Statistica Sinica Preprint No: SS-2022-0277

Title	Optimal Subsampling for Multinomial Logistic Models With Big Data
Manuscript ID	SS-2022-0277
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0277
Complete List of Authors	Zhiqiang Ye, Jun Yu and Mingyao Ai
Corresponding Authors	Mingyao Ai
E-mails	myai@math.pku.edu.cn
Notice: Accepted version subject to English editing.	

OPTIMAL SUBSAMPLING FOR MULTINOMIAL LOGISTIC MODELS WITH BIG DATA

Zhiqiang Ye[1], Jun Yu[2], Mingyao Ai[1]

LMAM, School of Mathematical Sciences and Center for Statistical Science, Peking University[1]

School of Mathematics and Statistics, Beijing Institute of Technology[2]

Abstract: To model categorical responses, multinomial logistic regressions with different links and parameter restrictions have widely been adopted based on the relationships among different categories. In this paper, a unified Poisson subsampling method is proposed to approximate efficiently the maximum likelihood estimator for regression parameters when big data are encountered. The asymptotic normality of the estimator generated from the Poisson subsample is established. Based on the derived asymptotic variance, optimal subsampling probabilities are given according to the A-optimality criterion. To mitigate the burden on the calculation of optimal subsampling probabilities, a random projection based procedure is applied. For practical implementation, some robustness issues including model misspecification and full data with possible outliers are further discussed with theoretical backups. The advantages of the proposed methods are illustrated through numerical studies on both simulated and real datasets.

Key words and phrases: Categorical data; Johnson-Lindenstrauss transform; Poisson subsampling; Randomized Hadamard transform.

1. Introduction

Extremely large datasets are ubiquitous due to the rapid development of science and technologies. Volume is one of the key concepts associated with big data. Specifically, the quantity of generated and stored data in a big data era is usually larger than terabytes and petabytes. Therefore, it is a common challenge on extracting useful information from massive datasets with limited computing resources.

Many statistical methods, which focus on drawing an inference based on a big data set with a fixed computational budget, have been developed up to now. Subsampling is one of the most popular techniques of achieving a good balance between computational complexity and statistical efficiency. Extensive researches show the great success of subsampling in dealing with massive data in various fields. For example, uniform subsampling was used in Drineas et al. (2011) to approximate ordinary least square estimators in linear regressions. To further improve statistical accuracy, some non-uniform subsampling strategies such as leverage score subsampling (Ma et al., 2015, 2020), volume subsampling (Dereziński et al., 2018), and information-based optimal subdata selection (Wang et al., 2019) are proposed to address this issue without increasing too much computational costs compared with uniform subsampling. To accommodate the variety types of responses, the local case-control subsampling (Fithian et al., 2014) and D-optimal based subdata selection (Cheng et al., 2020) are proposed for binary classifications. The subsampling method motivated by A-optimality criterion and its variants are developed for generalized linear models, quantile regressions, and additive hazards models. Important works include but not limited to Wang et al. (2018), Yu et al. (2022), Ai et al. (2021), Wang and Ma (2021), Zuo et al. (2021). A literature review can be found in Yu et al. (2023).

Subsample based classifiers gain a lot of attention from data scientists since classification is one of the most important tasks for big data analysis. Learning a classifier based on subsample is usually regarded as an effective way when dealing with massive data. To the best of our knowledge, most of the existing works focus on binary classification problems. Typical examples include but not limited to local case control subsampling (Fithian et al., 2014) and optimal subsampling motivated by A-optimality criterion (Wang et al., 2018; Wang, 2019) for logistic regressions.

Contrarily, systematic approaches for modeling multi-class categorical responses based on subsampling techniques are still elusive due to the complexity brought by the different kinds of order relations. To address the order relation, a multinomial distribution with specific link functions is widely adopted in practice. Examples include baseline-category logit models for nominal responses, cumulative logit models and adjacent-categories logit models for ordinal responses, continuation-ratio logit models for hierarchical responses. See Chapter 6 of Agresti (2019) for a comprehensive discussion. Moreover, different parameter restrictions are added to capture the relationships between the responses and the explanatory variables. For example, McCullagh (1980) proposed a model in which all the parameters are the same across different categories, except for the intercept. The underlying logic is that all responses can be regarded as a partition of an underlying continuous variable. For the case where there is no reference to an underlying continuous variable, it is reasonable to allow parameters to change across categories. Peterson and Harrell (1990) suggested applying the partial proportional odds in which a subset of explanatory variables are assumed common in all categories, and some special explanatory variables are used in certain categories only. This is quite different from the logistic regression for binary responses and the softmax regression for multi-class responses.

To the best of our knowledge, few works systematically studied the multinomial logistic models in the presence of different links and complex restrictions on the parameters. The most relevant work is given by Yao and Wang (2019), which only considered the baseline-category link with no common parameters among different categories. The optimal subsampling approaches for multinomial logistic models with different links and the partial proportional odds are studied here to fill this gap. Our main contributions to the subsampling technique are three folds. Firstly, we establish a unified optimal subsampling procedure for all four aforementioned links under the partial proportional odds assumption, which clearly covers the method in Yao and Wang (2019) as a special case. Secondly, to further accelerate the subsampling algorithm, the Johnson-Lindenstrauss transform and the subsampled randomized Hadamard transform (Drineas et al., 2012) are successively adopted to approximate the optimal subsampling probabilities. The proposed algorithm runs in $O(Nd\log N)$ time, as opposed to the $O(Nd^2)$ time required by the direct calculation, where N is the size of the data, and d is the number of parameters. An error bound of such an approximation is given accordingly. Thirdly, we also carefully evaluate the performance under the scenario where the specified model is incorrect and there are some outliers in the full data. Both analytical and numerical studies show that the subsample based estimator still converges to the maximum likelihood estimator of the full data under the postulated model. A stratification subsampling procedure is proposed to mitigate the influence brought by the outliers.

The rest of the paper is organized as follows. In Section 2, we briefly introduce multinomial logistic models and a general Poisson subsampling framework. The asymptotic normality of the estimator based on the subsample is derived. Section 3 presents optimal subsampling probabilities according to the A-optimality criterion. Randomized approximation algorithms are applied to further reduce the time in calculating subsampling probabilities. In Section 4, some practical issues to implement the optimal subsampling procedures and theoretical justifications are considered. Section 5 further discusses some robustness issues including model misspecification and possible outliers in massive data sets. Simulation studies and real data analyses are provided in Section 6. Section 7 concludes this paper. All the proofs and additional simulation results are delegated to the Supplementary Material.

2. Preliminaries

2.1 Multinomial Logistic Models

Let $\mathcal{F}_N = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ be a sequence of independent and identically distributed random variables with response $y_i \in \{1, \ldots, J\}(J \geq 2)$. The response y_i can be regarded as a random variable coming from a multinomial distribution with

$$f(y_i|\boldsymbol{x}_i) = \prod_{j=1}^{J} \pi_{ij}^{\mathbb{I}(y_i=j)}(\boldsymbol{\beta}),$$

where $\pi_{ij}(\boldsymbol{\beta})$ is the probability that the response is $j(j = 1, \ldots, J)$ under the covariate $\boldsymbol{x}_i$, $\boldsymbol{\beta}$ is the unknown parameter vector, and $\mathbb{I}(\cdot)$ is an indicator function.

In order to characterize the relationships among the response probabilities, the multinomial logistic models with baseline-category, cumulative, adjacent-categories, and continuation-ratio link under the partial proportional odds assumption are briefly introduced as follows:

$$\log\left(\frac{\pi_{i1}(\boldsymbol{\beta})}{\pi_{iJ}(\boldsymbol{\beta})}\right) = \boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j, \quad \text{baseline-category,} \tag{2.1}$$

$$\log\left(\frac{\pi_{i1}(\boldsymbol{\beta}) + \cdots + \pi_{ij}(\boldsymbol{\beta})}{\pi_{i,j+1}(\boldsymbol{\beta}) + \cdots + \pi_{iJ}(\boldsymbol{\beta})}\right) = \boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j, \quad \text{cumulative,} \tag{2.2}$$

$$\log\left(\frac{\pi_{ij}(\boldsymbol{\beta})}{\pi_{i,j+1}(\boldsymbol{\beta})}\right) = \boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j, \quad \text{adjacent-categories,} \tag{2.3}$$

$$\log\left(\frac{\pi_{ij}(\boldsymbol{\beta})}{\pi_{i,j+1}(\boldsymbol{\beta}) + \cdots + \pi_{iJ}(\boldsymbol{\beta})}\right) = \boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j, \quad \text{continuation-ratio,} \tag{2.4}$$

for $i = 1, \ldots, N, j = 1, \ldots, J-1$, where $\boldsymbol{x}_{i(0)} \in \mathbb{R}^{d_0}$ stands for the common predictors among all categories, and $\boldsymbol{x}_{i(j)} \in \mathbb{R}^{d_j}(j = 1, \ldots, J-1)$ stands for the individual predictors belonging to the j-th category only. The corresponding $\boldsymbol{\beta}_j \in \mathbb{R}^{d_j}$ are parameters of interest. For notation simplicity, denote $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^T, \ldots, \boldsymbol{\beta}_{J-1}^T)^T \in \mathbb{R}^d$ with $d = d_0 + \cdots + d_{J-1}$. It is worth mentioning that both $\boldsymbol{x}_{i(0)}$ and $\boldsymbol{x}_{i(j)}$ are contained in the explanatory variable $\boldsymbol{x}_i$, and $\boldsymbol{x}_{i(j)}$ may contain none or some variables in $\boldsymbol{x}_{i(0)}$.

