Statistica Sinica Preprint No: SS-2018-0439	
Title	OPTIMAL SUBSAMPLING ALGORITHMS FOR BIG DATA REGRESSIONS
Manuscript ID	SS-2018-0439
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0439
Complete List of Authors	Mingyao Ai Jun Yu Huiming Zhang and HaiYing Wang
Corresponding Author	HaiYing Wang
E-mail	haiying.wang@uconn.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

Statistica Sinica

OPTIMAL SUBSAMPLING ALGORITHMS FOR BIG DATA REGRESSIONS

Mingyao Ai[1], Jun Yu[2], Huiming Zhang[1], HaiYing Wang[3]

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

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

Department of Statistics, University of Connecticut [3]

Abstract: In order to quickly approximate maximum likelihood estimators from massive data, this study examines the **o**ptimal **s**ubsampling **m**ethod under the **A**-optimality **c**riterion (OSMAC) for generalized linear models. The consistency and asymptotic normality of the estimator from a general subsampling algorithm are established, and optimal subsampling probabilities under the A- and L-optimality criteria are derived. Furthermore, using Frobenius-norm matrix concentration inequalities, the finite-sample properties of the subsample estimator based on optimal subsampling probabilities are also derived. Because the optimal subsampling probabilities depend on the full data estimate, an adaptive two-step algorithm is developed. The asymptotic normality and optimality of the estimator from this adaptive algorithm are established. The proposed methods are illustrated and evaluated using numerical experiments on simulated and real data sets.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

Key words and phrases: generalized linear models; massive data; matrix concentration inequality.

1. Introduction

Today, massive data sets are ubiquitous in many scientific fields and practices, including astronomy, economics, and industrial problems. Extracting useful information from these large data sets is a core challenge in areas such as computer science, machine learning, statistics, and, as a result, has attracted much attention. However, computational limitations still exist, owing to rapid growth in the volume of data. Subsampling is a popular technique for extracting useful information from massive data. Therefore, this study develops optimal subsampling strategies for generalized linear models (GLMs). Typically, the maximum likelihood estimators (MLEs) are found numerically by using the Newton–Raphson method. However, fitting a GLM on massive data is not an easy task using the iterative Newton–Raphson method, requiring $O(p^2n)$ time in each iteration of the optimization procedure.

Subsampling provides an efficient way to solve this problem (e.g., see Drineas et al., 2006) because it essentially reduces the volume of the data. Drineas et al. (2011) proposed performing a randomized Hadamard trans-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

form on the data and then using the uniform subsampling to take random subsamples to approximate the ordinary least squares estimators in linear regression models. Ma et al. (2015) and Ma and Sun (2015) developed an effective subsampling method for linear regression models that uses normalized statistical leverage scores of the covariate matrix as nonuniform subsampling probabilities. Jia et al. (2014) studied leverage sampling for GLMs, based on generalized statistical leverage scores. Wang et al. (2018b) and Yao and Wang (2019) developed an optimal subsampling procedure to minimize the asymptotic mean squared error (MSE) of the resultant subsample-estimator, given the full data, based on A- or L-optimality criteria in the language of optimal design. Wang et al. (2019) proposed a new algorithm, called the information-based optimal subdata selection method, for linear regressions on big data. The basic idea is to select the most informative data points deterministically based on D-optimality, without relying on random subsampling. A divide-and-conquer version of the algorithm is presented in Wang (2019). Recent developments related to the big data subsampling method can be found in Wang et al. (2016).

Methodological investigations on subsampling methods with statistical guarantees for massive data regression are still limited when models are complex. To the best of our knowledge, most existing results concern lin-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

ear regression models, as in Ma et al. (2015) and Wang et al. (2019). The optimal subsampling methods in Wang et al. (2018b) and Yao and Wang (2019) are designed specifically for logistic and multinomial regression models, respectively. However, using only linear and logistic regressions is not sufficient, in practice (Czado and Munk, 2000). For example, we may need a Poisson or a negative binomial distribution for count data, or need a Gamma or an inverse Gaussian distribution for data with nonnegative responses. In addition, the aforementioned investigations do not consider the finite-sample properties of the subsampled estimators. We attempt to fill these gaps by deriving the optimal subsampling probabilities for GLMs, including those with noncanonical link functions, thus allowing a wide range of statistical models for a regression analysis. Furthermore, we derive the finite-sample upper bounds for the approximation errors, which can be used in practice to balance the subsample size and the prediction accuracy. Owing to the nonnatural link, our investigation is quite different from that of Wang et al. (2018b). For example, the Hessian matrix in the models considered in this study may be dependent on the responses.

The rest of this paper is organized as follows. Section 2 introduces the model setup and derives the asymptotic properties for the general subsampling estimator. Section 3 derives optimal subsampling strategies based on

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

A- and L-optimality criteria for GLMs. The finite-sample error bounds are also derived in this section. Section 4 designs a two-step algorithm to approximate the optimal subsampling procedure, and obtains the asymptotic properties of the resultant estimator. Section 5 demonstrates the proposes method using numerical simulations and real data.

2. Preliminaries

2.1 Models and Assumptions

Recall the definition of the one-parameter exponential family of distributions $f(y|\theta) = h(y)\exp(\theta y - \psi(\theta))$, for $\theta \in \Theta$, as in (5.50) of Efron and Hastie (2016), where θ is called the canonical parameter, and Θ is called the natural parameter space. Here $f(\cdot|\theta)$ is a probability density function for the continuous case, or a probability mass function for the discrete case; $h(\cdot)$ is a specific function that does not depend on θ; and the parameter space Θ is defined as $\Theta := \{\theta \in \mathbb{R} : \int h(x)\exp(\theta x)\mu(dx) < \infty\}$, with μ being the dominating measure. The exponential family includes most of the commonly used distributions, such as the normal, gamma, Poisson, and binomial distributions (see Efron and Hastie, 2016; Mccullagh and Nelder, 1989).

A key tactic for a generalized linear regression model is to express θ

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

in the form of a linear function of regression coefficients. Let $(\boldsymbol{x}, y)$ be a pair of random variables, where $y \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^p$. The generalized linear regression model assumes that the conditional distribution of y_i, given $\boldsymbol{x}_i$, is determined by $\theta_i = u(\boldsymbol{\beta}^T \boldsymbol{x}_i)$. Specifically for the exponential family, it assumes that the distribution of $y|\boldsymbol{x}$ is

$$f(y|\boldsymbol{\beta}, \boldsymbol{x}) = h(y) \exp(yu(\boldsymbol{\beta}^T \boldsymbol{x}_i) - \psi(u(\boldsymbol{\beta}^T \boldsymbol{x}_i))), \quad \text{with} \quad \boldsymbol{\beta}^T \boldsymbol{x} \in \Theta. \tag{2.1}$$

The problem of interest is to estimate the unknown $\boldsymbol{\beta}$ from the observed data. As a special case, when $u(t) = t$, the corresponding models are the so-called GLMs with canonical link functions. Typical examples include the logistic regression for binary data, and the Poisson regression for count data. A commonly used GLM with a noncanonical link function is the negative binomial regression (NBR), which is often used as an alternative to the Poisson regression when the data exhibit overdispersion. For this model, $u(t) = t - \log(\nu + e^t)$ and $\psi(u(t)) = \nu \log(\nu + e^t)$, for some size parameter ν.

2.2 General Subsampling Algorithm and its Asymptotic Properties

In this subsection, we present a general subsampling algorithm for GLMs and obtain some asymptotic results.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

To facilitate the presentation, denote the full data matrix by $\mathcal{F}_n = (\boldsymbol{X}, \boldsymbol{y})$, where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ is the covariate matrix, and $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is the response vector. In this paper, we assume that $(\boldsymbol{x}_i, y_i)$'s are generated independently from a GLM. Let S be a set of subsamples with r data points, and define the sampling distribution π_i for all data points $i = 1, 2, ...n$, as $\boldsymbol{\pi}$. Then, we have the following general subsampling algorithm:

1. Assign a sampling distribution $\boldsymbol{\pi}$ such that, in each draw, the ith element in the full data set $\mathcal{F}_n$ has the inclusion probability π_i.

2. Sample with replacement r times to form the subsample set $S := \{(y_i^*, \boldsymbol{x}_i^*, \pi_i^*), i = 1, \ldots, r\}$, where $\boldsymbol{x}_i^*$, y_i^*, and π_i^* denote the covariates, responses, and subsampling probabilities, respectively, in the subsample.

3. Based on the subsample set S, calculate the weighted log-likelihood estimator by maximizing the following function:

$$L^*(\boldsymbol{\beta}) = \frac{1}{r} \sum_{t=1}^{r} \frac{1}{\pi_i^*} [y_i^* u(\boldsymbol{\beta}^T \boldsymbol{x}_i^*) - \psi(u(\boldsymbol{\beta}^T \boldsymbol{x}_i^*))]. \tag{2.2}$$

An important feature of the above algorithm is that the subsample estimator is essentially a weighted MLE, where the corresponding weights are inverses of the subsampling probabilities. This is analogous to the

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

Hansen–Hurwitz estimator (Hansen and Hurwitz, 1943) in classic sampling techniques. For an overview see Särndal et al. (1992). Although Ma et al. (2015) showed that the unweighted subsample estimator is asymptotically unbiased for $\boldsymbol{\beta}$ in leveraging sampling, an unweighted subsample estimator, is in general, biased if the sampling distribution $\boldsymbol{\pi}$ depends on the responses. The inverse-probability weighting scheme removes this bias; thus we restrict our analysis to the weighted estimator.

