Statistica Sinica Preprint No: SS-2023-0274	
Title	Nearly Optimal Two-step Poisson Sampling and Empirical Likelihood Weighting Estimation for M-estimation with Big Data
Manuscript ID	SS-2023-0274
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0274
Complete List of Authors	Yan Fan, Yang Liu, Yukun Liu and Jing Qin
Corresponding Authors	Yukun Liu
E-mails	ykliu@sfs.ecnu.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

Statistica Sinica

NEARLY OPTIMAL TWO-STEP POISSON SAMPLING AND EMPIRICAL LIKELIHOOD WEIGHTING ESTIMATION FOR M-ESTIMATION WITH BIG DATA

Yan Fan [a], Yang Liu [b], Yukun Liu [c], and Jing Qin [d]

[a] *School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai, China*

[b] *School of Mathematical Sciences, Soochow University, Suzhou, China*

[c] *KLATASDS - MOE, School of Statistics, East China Normal University, Shanghai, China*

[d] *National Institute of Allergy and Infectious Diseases, National Institutes of Health, Frederick, USA*

Abstract:

Subsampling techniques can effectively reduce the computational costs of processing big data. Practical subsampling plans typically involve initial uniform sampling and refined sampling. Subsample-based big data inferences are generally built on the inverse probability weighting (IPW), which may be unstable and cannot incorporate auxiliary information. In this paper, we consider a two-step Poisson sampling, which combines an initial uniform sampling with a second Poisson sampling. Under this sampling plan, we propose an empirical likelihood weighting (ELW) estimation approach to an M-estimation parameter, and then

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

construct a nearly optimal two-step Poisson sampling plan based on the ELW method to improve estimation efficiency of IPW-based optimal subsamplings. Further, we derive methods for determining the smallest sample sizes with which the proposed sampling-and-estimation method produces estimators of guaranteed precision. Our ELW method overcomes the instability of IPW by circumventing the use of inverse probabilities, and utilizes auxiliary information including the size and certain sample moments of big data. We show that the proposed ELW method produces more efficient estimators than IPW, leading to more efficient optimal sampling plans and more economical sample sizes for a prespecified estimation precision. These advantages are confirmed through real data based simulations.

Key words and phrases: Big data; Two-step Poisson sampling; Empirical likelihood

1. Introduction

One of the most significant features of big data is its incredibly large volume, which poses serious challenges to its timely processing. Data analytics need to be performed efficiently so that the results are made available to users in a cost-effective and timely manner. A popular and efficient strategy for solving this problem is to draw small-scale subsamples from the big data (original sample) and make statistical inferences based on the subsamples (Drineas *et al.*, 2006, 2011). Compared with the original big data, the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

subsamples are usually much smaller, and so subsample-based inferences significantly reduce the required computational resources.

Subsample-based inferences for big data generally involve two fundamental issues: how to draw an effective subsample and how to make efficient statistical inferences based on the subsample. Regarding the first issue, it is generally accepted that carefully designed sampling probabilities make unequal probability samplings more efficient than simple random or uniform sampling. Many researchers have developed efficient or optimal sampling plans for frequently encountered parametric statistical problems, including linear regression models (Ma *et al.*, 2014), logistic regression (Fithian and Hastie, 2014; Wang et al., 2018; Wang, 2019), softmax regression (Yao et al. , 2023), generalized linear models (Ai *et al.*, 2021b), quantile regression (Ai *et al.*, 2021a; Fan *et al.*, 2021; Wang and Ma, 2021), and more general models (Shen *et al.*, 2021; Yu *et al.*, 2022).

For the second issue, subsample-based statistical inferences for big data are usually performed through inverse probability weighting (IPW), which leads to the Hansen–Hurwitz estimator (Hansen and Hurwitz, 1943) under sampling with replacement and to the Horvitz–Thompson estimator (Horvitz and Thompson, 1952) under sampling without replacement. However, the subsample-based IPW estimation procedure for big data

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

analysis suffers from two weaknesses. First, the IPW estimator can be highly unstable if there are extremely small probabilities, resulting in poor finite-sample performance of the accompanying asymptotic-normality-based inferences (Kang and Schafer, 2007); see the simulation results in the missing data context in Han (2014, 2016) and Chen and Haziza (2017). Second, the current IPW-based subsampling techniques do not incorporate auxiliary information to improve estimation efficiency. For example, the sample mean of some variables in a big dataset can be quickly calculated at little computational cost; this can be taken as auxiliary information when inferences are made based on a subsample.

In the case of big data, the optimal sampling depends on the statistical problem under study and the accompanying subsample-based estimation procedure. To consider both the generality and convenience of theoretical analysis and implementation, we focus on M-estimation problems with convex loss functions, and consider the use of Poisson sampling. Under Poisson sampling, samples are drawn independently according to Bernoulli experiments with prespecified success probabilities. Poisson sampling never draws replicate observations and its implementation is free from memory constraints (Yao and Wang, 2021). Popular examples of M-estimation problems with convex loss functions include linear regression,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

quantile regression, and many generalized linear regressions. The sampling probabilities of the ideal optimal samplings depend on the ideal parameter estimator from the big data itself. For the optimal sampling to be practically applicable, an initial sample is required to produce an initial estimate of the parameter of interest.

In this paper, we consider Poisson samplings for both the two samplings, and regard the whole sampling procedure as a two-step Poisson sampling. We makes three contributions to the literature of subsample-based big data analysis.

1. First, we develop an empirical likelihood weighting (ELW) estimation method for a two-step Poisson sample from big data, incorporating auxiliary information defined by estimating equations. The proposed estimation procedure not only overcomes the instability of the IPW by circumventing the use of inverse probabilities, but also achieves enhanced efficiency by incorporating auxiliary information. We show that, in theory, the proposed ELW estimator is asymptotically more efficient than the IPW estimator.

2. Second, we construct a nearly optimal two-step Poisson sampling plan by minimizing the upper bound of the asymptotic mean square error (MSE) of the proposed ELW estimator. The sample from the first

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

step is used to estimate the subsampling probabilities in the second step.

3. Third, we determine the minimal sample size needed so that the proposed nearly optimal sampling plan achieves the desired precision requirement in terms of MSE and absolute error. As the ELW estimator is more efficient than the IPW estimator, the proposed nearly optimal two-step Poisson sampling is expected to outperform existing optimal IPW-based subsampling plans.

The remainder of this paper is organized as follows. In Section 2, we introduce the ELW estimation procedure with auxiliary information under a general two-step Poisson sampling plan, and study the asymptotic behavior of the ELW estimator. In Section 3, we construct a nearly optimal two-step Poisson sampling plan and discuss its practical implementation. In Section 4, we derive the minimal sample size needed for the proposed estimator to meet a prespecified precision. Real data based simulation studies are reported in Section 5. Section 6 concludes with a discussion. All technical proofs are given in the supplementary material for clarity.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

2. Empirical likelihood weighting estimation

2.1 Setup and IPW

Suppose that the big data consist of N observations, $Z_1, \ldots, Z_N$, which are independent and identically distributed (i.i.d.) copies from a population Z with an unknown cumulative distribution function F. Parametric models indexed by a q-dimensional parameter θ are usually imposed to extract information from data. Let $\ell(z, \theta)$ be a user-specific convex loss function that quantifies the lack-of-fit of a parametric model indexed by a parameter θ based on an observation z. The average loss or risk function is $R(\theta) = \mathbb{E}\{\ell(Z, \theta)\} = \int \ell(z, \theta) dF(z)$. We define the parameter of interest θ_0 to be the risk minimizer (Shen *et al.*, 2021)

$$\theta_0 = \arg\min_{\theta} R(\theta). \tag{2.1}$$

This setup includes many common problems as special cases. When Z is a scalar, the true parameter value θ_0 is the mean or median of Z if $\ell(z, \theta) = (z-\theta)^2$ or $|z-\theta|$. When $Z = (Y, X^\top)^\top$, θ_0 may be the population-level regression coefficient in the generalized linear regression, least-squares regression, quantile regression, and expectile regression models under the specification of $\ell(z; \theta)$ given in Table 1 of the supplementary material.