Clearly, models with proportional odds and non-proportional odds assumptions are two special cases with $x_{i(j)} = 1$, $j = 1, \ldots, J-1$, and $x_{i(0)} = 0$, respectively. When $x_{i(0)} = 0$ and $\boldsymbol{x}_{i(1)} = \cdots = \boldsymbol{x}_{i(J-1)} = \boldsymbol{x}_i$, Model (2.1) turns to be the softmax regression whose subsampling strategies are considered in Yao and Wang (2019) and Han et al. (2020). More specifically, for the case $J = 2$, Models (2.1)-(2.4) become the well-known logistic regression, and the subsampling methods are studied in Wang et al. (2018).

The unknown parameter vector $\boldsymbol{\beta}$ can be estimated via the maximum likelihood method. The resultant estimator based on the full data, denoted by $\hat{\boldsymbol{\beta}}_{full}$, is the maximizer of the following log-likelihood function

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{j=1}^{J} \mathbb{I}(y_i = j) \log \pi_{ij}(\boldsymbol{\beta}). \tag{2.5}$$

2.2 General Poisson Subsampling Algorithm

Clearly, finding the maximizer of (2.5) usually meets the computational bottleneck when N is big. Subsampling is commonly adopted as a feasible solution, which will be introduced in the following.

Let p_i be the inclusion probability of the i-th data point for $i = 1, \ldots, N$. Note that $n = p_1 + \cdots + p_N$ is the expected subsample size. Denote S to be a set consisting of subsample points and their corresponding inclusion probabilities. The 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 $R_i \sim \text{Bernoulli}(p_i)$;
    if $R_i = 1$ then
        Update $S = S \cup \{(\boldsymbol{x}_i, y_i, p_i)\}$;
Estimation: Obtain $\hat{\boldsymbol{\beta}}_{sub}$ by maximizing the following weighted
likelihood function based on the subsample $S$,
    $\ell^*(\boldsymbol{\beta}) = \sum_{S} \frac{1}{p_i} \left( \sum_{j=1}^{J} \mathbb{I}(y_i = j) \log \pi_{ij}(\boldsymbol{\beta}) \right).$    (2.6)
```

The weighted scheme in (2.6) is typically essential to avoid the potential bias when estimating the log likelihood function on the full data. The reason for using the Poisson subsampling is to relax the memory constraint for massive data. Compared with the subsampling with replacement which requires storing all the subsampling probabilities in the memory, the Poisson subsampling can extract the subsample points by scanning the full data one by one. Although the selected subsample size n^* fluctuates in Algorithm 1, one can show that n^* keeps concentrating near its expectation n with high probability (Ai et al., 2021). Therefore, the computational cost is still under control. When dealing with big data, it is common that $n \ll N$. In such a situation, the computational cost on the subsample is far less than on the full data.

In order to establish the asymptotic results of subsampling estimators, the following regularity assumptions are required.

Assumption 1. The parameter vector $\boldsymbol{\beta}$ lies in a compact parameter space.

Assumption 2. The categorical probabilities $\pi_{ij}(\boldsymbol{\beta}) > 0$, for $i = 1, \ldots, N, j = 1, \ldots, J$. In addition, if Model (2.2) is adopted, we further assume that $\boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j - \boldsymbol{x}_{i(j-1)}^T\boldsymbol{\beta}_{j-1} > c_0$ holds for $i = 1, \ldots, N, j = 2, \ldots, J-1$, where $c_0 > 0$ is some constant.

Assumption 3. As $N \to \infty$, $-N^{-1}\partial^2\ell(\hat{\boldsymbol{\beta}}_{full})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T$ goes to a positive definite matrix in probability.

Assumption 4. The covariate has the finite fourth-order moments, i.e., $E\|\boldsymbol{x}_1\|^4 < \infty$.

Assumption 5. $\max_{1\le i\le N}(Np_i)^{-1} = O_P(n^{-1})$.

Assumption 1 is used to ensure the consistency of the full sample based estimator. Similar assumptions can be found in Newey and McFadden (1994). Assumption 2 is required to guarantee the logarithm is well defined. Note that the left-hand side of Model (2.2) is monotonous with respect to j. The condition $\boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j - \boldsymbol{x}_{i(j-1)}^T\boldsymbol{\beta}_{j-1} > c_0$ simply ensures to keep such monotone, and $c_0 > 0$ is to ensure that any two categories can be separated. Assumption 3 essentially requires that the observed information matrix is asymptotically non-singular, which indicates the maximum likelihood estimator is unique. Assumption 4 is a moment requirement. Similar assumptions are used in Wang et al. (2018), and Yao and Wang (2019). Assumption 5 restricts the weights in the subsample log-likelihood function, which can protect the estimation equation (2.6) from being dominated by data points with extremely small subsampling probabilities.

To establish the asymptotic normality of $\hat{\boldsymbol{\beta}}_{sub}$, we start with introducing some necessary notations. Let

$$M_N(\boldsymbol{\beta}) = \frac{1}{N}\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}, \quad V_{Nc}(\boldsymbol{\beta}) = \frac{1}{N^2}\sum_{i=1}^{N}\frac{1-p_i}{p_i}\boldsymbol{u}_i(\boldsymbol{\beta})\boldsymbol{u}_i^T(\boldsymbol{\beta}), \tag{2.7}$$

where $\boldsymbol{u}_i(\boldsymbol{\beta}) = \left(\partial \log \boldsymbol{\pi}_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T\right)^T \boldsymbol{\delta}_i$ with $\boldsymbol{\pi}_i(\boldsymbol{\beta}) = (\pi_{i1}(\boldsymbol{\beta}), \ldots, \pi_{iJ}(\boldsymbol{\beta}))^T$, $\boldsymbol{\delta}_i = (\mathbb{I}(y_i = 1), \ldots, \mathbb{I}(y_i = J))^T$.

Theorem 1. *Suppose Assumptions 1-5 hold. As $N \to \infty$ and $n \to \infty$, conditional on $\mathcal{F}_N$ in probability,*

$$V_N^{-1/2}\left(\hat{\boldsymbol{\beta}}_{sub} - \hat{\boldsymbol{\beta}}_{full}\right) \to N(\mathbf{0}, I_d),$$

in distribution, where $V_N = M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})V_{Nc}(\hat{\boldsymbol{\beta}}_{full})M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})$.

Remark 1. Note that $V_{Nc}(\hat{\boldsymbol{\beta}}_{full})$ can be decomposed into the following two parts,

$$V_{Nc}(\hat{\boldsymbol{\beta}}_{full}) = \frac{1}{N^2}\sum_{i=1}^{N}\frac{1}{p_i}\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i^T(\hat{\boldsymbol{\beta}}_{full}) - \frac{1}{N^2}\sum_{i=1}^{N}\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i^T(\hat{\boldsymbol{\beta}}_{full}).$$

When $n/N > 0$, it is easy to see that Poisson sampling leads to a smaller variance compared with the sampling with replacement for the softmax regression in Yao and Wang (2019). This is because Poisson sampling is a kind of sampling without replacement. It will be more efficient for the case with $n/N > 0$. When $n \ll N$, which is common in big data settings, the second part is a small order term compared with the first part. Thus the variance is mainly determined by subsampling strategies. Therefore, in this work, we focus on finding the optimal subsampling probabilities to mitigate the impact of the subsampling.

3. Optimal Poisson Subsampling and its Approximation

This section is devoted to minimizing the asymptotic mean squared error (AMSE) of $\hat{\boldsymbol{\beta}}_{sub}$ in approximating $\hat{\boldsymbol{\beta}}_{full}$, which corresponds to the A-optimality in the language of optimal designs (Pukelsheim, 2003).

From Theorem 1, the subsample based estimator is asymptotically unbiased. As a result, minimizing AMSE is equivalent to minimizing its asymptotic variance. For clarity, we use “MV” to denote the subsampling strategy such that the asymptotic **v**ariance is **m**inimized.

Theorem 2. *Define $\hbar_i^{MV} = \|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|, i = 1, \ldots, N$, and let $\hbar_{(1)}^{MV} \leq \hbar_{(2)}^{MV} \leq \cdots \leq \hbar_{(N)}^{MV}$ denote the order statistics of $\{\hbar_i^{MV}\}_{i=1}^N$. Assume that $\hbar_{(N-n)}^{MV} > 0$. The AMSE of $\hat{\boldsymbol{\beta}}_{sub}$, $tr(V)$, attains its minimum, if p_i's are chosen to be*