Let $\dot{\psi}(t)$ and $\ddot{\psi}(t)$ be the first and the second derivatives of $\psi(t)$, respectively. To characterize the asymptotic properties of the subsampled estimators, we require the following regularity assumptions:

(H.1): Assume that $\boldsymbol{\beta}^T\boldsymbol{x}$ lies in the interior of a compact set $K \in \Theta$ almost surely.

(H.2): The regression coefficient $\boldsymbol{\beta}$ is an inner point of the compact domain $\Lambda_B = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\| \leq B\}$, for some constant B.

(H.3): Central moments condition: $n^{-1}\sum_{i=1}^{n} |y_i - \dot{\psi}(u(\boldsymbol{\beta}^T\boldsymbol{x}_i))|^4 = O_P(1)$, for all $\boldsymbol{\beta} \in \Lambda_B$.

(H.4): As $n \to \infty$, the observed information matrix

$$\mathcal{J}_X := \tfrac{1}{n}\sum_{i=1}^{n} \{\ddot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T[\dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)) - y_i] + \ddot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))\dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T]\}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

goes to a positive-definite matrix in probability.

(H.5): Require that the full sample covariates have finite sixth-order moments; that is, $E\|\boldsymbol{x}_1\|^6 \leq \infty$.

(H.6): Assume $n^{-2}\sum_{i=1}^{n}\|\boldsymbol{x}_i\|^k/\pi_i = O_P(1)$, for $k = 2, 4$.

(H.7): For $\gamma = 0$ and some $\gamma > 0$, assume

$$\frac{1}{n^{2+\gamma}}\sum_{i=1}^{n}\frac{|y_i - \dot{\psi}_i(u(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T\boldsymbol{x}_i))|^{2+\gamma}\|\dot{u}(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\|^{2+\gamma}}{\pi_i^{1+\gamma}} = O_P(1).$$

Assumptions (H.1) and (H.2) are used in Clémencon et al. (2014). The set in (H.2) is also called the admissible set, which, is the premise for consistent estimators in GLMs with full data (see Fahrmeir and Kaufmann, 1985). These two assumptions ensure that $E(y_i|\boldsymbol{x}_i) < \infty$, for all i. Assumption (H.4) imposes a condition on the covariates to ensure that the MLE based on the full data set is consistent. To obtain the Bahadur representation of the subsampled estimator, (H.3) and (H.5) are needed. Assumptions (H.6) and (H.7) are moment conditions on the covariates and the subsampling probabilities. Assumption (H.7) is required by the Lindeberg-Feller central limit theorem. Specifically, for uniform subsampling with $\pi_i = n^{-1}$ or, more generally, when $\max_{i=1,\ldots,n}(n\pi_i)^{-1} = O_P(1)$, (H.7) is implied by $n^{-1}\sum_{i=1}^{n}|y_i - \dot{\psi}_i(u(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T\boldsymbol{x}_i))|^{2+\gamma}\|\dot{u}(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\|^{2+\gamma} = O_P(1)$,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

which is guaranteed by the condition $E|y|^{4+2\gamma} = O(1)$ when (H.1) and (H.5) are satisfied.

The theorem below presents the consistency of the estimator from the subsampling algorithm to the full-data MLE.

Theorem 1. *If Assumptions (H.1)–(H.6) hold, then as $n \to \infty$ and $r \to \infty$, $\tilde{\boldsymbol{\beta}}$ is consistent to $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ in conditional probability, given $\mathcal{F}_n$. Moreover, the rate of convergence is $r^{-1/2}$. That is, with probability approaching one, for any $\epsilon > 0$, there exist finite Δ_ϵ and r_ϵ, such that*

$$P(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}\| \geq r^{-1/2}\Delta_\epsilon | \mathcal{F}_n) < \epsilon, \tag{2.3}$$

for all $r > r_\epsilon$.

In addition to the consistency, we derive the asymptotic distribution of the approximation error, thus proving that the approximation error, $\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}$, is asymptotically normal in conditional distribution.

Theorem 2. *If Assumptions (H.1)–(H.7) hold, then as $n \to \infty$ and $r \to \infty$, conditional on $\mathcal{F}_n$ in probability,*

$$V^{-1/2}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}) \longrightarrow N(0, I) \tag{2.4}$$

in distribution, where $V = \mathcal{J}_X^{-1} V_c \mathcal{J}_X^{-1} = O_p(r^{-1})$ and

$$V_c = \frac{1}{rn^2} \sum_{i=1}^{n} \frac{\{y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T \boldsymbol{x}_i))\}^2 \dot{u}^2(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T \boldsymbol{x}_i) \boldsymbol{x}_i \boldsymbol{x}_i^T}{\pi_i}. \tag{2.5}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

3. Optimal Subsampling Strategies

In this section, we specify the subsampling distribution $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^n$, with theoretical backup.

3.1 Optimal Subsampling Strategies Based on Optimal Design Criteria

Based on the A-optimality criterion in the theory of experiment design (see Pukelsheim, 2006), optimal subsampling selects subsampling probabilities such that the asymptotic MSE of $\tilde{\boldsymbol{\beta}}$ is minimized. This idea was proposed in Wang et al. (2018b). Here, we say the resulting subsampling strategy is *mV-optimal*.

Theorem 3. *A subsampling strategy is mV-optimal if the subsampling probability is chosen such that*

$$\pi_i^{\mathrm{mV}} = \frac{|y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i))| \|\mathcal{J}_X^{-1} \dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i)\boldsymbol{x}_i\|}{\sum_{j=1}^n |y_j - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i))| \|\mathcal{J}_X^{-1} \dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_j)\boldsymbol{x}_j\|}, \ i = 1, 2, ..., n. \tag{3.6}$$

The optimal subsampling probability $\boldsymbol{\pi}^{\mathrm{mV}}$ has a meaningful interpretation from the viewpoint of the optimal design of experiments (Pukelsheim, 2006). Note that, under a mild condition, the "empirical information matrix" $\mathcal{J}_X^e = \frac{1}{n}\sum_{i=1}^n [y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i))]^2 \dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i)\boldsymbol{x}_i \boldsymbol{x}_i^T$ and $\mathcal{J}_X$ converge

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

to the same limit, namely, the Fisher information matrix of model (2.1). This means that $\mathcal{J}_X^e - \mathcal{J}_X = o_P(1)$. Thus, $\mathcal{J}_X$ can be replaced by $\mathcal{J}_X^e$ in $\boldsymbol{\pi}^{\mathrm{mV}}$, because Theorem 2 still holds if $\mathcal{J}_X$ is replaced by $\mathcal{J}_X^e$ in (2.5). Let $\eta_{\boldsymbol{x}_i} = [y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i))]^2 \dot{u}^2(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i)\boldsymbol{x}_i \boldsymbol{x}_i^T$ be the contribution of the ith observation to the empirical information matrix, and let $\mathcal{J}_{X\boldsymbol{x}_i\alpha}^e = (1-\alpha)\mathcal{J}_X^e + \alpha\eta_{\boldsymbol{x}_i}$, which can be interpreted as a movement of the information matrix in a direction determined by the ith observation. The directional derivative of $\mathrm{tr}(\mathcal{J}_X^{e\,-1})$ through the direction determined by the ith observation is $F_i = \lim_{\alpha\to 0+} \alpha^{-1}\{\mathrm{tr}(\mathcal{J}_X^{e\,-1}) - \mathrm{tr}(\mathcal{J}_{X\boldsymbol{x}_i\alpha}^{e\,-1})\}$. This directional derivative is used to measure the relative gain in estimation efficiency under the A-optimality after adding the ith observations to the sample. Thus, the optimal subsampling strategy prefers to select data points with large values of directional derivatives, that is, data points that will result in a larger gain under the A-optimality.

The optimal subsampling strategy derived from the *mV-optimality* criterion requires that we calculate $\|\mathcal{J}_X^{-1}\dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i)\boldsymbol{x}_i\|$, for $i = 1, 2, ..., n$, which takes $O(np^2)$ time. To reduce the computing time, Wang et al. (2018b) proposed a modified optimality criterion to minimize $\mathrm{tr}(V_c)$. This criterion is essentially the L-optimality criterion in optimal experimental design (see Pukelsheim, 2006), which aims to improve the estimation qual-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

ity of $\mathcal{J}_X\tilde{\boldsymbol{\beta}}$. It is easy to see that only $O(np)$ time is needed to calculate the optimal sampling probabilities. We say the resulting subsampling strategy is *mVc-optimal*.

