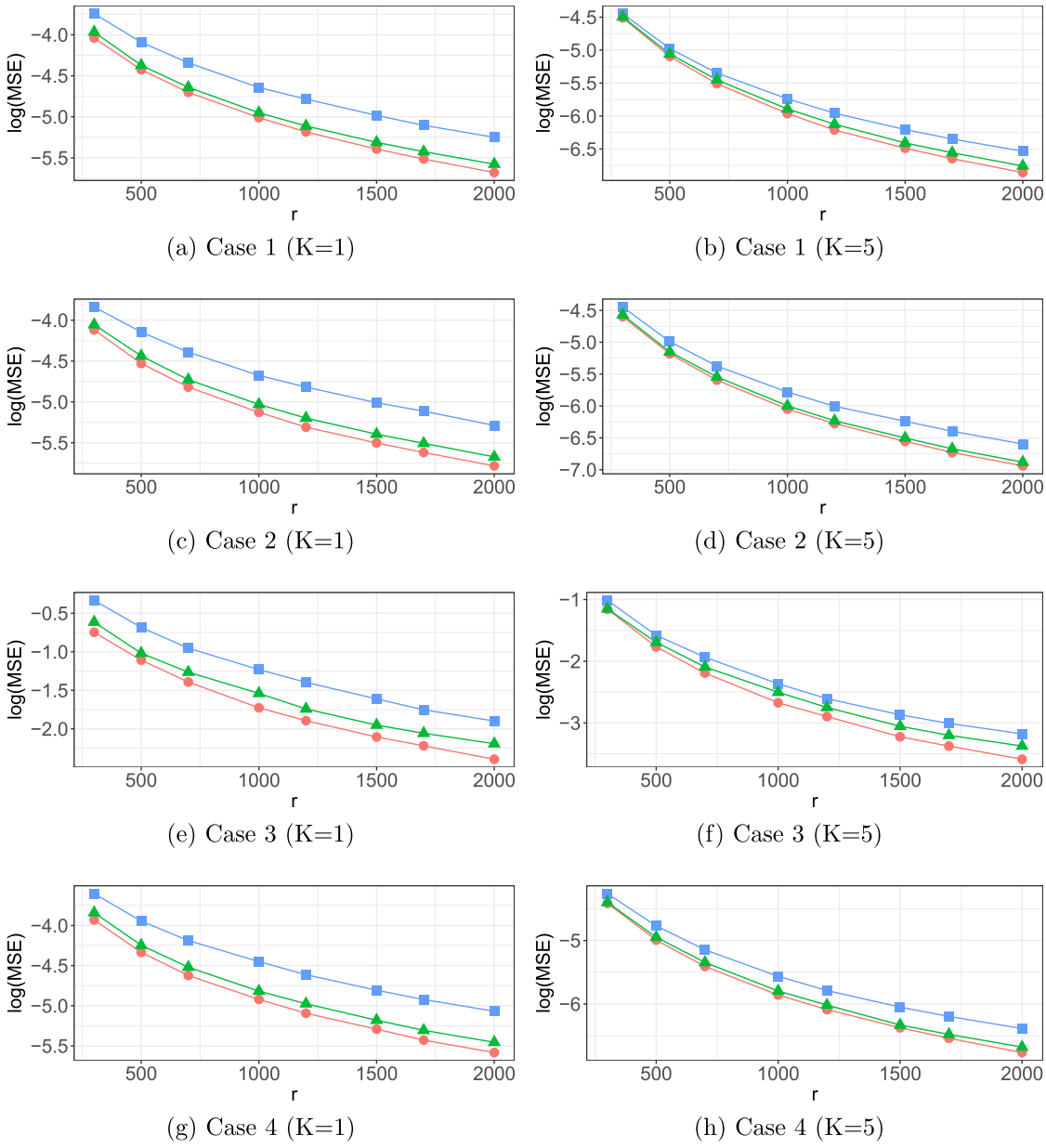


Figure 1. A graph showing the log of MSE with different r and K for different distributions of covariates based on MV (red circle), MVc (green triangle), and uniform subsampling (blue square) methods where $r_0 = 200$ and $\varrho = 0.2$.