Based on the big-data observations, $\hat{\theta}_N = \arg\min_\theta \sum_{i=1}^N \ell(Z_i, \theta)$ is

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

the ideal estimator of θ. For massive datasets, N can be so large that the direct calculation of $\hat{\theta}_N$ is formidable or practically infeasible. Subsampling techniques then come into play to reduce the computation costs. As discussed in the introduction, we consider the use of two-step Poisson sampling, where the first step is a Poisson sampling with an equal sampling probability and the second step is another Poisson sampling, but with generally unequal sampling probabilities. Let the unequal sampling probabilities in the second step be $\pi_i = \pi(Z_i)$, $i = 1, \ldots, N$, for some function $\pi(\cdot)$. The ideal sample sizes for both the Poisson samplings in the two-step Poisson sampling plan, n_{10} and $n_{20} = \sum_{i=1}^{N} \pi_i$, must be specified beforehand.

Since no information is available about the big data, in the first step, we choose to conduct a Poisson sampling with inclusion probabilities all equal to $\alpha_{10} = n_{10}/N$. For $1 \leq i \leq N$, denote the sampling result for Z_i as D_{i1}, which is equal to 1 for success and 0 otherwise. Datum Z_i is sampled in the first step if and only if $D_{i1} = 1$. The sample in the first step is used to produce an initial estimate of θ, which is then employed to determine the sampling probabilities in the second step. For now, we assume that the π_i are known. In the second step, we again conduct a Bernoulli experiment, but with success probability π_i for datum Z_i, and denote the result as D_{i2};

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

datum Z_i is sampled if and only if $D_{i2} = 1$. Finally, the resulting two-step Poisson sample can be written as $\{(D_i Z_i, D_{i1}, D_{i2}),\ i = 1, 2, \ldots, N\}$, where $D_i = I(D_{i1} + D_{i2} > 0)$ and $I(\cdot)$ is the indicator function.

Assumption 1. The N random vectors (Z_i, D_{i1}, D_{i2}) $(i = 1, \ldots, N)$ are i.i.d. copies of $(Z, D_{(1)}, D_{(2)})$. Suppose that the distribution $F(z)$ of Z is nondegenerate, $\mathbb{E}(D_{(1)}|Z) = \mathbb{E}(D_{(1)}) = \alpha_{10}$, $\mathbb{E}(D_{(2)}|Z) = \pi(Z)$, and $\alpha_{20} = \mathbb{E}(D_{(2)}) = \mathbb{E}\{\pi(Z)\}$.

Let $D = I(D_{(1)} + D_{(2)} > 0)$, where $D_{(1)}$ and $D_{(2)}$ are as defined in Assumption 1. Then, $\mathbb{E}(D) = 1 - \{1 - \mathbb{E}(D_1)\}\{1 - \mathbb{E}(D_2)\} = 1 - (1 - \alpha_{10})(1-\alpha_{20})$. For a given datum Z, the overall probability of being sampled is $\varphi(Z) = \mathbb{E}(D \mid Z) = 1 - (1 - \alpha_{10})\{1 - \pi(Z)\}$ under Assumption 1. Based on the two-step Poisson sample, the IPW estimator of θ is

$$\hat{\theta}_{\mathrm{IPW}} = \arg\min_{\theta} \hat{R}_{\mathrm{IPW}}(\theta) \equiv \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \frac{D_i}{\varphi(Z_i)} \ell(Z_i, \theta), \tag{2.2}$$

where $\hat{R}_{\mathrm{IPW}}(\theta)$ is the IPW estimator of the risk function $R(\theta)$.

Assumption 2. Suppose that $\ell(z, \theta)$ is a loss function that is convex with respect to θ, and that $\ell(z, \theta_0 + t) = \ell(z, \theta_0) + \dot{\ell}(z)^\top t + \xi(z, t)$ holds in a neighborhood of $t = 0$. Here, the function $\dot{\ell}(z)$ satisfies $\mathbb{E}\{\dot{\ell}(Z)\} = 0$ and $B_{\dot{\ell}\dot{\ell}} = \mathbb{E}\{\dot{\ell}(Z)\dot{\ell}^\top(Z)/\varphi(Z)\}$ is finite, and $\xi(z, t)$ satisfies $\mathbb{E}\{\xi(Z, t)\} =$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

$(1/2)t^\top Vt + o(\|t\|^2)$ and $\mathbb{E}\{\xi^2(Z,t)\} = o(\|t\|^2)$ for a positive-definite matrix V as $\|t\| \to 0$, where $\|\cdot\|$ denotes the Euclidean norm.

Assumption 2 is satisfied by many commonly-used regression models, such as the generalized linear regression, least-squares regression, quantile regression, and expectile regression models. Lemma 1 shows the asymptotic normality of $\hat{\theta}_{\mathrm{IPW}}$, which has been established under various settings; see, e.g., Wang et al. (2018), Ai *et al.* (2021a) and Shen *et al.* (2021).

Lemma 1. *Suppose that Assumptions 1 and 2 are satisfied and that* $\alpha_{10}, \alpha_{20} \in (0,1)$ *are fixed quantities. As* N *goes to infinity,* $\sqrt{N}(\hat{\theta}_{\mathrm{IPW}} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\mathrm{IPW}})$*, where* $\xrightarrow{d}$ *denotes "converges in distribution to" and* $\Sigma_{\mathrm{IPW}} = V^{-1} B_{\dot{\ell}\dot{\ell}} V^{-1}$.

As discussed in the introduction, if some probabilities $\varphi(Z_i)$ are too close to zero, $\hat{R}_{\mathrm{IPW}}(\theta)$ exhibits remarkable instability, making the resulting IPW estimator $\hat{\theta}_{\mathrm{IPW}}$ in (2.2) undesirably unstable. In the context of big data analysis, auxiliary information is often available. For example, the response mean of a big data sample can often be quickly calculated with little extra effort, and can be regarded as auxiliary information in subsample-based analysis. However, the estimation efficiency of the IPW method cannot be enhanced by incorporating auxiliary information. Based on optimal estimating function theory (Godambe, 1960), the score function

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

derived from the complete-data likelihood is optimal in the class of inverse weighting estimating functions (Qin, 2017, Section 5.2). This motivates us to consider the full-likelihood-based inference approach under the two-step Poisson sampling.

2.2 ELW estimation under two-step Poisson sampling

Given data $\{(D_i Z_i, D_{i1}, D_{i2}),\ i = 1, 2, \ldots, N\}$, the full likelihood is

$$\binom{N}{n} \prod_{i=1}^{N} \left[\{\varphi(Z_i) dF(Z_i)\}^{D_i} \cdot (1-\alpha)^{1-D_i}\right], \tag{2.3}$$

where $n = \sum_{i=1}^{N} D_i$ and $\alpha = \mathbb{E}(D) = \int \varphi(z) dF(z)$ is the marginal probability of observing a value of Z. The true value of α is $\alpha_0 = 1 - (1-\alpha_{10})(1-\alpha_{20})$ under Assumption 1. We use the EL method (Owen, 1988) to handle $F(z)$. Letting $p_i = dF(Z_i)$, the full log-likelihood becomes

$$\sum_{i=1}^{N} [D_i \log(p_i) + D_i \log\{\varphi(Z_i)\} + (1 - D_i) \log(1-\alpha)], \tag{2.4}$$

where the feasible p_i satisfy $p_i \geq 0$, $\sum_{i=1}^{N} p_i = 1$, and

$$\sum_{i=1}^{N} p_i \{\varphi(Z_i) - \alpha\} = 0. \tag{2.5}$$

The previous equation follows from $\alpha = \int \varphi(z) dF(z)$. The Z_i with $D_i = 0$ are not observed. Although appearing in the expression of the above likelihood, they do not actually contribute to the likelihood. The expression

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

of the empirical log-likelihood implies that only those p_i with $D_i = 1$ make a contribution to the likelihood.