Theorem 4. *A subsampling strategy is mVc-optimal if the subsampling probability is chosen such that*

$$\pi_i^{\mathrm{mVc}} = \frac{|y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i))|\|\dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\|}{\sum_{j=1}^n |y_j - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_j))|\|\dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_j)\boldsymbol{x}_j\|}, \ i = 1, 2, ..., n. \quad (3.7)$$

Note that in order to calculate $\|\mathcal{J}_X^{-1}\dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\|$, for $i = 1, 2, ..., n$, we need $O(np^2)$ time, but we only need $O(np)$ time to evaluate $\|\dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\|$. Here, $\mathcal{J}_X$ and V_c are nonnegative definite, and $V = \mathcal{J}_X^{-1}V_c\mathcal{J}_X^{-1}$. Simple matrix algebra yields $\mathrm{tr}(V) = \mathrm{tr}(V_c\mathcal{J}_X^{-2}) \leq \sigma_{\max}(\mathcal{J}_X^{-2})\mathrm{tr}(V_c)$, where $\sigma_{\max}(A)$ denotes the maximum singular value of matrix A. Because $\sigma_{\max}(\mathcal{J}_X^{-2})$ does not depend on $\boldsymbol{\pi}$, minimizing $\mathrm{tr}(V_c)$ minimizes an upper bound of $\mathrm{tr}(V)$. In fact, for two given subsampling strategies $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$, if $V_c(\boldsymbol{\pi}^{(1)}) \leq V_c(\boldsymbol{\pi}^{(2)})$ in the sense of Loewner-ordering, then it follows that $V(\boldsymbol{\pi}^{(1)}) \leq V(\boldsymbol{\pi}^{(2)})$. Thus, the alternative optimality criterion greatly reduces the computing time, without losing too much in terms of estimation accuracy.

The score function for the log-likelihood means that π_i^{mVc} in Theorem 4 is proportional to $\|\{y_i - \dot{\psi}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\}\dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\|$, the norms of the gradients of the log-likelihood at individual data points, evaluated at the

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

full-data MLE. Here, we are trying to identify the subsample that best approximates the full data score function at the full-data MLE.

We now illustrate Theorem 3 and Theorem 4 using some commonly used GLMs. Note that $u(\cdot)$ is the identity function for GLMs with natural link functions, such as the logistic and Poisson regressions. For the logistic regression,

$$\pi_i^{\mathrm{mV}} = \frac{|y_i - p_i| \left\|\mathcal{J}_X^{-1}\boldsymbol{x}_i\right\|}{\sum_{j=1}^n |y_j - p_j| \left\|\mathcal{J}_X^{-1}\boldsymbol{x}_j\right\|}, \qquad \pi_i^{\mathrm{mVc}} = \frac{|y_i - p_i| \left\|\boldsymbol{x}_i\right\|}{\sum_{j=1}^n |y_j - p_j| \left\|\boldsymbol{x}_j\right\|},$$

with $p_i = \exp(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)/\{1+\exp(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)\}$ and $\mathcal{J}_X = n^{-1}\sum_{k=1}^n p_k(1-p_k)\boldsymbol{x}_k\boldsymbol{x}_k^T$. These are the same as the results in Wang et al. (2018b). For the Poisson regression,

$$\pi_i^{\mathrm{mV}} = \frac{|y_i - \lambda_i| \left\|\mathcal{J}_X^{-1}\boldsymbol{x}_i\right\|}{\sum_{j=1}^n |y_j - \lambda_j| \left\|\mathcal{J}_X^{-1}\boldsymbol{x}_j\right\|}, \qquad \pi_i^{\mathrm{mVc}} = \frac{|y_i - \lambda_i| \left\|\boldsymbol{x}_i\right\|}{\sum_{j=1}^n |y_j - \lambda_j| \left\|\boldsymbol{x}_j\right\|},$$

with $\lambda_i = \exp(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)$ and $\mathcal{J}_X = n^{-1}\sum_{k=1}^n \exp(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_k)\boldsymbol{x}_k\boldsymbol{x}_k^T$. The NBR does not have a canonical link function, and the conditional distribution of the response is modeled by the two-parameter distribution

$$f(y_i|\nu,\mu_i) = \frac{\Gamma(\nu+y_i)}{\Gamma(\nu)y_i!}\left(\frac{\mu_i}{\nu+\mu_i}\right)^{y_i}\left(\frac{\nu}{\nu+\mu_i}\right)^{\nu}, \quad i = 1,2,\ldots,n,$$

where the size parameter ν can be estimated as a nuisance parameter. The optimal subsampling probabilities for NBR with size parameter ν are

$$\pi_i^{\mathrm{mV}} = \frac{|y_i-\mu_i|\left\|\mathcal{J}_X^{-1}\frac{\nu\boldsymbol{x}_i}{\nu+\mu_i}\right\|}{\sum_{j=1}^n |y_j-\mu_j|\left\|\mathcal{J}_X^{-1}\frac{\nu\boldsymbol{x}_j}{\nu+\mu_j}\right\|}, \qquad \pi_i^{\mathrm{mVc}} = \frac{|y_i-\mu_i|\left\|\frac{\nu\boldsymbol{x}_i}{\nu+\mu_i}\right\|}{\sum_{j=1}^n |y_j-\mu_j|\left\|\frac{\nu\boldsymbol{x}_i}{\nu+\mu_j}\right\|},$$

with $\mu_i = \exp(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\boldsymbol{x}_i)$ and $\mathcal{J}_X = n^{-1}\sum_{k=1}^n\{\nu(\nu+y_i)\mu_i\}/(\nu+\mu_i)^2\boldsymbol{x}_k\boldsymbol{x}_k^T$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

3.2 Non-asymptotic Properties

Here, we derive some finite-sample properties of the subsample estimators based on the optimal subsampling probabilities $\boldsymbol{\pi}^{\mathrm{mV}}$ and $\boldsymbol{\pi}^{\mathrm{mVc}}$. The results are presented in the form of the excess risk when approximating the mean responses and they hold for fixed r and n, without requiring any quantity to go to infinity. These results may identify the factors that affect the approximation accuracy.

Because $\dot{\psi}(u(\boldsymbol{x}_i^T\boldsymbol{\beta}))$ is the conditional expectation of the response y_i, given $\boldsymbol{x}_i$, we aim to characterize the quantity of $\tilde{\boldsymbol{\beta}}$ in the prediction by examining $\|\dot{\psi}(u(\boldsymbol{X}_d^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})) - \dot{\psi}(u(\boldsymbol{X}_d^T\tilde{\boldsymbol{\beta}}))\|$. This quantity is the distance between the estimated conditional mean responses based on the full-data, and that based on the subsamples. Intuitively, it measures the goodness of fit when using a subsample estimator to predict the mean responses. Note that we can always improve the accuracy of the estimator by increasing the subsample size r. Here, we examine the effects of different quantities such as the covariate matrix, data dimension, and the effect of subsample size r, on approximation accuracy.

Let $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ be the maximum and minimum nonzero singular values, respectively, of matrix A, where $\kappa(A) := \sigma_{\max}(A)/\sigma_{\min}(A)$. Denote $\dot{\psi}(u(\boldsymbol{X}^T\boldsymbol{\beta}))$, a vector with the ith element equals to $\dot{\psi}(u(\boldsymbol{x}_i^T\boldsymbol{\beta}))$,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

and define $\dot{u}(\boldsymbol{X}^T\boldsymbol{\beta}) := \text{diag}\{\dot{u}(\boldsymbol{x}_1^T\boldsymbol{\beta}), \cdots, \dot{u}(\boldsymbol{x}_n^T\boldsymbol{\beta})\}$. For the estimator $\tilde{\boldsymbol{\beta}}$ obtained from the algorithm in Section 2 based on the subsampling probabilities, $\boldsymbol{\pi}^{\text{mV}}$ and $\boldsymbol{\pi}^{\text{mVc}}$, the following theorem holds.

Theorem 5. *Let $\tilde{\boldsymbol{X}}$ denote the design matrix consisting of the subsample covariates, with each sampled element rescaled by $1/\sqrt{r\pi_i^*}$. Assume that $\sigma^2_{\min}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{X}}) \geq 0.5\sigma^2_{\min}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X})$, and both $\sigma_{\max}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X})/\sqrt{n}$ and $\sigma_{\min}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X})/\sqrt{n}$ are bounded. For any given $\epsilon \in (0, 1/3)$, with probability at least $1-\epsilon$, we have*

$$\begin{aligned}&\|\dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\text{MLE}})) - \dot{\psi}(u(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}}))\| \\ &\leq 2C_{\dot{u}}[1 + \frac{4\alpha\sqrt{\log(1/\epsilon)}}{\sqrt{r}}]\sqrt{p}\kappa^2(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X})\|[\boldsymbol{y} - \dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\text{MLE}}))]\|, \quad (3.8)\end{aligned}$$

where $\alpha = \kappa(\mathcal{J}_X^{-1})$ for $\boldsymbol{\pi}^{\text{mV}}$, $\alpha = 1$ for $\boldsymbol{\pi}^{\text{mVc}}$, and $C_{\dot{u}} = \sup_{r\in K\subset\Theta} |\dot{u}(r)|$.