Case 4. This scenario is the same as Case 2 except that $x_{ij} \overset{\text{iid}}{\sim} U([-1, 1])$ for $j = 6, 7$. For this case, the supports for different covariates are not all the same.

In the following, we evaluate the performance of Algorithm 3 based on MV and MVc subsampling probabilities with partition number $K = 1$ and $K = 5$. Note that Algorithm 3 with $K = 1$ and Algorithm 2 differ only in the way to incorporate pilot sample information, so their performances are similar. Results of uniform subsampling are also calculated for comparisons.

We fix $r_0 = 200$ and $\varrho = 0.2$, and choose r to be 300, 500, 700, 1000, 1200, 1500, 1700, and 2000. Since the uniform subsampling probability does not depend on unknown parameters and no pilot subsamples are required, it is implemented with subsample size $r + r_0$ for fair comparisons.

Figure 1 gives the simulation results. It is seen that for the four datasets, subsampling methods based on MV and MVc always result in smaller empirical MSEs compared with the uniform subsampling, which agrees with the theoretical results in Section 3. The MSEs for all subsampling methods decrease as r increases, which confirms the theoretical result on consistency of the subsampling methods.

Next, we will explore the effect of different ϱ with fixed r_0 and r. The results are given in Figure 2 with $r_0 = 200$, and r = 1200 and 1500. It is clear to see that the subsampling method outperforms the uniform subsampling method when $\varrho \in [0.01, 0.99]$. When ϱ is close to 1, the performances of $\tilde{\boldsymbol{p}}^{\text{sos}}$ are similar to that of the uniform subsampling. The two-step approach works the best when ϱ is around 0.25. This implies that the shrinkage estimator effectively protect the weighted estimating equation from data points with $|y_i - \psi(\tilde{\boldsymbol{\beta}}_0^T \boldsymbol{x}_i)|$ close

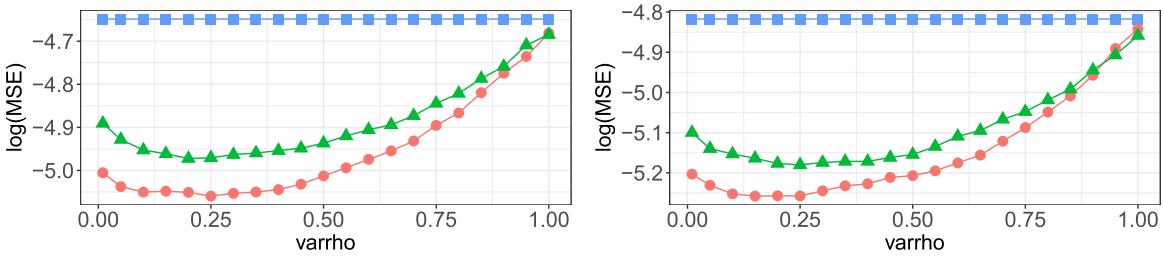


Figure 2. Log MSEs for Case 4 with different ϱ and a fixed $r_0 = 200$ based on MV (red circle), MVc (green triangle), and uniform subsampling (blue square) methods.

Table 1. MSE for different expected size r under varying subsampling strategy with $r_0 = 2000$ and $\varrho = 0.2$ on Case 4.

Method	r/N =0.01	r/N =0.1	r/N =0.3	r/N =0.5	r/N =0.7
UNIF	1.75E−03	1.93E−04	5.03E−05	2.14E−05	9.21E−06
MV with $M = \infty$	1.18E−03	1.12E−04	2.35E−05	8.35E−06	3.64E−06
MV with $M = Q$	1.19E−03	1.11E−04	2.54E−05	8.19E−06	1.57E−06
MV with $M = E$	1.21E−03	1.16E−04	2.29E−05	7.55E−06	2.36E−06
MVc with $M = \infty$	1.32E−03	1.22E−04	2.82E−05	1.09E−05	5.26E−06
MVc with $M = Q$	1.28E−03	1.22E−04	2.72E−05	9.65E−06	2.12E−06
MVc with $M = E$	1.35E−03	1.26E−04	2.74E−05	8.76E−06	2.67E−06

NOTE: Here Q is the $(1 - r/(2n))$th quantile of $\{\hbar_i^{*\mathrm{MV}}\}_{i=1}^{r_0}$ or $\{\hbar_i^{*\mathrm{MVc}}\}_{i=1}^{r_0}$, and E is calculated according to the formula for M in Theorem 3 or 4 with $\hat{\boldsymbol{\beta}}_{\mathrm{QLE}}$ replaced by $\tilde{\boldsymbol{\beta}}_0$.