If we take α to be an unknown parameter, the maximum point of (2.4) under the constraints $p_i \geq 0$, $\sum_{i=1}^{N} p_i = 1$, and (2.5) is always well defined (Liu and Fan, 2023) if there are at least two different values in $\{\varphi(Z_i) : D_i = 1, i = 1, 2, \ldots, N\}$ (or, equivalently, $\{\pi(Z_i) : D_i = 1, i = 1, 2, \ldots, N\}$). Liu and Fan (2023) took the resulting p_i, say $\tilde{p}_i$, as the weights and proposed a biased-sample EL weighting estimation method that serves the same purpose as IPW, but overcomes the problem of instability.

Under the two Poisson samplings in the two-step Poisson sampling, the true parameter values α_{10} and α_{20} need to be prespecified prior to their implementation, so that $\alpha_0 = 1 - (1 - \alpha_{10})(1 - \alpha_{20})$ is known a priori. Unlike Liu and Fan (2023), we make full use of this and other auxiliary information to improve the efficiency of the resulting point estimator of θ. The feasible p_i should satisfy

$$\sum_{i=1}^{N} p_i\{\varphi(Z_i) - \alpha_0\} = 0. \tag{2.6}$$

In addition, for massive datasets, although solving the optimization problem $\min \sum_{i=1}^{N} \ell(Z_i, \theta)$ is complicated and time-consuming, the big data sample mean $\sum_{i=1}^{N} Z_i/N$ or other sample moments can be calculated relatively easily. This can be taken as auxiliary information when we make statistical

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

inferences about the big data based on a subsample. Suppose that $\bar{h} = (1/N)\sum_{i=1}^{N} h(Z_i)$ is available for some function h, which may be vector-valued. For convenience, we assume that $\mathbb{E}\{h(Z)\} = 0$ is known. This can be formulated as one more estimating equation:

$$\sum_{i=1}^{N} p_i h(Z_i) = 0. \tag{2.7}$$

In summary, we recommend estimating the p_i by the maximum EL estimator, which is the maximizer of the empirical log-likelihood (2.4) under the constraints $p_i \geq 0$, $\sum_{i=1}^{N} p_i = 1$, (2.6), and (2.7). By the Lagrange multiplier method, we have

$$\hat{p}_i = \frac{1}{\sum_{j=1}^{N} D_j} \cdot \frac{D_i}{1 + \hat{\lambda}^\top h_e(Z_i)}, \tag{2.8}$$

where $h_e(Z) = (\varphi(Z) - \alpha_0,\ h^\top(Z))^\top$ and $\hat{\lambda}$ is the solution to $\sum_{i=1}^{N} D_i h_e(Z_i)/\{1 + \hat{\lambda}^\top h_e(Z_i)\} = 0$. There is a close relationship between $\hat{p}_i$ and the IPW weights $D_i/\{N\varphi(Z_i)\}$, namely

$$\hat{p}_i = \frac{D_i}{N\varphi(Z_i)} + \frac{D_i}{N\varphi(Z_i)} \left\{ O_p(N^{-1/2}) + \frac{\|h_e(Z_i)\|}{\varphi(Z_i)} O_p(N^{-1/2}) \right\}.$$

See our proof of Theorem 1.

Given $\hat{p}_i$, we propose to estimate θ by the ELW estimator

$$\hat{\theta}_{\mathrm{ELW}} = \arg\min_{\theta} \hat{R}_{\mathrm{ELW}}(\theta) \equiv \arg\min_{\theta} \sum_{i=1}^{N} D_i \hat{p}_i \ell(Z_i, \theta) \tag{2.9}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

where $\hat{R}_{\mathrm{ELW}}(\theta)$ is the ELW estimator of the risk function $R(\theta)$. If the loss function $\ell(z,\theta)$ is differentiable with respect to θ for almost all z, an alternative ELW estimator of θ can be obtained by maximizing the empirical log-likelihood (2.4) under the constraints $p_i \geq 0$, $\sum_{i=1}^{N} p_i = 1$ with (2.6), (2.7), and $\sum_{i=1}^{N} p_i \partial\ell(Z_i,\theta)/\partial\theta = 0$. Because the dimensions of θ and $\partial\ell(Z_i,\theta)/\partial\theta$ are the same, the resulting maximum EL estimator is exactly equal to $\hat{\theta}_{\mathrm{ELW}}$.

Theorem 1. *Suppose that Assumptions 1 and 2 hold,* $B_{hh} = \mathbb{E}\{h_e(Z)h_e^{\top}(Z)/\varphi(Z)\}$ *is positive-definite, and* $\alpha_{10}, \alpha_{20} \in (0,1)$ *are fixed and known. As* N *goes to infinity, (a)* $\hat{\theta}_{\mathrm{ELW}}$ *is consistent with* θ_0 *and* $\sqrt{N}(\hat{\theta}_{\mathrm{ELW}} - \theta_0) = -V^{-1} \cdot N^{1/2} \sum_{i=1}^{N} \hat{p}_i \dot{\ell}(Z_i) + o_p(1)$*; (b)* $\sqrt{N}(\hat{\theta}_{\mathrm{ELW}} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\mathrm{ELW}})$ *with* $\Sigma_{\mathrm{ELW}} = V^{-1}(B_{\dot{\ell}\dot{\ell}} - B_{\dot{\ell}h}B_{hh}^{-1}B_{\dot{\ell}h}^{\top})V^{-1}$*, where* $B_{\dot{\ell}h} = \mathbb{E}\{\dot{\ell}(Z)h_e^{\top}(Z)/\varphi(Z)\}$ *and* $B_{\dot{\ell}\dot{\ell}} = \mathbb{E}\{\dot{\ell}(Z)\dot{\ell}^{\top}(Z)/\varphi(Z)\}$*; (c) If the auxiliary information defined by* (2.7) *is ignored, then* $\sqrt{N}(\hat{\theta}_{\mathrm{ELW}} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\mathrm{ELW0}})$*, where* $\Sigma_{\mathrm{ELW0}} = V^{-1}\{B_{\dot{\ell}\dot{\ell}} - (B_{\dot{\ell}1}B_{\dot{\ell}1}^{\top})/(B_{11} - \alpha_0^{-1})\}V^{-1}$ *with* $B_{\dot{\ell}1} = \mathbb{E}\{\dot{\ell}(Z)/\varphi(Z)\}$ *and* $B_{11} = \mathbb{E}\{1/\varphi(Z)\}$.

The results in Theorem 1 hold with $h(Z) = g(Z) - \frac{1}{N}\sum_{i=1}^{N} g(Z_i)$ for any completely known function $g(Z)$. Because $\Sigma_{\mathrm{IPW}} - \Sigma_{\mathrm{ELW}} = V^{-1}B_{\dot{\ell}h}B_{hh}^{-1}B_{\dot{\ell}h}^{\top}V^{-1}$ is nonnegative-definite, the ELW estimator is asymptotically more efficient than the IPW estimator. Roughly speaking,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

$\Sigma_{\mathrm{IPW}} - \Sigma_{\mathrm{ELW}} = V^{-1}B_{\dot{\ell}h}B_{hh}^{-1}B_{\dot{\ell}h}^{\top}V^{-1}$ is the projection of $V^{-1}\dot{\ell}(Z)$ onto the orthogonal complement of the augmentation space (Tsiatis, 2006) consisting of $h_e(Z)$. This indicates that as the dimension of h increases or equivalently more auxiliary information is incorporated, the ELW estimator is more efficient. Even so, its calculation burden becomes heavier. When the dimension of h is fixed, it may be desirable to determine the optimal choice for h by maximizing $V^{-1}B_{\dot{\ell}h}B_{hh}^{-1}B_{\dot{\ell}h}^{\top}V^{-1}$ in some sense. However, the optimal choice of h depends not only on the underlying criterion but also on unknown quantities, making it not necessary or useful in practice. We choose h to be $Y - N^{-1}\sum_{i=1}^{N} Z_i$ in our numerical studies for convenience.

