Sample Size Determination For Multidimensional Parameters And The A-Optimal Subsampling In A Big Data Linear Regression Model

A Preprint

Sheng Zhang[*], Fei Tan[*], and Hanxiang Peng[*]

November 9, 2023

Abstract

To fast approximate the least squares estimator (LSE) efficiently in a Big Data linear regression by a subsampling LSE, numerous optimal sampling distributions are derived based on the criterion of minimizing the sum of the component variances of the subsampling LSE. We discuss truncation of the distributions, and construct the Scoring Algorithm with far less running time for implementing the subsampling LSE than for the full-sample LSE. The subsampling LSE is proved to be almost surely asymptotically normal for an arbitrary sampling distribution under suitable conditions. Motivated by subsampling and data-splitting in machine learning, sample size determination for multidimensional parameters is investigated. We conduct a comprehensive evaluation of our proposed approach through various numerical studies and compare it with the uniform sampling. Our results in both simulated and real data indicate that our approach substantially outperforms the uniform and the Algorithm significantly reduces the computational time required for implementing the full-sample LSE.

***K*eywords** Asymptotic normality; Least squares estimator; Big data; Optimal sampling; Sample size determination

1 Introduction

In a linear regression model, the response y_i and covariate vector $\mathbf{x}_i$ satisfy

$$y_i = \boldsymbol{\beta}^\top \mathbf{x}_i + \varepsilon_i, \quad i = 1, \dots, n, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown parameter and $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically (i.i.d.) random errors with zero mean and finite positive variance $\sigma^2 = \mathrm{Var}(\varepsilon_i)$. Assume that $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ is a nonrandom $n \times p$ matrix of full rank p.

The parameter vector $\boldsymbol{\beta}$ can be estimated by the ordinary least squares estimator (LSE) $\hat{\boldsymbol{\beta}}_{\mathrm{ols}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$, where $\mathbf{y} = (y_1, \dots, y_n)^\top$. Consider the case of data of massive size in which $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$ is not available. One may draw a subsample $(\mathbf{X}^*, \mathbf{y}^*)$ of small size $r << n$ using a sampling distribution $\boldsymbol{\pi}_n = (\pi_1, \dots, \pi_n)$ as a surrogate for the full sample, and calculate the subsampling weighted LSE $\hat{\boldsymbol{\beta}}^*_r$ to approximate $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$,

$$\hat{\boldsymbol{\beta}}^*_r = (\mathbf{X}^{*\top}\mathbf{W}^*\mathbf{X}^*)^{-1}\mathbf{X}^{*\top}\mathbf{W}^*\mathbf{y}^*. \tag{2}$$

where $\mathbf{W}^* = \mathrm{diag}(1/r\boldsymbol{\pi}^*)$ is the diagonal matrix with $\boldsymbol{\pi}^*$ equal to the vector of the corresponding sampling probabilities. Here we adopt the componentwise division $\mathbf{a}/\mathbf{b} = (a_1/b_1, \dots, a_n/b_n)^\top$ for vectors $\mathbf{a}, \mathbf{b}$. This is a Hansen-Hurwitz estimator and could also be viewed as a weighted bootstrap estimator based on the subsample. Full sample weighted bootstrap estimators were well studied in the literature, see the monograph by Barbe and Bertail (1995)[2].

Over the past two decades, there have been considerable progresses on subsampling, see Liang, *et al.* (2013)[11], Kleiner, *et al.* (2014)[9], Wang, *et al.* (2015)[20], Wang, *et al.* (2019)[19] among others. Algorithms for fast computing

[*] Department of Mathematical Sciences, IUPUI, 402 N Blackford St., LD 270, Indianapolis, IN 46202, USA.
Email: shezhang@iu.edu, feitan@iu.edu, and hanxpeng@iu.edu.

the LSE were constructed, see the monograph by Mahoney (2011)[14] and the references therein. A key feature of these results is the nonuniform sampling. While these results were mainly focused on the algorithmic properties, we shall be concerned with statistical inference. Zhu, *et al.* (2015)[22] pioneered in this aspect and their work is influential in our work. They obtained several A-optimal distributions and proved asymptotic normality. We give the A-optimal distributions for approximating a smooth function $\mathbf{g}(\hat{\boldsymbol{\beta}}_{\text{ols}})$ of $\hat{\boldsymbol{\beta}}_{\text{ols}}$ (the choice of $\mathbf{g}(\hat{\boldsymbol{\beta}}_{\text{ols}}) = \mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{ols}} = \mathbf{X}^\top\mathbf{y}$ yields their results), and prove an almost sure asymptotic normality result.