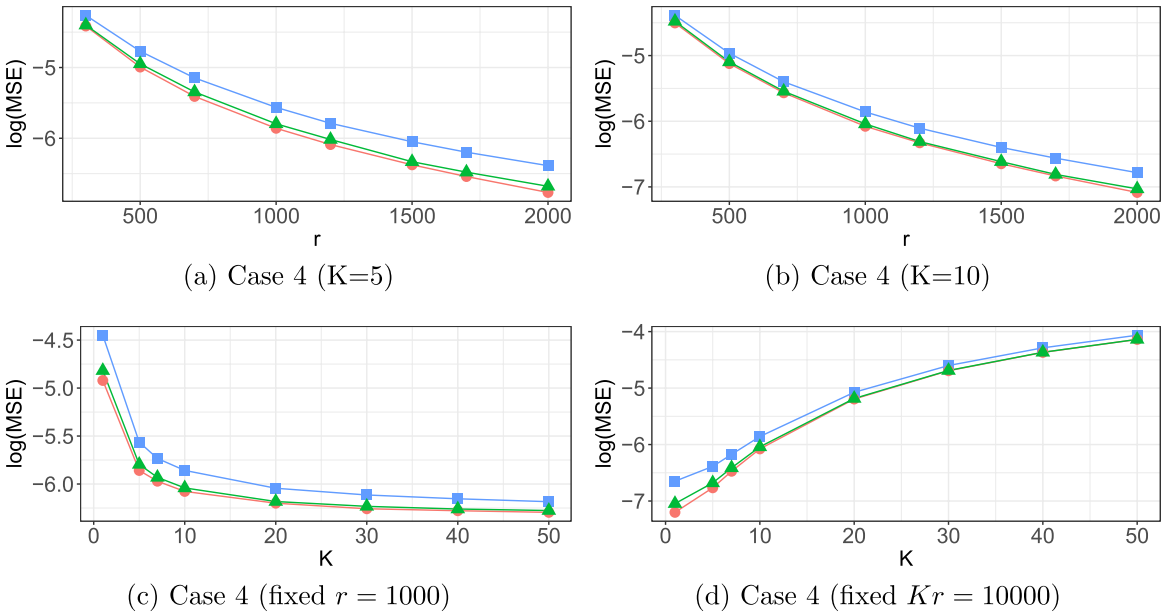


Figure 3. Log MSEs for different combination of r and K with $r_0 = 200$ and $\varrho = 0.2$ based on MV (red circle), MVc (green triangle), and uniform subsampling (blue square) methods.

to zero. We only present the performance of Case 4 here because results for all other cases are similar.

To see the effects of M in $\boldsymbol{p}^{\mathrm{sos}}$, we compare the choice of $M = \infty$ with another two choices: (1) M is approximated by the $(1 - r/(2n))$th quantile of $\{\hbar_i^{*\mathrm{MV}}\}_{i=1}^{r_0}$ or $\{\hbar_i^{*\mathrm{MVc}}\}_{i=1}^{r_0}$ calculated from pilot subsample set (denote this choice as $M = Q$), and (2) M is calculated according to the formulas in Theorem 3 or 4 except that $\hat{\boldsymbol{\beta}}_{\mathrm{QLE}}$ is replaced by $\tilde{\boldsymbol{\beta}}_0$ (denote this choice as $M = E$). we consider different values of r/N with choices of 0.01, 0.1, 0.3, 0.5, and 0.7, and report results in Table 1. When $r/N \leq 0.3$, the choice $M = \infty$ has comparable results as the choice $M = E$ (calculating M from the full). When $r/N \geq 0.5$, the choice $M = Q$ (using a quantile from the pilot subsample) still produce satisfactory results. Thus, the MSE is not very sensitive to the choice of M. In the big data subsampling scheme, since it is typical that $r \ll N$, we can simply use $M = \infty$.