Theorem 5 indicates that the accuracy increases with the subsample size r, which agrees with the results in Theorem 1. In addition, it enables us to examine the effects of various quantities such as the covariate matrix, data dimension, and the effect of subsample size r, on the approximation accuracy. Heuristically, the condition number of $\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}$ measures the collinearity of the covariates in the full-data covariate matrix, p shows the curse of dimensionality, and $\|\boldsymbol{y} - \dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\text{MLE}}))\|$ measures the goodness of fit of the underlying model on the full data.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

The result in (3.8) also indicates that we should choose $r \propto p$ to control the error bound; hence, it seems reasonable to choose the subsample size as $r = cp$. This agrees with the recommendation by Chapman et al. (1994) and Loeppky et al. (2009) of choosing a sample size as large as 10 times number of covariates for designed experiments. However, in such designed experiments, the covariate matrices are often orthogonal, or close to orthogonal, in which case, $\kappa(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X})$ is equal or close to one. Here, we consider that the full data may not be obtained from well-designed experiments, in which case, $\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}$ may vary substantially. Thus, $\kappa(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X})$ should also be considered when determining the required subsample size for a given level of prediction accuracy.

The constant 0.5 in Theorem 5's condition $\sigma^2_{\min}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{X}}) \geq 0.5\sigma^2_{\min}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X})$ can be replaced by any constant between 0 and 1. Here, we follow the setting of Drineas et al. (2011), and choose 0.5 for convenience. This condition indicates that the rank of $\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{X}}$ is the same as that of $\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}$. Further details and interpretations about this condition can be found in Mahoney (2012).

Using a similar argument to that in the proof of Theorem 5, we prove that this condition holds with high probability.

Theorem 6. *Let $\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{X}}$ denote the design matrix consisting of sub-*

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

samples, with each sampled element rescaled by $1/\sqrt{r\pi_i^*}$. *Assume that* $|y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T\boldsymbol{x}_i))|\|\dot{u}(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T\boldsymbol{x}_i)\boldsymbol{x}_i\| \geq \gamma\|\boldsymbol{x}_i\|$, *for all* i, *and that* $\sigma_{\max}(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X})/\sqrt{n}$, *and* $\sigma_{\min}(\dot{u}(\boldsymbol{X}^T\boldsymbol{\beta})\boldsymbol{X})/\sqrt{n}$ *are bounded. For any given* $\epsilon \in (0, 1/3)$, *let* $c_d \leq 1$ *be a constant depending on* $\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}$, $C_{\dot{u}} = \sup_{r\in K\subset\Theta} |\dot{u}(r)|$, *and* $r > 64c_d^2C_{\dot{u}}^2\log(1/\epsilon)\sigma_{\max}^4(\boldsymbol{X})p^2/(\alpha^2\delta^2\sigma_{\min}^4(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}))$, *where* δ *is some constant depending on* γ, *and* $\|\boldsymbol{y} - \dot{\psi}(u(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{\text{MLE}}))\|$. *Then, with probability at least* $1-\epsilon$:

$$\sigma_{\min}^2(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{X}}) \geq 0.5\sigma_{\min}^2(\dot{u}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\boldsymbol{X}),$$

where $\alpha = \kappa(\mathcal{J}_X^{-1})$ *for* $\boldsymbol{\pi}^{\text{mV}}$ *and* $\alpha = 1$ *for* $\boldsymbol{\pi}^{\text{mVc}}$.

4. Practical Consideration and Implementation

For practical implementation, the optimal subsampling probabilities $\{\pi_i^{\text{mV}} : i = 1, \ldots, n\}$ and $\{\pi_i^{\text{mVc}} : i = 1, \ldots, n\}$ cannot be used directly, because they depend on the unknown full-data MLE, $\hat{\boldsymbol{\beta}}_{\text{MLE}}$. As suggested in Wang et al. (2018b), in order to calculate $\boldsymbol{\pi}^{\text{mV}}$ or $\boldsymbol{\pi}^{\text{mVc}}$, a pilot estimator of $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ has to be used. Let $\tilde{\boldsymbol{\beta}}_0$ be a pilot estimator based on a subsample of size r_0. This can be used in place of $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ in $\boldsymbol{\pi}^{\text{mV}}$ or $\boldsymbol{\pi}^{\text{mVc}}$, which then can be used to derive more informative subsamples.

From the expression of $\boldsymbol{\pi}^{\text{mV}}$ or $\boldsymbol{\pi}^{\text{mVc}}$, the approximated optimal subsampling probabilities are both proportional to $|y_i - \dot{\psi}(u(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_i))|$. Thus,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

a data point with $y_i \approx \dot{\psi}(u(\tilde{\boldsymbol{\beta}}_0^T \boldsymbol{x}_i))$ has a very small probability of being selected, and the data point with $y_i = \dot{\psi}(u(\tilde{\boldsymbol{\beta}}_0^T \boldsymbol{x}_i))$ will never be included in a subsample. On the other hand, if these data points are included in the subsample, they may dominate the weighted log-likelihood function in (2.2). As a result, the subsample estimator may be sensitive to these data points. Ma et al. (2015) also noticed that some extremely small subsampling probabilities may inflate the variance of the subsampling estimator in the context of leveraging sampling.

To protect the weighted log-likelihood function from being inflated by these data points in practice, we propose setting a threshold, say δ, for $|y_i - \dot{\psi}(u(\tilde{\boldsymbol{\beta}}_0^T \boldsymbol{x}_i))|$; that is, use $\max\{|y_i - \dot{\psi}(u(\tilde{\boldsymbol{\beta}}_0^T \boldsymbol{x}_i))|, \delta\}$ in place of $|y_i - \dot{\psi}(u(\tilde{\boldsymbol{\beta}}_0^T \boldsymbol{x}_i))|$. Here, δ is a small positive number, say 10^{-6}. Setting a threshold δ in the subsampling probabilities truncates the weights of the subsample weighted log-likelihood. Truncating the weight function is commonly used in practice to ensure a robust estimation. Note that, in practice, an intercept should always be included in the model, so it is typical that $\|\dot{u}(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T \boldsymbol{x}_i)\boldsymbol{x}_i\|$ and $\|\mathcal{J}_X^{-1}\dot{u}(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T \boldsymbol{x}_i)\boldsymbol{x}_i\|$ are bounded away from zero, and do not need a threshold. Let $\tilde{V}$ be the version of V with $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ substituted

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

by $\tilde{\boldsymbol{\beta}}_0$. It can be shown that

$$\mathrm{tr}(\tilde{V}) \leq \mathrm{tr}(\tilde{V}^{\delta}) \leq \mathrm{tr}(\tilde{V}) + \frac{\delta^2}{n^2 r} \sum_{i=1}^{n} \frac{1}{\pi_i} \|\tilde{\mathcal{J}}_X^{-1} \dot{u}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T \boldsymbol{x}_i) \boldsymbol{x}_i\|^2.$$

Thus, minimizing $\mathrm{tr}(\tilde{V}^{\delta})$ is close to minimizing $\mathrm{tr}(\tilde{V})$ if δ is sufficiently small. The threshold δ makes our subsampling estimator more robust, without compromising too much on estimation efficiency. Here, we can also approximate $\mathcal{J}_X$ using the pilot sample. Specifically, the $\mathcal{J}_X$ in mV is approximated by $\tilde{\mathcal{J}}_X = (r_0)^{-1} \sum_{i=1}^{r_0} \{\ddot{u}(\tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i^*) \boldsymbol{x}_i^* \boldsymbol{x}_i^{*T} [\dot{\psi}(u(\tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i{}^*)) - y_i{}^*] + \ddot{\psi}(u(\tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i{}^*)) \dot{u}^2(\tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i{}^*) \boldsymbol{x}_i{}^* \boldsymbol{x}_i^{*T}]\}$, based on the first-stage subsamples $\{(\boldsymbol{x}_i^*, y_i^*) : i = 1, \ldots, r_0\}$.

For transparent presentation, we combine the aforementioned practical considerations in the following two-step algorithm:

1. Run the general subsampling algorithm with $\boldsymbol{\pi} = \boldsymbol{\pi}^{\mathrm{UNIF}}$ and $r = r_0$ to obtain the pilot subsample set $\tilde{S}_{r_0}$ and a pilot estimator $\tilde{\boldsymbol{\beta}}_0$.

2. Use $\tilde{\boldsymbol{\beta}}_0$ to calculate the approximated subsampling probabilities $\tilde{\boldsymbol{\pi}}^{\mathrm{opt}} = \{\tilde{\pi}_i^{\mathrm{mV}}\}_{i=1}^n$ or $\tilde{\boldsymbol{\pi}}^{\mathrm{opt}} = \{\tilde{\pi}_i^{\mathrm{mVc}}\}_{i=1}^n$, where $\tilde{\pi}_i^{\mathrm{mV}}$ is proportional to $\max(|y_i - \dot{\psi}(u(\tilde{\boldsymbol{\beta}}_0^T \boldsymbol{x}_i))|, \delta) \|\tilde{\mathcal{J}}_X^{-1} \dot{u}(\tilde{\boldsymbol{\beta}}_0^T \boldsymbol{x}_i) \boldsymbol{x}_i\|$s and $\tilde{\pi}_i^{\mathrm{mVc}}$ is proportional to $\max(|y_i - \dot{\psi}(u(\tilde{\boldsymbol{\beta}}_0^T \boldsymbol{x}_i))|, \delta) \|\dot{u}(\tilde{\boldsymbol{\beta}}_0^T \boldsymbol{x}_i) \boldsymbol{x}_i\|$.