The statistical leverage scores based distribution $\boldsymbol{\ell}$ has played a central role in the development of randomized matrix algorithms, see e e.g. Candés and Tao (2009)[3]; Drineas *et al.* (2012)[7]; Ma and Sun (2014)[12]; Ma, *et al.* (2015)[13]; Xu, *et al.* (2016)[21]. Interestingly, $\boldsymbol{\ell}$ and the A-optimal distribution $\hat{\boldsymbol{\pi}}_2$ draw data points in a totally opposite way. Specifically, the former draws points close to the regression hyperplane, whereas the latter does away from the hyperplane.

While classic methods compute the LSE $\hat{\boldsymbol{\beta}}_{\text{ols}}$ in $O(np^2)$ time, randomized methods usually take $o(np^2)$ time. Typically, the bottleneck is to compute the appropriate sampling distributions, and the A-optimal distributions fall in with this category. As the LSE $\hat{\boldsymbol{\beta}}_{\text{ols}}$ and $\boldsymbol{\ell}$ are fundamental and ubiquitous, there have been developed randomized algorithms on rapidly approximating them, see e.g. Drineas, *et al.* (2006)[6]. These algorithms can be utilized to fast compute the optimal distributions. In the spirit of the scoring method for improving estimation efficiency, we construct the Scoring Algorithm in Fig. 2 with running time $O(rp^2)$ where $r << n$. Our extensive simulations indicated that the algorithm worked particularly well.

It is obvious that a suitable subsample size is critical for obtaining a desired result within a desired peroid of time. Sample size determination (SSD) for scalar parameters is a melody. In this article, we introduce SSD for multidimensional parameters and study its numerical properties through simulations and real data. The result may be useful for data spliting in machine learning.

The article is organized as follows. In Section 2, we define SSD for multidimensional parameters and proivde the formulas. In Section 3, we present an asymptotic normality result, give the A-optimal distributiogns, construct the Scoring Algorithm, and discuss truncation and the raltationship between the leverage-scores- based distribution and the A-optimal distributions. Simulations and real data applications are reported in Section 4. The ASN is proved in Section 5.

2 SSD for Multidimensional Parameters

Let P be a probability measure on some measurable space. Let m the volume measure on $\mathbb{R}^p$. Consider a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$, and a random region $\mathbb{R}_n$ on $\mathbb{R}^p$. Given $\epsilon > 0$ and $\alpha \in (0, 1)$, we seek a minimum sample size n such that at the level $1-\alpha$ of confidence, $\mathbb{R}_n$ catches $\boldsymbol{\theta}$ within the "range of error" (ROE) ϵ, that is, $m(\mathbb{R}_n) \leq \epsilon$. Let $\boldsymbol{\theta}_0$ denote the true value of parameter.

Definition 1. *Given $\epsilon > 0$ and $\alpha \in (0, 1)$, the sample size with the ROE $\epsilon > 0$ at the $1-\alpha$ level of confidence is defined as*

$$n(\epsilon, \alpha) = \min\left\{n : \ P(\boldsymbol{\theta}_0 \in \mathbb{R}_n, \, m(\mathbb{R}_n) \leq \epsilon^p) \geq 1-\alpha\right\}.$$

Analogous to selecting bootstrapping sample sizes, both α and ϵ must be appropriately chosen in which ϵ is critical. We now give two examples, using the following two-step method.

Step 1 Construct a $1-\alpha$ level confidence region $\mathbb{R}_n$ for $\boldsymbol{\theta}$.

Step 2 Find the minimum sample size n such that the ROE is ϵ, that is, $m(\mathbb{R}_n) \leq \epsilon^p$.