To have a closer look at the effect of K, we implement Algorithm 3 with fixed partition number $K = 5$ or $K = 10$ and changing r with choices of 300, 500, 700, 1000, 1200, 1500, 1700, and 2000. We also consider the cases where r and Kr are fixed.The results for Case 4 are reported in Figure 3 with $r_0 = 200$ and $\varrho = 0.2$. For comparisons, the uniform subsampling is also implemented through Algorithm 3 with $\tilde{\boldsymbol{p}}^{\mathrm{sos}}$ replaced by $\boldsymbol{p}^{\mathrm{UNIF}}$. Figure 3 shows that the subsampling method outperforms the uniform subsampling method for both $K = 5$ and $K = 10$. If r is fixed, the aggregate estimator approximates $\hat{\boldsymbol{\beta}}_{\mathrm{QLE}}$ better when K is larger since more data are involved in each subsample set. However, when Kr is fixed,

Table 2. Empirical coverage probabilities and average lengths of 95% confidence intervals for β_2 with $r_0 = 200$ and $\varrho = 0.2$.

			MV		MVc		UNIF	
		r	Coverage	Length	Coverage	Length	Coverage	Length
Case 1	$k = 1$	1000	0.950	0.1867	0.949	0.1880	0.944	0.1932
		1500	0.949	0.1829	0.946	0.1837	0.945	0.1871
	$k = 5$	1000	0.947	0.1774	0.944	0.1776	0.931	0.1783
		1500	0.951	0.1766	0.946	0.1767	0.935	0.1771
Case 2	$k = 1$	1000	0.935	0.1619	0.944	0.1638	0.927	0.1680
		1500	0.940	0.1587	0.937	0.1600	0.934	0.1627
	$k = 5$	1000	0.945	0.1542	0.934	0.1546	0.936	0.1550
		1500	0.938	0.1536	0.938	0.1538	0.932	0.1539
Case 3	$k = 1$	1000	0.957	1.8103	0.956	1.8504	0.947	1.8961
		1500	0.954	1.7788	0.956	1.8058	0.941	1.8368
	$k = 5$	1000	0.956	1.7344	0.951	1.7418	0.936	1.7503
		1500	0.951	1.7280	0.951	1.7325	0.945	1.7374
Case 4	$k = 1$	1000	0.935	0.1949	0.928	0.1977	0.935	0.2034
		1500	0.927	0.1913	0.933	0.1932	0.936	0.1970
	$k = 5$	1000	0.928	0.1862	0.928	0.1865	0.949	0.1877
		1500	0.930	0.1854	0.927	0.1856	0.946	0.1864

as K increases, the performance of the aggregate estimator deteriorates.

Now we evaluate the performance of the proposed subsampling method for statistical inference under different values of r and K. As an example, we take β_2 as the parameter of interest and construct 95% confidence intervals for it. The estimator given in (25) is used to estimate the variance-covariance matrices based on selected subsamples. Table 2 reports empirical coverage probabilities and average lengths over the four synthetic datasets with $r_0 = 200$ and $\varrho = 0.2$. It is clear that MV and MVc based subsampling methods have similar performances and they are uniformly better than the uniform subsampling method. As r or K increases, lengths of confidence intervals decrease. The 95% confidence intervals in Case 3 are longer than those in other cases with the same subsample sizes. This coincides with the aforementioned results.

Additional simulation results on both estimation efficiency and computational efficiency with larger full data sizes and higher dimensions are available in the supplementary materials.

5.2. Citation Number Dataset

The number of citations is an important factor about the quality of a research paper, and it is of interest to most of the researchers in every field. As a result, study of paper citations itself has become an interesting research topic. In this example, we applied the proposed method to a real dataset about over four million papers associated with abstract, authors, year, venue, title, type and citation numbers (Tang et al. 2008). The dataset is available at *https://www.aminer.cn/citation*, and our goal is to model the number of citations using features extracted from the text information about the articles.