Note that $\Sigma_{\mathrm{IPW}} - \Sigma_{\mathrm{ELW0}} = V^{-1}B_{\dot{\ell}1}B_{\dot{\ell}1}^{\top}V^{-1}/(B_{11} - \alpha_0^{-1})$ is nonnegative-definite because

$$B_{11} - \alpha_0^{-1} = \frac{1}{\alpha_0^2}\mathbb{E}\left[\frac{\{\varphi(Z) - \alpha_0\}^2}{\varphi(Z)}\right] > 0,$$

Therefore, the efficiency gain of the ELW estimator over the IPW estimator remains even if we ignore constraint (2.7), or if no auxiliary information is incorporated in the ELW estimator. It can be verified that $\Sigma_{\mathrm{ELW0}} - \Sigma_{\mathrm{ELW}} = V^{-1}\{B_{\dot{\ell}h}B_{hh}^{-1}B_{\dot{\ell}h}^{\top} - (B_{\dot{\ell}1}B_{\dot{\ell}1}^{\top})/(B_{11} - \alpha_0^{-1})\}V^{-1}$ is nonnegative-definite, which again indicates that incorporating auxiliary information enhances the efficiency of the proposed ELW estimator.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

2.3 The case with negligible sampling fraction

Thus far, we have assumed that the overall sampling fraction of the big data is nonnegligible, i.e. $\alpha_0 \in (0, 1)$. When the volume of the big data is huge, it is reasonable to assume that the sampling fraction may be negligible.

Assumption 3. Suppose that $\pi(z)$ depends on N and is written as $\pi_N(z)$, there exist a positive sequence $\{b_N\}_{N=1}^{\infty}$, a positive function $0 < \pi_*(Z) \leq 1$, and a positive constant α_{1*} such that $b_N \to \infty$, $b_N/N \to 0$, $b_N\pi_N(Z) \to \pi_*(Z)$, and $b_N\alpha_{10} \to \alpha_{1*}$ as $N \to \infty$.

Under Assumption 3, we have $b_N\alpha_{20} = \mathbb{E}\{b_N\pi_N(Z)\} \to \alpha_{2*} = \mathbb{E}\{\pi_*(Z)\}$ as $N \to \infty$. Define $\alpha_0 = \alpha_{10}+\alpha_{20}$ and $\varphi(Z) = \alpha_{10}+\pi_N(Z)$. Then, $b_N\alpha_0$ and $b_N\varphi(Z)$ converge to $\alpha_* = \alpha_{1*} + \alpha_{2*}$ and $\varphi_*(Z) = \alpha_{1*} + \pi_*(Z)$, respectively. Because α_{10} and the $\pi_N(Z_i)$ are prespecified, the log-likelihood (2.4) under Assumption 3, up to a constant not depending on the unknown parameters p_i, is equal to $\sum_{i=1}^{N} D_i \log(p_i)$. The proposed ELW estimator $\hat{\theta}_{\text{ELW}}$ is still defined as (2.9) with $\pi_N(Z)$ in place of $\pi(Z)$.

Theorem 2. *Let $h_{e*}(Z) = (\varphi_*(Z) - \alpha_*, h^\top(Z))^\top$. Suppose that Assumptions 1–3 hold, the distribution of Z is nondegenerate, and that $C_{hh*} = \mathbb{E}\{h_{e*}(Z)h_{e*}^\top(Z)/\varphi_*(Z)\}$ is positive-definite. As N goes to infinity, $\sqrt{N/b_N}(\hat{\theta}_{\text{ELW}} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\text{ELW}*})$ and $\sqrt{N/b_N}(\hat{\theta}_{\text{IPW}} - \theta_0) \xrightarrow{d}$*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

$\mathcal{N}(0, \Sigma_{\mathrm{IPW}*})$, *where* $\Sigma_{\mathrm{ELW}*} = V^{-1}(C_{\dot{\ell}\dot{\ell}*} - C_{\dot{\ell}h*}C_{hh*}^{-1}C_{\dot{\ell}h*}^{\top})V^{-1}$ *and* $\Sigma_{\mathrm{IPW}*} = V^{-1}C_{\dot{\ell}\dot{\ell}*}V^{-1}$ *with* $C_{\dot{\ell}h*} = \mathbb{E}\{\dot{\ell}(Z)h_{e*}^{\top}(Z)/\varphi_*(Z)\}$ *and* $C_{\dot{\ell}\dot{\ell}*} = \mathbb{E}\{\dot{\ell}(Z)\dot{\ell}^{\top}(Z)/\varphi_*(Z)\}$.

Theorem 2 indicates that even as the sampling fraction tends to zero, both the ELW and IPW estimators are consistent at the rate $\sqrt{N/b_N}$, a lower rate than $\sqrt{N}$, and our ELW estimator is still asymptotically more efficient than the IPW estimator. Although the asymptotic results here are slightly different from those in Lemma 1 and Theorem 1, the variances of $\hat{\theta}_{\mathrm{ELW}}$ and $\hat{\theta}_{\mathrm{IPW}}$ can always be approximated by Σ_{ELW}/N and Σ_{IPW}/N, respectively.

3. Optimal two-step Poisson sampling plan

The asymptotic efficiency of subsample-based statistical inferences depends critically on the underlying subsampling plan. Carefully chosen sampling plans can lead to remarkable efficiency gains over uniform sampling, which motivates optimal subsampling for big data.

3.1 Ideal optimal sampling plan

MSE is a popular evaluation criterion for the performance of a point estimator. For a vector-valued parameter θ, the MSE of an estimator $\hat{\theta}$ is defined as $\mathbb{M}\mathrm{se}(\theta) = \mathbb{E}(\|\hat{\theta} - \theta\|^2)$. According to Theorem 1, the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

proposed ELW estimator has no asymptotic bias, therefore its asymptotic MSE is equal to the trace of asymptotic variance. For a constant matrix Q, Theorem 1 implies that N times the asymptotic MSE of $Q\hat{\theta}_{\mathrm{ELW}}$ is approximated by

$$N \times \mathbb{M}\mathrm{se}(Q\hat{\theta}_{\mathrm{ELW}}) \approx \mathbf{tr}\{QV^{-1}(B_{\dot{\ell}\dot{\ell}} - B_{\dot{\ell}h}B_{hh}^{-1}B_{\dot{\ell}h}^{\top})V^{-1}Q^{\top}\}.$$

According to Theorem 2, this approximation still holds when the sampling fraction is negligible. The MSEs with $Q = I$ and V correspond to the A- and L-optimality criteria, respectively. When $Q = V$, the MSE criterion is independent of V, and hence has much practical convenience. However, $Q = I$ is preferred when we are more interested in the efficiency of the ELW estimator itself.

Because $h_e(Z) = (\varphi(Z) - \alpha_0, h^{\top}(Z))^{\top}$, $\mathbb{E}\{\dot{\ell}(Z, \theta_0)\} = 0$, and $\mathbb{E}\{h(Z)\} = 0$, we have

$$B_{\dot{\ell}h} = \mathbb{E}\left\{\frac{\dot{\ell}(Z,\theta_0)(-\alpha_0, h^{\top}(Z))}{\varphi(Z)}\right\}, \quad B_{hh} = \mathbb{E}\left[\frac{\{(-\alpha_0, h^{\top}(Z))^{\top}\}^{\otimes 2}}{\varphi(Z)}\right] - \alpha_0 e_1^{\otimes 2},$$

where e_1 is a unit vector in which the first component is 1. Let $\pi = (\pi_1, \ldots, \pi_N)$ with $\pi_i = \pi(Z_i)$, and $\varphi = (\varphi_1, \ldots, \varphi_N)$, where $\varphi_i = 1 - (1 - \alpha_{10})(1 - \pi_i)$. Given π, natural "estimators" of $B_{\dot{\ell}\dot{\ell}}, B_{\dot{\ell}h}$, and B_{hh} are

$$\hat{B}_{\dot{\ell}\dot{\ell}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\{\dot{\ell}(Z_i,\theta_0)\}^{\otimes 2}}{\varphi_i}, \; \hat{B}_{\dot{\ell}h} = \frac{1}{N}\sum_{i=1}^{N}\frac{\dot{\ell}(Z_i,\theta_0)b_i^{\top}}{\varphi_i}, \; \hat{B}_{hh} = \frac{1}{N}\sum_{i=1}^{N}\frac{b_i^{\otimes 2}}{\varphi_i} - \alpha_0 e_1^{\otimes 2},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

where $b_i = (-\alpha_0, h^\top(Z_i))^\top$ for $i = 1, \ldots, N$. Accordingly, a natural consistent "estimator" of $N \times \mathbb{M}\text{se}(\hat{\theta}_{\text{ELW}})$ is

$$H_*(\varphi) = \mathbf{tr}\{QV^{-1}(\hat{B}_{\dot{\ell}\dot{\ell}} - \hat{B}_{\dot{\ell}h}\hat{B}_{hh}^{-1}\hat{B}_{\dot{\ell}h}^\top)V^{-1}Q^\top\}.$$

Because there is a one-to-one map from π to φ, determining the optimal sampling plan π is equivalent to determining the optimal φ. The optimal φ in terms of parameter estimation accuracy is the solution to

$$\min_{\varphi} H_*(\varphi) \quad \text{s.t.} \quad \sum_{i=1}^{N} \varphi_i = N\alpha_0,\ \alpha_{10} < \varphi_i < 1 \text{ for } i = 1, \ldots, N. \tag{3.1}$$

Unfortunately, there is no closed-form solution to problem (3.1), which makes it impractical and motivates us to derive a nearly optimal sampling plan using several techniques.