$$p_i^{MV} = n\frac{\hbar_i^{MV} \wedge M}{\sum_{j=1}^N \left(\hbar_j^{MV} \wedge M\right)}, \tag{3.8}$$

where $a \wedge b = \min(a, b)$, $M = (n-k)^{-1}\sum_{i=1}^{N-k} \hbar_{(i)}^{MV}$, with

$$k = \min\left\{ s \,\middle|\, 0 \leq s \leq n, (n-s)\hbar_{(N-s)}^{MV} < \sum_{i=1}^{N-s} \hbar_{(i)}^{MV} \right\}.$$

Remark 2. One can decompose $\hbar_i^{MV}$ into two parts according to the expression of $\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})$. More precisely, $\hbar_i^{MV} = \|M_N^{-1}\frac{\partial \boldsymbol{\pi}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} D\|$ with D being a diagonal matrix whose diagonal entries are $\mathbb{I}(y_i = 1)/\pi_{i,1}(\hat{\boldsymbol{\beta}}_{full}), \ldots, \mathbb{I}(y_i = J)/\pi_{i,J}(\hat{\boldsymbol{\beta}}_{full})$. Clearly, the $\mathbb{I}(y_i = j)/\pi_{i,j}(\hat{\boldsymbol{\beta}}_{full})$ describes the concordance between observation and prediction. If a point is easy to be correctly predicted, it has less chance to be included in the subdata set. Note that the $M_N^{-1}\partial \boldsymbol{\pi}_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^T$ only depends on the information of covariates and it is used to construct optimal designs for the multinomial logistic regressions. See Bu et al. (2020); Ai et al. (2023) as examples. From the explicit forms provided in the Supplementary Material, one can expect that a point close to the origin is unlikely to be sampled since it contains little information on the slope parameters (Wang et al., 2019; Yu et al., 2023). To further take a close look at the proposed subsampling probabilities, we simplify the sampling probability under a logistic regression when $n/N \to 0$. Simple calculation yields that $p_i^{MV} \propto |y_i - \pi_{i,1}(\hat{\boldsymbol{\beta}}_{full})|\|M_N^{-1}\boldsymbol{x}_i\|$. The first term $|y_i - \pi_{i,1}(\hat{\boldsymbol{\beta}}_{full})|$ is the same as the local case-control subsampling (Fithian et al., 2014), which is useful in dealing with imbalanced data.

Remark 3. The M is the minimum number such that $nN^{-1}(\hbar_i^{MV} \wedge M) \leq N^{-1}\sum_{i=1}^{N}(\hbar_i^{MV} \wedge M)$. The exact calculation of M requires additional $O(n+r\log r)$ times (Yu et al., 2022). Fortunately, when $n/N \to 0$, one can expect $0 < N^{-1}\sum_{i=1}^{N}(\hbar_i^{MV})$ almost surely, so M can be dropped, i.e., $M = \infty$. In such cases, it is equivalent to selecting $p_i^{MV} = n\hbar_i^{MV}/\sum_{j=1}^{N}\hbar_j^{MV}$.

Calculating $M_N(\hat{\boldsymbol{\beta}}_{full})$ and $\hbar_i^{MV}$ is not easy. More precisely, $O(Nd^2)$ time is required to obtain $M_N(\hat{\boldsymbol{\beta}}_{full})$. Additional $O(Nd^2)$ time is needed in deriving $\hbar_i^{MV}$, for $i = 1, \ldots, N$, based on $M_N(\hat{\boldsymbol{\beta}}_{full})$, which is calculated in the previous step. To further accelerate the algorithm, $\hbar_i^{MV}$ is usually replaced by $\|\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|$ (Wang et al., 2018). Consequently, $O(Nd)$ time is enough to obtain the subsampling probabilities. However, such subsampling probabilities usually do not lead to the smallest AMSE. In fact, it minimizes the AMSE of $M_N(\hat{\boldsymbol{\beta}}_{full})\hat{\boldsymbol{\beta}}_{sub}$, which is not of interest compared with minimizing the AMSE of $\hat{\boldsymbol{\beta}}_{sub}$ in practice.

To inherit the optimality of p_i^{MV}, an efficient algorithm for approximating $\hbar_i^{MV}$ is developed to reduce the computing time in the rest of this section.

Firstly, we focus on the fast approximation of the matrix $M_N(\hat{\boldsymbol{\beta}}_{full})$. Recall that $M_N(\hat{\boldsymbol{\beta}}_{full}) = N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{J}\mathbb{I}(y_i = j)\partial^2\log\pi_{ij}(\hat{\boldsymbol{\beta}}_{full})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T$. We begin with dealing with the $d \times d$ matrix $N^{-1}\partial^2\log\pi_{ij}(\hat{\boldsymbol{\beta}}_{full})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T$. Simple calculation yields

$$\frac{1}{N}\frac{\partial^2 \log \pi_{ij}(\hat{\boldsymbol{\beta}}_{full})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = \begin{pmatrix} c_{00}^{(i,j)}\boldsymbol{x}_{i(0)}\boldsymbol{x}_{i(0)}^T & \cdots & c_{0,J-1}^{(i,j)}\boldsymbol{x}_{i(0)}\boldsymbol{x}_{i(J-1)}^T \\ c_{10}^{(i,j)}\boldsymbol{x}_{i(1)}\boldsymbol{x}_{i(0)}^T & \cdots & c_{1,J-1}^{(i,j)}\boldsymbol{x}_{i(1)}\boldsymbol{x}_{i(J-1)}^T \\ \vdots & & \vdots \\ c_{J-1,0}^{(i,j)}\boldsymbol{x}_{i(J-1)}\boldsymbol{x}_{i(0)}^T & \cdots & c_{J-1,J-1}^{(i,j)}\boldsymbol{x}_{i(J-1)}\boldsymbol{x}_{i(J-1)}^T \end{pmatrix}, \quad (3.9)$$

for $i = 1, \dots, N, j = 1, \dots, J$. The derivation of Equation (3.9) and the details of calculating $\{c_{kl}^{(i,j)}\}_{0\le k,l\le J-1}$ are deferred to the Supplementary Material for the sake of brevity. Since $c_{kl}^{(i,j)}\boldsymbol{x}_{i(k)}\boldsymbol{x}_{i(l)}^T$'s have a similar structure for all k, l, without loss of generality, we choose the upper left corner block of $M_N(\hat{\boldsymbol{\beta}}_{full})$ as an example to illustrate the approximation method. Let $X_0 = (\boldsymbol{x}_{1(0)}, \dots, \boldsymbol{x}_{N(0)})^T$ and $C_{00} = \operatorname{diag}\left(c_{00}^{(1,y_1)}, \dots, c_{00}^{(N,y_N)}\right)$. Then $\sum_{i=1}^N c_{00}^{(i,y_i)}\boldsymbol{x}_{i(0)}\boldsymbol{x}_{i(0)}^T$ can be written as $X_0^T C_{00} X_0$ in the matrix form. Consequently, it is natural to use a Fast Johnson-Lindenstrauss Transform (FJLT) (Ailon and Chazelle, 2006) to approximate $X_0^T C_{00} X_0$.

One can use a subsampled randomized Hadamard transform (Drineas et al., 2012) to construct an FJLT with high probability. For simplicity, we assume that N is a power of two, then the approximation of $X_0^T C_{00} X_0$ can be obtained through the following steps.

(i) Construct an $N \times N$ matrix of the Hadamard transform recursively. Let $H_1 = (1)$ be a 1×1 matrix, and

$$H_{k+1} = \begin{pmatrix} H_k & H_k \\ H_k & -H_k \end{pmatrix}, k = 1, \dots, \log_2(N/2),$$

then $H_{\log_2 N}$ is the $N \times N$ matrix of the Hadamard transform.

(ii) Let $D \in \mathbb{R}^{N\times N}$ be a random diagonal matrix with independent diagonal entries $D_{ii} = 1$ with probability $1/2$ and $D_{ii} = -1$ with probability $1/2$, $i = 1, \ldots, N$.

(iii) Let $S = (\boldsymbol{e}_{j_1}, \ldots, \boldsymbol{e}_{j_{r_1}})$, where $\boldsymbol{e}_j \in \mathbb{R}^N$ is a standard unit vector with j-th element being one, and $j_1, \ldots, j_{r_1}$ are randomly sampled from $\{1, \ldots, N\}$.

(iv) Then $X_0^T C_{00} X_0$ is approximated by $(T_1X_0)^T T_1(C_{00}X_0)$ with $T_1 = S^T HD/\sqrt{N}$.

Remark 4. For the FJLT $T_1 \in \mathbb{R}^{r_1\times N}$ and an N dimensional vector $\boldsymbol{z}$, the time complexity for performing $T_1\boldsymbol{z}$ is $O(N \log r_1)$ (Drineas et al., 2012). Note that C_{00} is a diagonal matrix. It is clear to see that only $O(Nd_0 \log r_1)$ time is needed to calculate T_1X_0 and $T_1(C_{00}X_0)$. As suggested in Lemma 6 of Drineas et al. (2012), $r_1 = O(d_0 \log N \log(d_0 \log N))$. Clearly, N is much larger than r_1 and d_0. Thus calculating $(T_1X_0)^T T_1(C_{00}X_0)$ only takes $O(Nd_0 \log r_1)$ time, which is much less than the time required to calculate $X_0^T C_{00}X_0$. For the rest of $M_N(\hat{\boldsymbol{\beta}}_{full})$, the same method can be used. Thus we only need $O(NdJ \log r_1)$ time to approximate $M_N(\hat{\boldsymbol{\beta}}_{full})$. The approximation is denoted as $\widetilde{M_N}(\hat{\boldsymbol{\beta}}_{full})$ for notation simplicity.