3. Sample with replacement r times based on $\tilde{\boldsymbol{\pi}}^{\mathrm{opt}}$ to obtain the subsample set $S_{r^*} := \tilde{S}_{r_0} \cup \{(y_i^*, \boldsymbol{x}_i^*, \tilde{\pi}_i^*), i = 1, \ldots, r\}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

4. Maximize the following weighted log-likelihood function to obtain the estimator $\breve{\boldsymbol{\beta}}$:

$$L^*(\boldsymbol{\beta}) = \frac{1}{r + r_0} \sum_{i \in S_{r^*}} \frac{1}{\tilde{\pi}_i^*} [y_i^* u(\boldsymbol{\beta}^T \boldsymbol{x}_i^*) - \psi(u\boldsymbol{\beta}^T \boldsymbol{x}_i^*))]. \tag{4.9}$$

The following theorems describe the asymptotic properties of $\breve{\boldsymbol{\beta}}$.

Theorem 7. *Under Assumptions (H.1)–(H.5), if* $r_0 r^{-1} \to 0$ *as* $r_0 \to \infty, r \to \infty$, *and* $n \to \infty$, *then for the estimator* $\breve{\boldsymbol{\beta}}$ *obtained from the two-step algorithm, with probability approaching one, for any* $\epsilon > 0$, *there exist finite* Δ_ϵ *and* r_ϵ, *such that*

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

for all $r > r_\epsilon$.

The asymptotic normality is presented in the following theorem.

Theorem 8. *Under assumptions (H.1)–(H.5), if* $r_0 r^{-1} \to 0$, *then for the estimator obtained from the two-step algorithm, as* $r_0 \to \infty$, $r \to \infty$, *and* $n \to \infty$, *conditional on* $\mathcal{F}_n$,

$$V_{opt}^{-1/2} (\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}) \to N(0, I), \tag{4.10}$$

where $V_{opt} = \mathcal{J}_X^{-1} V_{c,opt} \mathcal{J}_X^{-1}$;

$$V_{c,opt} = \frac{1}{r} \frac{1}{n} \sum_{i=1}^{n} \frac{\{y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T \boldsymbol{x}_i))\}^2 \dot{u}^2(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T \boldsymbol{x}_i) \boldsymbol{x}_i \boldsymbol{x}_i^T}{\max(|y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T \boldsymbol{x}_i))|, \delta) \|\dot{u}(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T \boldsymbol{x}_i) \boldsymbol{x}_i\|} \tag{4.11}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

$$\times \frac{1}{n}\sum_{i=1}^{n} \max(|y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T \boldsymbol{x}_i))|, \delta)\|\dot{u}(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T \boldsymbol{x}_i)\boldsymbol{x}_i\|$$

when subsampling probabilities based on $\tilde{\pi}_i^{\text{mVc}}$*, and*

$$V_{c,opt} = \frac{1}{r}\frac{1}{n}\sum_{i=1}^{n} \frac{\{y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T \boldsymbol{x}_i))\}^2 \dot{u}^2(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T \boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T}{\max(|y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T \boldsymbol{x}_i))|, \delta)\|\mathcal{J}_X^{-1}\dot{u}(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T \boldsymbol{x}_i)\boldsymbol{x}_i\|}$$
$$\times \frac{1}{n}\sum_{i=1}^{n} \max(|y_i - \dot{\psi}(u(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T \boldsymbol{x}_i))|, \delta)\|\mathcal{J}_X^{-1}\dot{u}(\hat{\boldsymbol{\beta}}_{\text{MLE}}^T \boldsymbol{x}_i)\boldsymbol{x}_i\|$$

when subsampling probabilities based on $\tilde{\pi}_i^{\text{mV}}$*.*

In order to obtain the standard error of the corresponding estimator, we estimate the variance-covariance matrix of $\breve{\boldsymbol{\beta}}$ by $\breve{V} = \breve{\mathcal{J}}_X^{-1}\breve{V}_c\breve{\mathcal{J}}_X^{-1}$, where

$$\breve{\mathcal{J}}_X = \frac{1}{n(r_0+r)}\times$$
$$\Bigg\{\sum_{i=1}^{r_0} \frac{\ddot{u}(\breve{\boldsymbol{\beta}}^T\boldsymbol{x}_i^*)\boldsymbol{x}_i^*\boldsymbol{x}_i^T[\dot{\psi}(u(\breve{\boldsymbol{\beta}}^T\boldsymbol{x}_i^*)) - y_i^*] + \ddot{\psi}(u(\breve{\boldsymbol{\beta}}^T\boldsymbol{x}_i^*))\dot{u}^2(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_i^*)\boldsymbol{x}_i^*\boldsymbol{x}_i^T}{\pi_{i0}^*}$$
$$+\sum_{s=1}^{r} \frac{\ddot{u}(\breve{\boldsymbol{\beta}}^T\boldsymbol{x}_s^*)\boldsymbol{x}_s^*\boldsymbol{x}_s^T[\dot{\psi}(u(\breve{\boldsymbol{\beta}}^T\boldsymbol{x}_s^*)) - y_s^*] + \ddot{\psi}(u(\breve{\boldsymbol{\beta}}^T\boldsymbol{x}_s^*))\dot{u}^2(\breve{\boldsymbol{\beta}}^T\boldsymbol{x}_s^*)\boldsymbol{x}_s\boldsymbol{x}_s^T}{\tilde{\pi}_s^*}\Bigg\},$$
$$\breve{V}_c = \frac{1}{n^2(r_0+r)^2}\Bigg\{\sum_{i=1}^{r_0} \frac{\{y_i - \dot{\psi}(u(\breve{\boldsymbol{\beta}}^T\boldsymbol{x}_i^*))\}^2\dot{u}^2(\breve{\boldsymbol{\beta}}^T\boldsymbol{x}_i^*)\boldsymbol{x}_i^*(\boldsymbol{x}_i^*)^T}{(\tilde{\pi}_{i0}^*)^2}$$
$$+\sum_{i=1}^{r} \frac{\{y_i^* - \dot{\psi}(u(\breve{\boldsymbol{\beta}}^T\boldsymbol{x}_i^*))\}^2\dot{u}^2(\breve{\boldsymbol{\beta}}^T\boldsymbol{x}_i^*)\boldsymbol{x}_i^*(\boldsymbol{x}_i^*)^T}{(\tilde{\pi}_i^*)^2}\Bigg\},$$

π_{i0}^* is the subsampling probability used in the first stage, and $\tilde{\pi}_i^* = \tilde{\pi}_i^{\text{mV}*}$ or $\tilde{\pi}_i^{\text{mVc}*}$, for $i = 1, \ldots, r$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

5. Numerical Studies

5.1 Simulation Studies

In this section, we use simulations to evaluate the finite-sample performance of the proposed method for a Poisson regression and a NBR. Computations are performed in `R` (R Core Team, 2018). The performance of a sampling strategy $\boldsymbol{\pi}$ is evaluated using the empirical mean squared error (eMSE) of the resultant estimator: $\text{eMSE} = K^{-1}\sum_{k=1}^{K}\|\boldsymbol{\beta}_{\boldsymbol{\pi}}^{(k)} - \hat{\boldsymbol{\beta}}_{\text{MLE}}\|$, where $\boldsymbol{\beta}_{\boldsymbol{\pi}}^{(k)}$ is the estimator from the kth subsample with subsampling probability $\boldsymbol{\pi}$, and $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ is the MLE calculated from the whole data set. We set $K = 1000$ throughout this section.

Poisson regression. Full data of size $n = 10,000$ are generated from model $y|\boldsymbol{x} \sim \mathcal{P}(\exp(\boldsymbol{\beta}^T\boldsymbol{x}))$, where the true value of $\boldsymbol{\beta}$ is a 7×1 vector of 0.5. We consider the following four cases to generate the covariates $\boldsymbol{x}_i = (x_{i1}, ..., x_{i7})^T$.

Case 1: The seven covariates are independent and identically distributed (i.i.d) from the standard uniform distribution, namely, $x_{ij} \overset{\text{i.i.d}}{\sim} U([0,1])$, for $j = 1, ..., 7$.

Case 2: The first two covariates are highly correlated. Specifically, $x_{ij} \overset{\text{i.i.d}}{\sim} U([0,1])$, for all j except that $x_{i2} = x_{i1} + \varepsilon_i$, with $\varepsilon_i \overset{\text{i.i.d}}{\sim} U([0, 0.1])$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

For this setup, the correlation coefficient between the first two covariates is about 0.8.