Let $a_i(\theta_0) = QV^{-1}\dot{\ell}(Z_i, \theta_0)$. First, we replace problem (3.1) by

$$\min_{\varphi} H(\varphi) \quad \text{s.t.} \quad \sum_{i=1}^{N} \varphi_i = N\alpha_0,\ \alpha_{10} < \varphi_i < 1 \text{ for } i = 1, \ldots, N. \tag{3.2}$$

where

$$H(\varphi) = \frac{1}{N}\sum_{i=1}^{N} \frac{\|a_i(\theta_0)\|^2}{\varphi_i} - \frac{1}{N}\mathbf{tr}\left[\left\{\sum_{i=1}^{N} \frac{a_i(\theta_0)b_i^\top}{\varphi_i}\right\}\left(\sum_{i=1}^{N} \frac{b_i b_i^\top}{\varphi_i}\right)^{-1}\left\{\sum_{i=1}^{N} \frac{b_i a_i^\top(\theta_0)}{\varphi_i}\right\}\right].$$

The fact that $H_*(\varphi) \le H(\varphi)$ holds for any φ implies that a sampling plan with a small $H(\varphi)$ somehow leads to a small $H_*(\varphi)$. The solution to problem (3.2) is a nearly optimal φ and hence produces a nearly optimal sampling plan.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

Second, we transform the optimization problem (3.2) to an equivalent constrained optimization problem. Define $H_1(\varphi, K) = (1/N)\sum_{i=1}^{N} \|a_i(\theta_0) - Kb_i\|^2/\varphi_i$, where K is a matrix of the same dimensions as $a_i(\theta_0)b_i^\top$. Clearly, $H(\varphi) = \min_K H_1(\varphi, K)$ for any fixed φ. Because $H_1(\varphi, K)$ is a convex function of (φ, K), it follows that

$$\min_{\varphi} H(\varphi) = \min_{\varphi} \min_{K} H_1(\varphi, K) = \min_{\varphi, K} H_1(\varphi, K),$$

and that the solution to (3.2) can be obtained by solving

$$\min_{\varphi, K} H_1(\varphi, K) \quad \text{s.t.} \quad \sum_{i=1}^{N} \varphi_i = N\alpha_0,\ \alpha_{10} < \varphi_i < 1 \text{ for } i = 1, \ldots, N. \quad (3.3)$$

If we retain only the equality constraint, then

$$\min_{\varphi, K} H_1(\varphi, K) = \min_{K}\{\min_{\varphi} H_1(\varphi, K)\} = \min_{K} \frac{\{H_2(K)\}^2}{N^2\alpha_0} = \frac{\{\min_K H_2(K)\}^2}{N^2\alpha_0},$$

where $H_2(K) = \sum_{i=1}^{N} \|a_i(\theta_0) - Kb_i\|$. Denote $\hat{K} = \arg\min_K H_2(K)$. In this situation, a nearly optimal φ is $\hat{\varphi} = (\hat{\varphi}_1, \ldots, \hat{\varphi}_N)$ with

$$\hat{\varphi}_i = \alpha_0 \cdot \frac{\|a_i(\theta_0) - \hat{K}b_i\|}{N^{-1}\sum_{j=1}^{N} \|a_j(\theta_0) - \hat{K}b_j\|}. \quad (3.4)$$

Note that $(\hat{\varphi}, \hat{K})$ is generally different from $(\hat{\varphi}_*, \hat{K}_*)$, which is the minimizer of problem (3.3), because the optimization problems with and without the inequality constraint $\alpha_{10} < \varphi_i < 1$ $(i = 1, \ldots, N)$ are not equivalent. From a practical perspective, we propose to take $\hat{K}$ as an approximation of $\hat{K}_*$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

and adopt the optimal sampling plan with φ solving

$$\min_{\varphi} H_1(\varphi, \hat{K}) \quad \text{s.t.} \quad \sum_{i=1}^{N} \varphi_i = N\alpha_0,\ \alpha_{10} < \varphi_i < 1 \text{ for } i = 1, \ldots, N. \tag{3.5}$$

By the Karush–Kuhn–Tucker condition, the solution to (3.5) is

$$\hat{\varphi}_{ei} = \max\left[\alpha_{10},\ \min\left\{\hat{\gamma} \cdot \frac{\|a_i(\theta_0) - \hat{K}b_i\|}{N^{-1}\sum_{j=1}^{N}\|a_j(\theta_0) - \hat{K}b_j\|},\ 1\right\}\right], \tag{3.6}$$

where the subscript "e" denotes that $\hat{\varphi}_{ei}$ is the "exact" solution to (3.5), and $\hat{\gamma} > 0$ is the solution to

$$N^{-1}\sum_{i=1}^{N} \max\left[\alpha_{10},\ \min\left\{\gamma \cdot \frac{\|a_i(\theta_0) - \hat{K}b_i\|}{N^{-1}\sum_{j=1}^{N}\|a_j(\theta_0) - \hat{K}b_j\|},\ 1\right\}\right] = \alpha_0.$$

3.2 Practical considerations

The sampling plans with $\hat{\varphi}_i$ and $\hat{\varphi}_{ei}$ are not practically applicable, because both of them depend on θ_0, which needs to be estimated beforehand. To this end, the convention is to draw an initial sample, say $\{\tilde{z}_i = (\tilde{y}_i, \tilde{x}_i^\top)^\top : i = 1, \ldots, n_1\}$, by uniformly sampling from the big data being studied. The first step in our two-step Poisson sampling plays exactly the same role. Let $\tilde{\theta} = \arg\min_\theta \sum_{i=1}^{n_1} \ell(\tilde{z}_i, \theta)$ and $\tilde{V}$ be a consistent estimator of V based on the first-step sample. Denote $\tilde{K} = \arg\min \sum_{i=1}^{n_1} \|\tilde{a}_i - K\tilde{b}_i\|$, where $\tilde{b}_i = (-\alpha_0, h^\top(\tilde{z}_i))^\top$, and $\tilde{a}_i = \tilde{V}^{-1}\dot{\ell}(\tilde{z}_i, \tilde{\theta})$ in the A-optimality criterion or $\tilde{a}_i = \dot{\ell}(\tilde{z}_i, \tilde{\theta})$ in the L-optimality criterion. Calculating $\tilde{K}$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

may be computationally intensive, and so we use the least-squares estimate $\tilde{K} = \left(\sum_{k=1}^{n_1} \tilde{a}_k \tilde{b}_k^\top\right)\left(\sum_{j=1}^{n_1} \tilde{b}_j \tilde{b}_j^\top\right)^{-1}$ instead. Define

$$\tilde{\varphi}_{ei} = \max\left\{\alpha_{10},\ \min\left(\tilde{\gamma}\cdot\frac{\|a_i - \tilde{K}b_i\|}{n_1^{-1}\sum_{j=1}^{n_1}\|\tilde{a}_j - \tilde{K}\tilde{b}_j\|},\ 1\right)\right\},\ 1 \le i \le N, \quad (3.7)$$

where $\tilde{\gamma} > 0$ is the smallest solution to

$$n_1^{-1}\sum_{i=1}^{n_1}\max\left\{\alpha_{10},\ \min\left(\gamma\cdot\frac{\|\tilde{a}_i - \tilde{K}\tilde{b}_i\|}{n_1^{-1}\sum_{j=1}^{n_1}\|\tilde{a}_j - \tilde{K}\tilde{b}_j\|},\ 1\right)\right\} = \alpha_0.$$

Our recommended sampling plan for the second step is $\tilde{\pi} = (\tilde{\pi}_1, \ldots, \tilde{\pi}_N)$ with

$$\tilde{\pi}_i = (\tilde{\varphi}_{ei} - \alpha_{10})/(1 - \alpha_{10}), \quad (3.8)$$

where $\alpha_{10} \in (0, 1)$ is the known sampling fraction of the first-step sampling.