The original dataset is in text format, and we extract the following numerical features to characterize each article. First, the number of years between the year the paper was published and the year of 2018 (x_1). This feature describes the time effect since the citation numbers are nondecreasing in x_1. We categorize the length of the abstract of each paper into detail/brief/non-present status, and bring two indicator variables to denote them. Specifically, $x_2 = 1$ if the paper has an abstract with more than 100 words and $x_2 = 0$ otherwise; and $x_3 = 1$ if the paper has an abstract with less than 100 words and $x_3 = 0$ otherwise. Similarly, we characterize the length of the title for each paper by defining $x_4 = 1$ if the title contains more than 10 words and $x_4 = 0$ otherwise. We also consider the publication type, and use $x_5 = 1$ to denote journal papers and $x_5 = 0$ for the rest of papers. To measure the influence of the journal or publisher, we use the newest SJR score (x_6) provided by *https://www.scimagojr.com*. We also consider the SJR ranking and let $x_7 = 1$ for journals or publishers that are marked with "Q1" and let $x_7 = 0$ otherwise. The author information of each paper is also taken into account. We define x_8 as the average number of author publications for each paper, which is calculated by dividing the total number of publications from the author(s) of the paper before 2018 by the total number of author(s) in the paper. We remove all the incomplete cases in the dataset, and there are $n = 2{,}803{,}027$ data points after the data cleaning.

To describe the relationship between the number of citations and the aforementioned features, a Poisson regression is used. The estimated mean model from the quasi-likelihood estimator based on the full dataset is given as below:

$$E(Y|X) = \exp(1.32 + 0.39x_1 + 1.44x_2 + 1.09x_3 - 0.26x_4 + 0.03x_5 + 0.20x_6 + 0.55x_7 + 0.21x_8).$$

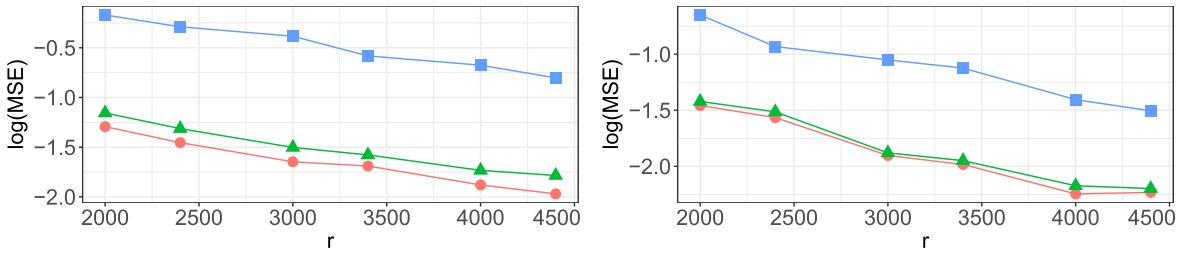
From the fitted model, we have the following findings. (1) A detailed abstract helps to attract more citations while a detailed title may not be popular among scholars. This may be because follow-up papers often have longer titles compared with the original paper, but they usually gain less attentions. (2) The publication type is not critical to receive high number of citations comparing with other factors. (3) SJR ranking is critical to receive higher number of citations. This is because the paper published in high quality publishers usually receive more attentions. (4) More productive authors gain more citations since they may have more influences.

To assess the performance of the proposed method in approximating the full data estimates, we apply them on the citation data for 1000 times and report the averages of parameter

Table 3. Average estimates for the citation number dataset from the proposed methods with $\varrho = 0.2$, $r_0 = 800$, and $r = 4400$.