Case 3: This case is the same as the second, except that $\varepsilon_i \overset{\text{i.i.d}}{\sim} U([0,1])$. For this case, the correlation between the first two covariates is close to 0.5.

Case 4: This case is the same as the third, except that $x_{ij} \overset{\text{i.i.d}}{\sim} U([-1,1])$, for $j = 6, 7$. For this case, the bounds for each covariates are not all the same.

We consider both $\tilde{\pi}_i^{\text{mV}}$ and $\tilde{\pi}_i^{\text{mVc}}$, and choose $\delta = 10^{-6}$. For comparison, we also consider uniform subsampling with $\pi_i = 1/n$ for all i, and the leverage subsampling strategy in Ma et al. (2015), in which $\pi_i = h_i / \sum_{j=1}^{n} h_i = h_i/p$, with $h_i = \boldsymbol{x}_i(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$. Here h_i is the leverage score for the linear regression. For GLMs, the leverage scores are defined by using the adjusted covariate matrix, namely, $\tilde{h}_i = \tilde{\boldsymbol{x}}_i(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{x}}_i$, where $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}_1, ..., \tilde{\boldsymbol{x}}_n)^T$, $\tilde{\boldsymbol{x}}_i = \sqrt{-E\{\partial^2 \log f(y_i|\tilde{\theta}_i)/\partial\theta^2\}}\boldsymbol{x}_i$, and $\tilde{\theta}_i = \tilde{\boldsymbol{\beta}}^T\boldsymbol{x}_i$, with an initial estimate $\tilde{\boldsymbol{\beta}}_0$ (see Lee, 1987). In this example, simple algebra yields $\tilde{\boldsymbol{x}}_i = \sqrt{\exp(\tilde{\boldsymbol{\beta}}_0^T\boldsymbol{x}_i)}\boldsymbol{x}_i$. For the leverage score subsampling, we considered both h_i and $\tilde{h}_i$. We compare the following methods: UNIF, uniform subsample; mV, $\pi_i = \tilde{\pi}_i^{\text{mV}}$; mVc, $\pi_i = \tilde{\pi}_i^{\text{mVc}}$; Lev, leverage sampling based

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

on h_i; and Lev-A, adjusted leverage sampling based on $\tilde{h}_i$.

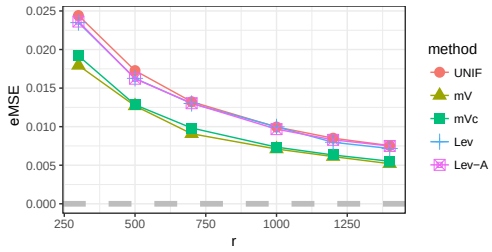
We first consider the case in which the first step sample size is fixed. We let $r_0 = 200$, and the second step sample size r be 300, 500, 700, 1000, 1200, and 1400. When subsampling probabilities that do not depend on unknown parameters, these are implemented with a subsample size $r + r_0$, for fair comparisons.

Figure 1 shows the eMSEs. For all four data sets, the subsampling methods based on $\tilde{\boldsymbol{\pi}}^{\mathrm{mV}}$ and $\tilde{\boldsymbol{\pi}}^{\mathrm{mVc}}$ always result in a smaller eMSE than that of the uniform subsampling, which agrees with the theoretical result that they aim to minimize the asymptotic eMSEs of the resultant estimator. If the components of $\boldsymbol{x}$ are independent, $\tilde{\boldsymbol{\pi}}^{\mathrm{mV}}$ and $\tilde{\boldsymbol{\pi}}^{\mathrm{mVc}}$ exhibit similar performance. However, they may perform differently if some covariates are highly correlated because $\tilde{\boldsymbol{\pi}}^{\mathrm{mVc}}$ reduces the impact of the data correlation structure, because we replaced $\|\tilde{\mathcal{J}}_X^{-1}\boldsymbol{x}_i\|^2$ in $\tilde{\boldsymbol{\pi}}^{\mathrm{mV}}$ with $\|\boldsymbol{x}_i\|^2$ in $\tilde{\boldsymbol{\pi}}^{\mathrm{mVc}}$.

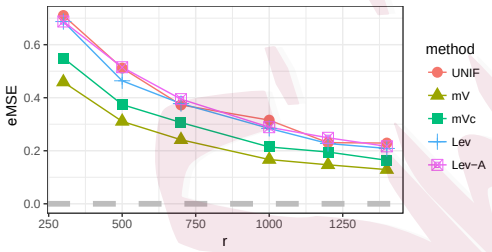
For Cases 1, 3, and 4, the eMSEs are small. This is because the condition number of $\boldsymbol{X}_d$ is quite small (≈ 5), and a small subsample size $r = 100$ produces satisfactory results. However, for Case 2, the condition number is large (≈ 40); therefore, a larger subsample size is needed to approximate $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ accurately. This agrees with the conclusion in Theorem 5.

Theorem 8 also enables inferences on $\boldsymbol{\beta}$. Note that in the subsampling

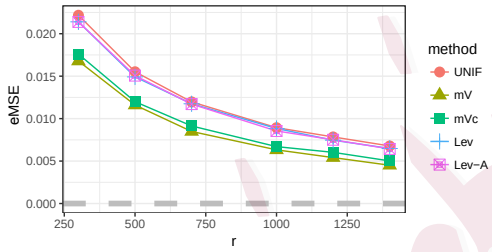
Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439



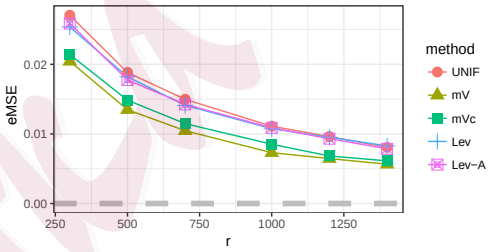
(a) Case 1 (independent covariates)



(b) Case 2 (highly correlated covariates)



(c) Case 3 (weakly correlated covariates)



(d) Case 4 (unequal bounds of covariates)

Figure 1: The eMSEs for a Poisson regression with different second step subsample size r and a fixed first step subsample size $r_0 = 200$. The distributions of the covariates are listed at the beginning of Section 5.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

setting, r is much smaller than the full data size n. If $r = o(n)$, then $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ in Theorem 8 can be replaced by the true parameter. As an example, we take β_2 as a parameter of interest and construct 95% confidence intervals for it. Here the estimator given by $\breve{V} = \breve{\mathcal{J}}_X^{-1}\breve{V}_c\breve{\mathcal{J}}_X^{-1}$ is used to estimate the variance-covariance matrices based on selected subsamples. For comparison, the uniform subsampling method is also implemented.

Table 1 reports the empirical coverage probabilities and average lengths for the Poisson regression model over the four synthetic data sets, with the first step subsample size fixed at $r_0 = 200$. It is clear that $\tilde{\boldsymbol{\pi}}^{\text{mV}}$ and $\tilde{\boldsymbol{\pi}}^{\text{mVc}}$ exhibit similar performance and are uniformly better than the uniform subsampling method. As r increases, the lengths of the confidence intervals decrease uniformly, which echoes the results of Theorem 8. The confidence intervals in Case 2 are longer than those in other cases with the same subsample sizes. This is because the condition number of $\boldsymbol{X}_d$ in Case 2 is bigger than that of $\boldsymbol{X}_d$ in other cases. This indicates that we should select a larger subsample when the condition number of the full data set is bigger, which echoes the results discussed in Section 3.2.

Negative Binomial Regression. Next, we perform a simulation for the negative binomial regression with $n = 100,000$; the results are summarized in Figure 2. Here, we assume $y_i|\boldsymbol{x}_i \sim \text{NB}(\mu_{\text{i}}, \nu)$, where $\mu_i = \exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

Table 1: Empirical coverage probabilities and average lengths of confidence intervals for β_2. The first step subsample size is fixed at $r_0 = 200$.

	method	mV		mVc		UNIF	
	r	Coverage	Length	Coverage	Length	Coverage	Length
case 1	300	0.954	0.2037	0.955	0.2066	0.952	0.2275
	500	0.954	0.1684	0.945	0.1713	0.942	0.1924
	1000	0.946	0.1254	0.938	0.1281	0.953	0.1471
case 2	300	0.961	1.9067	0.946	2.0776	0.950	2.2549
	500	0.958	1.5470	0.948	1.7263	0.947	1.9082
	1000	0.954	1.1379	0.948	1.2919	0.945	1.4559
case 3	300	0.959	0.1770	0.953	0.1816	0.939	0.2000
	500	0.942	0.1451	0.949	0.1507	0.942	0.1693
	1000	0.954	0.1082	0.954	0.1132	0.939	0.1291
case 4	300	0.955	0.2097	0.951	0.2179	0.953	0.2402
	500	0.951	0.1721	0.956	0.1803	0.942	0.2033
	1000	0.957	0.1276	0.960	0.1347	0.943	0.1552

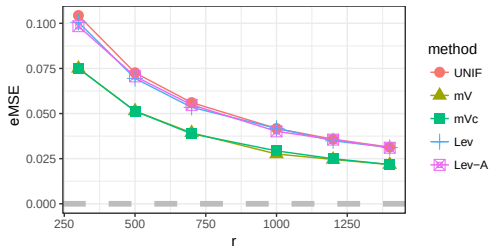
Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

and $\nu = 2$. The other simulation settings are the same as the Poisson regression example. Note that, compared with the Poison regression, the eMSEs are lager for the NBR when r is the same. This agrees with Theorem 5, because $C_{\dot{u}} > 1$ for NBR. The result for the 95% confidence intervals of β_2 are reported in Table 2.