4. Sample size determination

For a given subsample, the performance of the IPW and ELW estimators depends not only on the underlying sampling plan, but also on the size of the subsample. If the size n or the ideal size n_0 of a two-step Poisson subsample is too small, the resulting estimator will be so unstable that it does not make any sense. When the (optimal) sampling plan is fixed, it is necessary to specify the subsample size that guarantees the resulting estimate meets a certain precision requirement. To the best of our knowledge, this issue

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

has never been discussed in the literature of subsampling for big data. We address the issue of determining n_0 under each of the following two precision requirements on $\hat{\theta}_{\text{ELW}}$: (R1) The asymptotic MSE of $\hat{\theta}_{\text{ELW}}$ is no greater than a prespecified positive constant C_0, i.e., $\mathbb{M}\text{se}(\hat{\theta}_{\text{ELW}}) \leq C_0$. (R2) The absolute error of $\hat{\theta}_{\text{ELW}}$ is no greater than a critical value $d_0 > 0$ at the confidence level $(1-a)$, i.e.,

$$P(\|\hat{\theta}_{\text{ELW}} - \theta_0\| \leq d_0) \geq 1 - a. \tag{4.1}$$

We assume that the sample fraction $\alpha_{10} > 0$ of the first sample is known, but that for the second sample α_{20} is unknown. Because $n_0/N = \alpha_0 = 1 - (1-\alpha_{10})(1-\alpha_{20})$, when the (optimal) sampling plan is fixed, determining α_{20} is equivalent to determining n_0. Recall that a nearly optimal subsampling plan can be approximated by (3.4) or $\tilde{\varphi}_* = (\tilde{\varphi}_{*1}, \ldots, \tilde{\varphi}_{*N})$, where $\tilde{\varphi}_{*i}$ is $\hat{\varphi}_i$ with $\hat{K}$ replaced by $\tilde{K}$. With the sampling plan $\tilde{\varphi}_*$, an upper bound for the asymptotic MSE of $\hat{\theta}_{\text{ELW}}$ is

$$\frac{H(\tilde{\varphi}_*)}{N} = \frac{1}{N^3\alpha_0}\left\{\sum_{j=1}^{N}\|a_j(\theta_0) - \tilde{K}b_j\|\right\}^2 = \frac{1}{n_0}\left\{\frac{1}{N}\sum_{j=1}^{N}\|a_j(\theta_0) - \tilde{K}b_j\|\right\}^2,$$

which can be estimated by $n_0^{-1}(n_1^{-1}\sum_{j=1}^{n_1}\|\tilde{a}_j - \tilde{K}\tilde{b}_j\|)^2$ based on a pilot sample of size $n_1 \approx N\alpha_{10}$. Under requirement (R1), a sufficient approximation is to constrain $n_0^{-1}(n_1^{-1}\sum_{j=1}^{n_1}\|\tilde{a}_j - \tilde{K}\tilde{b}_j\|)^2 \leq C_0$. Note that the elements of $\tilde{K}$ and $\widetilde{b}_j$ contain the unknown parameter $\alpha_0 = n_0/N$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

Therefore, the minimal sample size n_0 that satisfies requirement (R1) should be the solution to

$$n_0 = \frac{1}{C_0}\left(\frac{1}{n_1}\sum_{j=1}^{n_1}\|\tilde{a}_j - \tilde{K}\tilde{b}_j\|\right)^2. \tag{4.2}$$

This is our first recommended sample size determination method, which we denote as M1 for convenience.

To determine the sample size under requirement (R2), note that the inequality $\|\hat{\theta}_{\mathrm{ELW}} - \theta_0\| \leq d_0$ is equivalent to $\zeta^\top \Sigma_{\mathrm{ELW}}\zeta \leq Nd_0^2$, where $\zeta = \sqrt{N}\Sigma_{\mathrm{ELW}}^{-1/2}(\hat{\theta}_{\mathrm{ELW}} - \theta_0)$ approximately follows the q-dimensional standard normal distribution, where q is the dimension of θ. The distribution of $\zeta^\top \Sigma_{\mathrm{ELW}}\zeta$ can be further approximated by a weighted chi-square distribution of $\sum_{k=1}^{q}\lambda_k\zeta_k^2$, where the λ_k are the eigenvalues of Σ_{ELW} and the ζ_k are i.i.d. standard normal random variables. According to Kim *et al.* (2006)[Lemma 2, page 453], the cumulative distribution of $\sum_{k=1}^{p}\lambda_k\zeta_k^2$ can be approximated by that of $\nu^{-1}\chi^2_{\nu_*}$, where $\nu = \sum_{k=1}^{q}\lambda_k / \sum_{j=1}^{q}\lambda_j^2$ and $\nu_* = (\sum_{k=1}^{q}\lambda_k)^2 / \sum_{j=1}^{q}\lambda_j^2$. It follows that $P(\|\hat{\theta}_{\mathrm{ELW}} - \theta_0\| \leq d_0) \approx P(\chi^2_{\nu_*} \leq \nu N d_0^2)$, which together with (4.1) implies the approximation $\nu N d_0^2 = \chi^2_{\nu_*}(1-a)$, where $\chi^2_{\nu_*}(1-a)$ is the $(1-a)$th quantile of the chi-square distribution with ν_* degrees of freedom.

Moreover, ν and ν_* are approximately equal to $\tilde{\nu} = \sum_{k=1}^{q}\tilde{\lambda}_k / \sum_{j=1}^{q}\tilde{\lambda}_j^2$ and $\tilde{\nu}_* = (\sum_{k=1}^{q}\tilde{\lambda}_k)^2 / \sum_{j=1}^{q}\tilde{\lambda}_j^2$, respectively, where the $\tilde{\lambda}_k$ are the eigenval-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

ues of $\tilde{\Sigma}_{\mathrm{ELW}} = \tilde{V}^{-1}(\tilde{B}_{\dot{\ell}\dot{\ell}} - \tilde{B}_{\dot{\ell}h}\tilde{B}_{hh}^{-1}\tilde{B}_{\dot{\ell}h}^{\top})\tilde{V}^{-1}$. Herein, $\tilde{B}_{\dot{\ell}\dot{\ell}}$, $\tilde{B}_{\dot{\ell}h}$, and $\tilde{B}_{hh}$ are the sample-mean estimates of $B_{\dot{\ell}\dot{\ell}}$, $B_{\dot{\ell}h}$, and B_{hh} based on the first sample. Because Σ_{ELW} (and hence λ_k) depends on $\alpha_0 = n_0/N$, so do $\tilde{\Sigma}_{\mathrm{ELW}}$, $\tilde{\lambda}_k$, and $\tilde{\nu}$. We denote $\tilde{\nu}$ and $\tilde{\nu}_*$ by $\tilde{\nu}(n_0)$ and $\tilde{\nu}_*(n_0)$, respectively, to highlight this dependence. Our recommended sample size n_0 under requirement (R2), denoted as M2, is the root of

$$\tilde{\nu}(n_0) = \nu N d_0^2 = \chi^2_{\tilde{\nu}_*(n_0)}(1-a). \tag{4.3}$$

5. Numerical results

In this section, we investigate the finite-sample performance of the proposed ELW estimation and sampling strategy as well as the sample size determination method by analyzing two real datasets: a bike sharing dataset and a hospital length-of-stay dataset.