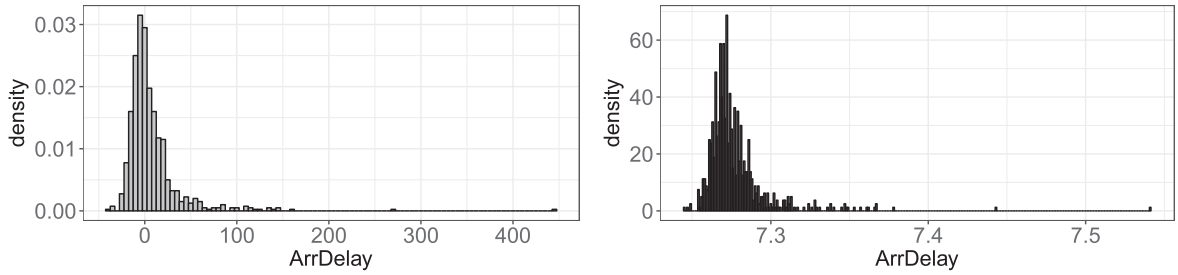
	$K = 1$			$K = 5$		
	UNIF	MV	MVc	UNIF	MV	MVc
β_0	1.29 (0.339)	1.35 (0.183)	1.35 (0.213)	1.47 (0.221)	1.44 (0.153)	1.44 (0.159)
β_1	0.41 (0.054)	0.39 (0.027)	0.39 (0.023)	0.38 (0.039)	0.38 (0.027)	0.38 (0.026)
β_2	1.46 (0.333)	1.42 (0.182)	1.42 (0.208)	1.37 (0.218)	1.40 (0.145)	1.40 (0.149)
β_3	1.11 (0.365)	1.07 (0.190)	1.07 (0.218)	1.02 (0.240)	1.05 (0.160)	1.05 (0.162)
β_4	−0.25 (0.139)	−0.26 (0.100)	−0.26 (0.083)	−0.25 (0.094)	−0.25 (0.071)	−0.25 (0.069)
β_5	0.03 (0.159)	0.03 (0.103)	0.03 (0.099)	0.04 (0.103)	0.04 (0.073)	0.04 (0.073)
β_6	0.21 (0.051)	0.21 (0.022)	0.20 (0.017)	0.21 (0.035)	0.21 (0.020)	0.21 (0.019)
β_7	0.54 (0.185)	0.55 (0.114)	0.55 (0.111)	0.54 (0.117)	0.55 (0.093)	0.55 (0.094)
β_8	0.22 (0.044)	0.21 (0.023)	0.21 (0.017)	0.21 (0.027)	0.21 (0.017)	0.21 (0.016)

NOTE: In the table $\beta_1, \ldots, \beta_8$ are the regression coefficients for $x_1, \ldots, x_8$, respectively, and β_0 is the intercept coefficient. The numbers in the parentheses are the empirical standard errors.



(a) Log MSEs (K=1) (b) Log MSEs (K=5)

Figure 4. A graph showing the log of MSEs for the citation number dataset with $r_0 = 400$ and different r and partition number K based on MV (red circle), MVc (green triangle), and uniform subsampling (blue square) methods.



(a) Actual Delays (b) Log Actual Delays

Figure 5. Distribution of actual delays and log-transformed actual delays based on the pilot samples ($r_0 = 800$).

estimates along with the empirical standard errors in Table 3. The uniform subsampling method is also implemented for comparison. In this table, $\varrho = 0.2$, $r_0 = 800$, and $r = 4400$. It is seen that all subsampling methods produce average estimates that are close to the full data estimates. However, the proposed methods have significantly smaller empirical standard errors.

Similar to the simulation studies, we also compare our methods with the uniform subsampling method with various sampling budget r varying from 2000 to 4400 and r_0 being fixed at 800. Figure 4 shows the results on the empirical MSE. We see that MV and MVc perform similarly and they both dominate the uniform sampling method. This pattern is similar to that in the simulation studies.

5.3. Airline On-Time and Delay Dataset

To track the on-time performance of domestic flights operated by large air carriers, information about on-time, delayed, canceled, and diverted flights have been collected since October 1987. The full dataset contains 123,534,969 records (~11 GB) which is available on *http://stat-computing.org/dataexpo/2009/the-data.html*. One purpose for analyzing this dataset is to build a model for airlines delays. We first plot the histogram of actual arrive delays based on the pilot sample and notice a very large discrepancy from normality. The distribution of actual delays are extremely skewed and heavy-tailed (see Figure 5(a)).