Remark 5. Since the Hadamard transform only exists for the case that N is a power of two, the aforementioned algorithm only works for some specific sample size N. To relax the constraint, a naive method is to randomly drop out some data points until the data size is a power of two. After this operation, the difference between $\hat{\boldsymbol{\beta}}_{full}$ and the MLE based on the drop out data $\hat{\boldsymbol{\beta}}_{drop}$ is $O_P(N^{-1/2})$ since both two estimator are $\sqrt{N}$ consistent estimators of the true parameters. The difference is much smaller than the difference between $\hat{\boldsymbol{\beta}}_{sub}$ and $\hat{\boldsymbol{\beta}}_{full}$. Thus we can ignore the impact of this operation.

Another possible solution is to use a two-stage sampling procedure. To be precise, one can apply a simple random sampling to downsize the full sample size at first and then use the optimal subsampling procedure on the subsample points obtained at the first stage. Mathematically, let T_2 be a $r_2 \times N$ random matrix whose rows are chosen randomly with replacement from the rows of $\sqrt{N/r_2}I_N$, where I_N is an $N \times N$ identity matrix. Then $M_N(\hat{\boldsymbol{\beta}}_{full})$ can be approximated by the same technique of $\widetilde{M_N}(\hat{\boldsymbol{\beta}}_{full})$ except T_1 is replaced by T_2. For notation simplicity, the resultant approximation is denoted as $\widehat{M_N}(\hat{\boldsymbol{\beta}}_{full})$. Clearly, the computational cost is $O(r_2d^2)$.

Note that we still need $O(Nd^2)$ time to calculate $\{\|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|\}_{i=1}^N$ even $M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})$ is replaced by $\widetilde{M}_N^{-1}(\hat{\boldsymbol{\beta}}_{full})$ or $\widehat{M}_N^{-1}(\hat{\boldsymbol{\beta}}_{full})$. Now, we focus on accelerating the calculation on $\{\|\widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|\}_{i=1}^N$. We omit the case for $\widetilde{M_N}(\hat{\boldsymbol{\beta}}_{full})$ due to its similarity.

To fast approximate $\{\|\widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|\}_{i=1}^N$ without losing too much accuracy, we employ Johnson-Lindenstrauss Transform (JLT) (Achlioptas, 2003). To be precise, $\|\widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|$ is approximated by $\|T_3\widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|$, where $T_3 \in \mathbb{R}^{r_3 \times d}$ is a JLT, whose (i, j)-th entry, $T_{3,ij}$, is chosen independently from the following distribution:

$$T_{3,ij} = \begin{cases} \sqrt{3/r_3} & \text{with probability } 1/6, \\ 0 & \text{with probability } 2/3, \\ -\sqrt{3/r_3} & \text{with probability } 1/6, \end{cases}$$

with $r_3 < d$. One can see that only $O(Nr_3d)$ time is required in calculating $\{\|T_3\widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|\}_{i=1}^N$. In practice, r_3 has the same order as $\log N$ (Achlioptas, 2003, Theorem 1.1). When d is not too small, the approximation will faster than calculating $\{\|\widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|\}_{i=1}^N$ directly.

The following theorem states the approximation accuracy of the subsampling probability.

Theorem 3. *Assume* $\nu_1, \nu_2 \in (0, 1/3)$, $r_3 \geq 48\log N - 24\log \nu_1$, *then with probability at least* $(1-\nu_1)(1-\nu_2)$, *conditional on* $\mathcal{F}_N$, *we have*

$$\begin{aligned} &\left| \|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\| - \|T_3\widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\| \right| \\ \leq & \frac{\|\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|}{\gamma\lambda_{min}^2(M_N(\hat{\boldsymbol{\beta}}_{full}))}\sqrt{\frac{c_1}{r_2\nu_2}} + \frac{\|\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|}{\gamma\lambda_{min}(M_N(\hat{\boldsymbol{\beta}}_{full}))}\sqrt{\frac{c_2}{r_3}}\mathbb{I}(r_3 < d), \end{aligned}$$

for some constants γ, c_1, c_2, *which depend on* d, ν_1, *and* $N^{-1}\sum_{i=1}^N \|\boldsymbol{x}_i\|^4$ *only, where* $\lambda_{min}(\cdot)$ *denotes the minimal eigenvalue of the corresponding matrix.*

Theorem 3 gives us a guide for the trade-off between the computational time and the approximation accuracy. Clearly, the approximation accuracy will improve as r_2 and r_3 increase. Note that the computational cost will be increased linearly as r_2 and r_3 grow up, while the approximation error decreasing is much slower than $1/r_2$ and $1/r_3$. The property of diminishing marginal utility for increasing r_2, r_3 can be seen in Theorem 3. Therefore, the careful choice of r_2, r_3 can achieve the balance between computational complexity and statistical accuracy. As shown in Section 6, with proper choice of r_2, r_3, such an approximation can save about 60% of the time without sacrificing too much accuracy.

4. Practical Implementation

The optimal subsampling probabilities p_i^{MV}'s derived in Theorem 2 and their approximations cannot be applied directly since the probabilities depend on $\hat{\boldsymbol{\beta}}_{full}$ which is unknown in practice. As suggested in Wang et al. (2018), $\hat{\boldsymbol{\beta}}_{full}$ is usually approximated by a pilot estimator, say $\hat{\boldsymbol{\beta}}_{pilot}$. To obtain $\hat{\boldsymbol{\beta}}_{pilot}$, we can draw a small pilot sample via uniform subsampling or other subsampling approaches satisfying Assumption 5. Therefore, the proposed method is computationally feasible to implement.

When some subsampling probability p_i^{MV} is small, the weighted likelihood function may be sensitive to the data point $(\boldsymbol{x}_i, y_i)$ if it is included in the subsample. Ma et al. (2015) proposed a shrinkage-based subsampling method to make the estimators more stable and robust. To be specific, we use the following subsampling probabilities,

$$\breve{p}_i = (1-\rho)\frac{n\left\|M_N^{-1}(\hat{\boldsymbol{\beta}}_{pilot})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{pilot})\right\|}{\sum_{i=1}^N\left\|M_N^{-1}(\hat{\boldsymbol{\beta}}_{pilot})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{pilot})\right\|} + \rho\frac{n}{N}, \tag{4.10}$$

where $\rho \in (0,1)$. Clearly, the $\breve{p}_i$ is a linear combination of the optimal probability and the uniform subsampling, thereby it is natural to obtain the benefits of each. Precisely, involving uniform subsampling can detect the departure of model misspecification in the design region with low subsampling probabilities. Moreover, with a fraction of data points with low probabilities, the resultant estimator will be much more stable and not particularly susceptible to outliers.

Clearly, the approximation techniques introduced in Section 3 are still valid with replacing the $\hat{\boldsymbol{\beta}}_{full}$ with a pilot estimator. Therefore, we can use $T_3\widetilde{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{pilot})$ or $T_3\widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{pilot})$ to instead $M_N^{-1}(\hat{\boldsymbol{\beta}}_{pilot})$. From Theorem 3, one can expect that such approximations speed up the computation remarkably without scarifies too much accuracy. As stated in Remark 3, we simply fetch $M = +\infty$ to accommodate big data environments. To prevent the case that some $\breve{p}_i$'s are larger than one, we use $\breve{p}_i \wedge 1$'s in practice, where $a \wedge b$ stands for the minimal number between a and b. For clear transparency, we summarize the two-step algorithm in Algorithm 2.

The asymptotic normality of $\hat{\boldsymbol{\beta}}_{ts}$ obtained by Algorithm 2 is derived as follows.

Theorem 4. *Suppose Assumptions 1-4 hold and $n_0 n^{-1/2} \to 0$. As $N \to \infty$, $n \to \infty$, and $n_0 \to \infty$, conditional on $\mathcal{F}_N$ in probability,*

$$\grave{V}^{-1/2}\left(\hat{\boldsymbol{\beta}}_{ts} - \hat{\boldsymbol{\beta}}_{full}\right) \to N(\mathbf{0}, I_d),$$