5.1 Methods under comparison

We use ELW and ELWAI to denote the proposed ELW estimation methods without and with auxiliary information, where the full-data response mean is taken as auxiliary information. We compare their performance with the two-step IPW estimation method of Yu *et al.* (2022), where the resulting estimator is an inverse-variance weighting average of the IPW estimator

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

based on the second-step sample and the pilot estimator based on the first-step sample. As a benchmark, we also consider the one-step estimator based only on the first-step uniform sample.

We use ELWS and ELWAIS to represent the proposed nearly optimal two-step Poisson sampling plans corresponding to the ELW and ELWAI estimation methods, respectively. Specifically, ELWAIS refers to the sampling plan with the second-step subsampling probabilities being (3.8), which incorporates auxiliary information, while ELWS refers to the counterpart without utilizing auxiliary information. We compare them with the optimal sampling plan (IPWS) of Yu *et al.* (2022, equation (21)), which is derived based on their IPW method. The second-step subsampling probabilities of IPWS is

$$\tilde{\pi}_i = \frac{n_{20}}{N} \cdot \frac{(1-\varrho)|y_i - \exp(x_i^\top \tilde{\theta})|h(x_i)}{n_1^{-1}\sum_{j=1}^{n_1}|\tilde{y}_j - \exp(\tilde{x}_j^\top \tilde{\theta})|h(\tilde{x}_j)} + \frac{\varrho n_{20}}{N}, \quad i = 1, \ldots, N, \quad (5.1)$$

where $\tilde{\theta}$ is the pilot estimator based on the first step sample $\{(\tilde{x}_j, \tilde{y}_j) : j = 1, \ldots, n_1\}$ and n_{20} is the average size of the second step sample. Note that a shrinkage technique (Ma *et al.*, 2014) is used with a tuning parameter $\varrho \in [0, 1]$ when calculating the IPWS optimal sampling probabilities. When $\varrho = 0$, the function $h(x) = \|x\|$ and $\|\tilde{V}^{-1}x\|$ with $\tilde{V} = n_1^{-1}\sum_{j=1}^{n_1} e^{\tilde{x}_j^\top \tilde{\theta}} \tilde{x}_j^{\otimes 2}$ correspond to the L- and A-optimality criteria of the classical IPW method, respectively. We fix $\varrho = 0.2$ for consistency with the setup of Yu *et al.*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

(2022). As estimation efficiency is of primary concern, we consider only optimal subsampling probabilities under the A-optimality criterion. The results under the L-optimality criterion are similar and omitted to save space.

5.2 Data description

The bike sharing dataset consists of N =17,379 observations and is available from the UCI Machine Learning Repository (`https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset`). We are interested in the problem of how the number of bikes rented hourly (Y) is influenced by four covariates: a binary variable (X_1) indicating whether a certain day is a working day or not, temperature (X_2), humidity (X_3), and windspeed (X_4). The hospital length-of-stay dataset coinsists of N =100,000 observations on patients admitted into hospital and is available from the Microsoft Machine Learning Services (`https://microsoft.github.io/r-server-hospital-length-of-stay/`). The problem of interest is to investigate how the length of stay within hospital (Y) is influenced by two covariates: readmission count and the number of symptoms including renal disease, asthma, iron deficiency, pneumonia, substance dependence, fibrosis, malnutrituion, blood disorder, depression, major psychological disorder,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

and other psychological disorder during encounter. As the readmission count data are categorized into six groups with levels 0, 1, 2, 3, 4, and 5+, we encode readmission count as five dummy variables X_1–X_5, indicating the non-zero levels, and encode number of symptoms as X_6. For each dataset, we model the responses given the covariates by Poisson regression, which is widely used for count data modelling. To eliminate the influence of scales of different variables, we centralize and standardize the covariates in both datasets.

5.3 Estimation and sampling comparisons

From each of the two big datasets, we generate 5000 samples by each of the IPWS, ELWS, and ELWAIS sampling plans and then calculate the IPW, ELW, and ELWAI estimates of the regression coefficient θ based on each sample. We set the ideal sample size n_{10} to 200 in the first step, and set the ideal sample size n_{20} to 300, 500, and 1000, respectively, in the second step. For fair comparisons, we set the size of the uniform sample for the one-step estimator to be the same as the overall sample sizes (namely, $n_{10} + n_{20}$). We evaluate the estimation performance of the one-step, IPW, ELW, and ELWAI estimators in terms of the empirical MSE

$$\text{MSE} = \frac{1}{5000} \sum_{b=1}^{5000} \|\breve{\theta}_b - \hat{\theta}_N\|^2, \tag{5.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

where $\breve{\theta}_b$ is a generic subsample-based estimate in the bth repetition and $\hat{\theta}_N$ is the full-data-based estimate of regression coefficients reported in Table 2 of the supplementary material. Figure 1 displays the logarithm of empirical MSE versus n_{20} for the one-step estimator under the uniform sampling plan and the IPW, ELW, and ELWAI estimators under the two-step sampling plans.

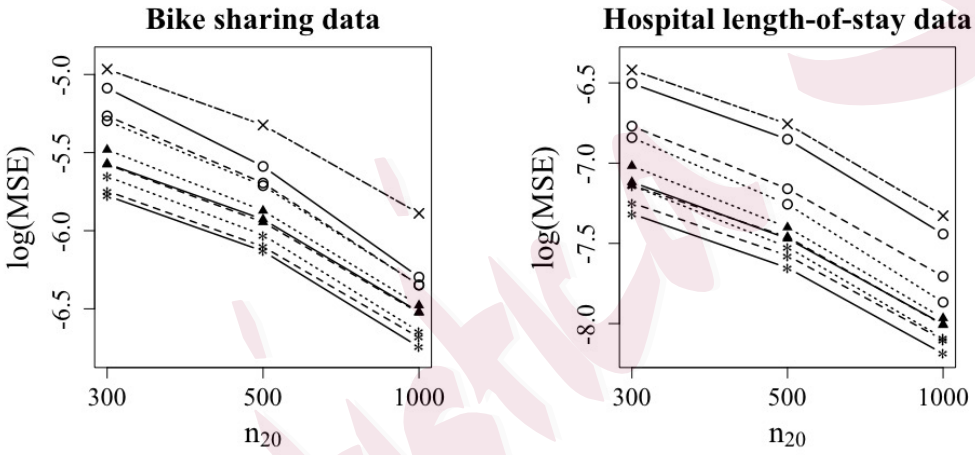

Figure 1: Plots of the logarithm of MSE versus n_{20} for the one-step ($\times$), IPW ($\circ$), ELW ($\blacktriangle$), and ELWAI ($*$) estimators when data were drawn by the uniform (dot-dashed line), IPWS (dotted line), ELWS (dashed line), and ELWAIS (solid line) sampling plans from the two real datasets.

We first examine the performance of the one-step, IPW, ELW, and ELWAI estimation methods from Figure 1. Clearly, all three two-step

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

estimators have much lower MSEs than the one-step estimator. This confirms that employing carefully designed non-uniform sampling strategies can enhance the efficiency of the resulting subsample-based estimators. Below, we focus on the the evaluation of the two-step subsampling and estimation methods. In terms of MSE, both ELW and ELWAI outperform IPW uniformly for all combinations of sampling plan, real dataset and n_{20}. This suggests that the proposed ELW estimation method always produce more reliable estimates than the IPW method regardless of whether auxiliary information is incorporated. Meanwhile the ELWAI estimator always has a uniformly smaller MSE than the ELW estimator in all cases, which confirms our finding below Theorem 1, namely incorporating auxiliary information does lead to higher estimation efficiency for the ELW method.

Next, we investigate the efficiency of the three two-step sampling plans: ELWS, ELWAIS and IPWS. Figure 1 shows that given each of the IPW, ELW and ELWAI estimation methods, the corresponding (nearly) optimal sampling plan leads to uniformly smaller MSEs than the other two sampling plans. For example, if we choose the ELW method to estimate the regression coefficient, ELWS leads to a more reliable estimator than IPWS and ELWAIS. This makes sense as each of the three sampling plans

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

is designed for the corresponding estimation method.