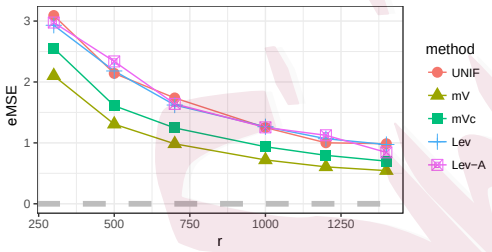
Now, we investigate the effect of different sample size allocations between the two steps. Because the Poisson regression and the NBR exhibit similar performance, we report the results for the Poisson regression only, for brevity. Here, we calculate the eMSEs for various proportions of the first step subsamples, with fixed total subsample sizes. The results are given in Figure 3, with total subsample size $r_0 + r = 800$ and 1200. Because the results are similar in all cases, we present the results for Case 4 only. Note that the two-step method outperforms the uniform subsampling method in all four cases, for both the Poisson regression and the NBR, when $r_0/r \in [0.1, 0.9]$. This indicates that the two-step approach is more efficient than the uniform subsampling. The two-step approach works best when r_0/r is around 0.2.

To explore the influence of δ in $\tilde{\pi}_i^{\mathrm{mV}}$ and $\tilde{\pi}_i^{\mathrm{mVc}}$, we calculate the eMSEs for various δ, ranging from 10^{-6} to 1, with fixed total subsample sizes. Because the results for the Poisson regression and the NBR are similar,

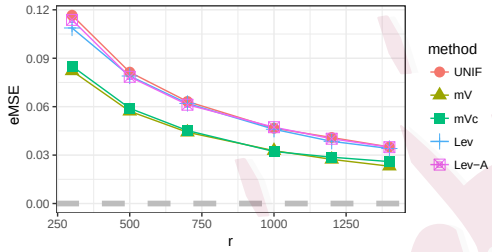
Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439



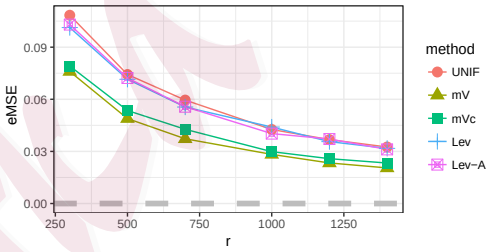
(a) Case 1 (independent covariates)



(b) Case 2 (highly correlated covariates)



(c) Case 3 (weakly correlated covariates)



(d) Case 4 (unequal bounds of covariates)

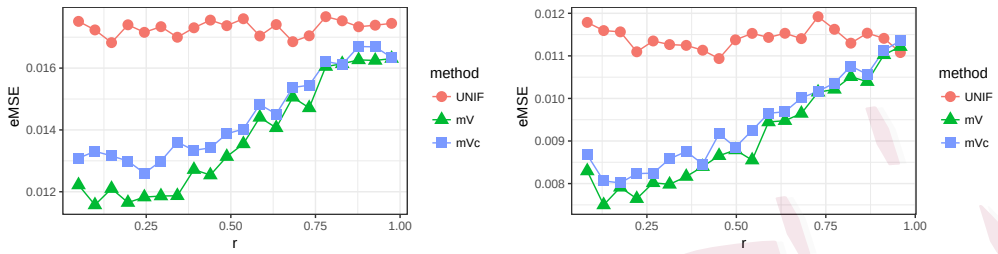
Figure 2: The eMSEs for the NBR with different second step subsample size r and a fixed first step subsample size $r_0 = 200$. The distributions of the covariates are listed at the beginning of Section 5.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

Table 2: Empirical coverage probabilities and average lengths of the confidence intervals for β_2 in the NBR with $\nu = 2$. The first step subsample size is fixed at $r_0 = 200$.

	method	mV		mVc		UNIF	
	r	Coverage	Length	Coverage	Length	Coverage	Length
case1	300	0.952	0.2122	0.955	0.2147	0.947	0.2354
	500	0.952	0.1758	0.954	0.1776	0.946	0.1991
	1000	0.951	0.1305	0.933	0.1331	0.940	0.1520
case2	300	0.947	2.0228	0.963	2.2160	0.943	2.3913
	500	0.953	1.6468	0.952	1.8423	0.946	2.0225
	1000	0.957	1.2065	0.947	1.3849	0.942	1.5439
case3	300	0.950	0.1878	0.950	0.1925	0.942	0.2110
	500	0.949	0.1546	0.954	0.1595	0.944	0.1786
	1000	0.953	0.1150	0.957	0.1197	0.943	0.1361
case4	300	0.956	0.2288	0.953	0.2366	0.953	0.2573
	500	0.968	0.1876	0.963	0.1956	0.936	0.2176
	1000	0.950	0.1396	0.952	0.1469	0.940	0.1662

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439



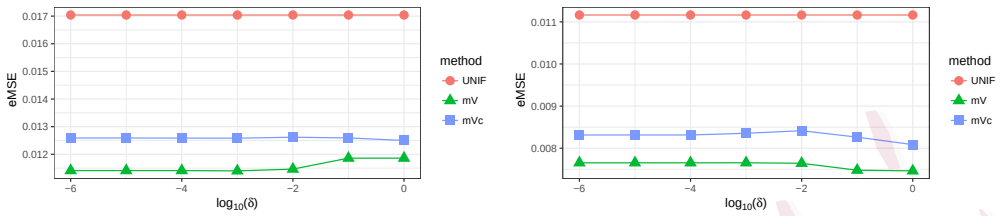
(a) Case 4 ($r_0 + r = 800$) (b) Case 4 ($r_0 + r = 1200$)

Figure 3: The eMSEs vs. the proportions of the first step subsample, with fixed total subsample sizes $r + r_0$, in the Poisson regression.

we report the results for the Poisson regression only. Figure 4 presents the results for Case 4, with a total subsample size $r_0 + r = 800$ and 1200. JFigure 4 shows that the eMSE is not sensitive to the choice of δ when δ is not large, say $\delta = 1$.

To evaluate the computational efficiency of the subsampling strategies, we record the computing time of each (uniform, π^{mV}, π^{mVc}, leverage score and adjusted leverage score), using the `Sys.time()` function in R to record the start and end times. Each subsampling strategy is evaluated 50 times. All methods are implemented in the R programming language. Computations are performed on a desktop computer running Windows 10, with an Intel I7 processor and 32 GB memory. Table 3 shows the results for Case 4

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439



(a) Case 4 ($r_0 + r = 800$) (b) Case 4 ($r_0 + r = 1200$)

Figure 4: The eMSEs vs δ ranging from 10^{-6} to 1 with fixed total subsample sizes $r + r_0$ in Poisson regression. Logarithm is taken on δ for better presentation.

with different r and a fixed $r_0 = 400$. The computing time for the full data set is also given for comparison.

It is not surprising to observe that the uniform subsampling algorithm requires the least computing time, because it does not require an additional step to calculate the subsampling probability. The algorithm based on $\boldsymbol{\pi}^{\mathrm{mV}}$ requires a longer computing time than that of the algorithm based on $\boldsymbol{\pi}^{\mathrm{mVc}}$, which agrees with the theoretical analysis in Section 4. The leverage score sampling takes nearly as long as the mV method, because the leverage scores are computed directly, by definition. Note that $p = 7$ is not sufficiently large to use the fast computing method of Drineas et al. (2011). For fairness, we also consider the case with $p = 80$, and $n = 100,000$, for which it is suitable

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

to use the fast computing method for the Lev and Lev-A methods. The first seven variables are generated as in Case 4, and the rest are generated independently from $U([0,1])$. Here, r_0 is also selected as 400, and the corresponding results are reported in Table 5. In order to see the estimation effects, we also present the eMSEs in Tables 4 and 6.

From Table 5, it is clear that the subsampling algorithms all take significantly less computing time than the full data approach does. The Lev and Lev-A methods are faster than the mV method because the fast algorithm runs in $O(pn\log n)$ time to obtain the subsampling probabilities, as opposed to the $O(p^2n)$ time required by the mV method. However, the mVc method is faster than the Lev and Lev-A methods, because the time complexity is just $O(pn)$ when computing the subsampling probabilities. As the dimension increases, the computational advantage of $\boldsymbol{\pi}^{\mathrm{mVc}}$ becomes even more significant.