```
Algorithm 2: Practical Two-Step Algorithm
    Pilot Subsampling: Run Algorithm 1 with $p_i = n_0/N, i = 1, \ldots, N$, or other subsampling approaches satisfying Assumption 5 to obtain the pilot subsample set $S_1$ and the pilot estimator $\hat{\boldsymbol{\beta}}_{pilot}$, where $n_0$ is the expected size of the pilot subsample.
    Initialization: $S_2 = \varnothing$;
    for $i = 1, \ldots, N$ do
        Generate $R_i \sim \text{Bernoulli}(1, p_i)$ with $p_i = \breve{p}_i \wedge 1$, where $\breve{p}_i$ is defined in (4.10);
        if $R_i = 1$ then
            Update $S_2 = S_2 \cup \{(\boldsymbol{x}_i, y_i, p_i)\}$
    Estimation: Find $\hat{\boldsymbol{\beta}}_{ts}$ to maximize the following weighted likelihood function based on the subsample $S_1$ and $S_2$, $n_0\ell_1^*(\boldsymbol{\beta}) + n\ell_2^*(\boldsymbol{\beta})$, where $\ell_1^*(\boldsymbol{\beta})$, $\ell_2^*(\boldsymbol{\beta})$ are defined in (2.6) with $p_i = n_0/N$, $p_i = \breve{p}_i \wedge 1$, respectively.
```

in distribution, where $\grave{V} = M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\grave{V}_{Nc}(\hat{\boldsymbol{\beta}}_{full})M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})$*, and*

$$\grave{V}_{Nc}(\hat{\boldsymbol{\beta}}_{full}) = \frac{1}{N^2}\sum_{i=1}^{N}\frac{1-(\grave{p}_i \wedge 1)}{\grave{p}_i \wedge 1}\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i^T(\hat{\boldsymbol{\beta}}_{full}),$$

with

$$\grave{p}_i = (1-\rho)\frac{n\left\|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\right\|}{\sum_{i=1}^{N}\left\|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\right\|} + \rho\frac{n}{N}.$$

Note that when $n_0 n^{-1/2} \to 0$, the contribution of the subsample at the first step can be ignored, the moment estimators of $M_N(\hat{\boldsymbol{\beta}}_{full})$ and $\grave{V}_{Nc}(\hat{\boldsymbol{\beta}}_{full})$ can be simply estimated by

$$\frac{1}{N}\sum_{S_2}\frac{1}{\breve{p}_i}\frac{\partial \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{ts})}{\partial \boldsymbol{\beta}^T}, \quad \text{and} \quad \frac{1}{N^2}\sum_{S_2}\frac{1}{\breve{p}_i}\left(\frac{1-\breve{p}_i}{\breve{p}_i}\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{ts})\boldsymbol{u}_i^T(\hat{\boldsymbol{\beta}}_{ts})\right),$$

respectively. Thus one can estimate $\grave{V}$ by "plug-in" the above moment estimators. Consequently, one can obtain the standard error and construct the confidence ellipsoid for the model parameters by applying Theorem 4.

5. Discussions on Robustness

In practical big data settings, robustness is important due to the following two reasons. Firstly, the postulated parametric model may sometimes be misspecified. Secondly, the full data set may contain some outliers. In this section, we will provide some discussions on robustness.

5.1 Model Robustness

Clearly, when the model is perfectly specified, Theorem 4 concludes that the subsample based estimator $\hat{\boldsymbol{\beta}}_{ts}$ is also a consistent estimator of the true parameter $\boldsymbol{\beta}_{true}$, since $\hat{\boldsymbol{\beta}}_{full}$ is a consistent estimator. The asymptotic result is still valid when $r/n \to 0$. In the following, we will derive the asymptotic behavior of $\hat{\boldsymbol{\beta}}_{ts}$ when the postulated model is misspecified.

Two kinds of model misspecifications oftentimes arise in subsampling procedures. The first is that the postulated model is incorrect when we plan the subsampling, while the model is correctly specified in the final estimation step.

This phenomenon is common when the dependence between input and output is misspecified initially and is corrected as the subdata points began to accumulate. The second is that the postulated model is incorrect from the subsampling step to the estimation step. It is also possible when some important features are missing in the full data.

For the first scenario, the following theorem states that $\hat{\boldsymbol{\beta}}_{ts}$ is still a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}_{true}$.

Theorem 5. *Suppose Assumptions 1-4 hold and the model in the estimation step is correctly specified. As $N \to \infty$ and $n \to \infty$, $\hat{\boldsymbol{\beta}}_{ts}$ obtained by Algorithm 2 is $\sqrt{n}$-consistent to $\boldsymbol{\beta}_{true}$ in probability, no matter the postulated model in designing the pilot subsampling probabilities is correct or not. That is, for any $\varepsilon > 0$, there exists a finite Δ_ε and n_ε such that*

$$P\left(\left\|\hat{\boldsymbol{\beta}}_{ts} - \boldsymbol{\beta}_{true}\right\| \geq n^{-1/2}\Delta_\varepsilon\right) < \varepsilon,$$

for all $n > n_\varepsilon$.

It is worth mentioning that Thoerem 2 no longer holds for this case. It is natural to see that the asymptotic variance is inflated compared with the optimal subsampling probabilities when the model is perfectly specified. This is the price we pay for the model misspecification.

For the second scenario, one cannot expect $\hat{\boldsymbol{\beta}}_{ts}$ to be a consistent estimator of $\boldsymbol{\beta}_{true}$. The following remark states that $\hat{\boldsymbol{\beta}}_{ts}$ is still a $\sqrt{n}$-consistent estimator of $\hat{\boldsymbol{\beta}}_{full}$ conditional on $\mathcal{F}_N$ with minimum asymptotic MSE.

Remark 6. Suppose Assumptions 1-4 hold. As $N \to \infty$ and $n \to \infty$, $\hat{\boldsymbol{\beta}}_{ts}$ obtained by Algorithm 2 is $\sqrt{n}$-consistent to $\hat{\boldsymbol{\beta}}_{full}$ conditional on $\mathcal{F}_N$ in probability.

As Chapter 5 of Van der Vaart (1998) pointed out that the $\hat{\boldsymbol{\beta}}_{full}$ will lead to the a unique distribution $F_n(\cdot, \hat{\boldsymbol{\beta}}_{full})$ in the family of misspecified multinomial distribution with the postulated link functions that has the smallest Kullback-Leibler divergence from the true underlying model. Thus, the fitted model with $\hat{\boldsymbol{\beta}}_{ts}$ can also be regarded as a useful model for prediction even if the postulated model is wrong.

5.2 Subsampling with Possible Outliers

In this subsection, we consider the subsampling when some outliers lie in the full data.

Along the same idea of “cook distance”, the outliers in this section refer to the strong influence points that negatively affect the regression coefficients estimation. First of all, we begin with analyzing how much the parameter estimator changes when the i-th observation is removed.

Lemma 1. *Let $\hat{\boldsymbol{\beta}}_{-i}$ denote the jackknife estimator of $\boldsymbol{\beta}$ with the i-th observation in the full data deleted. Assume that the number of outliers is finite. Under Assumptions 1–4, as $N \to \infty$, it follows that*

$$\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}}_{full} = -N^{-1} M_N^{-1}(\hat{\boldsymbol{\beta}}_{full}) \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full}) + o_P(N^{-1}). \tag{5.11}$$

Lemma 1 gives another view of the MV subsampling strategy. The selected subsample points are more likely to be the data points containing more information about the model and should be used for parameter estimation if all the data follow the same underlying model. However, if the full data contains some outliers, they may unfortunately be drawn into the subsample set.

It is worth mentioning that shrinkage probabilities in (4.10) not only down weight the inclusion probabilities of the outliers but also encourage the subsample to include some data points that have less influence on the estimation. Therefore, it is naturally more robust than we adopt the optimal subsampling probabilities only. When the number of outliers is relatively small, we can still adopt Algorithm 2 to achieve a barely satisfactory result.