Finally, it is natural to take an estimation method and the corresponding (nearly) optimal sampling plan as a toolkit for processing big data. We see that for all combinations of n_{20} and real dataset, the ELW estimation and sampling technique gives uniformly more efficient estimators than the IPW-based counterpart, and the ELWAI estimation and sampling technique uniformly outperforms the ELW-based counterpart. Again this demonstrates the superiority of the proposed ELW methods over the existing IPW method, and that of the ELWAI method over the ELW method by incorporating auxiliary information.

Besides estimation efficiency, computational cost is another important concern in the process of big data. Table 1 presents the average CPU times (in milliseconds) per subsample of the IPW, ELW and ELWAI estimation methods and the IPWS, ELWS and ELWAIS sampling methods. Regarding the three estimation methods, ELW is slower than IPW, and ELWAI is slower than ELW, as are expected. In particular, ELW spends about twice CPU time than IPW, and ELWAI spends more than four times CPU time than ELW in average when processing a subsample. Even so, the three sampling methods takes almost the same time per subsample. As the average estimation times are negligible compared with the average

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

sampling times, the overall estimation and sampling times of the proposed two methods are almost the same as that of the IPW-based method. Table 1 also reports the computation time for calculating the full-data-based estimates. We see that in the analyses of the bike sharing data and hospital length-of-stay data, the two-step sampling and estimation procedures can save much computation times than that based on the full data.

Table 1: Average CPU times (unit: millisecond) per subsample of the three estimation methods and the three sampling methods under comparison.

		Bike sharing data			Hospital length-of-stay data		
	n_{20}	300	500	1000	300	500	1000
Estimation	IPW	0.3	0.4	0.7	0.5	0.6	0.9
	ELW	0.6	0.6	1.0	1.1	1.2	1.5
	ELWAI	3.4	3.7	4.1	5.1	6.1	6.9
Sampling	IPWS	27.9	29.5	25.4	181.2	186.4	180.3
	ELWS	28.8	29.4	25.1	183.2	184.2	174.5
	ELWAIS	29.0	29.3	25.3	186.8	187.7	182.9
Full-data-based estimation		63.3			238.8		

In addition, we also conduct simulations to investigate the performances

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

of two sample size determination methods, M1 and M2. See the supplementary material. Our general finding is that these methods do produce desirable estimates with promised precision.

6. Discussion

The problem of optimal subsampling has a long-standing tradition within the field of survey sampling for inference regarding finite populations; see, e.g., Neyman (1938), Hajek (1959), Cassel *et al.* (1976), Brewer (1979), Isaki and Fuller (1982), and Bellhouse (1984). These works, however, are primarily concerned with linear estimators of scalar finite population characteristics. Stimulated by modern technological developments, the question of optimal subsample selection has attained renewed attention during the past few years for more complex inference problems, such as logistic regression for big data (Wang et al., 2018; Wang, 2019). However, the current optimal subsamplings for big data are built on the well-known Horvitz-Thompson estimator or the IPW estimator, which becomes unstable when some inclusion probabilities are close to zero. This motivates our optimal Poisson sampling for big data based on the ELW method.

Based on a two-step Poisson sample from a big dataset, we have developed an ELW estimation method for M-estimation problems. The

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

proposed approach not only overcomes the instability of the conventional IPW-based estimation method, but also improves the estimation efficiency by incorporating auxiliary information. A nearly optimal two-step Poisson sampling plan was constructed accordingly. Theoretically, the ELW method is asymptotically more efficient than the IPW method, which means that the proposed sampling and estimation method requires fewer samples to achieve the target estimation precision. Recently, Wang and Kim (2022) proposed a maximum sampled conditional likelihood method to overcome the instability of IPW. As built on a parametric model, their method suffers from model mis-specification. In contrast, our ELW method does not have this problem.

We assumed the convexity of the loss function in the M-estimation problem for technical convenience. Our ELW estimation method also applies to general M-estimation problems, general estimating equation problems (Qin and Lawless, 1994) and more general sampling plans including those with replacement. Further efforts may be needed to establish the asymptotic normality of the resulting point estimator, which is the foundation for constructing optimal sampling plans.

The two-step Poisson sampling we have considered consists of a pilot uniform sampling and a refined sampling. Under this sampling framework,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

we established two sample size determination methods under estimation precision requirements (R1) and (R2), respectively. These methods are new in the literature of optimal subsampling for big data. They may need to be modified when the parameter of interest is a smooth function of θ, such as $C\theta$ for a given matrix C, rather than θ itself. In addition, the current two-step Poisson sampling consists of only two subsampling processes, although this may be extended to multiple subsampling processes when needed.

Supplementary Material

The online supplementary material contains the proofs of Lemma 1 and Theorems 1–2, and additional simulation results.

Acknowledgements

This research is supported by the National Key R&D Program of China (2021YFA1000100 and 2021YFA1000101), the National Natural Science Foundation of China (11971300, 12101239, 12171157, 71931004), the Natural Science Foundation of Shanghai (19ZR1420900), the China Postdoctoral Science Foundation (Grant 2020M681220), and the 111 Project (B14019). The first two authors contributed equally to this paper. The second and third authors are co-corresponding authors.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

References

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

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

Bellhouse, D. R. (1984). A review of optimal designs in survey sampling. *Canadian Journal of Statistics*, **12**(1), 53–65.

Brewer, K. R. W. (1979). A class of robust sampling designs for large-scale surveys. *Journal of the American Statistical Association*, **74**(368), 911–915.

Cassel, C. M., Sarndal, C. E., and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, **63**(3), 615–620.

Chen, S., and Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*, **104**(2), 439–453.

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

Fan, Y., Liu, Y., and Zhu, L. (2021). Optimal subsampling for linear quantile regression models. *Canadian Journal of Statistics*, **49**(4), 1039–1057.

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

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, **31**, 1208–1212.

Hajek, J. (1959). Optimal strategy and other problems in probability sampling. *Casopis pro Pestovani Matematiky*, **84**(4), 387–423.

Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association*, **109**(507), 1159–1173.

Han, P. (2016). Combining inverse probability weighting and multiple imputation to improve robustness of estimation. *Scandinavian Journal of Statistics*, **43**(1), 246–260.

Hansen, M. H. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, **14**(4), 333–362.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**(260), 663–685.

Huber, P. J. (2011). Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

model. *Journal of the American Statistical Association*, **77**(377), 89–96.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, **22**(4), 523–539.

Kim, H.-Y., Gribbin, M. J., Muller, K. E., and Taylor, D. J. (2006). Analytic, computational, and approximate forms for ratios of noncentral and central gaussian quadratic forms. *Journal of Computational and Graphical Statistics*, **15**(2), 443–459.

Liu, Y. and Fan, Y. (2023). Biased-sample empirical likelihood weighting: An alternative to inverse probability weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **85**(1), 67–83.

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

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, **33**(201), 101–116.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**(2), 237–249.

Powell, J. L. (1990). Estimation of monotonic regression models under quantile restrictions. In *Nonparametric and Semiparametric Methods in Econometrics*. Cambridge University Press.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

Qin, J. (2017). *Biased Sampling, Over-Identified Parameter Problems and Beyond*. Singapore: Springer.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics*, **22**(1), 300–325.

Shen, X., Chen, K., and Yu, W. (2021). Surprise sampling: Improving and extending the local case-control sampling. *Electronic Journal of Statistics*, **15**(1), 2454–2482.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

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

Wang, H. and Kim J. K. (2022). Maximum sampled conditional likelihood for informative subsampling. *Journal of Machine Learning Research*, **23**, 1–50.

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

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

Yao, Y., Zou, J. , and Wang, H. (2023). Optimal poisson subsampling for softmax regression. *Journal of Systems Science & Complexity*, **36**(4), 1609–1625.

Yao, Y. and Wang, H. (2021). A review on optimal subsampling methods for massive datasets. *Journal of Data Science*, **19**(1), 151–172.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0274

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

School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai, China

E-mail: fanyan@suibe.edu.cn

School of Mathematical Sciences, Soochow University, Suzhou, China

E-mail: liuyangecnu@163.com

KLATASDS - MOE, School of Statistics, East China Normal University, Shanghai, China

E-mail: ykliu@sfs.ecnu.edu.cn

National Institute of Allergy and Infectious Diseases, National Institutes of Health, Frederick, USA

E-mail: jingqin@niaid.nih.gov