To extract useful information about arrive delays, we use linear regression, log-linear regression, and Gamma regression to model the relationship between arrive delays and other covariate variables: x_1, the distance between airports; x_2, day/night status (binary; 1 if departure between 7 a.m. and 6 p.m., 0 otherwise); x_3, weekend/weekday status (binary; 1 if departure occurred during the weekend, 0 otherwise); and x_4, departure delay status (binary; 1 if the delay is 15 min or more , 0 otherwise). Note that both log-linear and Gamma regression models are defined for nonnegative responses. Thus, we switch the locations of all the responses, that is, add 1440 to all the responses. Based on the pilot sample, the Bayesian information criterion values are 7312.672, −4407.750, and −4415.854 for linear regression, log-linear regression, and Gamma regression, respectively, which

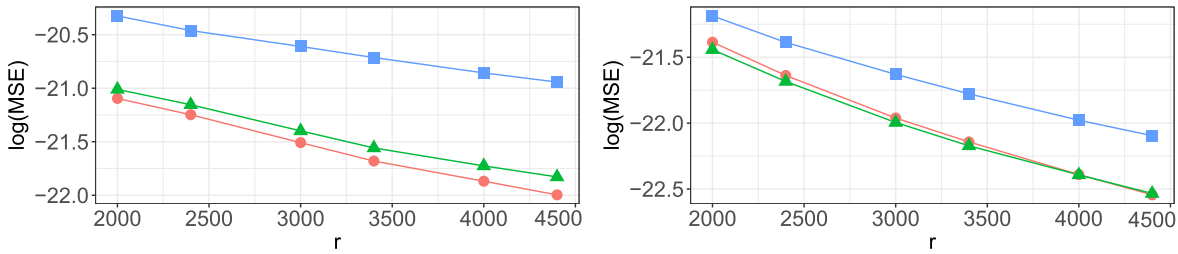


Figure 6. A graph showing the log of MSEs for the airline on-time and delay dataset with $r_0 = 800$, $\varrho = 0.2$ and different r and partition number K based on MV (red circle), MVc (green triangle), and uniform subsampling (blue square) methods.

implies that the posterior probability for Gamma regression model is around 0.98 in the view of Bayesian model averaging (see Neath and Cavanaugh 2012). Thus, we use Gamma regression for this case. In addition, we drop the NA values in the dataset. After data cleaning, we have $n = 119{,}793{,}199$ data points. Similar to the simulation studies, we also compare our method with the uniform subsampling method, and report the results under various sampling budget r varying from 2000 to 4400 with r_0 fixed at 800 in Figure 6. As expected, MV and MVc perform similarly and they both outperform the uniform sampling method.

6. Conclusion

In this article, we have derived the optimal Poisson subsampling probabilities for quasi-likelihood estimation, and developed a distributed optimal subsampling method. We have investigated the theoretical properties of the proposed methods and carried out extensive numerical experiments on simulated and real datasets to evaluate their practical performance. Both theoretical results and numerical results demonstrate the great potential of the proposed method in extracting useful information from massive datasets.

Supplementary Materials

The online supplementary materials present the detailed proofs of all results in the main text and additional simulations.

Acknowledgments

The authors are grateful to the co-editor, the associate editor, and three referees for their valuable comments and suggestions.

Funding

Yu's work was partially supported by Beijing Institute of Technology Research Fund Program for Young Scholars. Wang's work was partially supported by NSF grant 1812013. Ai's work was partially supported by NSFC grants 11671019 and LMEQF.

References

Berger, Y. G., and De La Riva Torres, O. (2016), "Empirical Likelihood Confidence Intervals for Complex Sampling Designs," *Journal of the Royal Statistical Society*, Series B, 78, 319–314. [267]

Breidt, F. J., and Opsomer, J. D. (2000), "Local Polynomial Regression Estimators in Survey Sampling," *The Annals of Statistics*, 28, 1026–1053. [267]

Chen, K., Hu, I., and Ying, Z. (1999), "Strong Consistency of Maximum Quasi-Likelihood Estimators in Generalized Linear Models With Fixed and Adaptive Designs," *The Annals of Statistics*, 27, 1155–1163. [266,267]

Chen, X. (2011), *Quasi Likelihood Method for Generalized Linear Model* (in Chinese), Hefei: Press of University of Science and Technology of China. [266,267]