When the number of outliers cannot be ignored compared with the sampling budget n, we propose to adopt the stratification subsampling according to $\{\|\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}}_{full}\|\}_{i=1}^{N}$ to make the balance between the information points and outliers. To be precise, define $\mathbb{S}(r)$ be the slab that $\mathbb{S}(r) = \{(y_i, \boldsymbol{x}_i) : \|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\| \le r\}$. The sample space can be divided into K layer $H_1, \ldots, H_K$ with pre-specified $r_1 < \ldots < r_{K-1}$, i.e.,

$$H_1 = \mathbb{S}(r_1), \ldots H_{K-1} = \mathbb{S}(r_{K-1})\backslash\mathbb{S}(r_{K-2}),\ H_K = \mathbb{R}^d\backslash\mathbb{S}(r_{K-1}).$$

Then we can conduct the subsampling on each layer.

Note that $p_i^{MV} \propto \|\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}}_{full}\| = \|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|$ when M in Theorem 2 is selected as ∞. As a result, the additional computational cost in dividing each data point into the different layers is only $O(N)$. Thus the computational cost is still affordable and the approximation methods introduced in Section 3 can also be applied to accelerate the sampling step. It is worth mentioning the pilot subsample contains no outliers with probability $(1 - o/N)^{n_0} \approx 1 - n_0 o/N$, where o is the number of outliers. Since o is finite and $n_0 \ll N$, thus we can still use the pilot estimator to approximate the $\hat{\boldsymbol{\beta}}_{full}$.

Therefore, for each layer, the subsampling step is the same as Algorithm 2 except the p_i's are the inclusion probabilities in each strata, and the estimation step is to find $\hat{\boldsymbol{\beta}}_{ts}$ which maximizes the following weighted likelihood function based on the sampled subdata set, i.e., $n_0 \ell_1^*(\boldsymbol{\beta}) + n_1 \ell_2^*(\boldsymbol{\beta}) + \ldots + n_K \ell_{K+1}^*(\boldsymbol{\beta})$, where $\ell_1^*(\boldsymbol{\beta})$, $\ell_2^*(\boldsymbol{\beta}), \ldots, \ell_{K+1}^*(\boldsymbol{\beta})$ are the subsample based log-likelihood for the pilot subdata and subdata in layers $H_1, \ldots, H_K$, respectively, and $n_1, \ldots, n_K$ are the subsample size of the corresponding layers. Here, the subsample size of each stratum is suggested to be proportional to the size of each layer. The r_K is recommended to satisfy the condition that the cardinal number of H_K is a little bit more than the number of outliers. In practice, we may use six sigma rule to decide r_K or simply use the quantile or the m-th largest values of $\{\|\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}}_{full}\|\}_{i=1}^N$ instead. Such tuning parameters can be regarded as a bet on outliers. Note that the objective function can be rewritten as $n_0/(n_0+n_1+\ldots+n_K)\ell_1^*(\boldsymbol{\beta}) + n_1/(n_0+n_1+\ldots+n_K)\ell_2^*(\boldsymbol{\beta}) + \ldots + n_K/(n_0 + n_1 + \ldots + n_K)\ell_{K+1}^*(\boldsymbol{\beta})$. Thus the contribution of outliers in the selected subsample to the subdata based likelihood is around $n_K/(n_0 + n_1 + \ldots + n_K)$, which will go to zero since the outliers are sufficiently small in the full data. The effect of outliers in the estimation step will naturally be mitigated. Clearly, the subdata in H_K can be ignored when $n_K/(n_0 + n_1 + \ldots + n_K)$ is sufficiently small.

6. Numerical Studies

In this section, we use some numerical examples to evaluate the performance of the methods proposed in Section 4. To evaluate the accuracy of the algorithms in approximating the full data maximum likelihood estimator, for each case, we repeat the implementation for K times and calculate the empirical mean squared error (MSE) of the resultant estimator: $K^{-1}\sum_{k=1}^{K}\|\hat{\boldsymbol{\beta}}_{\boldsymbol{p}}^{(k)} - \hat{\boldsymbol{\beta}}_{full}\|^2$, where $\hat{\boldsymbol{\beta}}_{\boldsymbol{p}}^{(k)}$ is the estimator with subsampling probability $\boldsymbol{p}$ from the k-th subsample. All the computations are performed using R. Throughout this section, we set $K = 1000$.

6.1 Simulations

In the following, we take Model (2.4) with $J = 3$ to illustrate our methods. The performance for the softmax regression is quite similar and we relegate it to the Supplementary Material. For reference, we list the model used in this subsection as follows:

$$\begin{aligned}\log\left(\frac{\pi_{i1}(\boldsymbol{\beta})}{\pi_{i2}(\boldsymbol{\beta}) + \pi_{i3}(\boldsymbol{\beta})}\right) &= \boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(1)}^T\boldsymbol{\beta}_1,\\ \log\left(\frac{\pi_{i2}(\boldsymbol{\beta})}{\pi_{i3}(\boldsymbol{\beta})}\right) &= \boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(2)}^T\boldsymbol{\beta}_2.\end{aligned} \tag{6.12}$$

Here we set $\boldsymbol{\beta}_0 = -0.5 \times \mathbf{1}_{10}$, $\boldsymbol{\beta}_1 = 0.5 \times \mathbf{1}_{10}$, $\boldsymbol{\beta}_2 = \mathbf{1}_{10}$, where $\mathbf{1}_{10}$ is a 10 dimensional all-ones vector. The corresponding covariate $\boldsymbol{x}_i = (\boldsymbol{x}_{i(0)}^T, \boldsymbol{x}_{i(1)}^T, \boldsymbol{x}_{i(2)}^T)^T$

with $N = 2^{16}$ is generated in the following scenarios.

Case 1. The $\boldsymbol{x}$ follows a multivariate normal distribution with mean $\mathbf{0}$, $N(\mathbf{0}_{30}, \Sigma)$, where Σ is a matrix with all diagonal elements equal to one and off-diagonal elements equal to 0.5.

Case 2. The $\boldsymbol{x}$ follows a mixture of two multivariate normal distributions with different means, $0.5N(\mathbf{1}_{30}, \Sigma) + 0.5N(-\mathbf{1}_{30}, \Sigma)$, where Σ is the same as in Case 1.

Case 3. The $\boldsymbol{x}$ follows a multivariate t distribution with degrees of freedom 3 and mean $\mathbf{0}$, $t_3(\mathbf{0}_{30}, \Sigma)/10$, where Σ is the same as in Case 1.

Case 4. The $\boldsymbol{x}$ follows a multivariate log-normal distribution $LN(\mathbf{0}_{30}, \Sigma)/10$, where Σ is the same as in Case 1.

The first three cases are symmetric with different distributions, whereas Case 4 is asymmetric. It is worth mentioning that in Cases 1, 2, and 4, Assumptions 2-4 are satisfied. However, in Case 3, Assumption 4 is not satisfied.

Now we evaluate the performance of the proposed methods, i.e., the optimal subsampling and its approximation, together with the uniform subsampling and the MVc subsampling. The MVc is L-optimal subsampling method which aims to minimize $\mathrm{tr}(V_{Nc}(\hat{\boldsymbol{\beta}}_{full}))$. Thus the resultant sampling probability is proportional to $\|\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|$. For fair comparisons, we assign all the subsampling probabilities to $(n_0 + n)/N$ for the uniform subsampling. The detailed subsampling methods considered in this section are listed in Table 1.

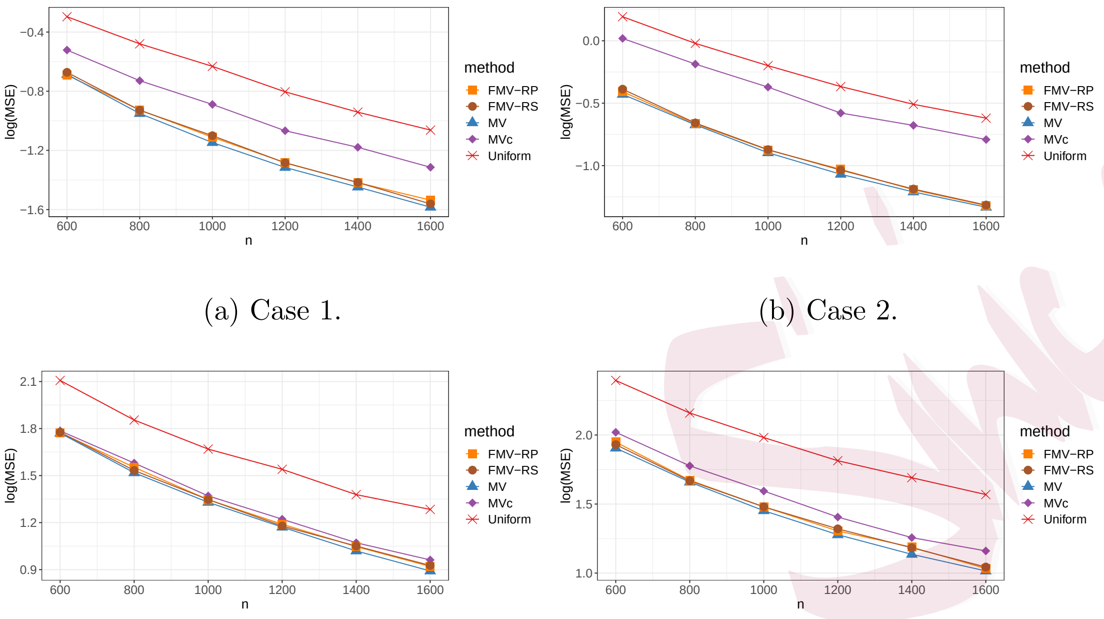
Table 1: Summary of different subsampling methods used in the numerical studies.

Name	Subsampling probabilities (p_i)
MV	$(1-\rho)\frac{n\left\|M_N^{-1}(\hat{\boldsymbol{\beta}}_{pilot})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{pilot})\right\|}{\sum_{i=1}^N\left\|M_N^{-1}(\hat{\boldsymbol{\beta}}_{pilot})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{pilot})\right\|}+\rho\frac{n}{N}$
FMV-RP	$(1-\rho)\frac{n\left\|T_3\widetilde{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{pilot})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{pilot})\right\|}{\sum_{i=1}^N\left\|T_3\widetilde{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{pilot})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{pilot})\right\|}+\rho\frac{n}{N}$
FMV-RS	$(1-\rho)\frac{n\left\|T_3\widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{pilot})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{pilot})\right\|}{\sum_{i=1}^N\left\|T_3\widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{pilot})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{pilot})\right\|}+\rho\frac{n}{N}$
MVc	$(1-\rho)\frac{n\left\|\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{pilot})\right\|}{\sum_{i=1}^N\left\|\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{pilot})\right\|}+\rho\frac{n}{N}$