5.2 Real Data Studies

In the following, we demonstrate the methods described in Section 4 by applying them to a data set from musicology. This data set contains 1,019,318 unique users' music play counts in the Echo Nest, which is available at http://labrosa.ee.columbia.edu/millionsong/tasteprofile. One of the chal-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

Table 3: Computing time (in seconds) for the Poisson regression in Case 4, with different r and fixed $r_0 = 400$.

r	FULL	UNIF	mV	mVc	Lev	Lev-A
1000	0.187	0.003	0.020	0.016	0.024	0.031
1500	0.195	0.005	0.022	0.017	0.022	0.033
2000	0.193	0.007	0.021	0.018	0.026	0.036
2500	0.194	0.004	0.027	0.022	0.024	0.036

Table 4: Empirical MSE for the Poisson regression in Table 3. The numbers in parentheses are standard errors.

r	UNIF	MV	MVc	Lev	Lev-A
1000	0.0091 (0.0065)	0.0064 (0.0041)	0.0088 (0.0051)	0.0088 (0.0065)	0.0095 (0.0068)
1500	0.0071 (0.0054)	0.0047 (0.0034)	0.0049 (0.0038)	0.0067 (0.0049)	0.0070 (0.0051)
2000	0.0056 (0.0043)	0.0037 (0.0026)	0.0040 (0.0031)	0.0054 (0.0041)	0.0054 (0.0040)
2500	0.0045 (0.0032)	0.0030 (0.0021)	0.0033 (0.0025)	0.0044 (0.0034)	0.0047 (0.0036)

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

Table 5: Computing time (in seconds) for the Poisson regression with $n = 100,000$, dimension $p = 80$, different values of r, and a fixed $r_0 = 400$.

r	FULL	UNIF	mV	mVc	Lev	Lev-A
1000	11.738	0.129	0.638	0.218	0.475	0.557
1500	11.659	0.163	0.689	0.253	0.514	0.595
2000	11.698	0.203	0.725	0.296	0.552	0.637
2500	12.005	0.240	0.777	0.339	0.602	0.681

Table 6: Empirical MSE for the Poisson regression in Table 5. The numbers in parentheses are standard errors.

r	UNIF	MV	MVc	Lev	Lev-A
1000	0.1003 (0.0174)	0.0786 (0.0135)	0.0782 (0.0136)	0.1011 (0.0172)	0.1021 (0.0192)
1500	0.0729 (0.0121)	0.0582 (0.0100)	0.0579 (0.0101)	0.0722 (0.0125)	0.0732 (0.0127)
2000	0.0562 (0.0095)	0.0472 (0.0085)	0.0470 (0.0085)	0.0565 (0.0094)	0.0577 (0.0099)
2500	0.0466 (0.0079)	0.0392 (0.0070)	0.0395 (0.0067)	0.0463 (0.0078)	0.0471 (0.0078)

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

lenges with this data set is to build a music recommendation system. As a basic step, it is interesting to predict the play counts using the song information collected in the Million Song Dataset (Bertin-Mahieux et al., 2011). Because the major mode and minor mode usually express different feelings, the play counts may perform differently under the two modes. Thus, we only focus on the major mode in this example. In addition to the mode of the music, the following six features are selected to describe the song characteristics: x_1, the duration of the track; x_2, the overall loudness of the song; x_3, tempo in BPM; x_4, artist hotnesss; x_5, song hotnesss; x_6, the hotness of the album, which is selected as the maximum value of the song hotnesss in the album. Here, x_1, x_2, and x_3 are features of a specified song, and x_4, x_5, and x_6 are features of the artist, audience, and album, respectively. The last three features are subjective assessments by The Echo Nest, and all are expressed on a scale between zero and one. Because the first three variables in the data set are on different scales, we normalize them first. In addition, we drop the `NA` values in the data set. After cleaning the data, we have $n = 205,032$ data points. As a first attempt to capture the relationship between the play counts and all regressors described above, we fit the basic Poisson regression model, and the results are shown in Figure 5a.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

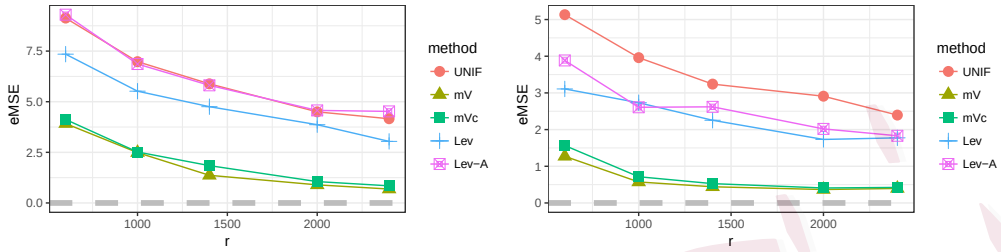
Another way of modeling count data is to use a NBR. For comparison, we also report the results from a NBR in Figure 5b, with the size parameter set as $\theta = 1.4$, which indicates overdispersion of the data.

Similarlly to the synthetic data sets, we compare our method with the uniform subsampling and leverage score subsampling methods, and report the results for r varying from 600 to 2800. The empirical MSEs are reported in Figure 5. It is clear that as r increases, the eMSE decreases quickly for all methods. Moreover, $\boldsymbol{\pi}^{\mathrm{mV}}$ and $\boldsymbol{\pi}^{\mathrm{mVc}}$ perform similarly, and are uniformly better than the uniform subsampling and leverage score subsampling methods for larger values of r. Note that the eMSE in the NBR is less than the Poisson regression. This may because the ratio of the squared Winsorized mean to the Winsorized variance of $\boldsymbol{y}$ is around 1.4, which implies the data is overdispersed. This echoes the results in Theorem 5, which advise us to include additional subsamples to improve the goodness of fit.

Supplementary Material

All technical proofs and additional simulation results are included in the online Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439



(a) Poisson Regression (b) Negative Binomial Regression

Figure 5: Empirical MSEs for different second step subsample size r with the first step subsample size being fixed at $r_0 = 400$.

Acknowledgements

The authors sincerely thank the editor, associate editor, and two referees for their valuable comments. The authors would like to thank Prof. Jinzhu Jia for his helpful suggestions and discussions. Ai's work was supported by NNSF of China grants 11671019 and LMEQF. Wang's work was partially supported by NSF grant 1812013 and an UConn REP grant.

References

Bertin-Mahieux, T., D. P. Ellis, B. Whitman, and P. Lamere (2011). The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

Chapman, W. L., W. J. Welch, K. P. Bowman, J. Sacks, and J. E. Walsh (1994). Arctic sea ice variability: Model sensitivities and a multidecadal simulation. *Journal of Geophysical Research Oceans 99*, 919–935.

Clémencon, S., P. Bertail, and E. Chautru (2014). Scaling up m-estimation via sampling designs: The horvitz-thompson stochastic gradient descent. In *Big Data (Big Data), 2014 IEEE International Conference on*, pp. 25–30. IEEE.

Czado, C. and A. Munk (2000). Noncanonical links in generalized linear models – when is the effort justified? *Journal of Statistical Planning and Inference 87*, 317 – 345.

Drineas, P., M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff (2011). Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research 13*, 3475–3506.

Drineas, P., M. W. Mahoney, and S. Muthukrishnan (2006). Sampling algorithms for l_2 regression and applications. *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, 1127–1136.

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

Efron, B. and T. Hastie (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. New York: Cambridge University Press.

Fahrmeir, L. and H. Kaufmann (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics 13*, 342–368.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

Hansen, M. H. and W. N. Hurwitz (1943). On the theory of sampling from finite populations. *Annals of the Rheumatic Diseases 14*, 2111–2118.

Jia, J., M. Michael, D. Petros, and Y. Bin (2014). Influence sampling for generalized linear models. In *Workshop Presentation: MMDS*.

Lee, A. H. (1987). Diagnostic displays for assessing leverage and influence in generalized linear models. *Australian Journal of Statistics 29*, 233–243.

Loeppky, J. L., J. Sacks, and W. J. Welch (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics 51*, 366–376.

Ma, P., M. W. Mahoney, and B. Yu (2015). A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research 16*, 861–919.

Ma, P. and X. Sun (2015). Leveraging for big data regression. *Wiley Interdisciplinary Reviews: Computational Statistics 7*, 70–76.

Mahoney, M. W. (2012). Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning 3*, 647–672.

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

Pukelsheim, F. (2006). *Optimal design of experiments*. Philadelphia: Society for Industrial and Applied Mathematics.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

Foundation for Statistical Computing.

Särndal, C. E., B. Swensson, and J. Wretman (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

Wang, C., M.-H. Chen, E. Schifano, J. Wu, and J. Yan (2016). Statistical methods and computing for big data. *Statistics and its interface 9*, 399.

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*, 393–405.

Wang, H., (2019). Divide-and-Conquer Information-Based Optimal Subdata Selection Algorithm. *Journal of Statistical Theory and Practice*, DOI: 10.1007/s42519-019-0048-5.

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

Yao, Y. and H. Wang (2019). Optimal subsampling for softmax regression. *Statistical Papers 60*, 235–249.

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

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

School of Mathematics and Statistics, Beijing Institute of Technology

E-mail: yujunbeta@bit.edu.cn

Department of Statistics, University of Connecticut

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0439

E-mail: haiying.wang@uconn.edu