Dhillon, P. S., Lu, Y., Foster, D., and Ungar, L. (2013), "New Subsampling Algorithms for Fast Least Squares Regression," in *International Conference on Neural Information Processing Systems*, pp. 360–368. [265]

Drineas, P., Mahoney, M. W., Muthukrishnan, S., and Sarlós, T. (2011), "Faster Least Squares Approximation," *Numerische Mathematik*, 117, 219–249. [265]

Duchi, J. C., Agarwal, A., and Wainwright, M. J. (2012), "Dual Averaging for Distributed Optimization: Convergence Analysis and Network Scaling," *IEEE Transactions on Automatic Control*, 57, 592–606. [265]

Fahrmeir, L., and Tutz, G. (2001), *Multivariate Statistical Modelling Based on Generalized Linear Models*, New York: Springer-Verlag. [266]

Jordan, M. I., Lee, J. D., and Yang, Y. (2019), "Communication-Efficient Distributed Statistical Inference," *Journal of the American Statistical Association*, 114, 668–681. [265]

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2015), "A Scalable Bootstrap for Massive Data," *Journal of the Royal Statistical Society*, Series B, 76, 795–816. [265]

Li, R., Lin, D. K., and Li, B. (2013), "Statistical Inference in Massive Data Sets," *Applied Stochastic Models in Business and Industry*, 29, 399–409. [265]

Lin, N., and Xi, R. (2011), "Aggregated Estimating Equation Estimation," *Statistics & Its Interface*, 1, 73–83. [265]

Ma, P., Mahoney, M. W., and Yu, B. (2015), "A Statistical Perspective on Algorithmic Leveraging," *Journal of Machine Learning Research*, 16, 861–919. [265,268]

Mahoney, M. W. (2012), "Randomized Algorithms for Matrices and Data," *Foundations and Trends® in Machine Learning*, 3, 647–672. [265]

Mccullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, Monographs on Statistics and Applied Probability (Vol. 37), London: Chapman & Hall. [266]

Neath, A. A., and Cavanaugh, J. E. (2012), "The Bayesian Information Criterion: Background, Derivation, and Applications," *Wiley Interdisciplinary Reviews: Computational Statistics*, 4, 199–203. [275]

Newey, W. K., and McFadden, D. (1994), "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics* (Vol. 4), eds. R. F. Engle and D. L. McFadden, Amsterdam: Elsevier, pp. 2111–2245. [267]

Pukelsheim, F. (2006), *Optimal Design of Experiments*, Philadelphia, PA: Society for Industrial and Applied Mathematics. [267,268]

Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2019), "Speeding Up MCMC by Efficient Data Subsampling," *Journal of the American Statistical Association*, 114, 831–843. [265]

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing, available at *https://www.R-project.org/*. [270]

Rao, C. R., Toutenburg, H., Shalabh and Heumann, C. (2007), *Linear Models and Generalizations: Least Squares and Alternatives* (3rd ed.), Berlin, Heidelberg: Springer Publishing Company, Inc. [267]

Särndal, C. E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer. [266]

Schifano, E. D., Wu, J., Wang, C., Yan, J., and Chen, M.-H. (2016), "Online Updating of Statistical Inference in the Big Data Setting," *Technometrics*, 58, 393–403. [265]

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. (2008), "Arnetminer: Extraction and Mining of Academic Social Networks," in *KDD'08*, pp. 990–998. [273]

Tzavelas, G. (1998), "A Note on the Uniqueness of the Quasi-Likelihood Estimator," *Statistics & Probability Letters*, 38, 125–130. [267]

van der Vaart, A. (1998), *Asymptotic Statistics*, New York: Cambridge University Press. [267]

Wang, H. Y., Yang, M., and Stufken, J. (2019), "Information-Based Optimal Subdata Selection for Big Data Linear Regression," *Journal of the American Statistical Association*, 114, 393–405. [265]

Wang, H. Y., Zhu, R., and Ma, P. (2018), "Optimal Subsampling for Large Sample Logistic Regression," *Journal of the American Statistical Association*, 113, 829–844. [265,267,268]