Uniform	$\frac{n_0+n}{N}$

MV means the optimal subsampling such that the asymptotic variance is minimized. FMV means "Fast MV-optimal subsampling probability approximation", RP stands for "Random Projection", and RS stands for "Random Subsampling". The T_3, $\widetilde{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{pilot})$, $\widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{pilot})$ are defined in Section 3. MVc means the optimal subsampling such that the trace of the $V_{Nc}(\hat{\boldsymbol{\beta}}_{full})$ is minimized. Since $\hat{\boldsymbol{\beta}}_{full}$ is unknown, it is replaced by $\hat{\boldsymbol{\beta}}_{pilot}$ in practice.

Here we set $n_0 = 400$, $\rho = 0.2$, $r_1 = r_2 = 5000$, $r_3 = 10$. To verify the consistency of the subsample based estimator, we choose the expected subsample size n to be 600, 800, 1000, 1200, 1400, and 1600. The simulation results are reported in Figure 1.

From Figure 1, one can see that subsampling methods based on MV, FMV-RP, FMV-RS, MVc always result in smaller empirical MSEs compared with the uniform subsampling. In addition, the MV method has the smallest empirical



(a) Case 1.

(b) Case 2.

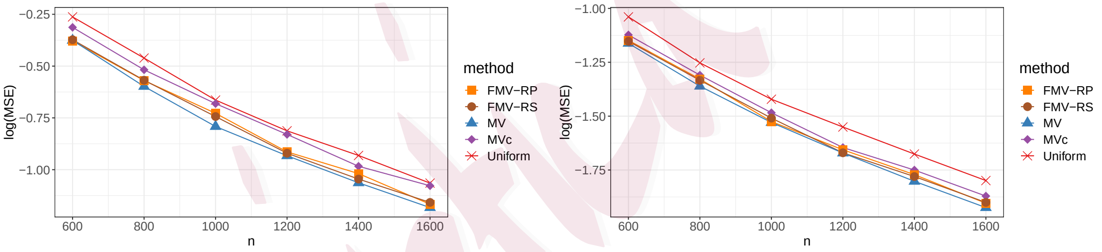
(c) Case 3.

(d) Case 4.

Figure 1: The log of MSE for Model (6.12) with different n based on MV, FMV-RP, FMV-RS, MVc, and Uniform methods, where $n_0 = 400$, $\rho = 0.2$, $r_1 = r_2 = 5000$, $r_3 = 10$.

MSEs and the MSEs for FMV-RP, FMV-RS are close to that for the MV method. Compared with the MVc, the approximation methods (FMV-RP, FMV-RS) have smaller empirical MSEs. This is because the MVc does not focus on minimizing the asymptotic MSE of $\hat{\boldsymbol{\beta}}_{sub}$ which echoes the discussions in Section 3. It is worth mentioning that the MSEs for all subsampling methods decrease as n increases, which confirms the theoretical result in Theorem 1.

We also consider the cases in which the postulated model is misspecified. Two scenarios are taken into account as we discussed in Section 5. In the first scenario, we focus on the case that only the postulated model used in designing the sampling probabilities is incorrect. Here we use Model (6.12) to generate the simulated data while a softmax regression is adopted in planning the sampling probabilities. Secondly, we further consider the case that the data is generated by Model (6.12) while the softmax regression is adopted to draw an inference on the full data. Since all the cases have similar performance, we only demonstrate the performances when the covariates are generated in Case 1. The results are reported in Figures 2 (a) and (b), respectively.



(a) Model misspecification in sampling step only.

(b) Model misspecification in both sampling and estimation step.

Figure 2: The log of MSE for the model misspecification scenarios with different n based on MV, FMV-RP, FMV-RS, MVc, and Uniform methods, where $n_0 = 400$, $\rho = 0.2$, $r_1 = r_2 = 5000$, $r_3 = 10$.

Now we consider the cases in which the full data is corrupted by various forms of outliers. The basic set of the full data that obey the underlying model is generated from Case 1, and two types of outliers are considered as follows.

$\mathcal{O}_1$: The $\boldsymbol{x}$ follows a uniform distribution from -11 to -10 for each dimension independently. The corresponding response is set to be two.

$\mathcal{O}_2$: The $\boldsymbol{x}$ follows a uniform distribution from 10 to 11 for each dimension independently. The corresponding response is set to be three.

The outliers in the full dataset is $\{(\boldsymbol{x}_i, y_i)\}_{i\in\mathcal{O}}$ with $\mathcal{O} = \mathcal{O}_1 \cup \mathcal{O}_2$, such that $|\mathcal{O}| = o$. Clearly, when the underlying model is correctly specified, the probability of the response being one, two, and three is around 0.6, 0, 0.4, respectively when $\boldsymbol{x}$ is generated as $\mathcal{O}_1$. For the case that $\boldsymbol{x}$ is generated as $\mathcal{O}_2$ and the underlying model is correctly specified, the probability for the response being one, two, and three is around 0.6, 0.4, 0, respectively. Thus, both $\mathcal{O}_1$ and $\mathcal{O}_2$ are outliers when the data is generated by Model (6.12).

Now we compare the stratification subsampling method proposed in Section 5 with MV, MVc, and Uniform subsampling methods. Here, the hyper-parameters of the MV, MVc is the same as Cases 1–4. The MV-S denotes the stratification subsampling method proposed in Section 5 with $K = 2$ and r_1 be the 99%-quantile of $\{\|M_N^{-1}(\hat{\boldsymbol{\beta}}_{pilot})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{pilot})\|\}_{i=1}^N$. To further remove the influence of the outliers on the full data MLE, we redefine the empirical mean squared error (MSE) of the resultant estimator as $K^{-1}\sum_{k=1}^K \|\hat{\boldsymbol{\beta}}_{\boldsymbol{p}}^{(k)} - \boldsymbol{\beta}_{true}\|^2$. The results are shown in Table 2.

From Table 2, it is clear that when the subsample size is small, the impact of the outliers are also relatively small since only a little part of the outliers will be

Table 2: Empirical MSE for different subsampling methods and subsample size under various number of the outliers in the full data set.

$\lvert\mathcal{O}\rvert$	method	600	800	1000	1200	1400	1600
65	MV	0.504	0.499	0.528	0.607	0.662	0.718
	MVc	0.540	0.545	0.570	0.622	0.690	0.780
	MV-S	**0.372**	**0.399**	**0.458**	**0.535**	**0.658**	**0.681**
	Uniform	0.590	0.631	0.652	0.719	0.801	0.925
327	MV	2.163	2.151	2.175	2.192	2.229	2.212
	MVc	2.162	2.237	2.209	2.289	2.336	2.303
	MV-S	**0.618**	**0.606**	**0.712**	**0.776**	**0.837**	**1.099**
	Uniform	2.134	2.138	2.140	2.150	2.271	2.215
655	MV	3.348	3.331	3.345	3.358	3.366	3.410
	MVc	3.436	3.450	3.415	3.459	3.434	3.488
	MV-S	**2.117**	**2.010**	**2.091**	**2.200**	**2.341**	**2.378**
	Uniform	3.269	3.351	3.282	3.306	3.426	3.413

included in the subsample set. The stratification method effectively reduces the impact of the outliers which makes the subsample based estimator more robust. As the number of outliers increases, the outliers are naturally more likely to influence both the pilot estimator and the final estimator. As a result, the MSE becomes worse compared with the cases of fewer outliers.

6.2 Real Data Analysis

We examined our methods on a real dataset about the on-time performance of flights operated by large air carriers. This data contains 5,819,079 unique flight information in 2015, which is available at https://www.kaggle.com/usdot/flight-delays.

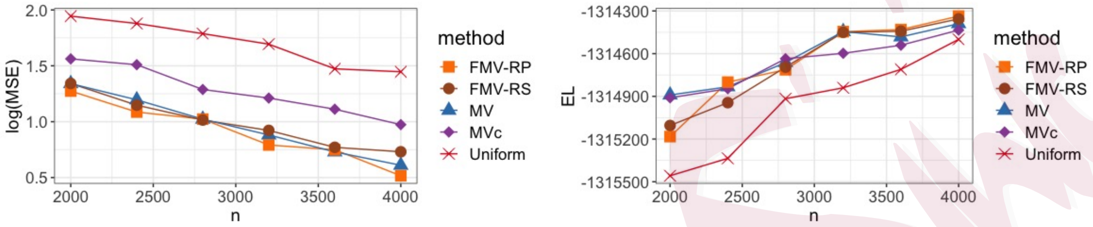
We use multinomial logistic regression to model the probability of late arrival (categorical; 1 for early arrival, 2 for a delay of less than five minutes, 3 for a delay between five and fifteen minutes, and 4 for a delay of more than fifteen minutes) as a function of the weekday/weekend status (x_1, binary; 1 if departure occurred during the weekday, 0 otherwise); day/night status (x_2, binary; 1 if departure occurred between 7 a.m. and 6 p.m., 0 otherwise); departure delay status (x_3, binary; 1 if the delay is five minutes or more, 0 otherwise); the planned time of flight (x_4, continuous); the distance between airports (x_5, continuous). The x_4 and x_5 are on a scale between 0 and 1. We drop the NA values in the dataset. In addition, to use the FMV-RP method, we consider the subsampling on the training dataset of size $N = 2^{22}(= 4,194,304)$, which is chosen randomly from the full data. The rest data are used for testing the prediction performance. The proportions of the four types are 67.11%, 9.42%, 10.85%, 12.62%, respectively.

In this work, we are interested in estimating the probability of further delay when the lower level delay in flight happens. Therefore, the continuation-ratio logit model is employed. Note that the delay levels are a partition of the arrival delay time which is a continuous variable. Thus we adopt the proportional odds assumption and build the following model:

$$\begin{aligned}\text{logit}\,(\pi_{i1}) &= -0.502 - 0.102x_{i1} + 0.039x_{i2} + 1.915x_{i3} + 1.831x_{i4} - 1.396x_{i5},\\ \text{logit}\left(\frac{\pi_{i2}}{\pi_{i2}+\pi_{i3}+\pi_{i4}}\right) &= -1.753 - 0.102x_{i1} + 0.039x_{i2} + 1.915x_{i3} + 1.831x_{i4} - 1.396x_{i5},\\ \text{logit}\left(\frac{\pi_{i3}}{\pi_{i3}+\pi_{i4}}\right) &= -0.651 - 0.102x_{i1} + 0.039x_{i2} + 1.915x_{i3} + 1.831x_{i4} - 1.396x_{i5},\end{aligned}$$

where logit$(\cdot)$ is the logistic function and all the coefficients are estimated from the full sample. To measure the prediction performance, we adopt the expected log-likelihood gain, which is widely used in the statistical literature and many other scientific disciplines. See Ando and chau Li (2017); McCoy et al. (2017) for details. To be precise, the expected log-likelihood gain is calculated by EL $= \sum_{i=1}^{n_{\text{test}}}\sum_{j=1}^{J}\mathbb{I}(y_i = j)\log\pi_{ij}(\hat{\boldsymbol{\beta}}_{\boldsymbol{p}}^{(k)})$, where n_{test} is the size of the test data. Since it is the likelihood of the test data, a larger EL corresponds to a better method.

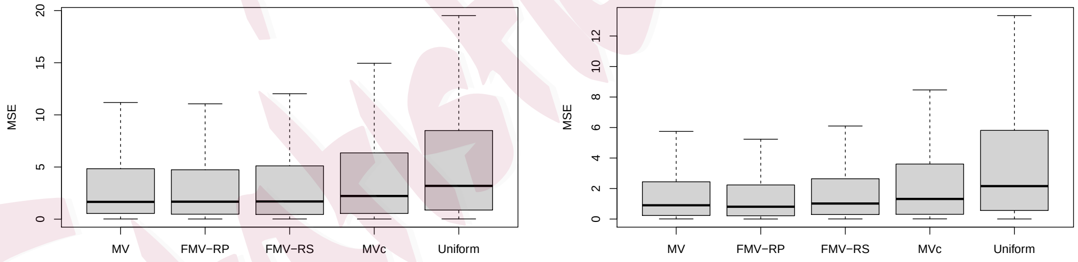
We also compared our methods with the MVc and uniform subsampling methods with various expected subsampling sizes from 2000 to 4000. We set $n_0 = 1000, \rho = 0.2$, $r_1 = r_2 = 20000$, $r_3 = 4$. Figure 3 presents the results. Clearly, MV, FMV-RP, and FMV-RS perform similarly and they all dominate the MVc and the uniform subsampling methods in both estimation and prediction. Figure 4 shows that our methods are more stable compared with the MVc and the uniform subsampling methods.



(a) log(MSE)

(b) EL

Figure 3: The log of MSE and EL based on MV, MVc, FMV-RP, FMV-RS, and Uniform methods, where $n_0 = 1000$, $\rho = 0.2$, $r_1 = r_2 = 20000$, $r_3 = 4$.



(a) $n = 2000$.

(b) $n = 4000$.

Figure 4: The MSE based on MV, MVc, FMV-RP, FMV-RS, and Uniform methods, where $n_0 = 1000$, $\rho = 0.2$, $r_1 = r_2 = 20000$, $r_3 = 4$.

7. Conclusion

In this paper, optimal Poisson subsampling algorithms for multinomial logistic models are derived, and the methods for fast approximation of the subsampling probabilities are also developed. The asymptotic normality of the subsample based estimator and the consistency of the approximations are shown in this work. A two-step algorithm is suggested for practical implementation. Some numerical experiments on simulated and real datasets are carried out to evaluate their practical performance. Both theoretical and numerical results demonstrate the great potential of the proposed methods in extracting useful information from massive datasets.

Before winding up this work it should be mentioned that we here only focus on the behavior of the subsample based estimator when the full dataset is given. As we know, the full data size may increase in the big data era. A typical example is high-speed data stream analysis. More precisely, the study of the asymptotic behavior when full data increases is also important and need to be investigated in the future research.

Supplementary Material

All technical proofs and additional simulation results are provided in the online Supplementary Material. The code of this work is available at `https://github.com/Quicy-PKU/OSMLM`.

Acknowledgments

The authors sincerely thank the editor, associate editor, and referees for their valuable comments and insightful suggestions, which lead to further improvement of this work. Both Ye and Yu contribute equally and are co-first authors. Ai is the corresponding author. Ai's work was supported by NSFC grants 12071014 and 12131001, National Key R&D Program of China 2020YFE0204200, and LMEQF. Yu's work was supported by NSFC grant 12001042, Beijing Municipal Natural Science Foundation No. 1232019, and the Beijing Institute of Technology research fund program for young scholars.

References

Achlioptas, D. (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences 66*(4), 671–687.

Agresti, A. (2019). *An Introduction to Categorical Data Analysis* (3rd ed.). John Wiley & Sons.

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

Ai, M., Z. Ye, and J. Yu (2023). Locally D-optimal designs for hierarchical response experiments. *Statistica Sinica 33*(1), 381–399.

Ai, M., J. Yu, H. Zhang, and H. Wang (2021). Optimal subsampling algorithms for big data regressions. *Statistica Sinica 31*(2), 749–772.

Ailon, N. and B. Chazelle (2006). Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing*, pp. 557–563. Association for Computing Machinery.

Ando, T. and K. chau Li (2017). A weight-relaxed model averaging approach for high-dimensional generalized linear models. *The Annals of Statistics 45*(6), 2654 – 2679.

Bu, X., D. Majumdar, and J. Yang (2020). D-optimal designs for multinomial logistic models. *The Annals of Statistics 48*(2), 983–1000.

Cheng, Q., H. Wang, and M. Yang (2020). Information-based optimal subdata selection for big data logistic regression. *Journal of Statistical Planning and Inference 209*, 112–122.

Dereziński, M., M. K. Warmuth, and D. Hsu (2018). Leveraged volume sampling for linear regression. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 2510–2519.

Drineas, P., M. Magdon-Ismail, M. W.Mahoney, and D. P. Woodruff (2012). Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research 13*(111), 3475–3506.

Drineas, P., M. W.Mahoney, S. Muthukrishnan, and T. Sarlós (2011). Faster least squares approximation. *Numerische Mathematik 117*(2), 219–249.

Fithian, W., T. Hastie, et al. (2014). Local case-control sampling: Efficient subsampling in imbalanced data sets. *The Annals of Statistics 42*(5), 1693–1724.

Han, L., K. M. Tan, T. Yang, and T. Zhang (2020). Local uncertainty sampling for large-scale multiclass logistic regression. *The Annals of Statistics 48*(3), 1770–1788.

Ma, P., M. W.Mahoney, and B. Yu (2015). A statistical perspective on algorithmic leveraging. *Journal* *of Machine Learning Research 16*(27), 861–911.

Ma, P., X. Zhang, X. Xing, J. Ma, and M. Mahoney (2020). Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. In *International Conference on Artificial Intelligence and Statistics*, pp. 1026–1035. PMLR.

McCoy, A. J., R. D. Oeffner, A. G. Wrobel, J. R. M. Ojala, K. Tryggvason, B. Lohkamp, and R. J. Read (2017). Ab initio solution of macromolecular crystal structures without direct methods. *Proceedings of the National Academy of Sciences 114*(14), 3637–3641.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 42*(2), 109–127.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. Volume 4 of *Handbook of Econometrics*, pp. 2111–2245. San Diego: Elsevier Science.

Peterson, B. and F. E. Harrell (1990). Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 39*(2), 205–217.

Pukelsheim, F. (2003). *Optimal Design of Experiments*. SIAM.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

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

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

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

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

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

Yu, J., M. Ai, and Z. Ye (2023). A review on design inspired subsampling for big data. *Statistical Papers*.

Yu, J., J. Liu, and H. Wang (2023). Information-based optimal subdata selection for non-linear models. *Statistical Papers*.

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

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

LMAM, School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing 100871, China.

E-mails: myai@math.pku.edu.cn, zqye@pku.edu.cn

School of Mathematics and Statistics, Beijing Institute of Technology, Beijing 100811, China.

E-mail: yujunbeta@bit.edu.cn