Example 1. (Ellipsoid) Let $\hat{\boldsymbol{\theta}}$ be an estimator of $\boldsymbol{\theta}_0 \in \Theta \subset \mathbb{R}^p$ with the (asymptotic) variance-covariance matrix Σ positive definite. Let $\mathbb{R}_n$ be the $1-\alpha$ level confidence ellipsoid centered at $\hat{\boldsymbol{\theta}}_n$, $\mathbb{R}_n = \{\boldsymbol{\theta} \in \Theta : \ T(\boldsymbol{\theta}) \leq q_\alpha(p)\}$, where $T(\boldsymbol{\theta}) = n(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \Sigma^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$, and $q_\alpha(p)$ denotes the upper α-percentile of the distribution of $T(\boldsymbol{\theta}_0)$, that is, $P(T(\boldsymbol{\theta}_0) > q_\alpha(p)) \leq \alpha$. By definition, the sample size is determined by

$$n_p(\epsilon, \alpha) = \min\left\{n : \ P(\boldsymbol{\theta}_0 \in \mathbb{R}_n, \, m(\mathbb{R}_n) \leq \epsilon^p) \geq 1-\alpha\right\}.$$

The volume of the ellipsoid is

$$m(\mathbb{R}_n) = \frac{n^{-p/2}\pi^{p/2}}{\Gamma(p/2+1)} q_\alpha^{p/2}(p) \prod_{d=1}^{p} \sqrt{\lambda_d},$$

where $\lambda_d, d = 1, \ldots, p$ are the eigenvalues of Σ. Solving $m(\mathbb{R}_n) \leq \epsilon$ for n yields the sample size $n_p(\epsilon, \alpha)$ with ROE ϵ at the $1 - \alpha$ level, given by

$$n_p(\epsilon, \alpha) = \frac{\pi q_\alpha(p)}{\Gamma^{2/p}(p/2+1)} \frac{\sqrt[p]{\det(\Sigma)}}{\epsilon^2}, \tag{3}$$

where $\det(\Sigma) = \prod_{d=1}^{p} \lambda_d$. Often Σ is unknown, one then uses an estimator $\hat{\Sigma}$ of it. For $p = 1$, as $\Gamma(3/2) = \sqrt{\pi}/2$, the sample size with ROE 2ϵ (margin of error (MOE) ϵ) at the $1 - \alpha$ level boils down to $n_1(\epsilon, \alpha) = q_\alpha(1)\sigma^2/\epsilon^2$, commonly found in textbooks. For large p, by Sterling's formula, $\Gamma^{2/p}(p/2+1) \approx (p/2e)(p\pi)^{1/p}$. A computationally easy formula is now given by

$$\tilde{n}_p(\epsilon, \alpha) = 2\pi e q_\alpha(p) \sqrt[p]{\det(\Sigma)/(p\pi)}/(p\epsilon^2). \tag{4}$$

Example 2. (Bonferroni) Consider the same problem as in Example 1, but now based on Bonferroni's method. We take $\mathbb{R}_n$ to be the p-dimensional $(1-\alpha)$-confidence hyperrectangle,

$$\mathbb{R}_n = \prod_{d=1}^{p} (\hat{\boldsymbol{\theta}}_{d,n} - z_{\alpha/(2p)}\sigma_d/\sqrt{n}, \quad \hat{\boldsymbol{\theta}}_{d,n} + z_{\alpha/(2p)}\sigma_d/\sqrt{n}),$$

where z_α denotes the upper α-percentile of the distribution of $\Sigma^{-1/2}\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\beta}_0)$, and $\hat{\boldsymbol{\theta}}_{d,n}$ and σ_d^2 denote the d-th component of $\hat{\boldsymbol{\theta}}_n$ and the d-th diagonal entry of Σ, respectively. As the volume of the hyperrectangle $\mathbb{R}_n$ is

$$m(\mathbb{R}_n) = 2^p n^{-p/2} z_{\alpha/(2p)}^p \sigma_1 \cdots \sigma_p,$$

solving $m(\mathbb{R}_n) \leq \epsilon^p$ about n yields the sample size,

$$n_p^{bon}(\epsilon, \alpha) = 4 z_{\alpha/(2p)}^2 \sigma_1^{2/p} \cdots \sigma_p^{2/p} \epsilon^{-2}. \tag{5}$$

For unknown parameters σ_d's, one uses estimates $\hat{\sigma}_d$'s of them.

Remark 1. *If $T_n(\boldsymbol{\theta}_0)$ has Chisquare distribution with p degrees of freedom, $\chi^2(p)$, (often approximately), then $q_\alpha(p) = \chi_\alpha^2(p)$, the upper α-percentile of $\chi^2(p)$. Similarly for Bonferroni, $q_\alpha = Z_\alpha$, the upper α-percentile of the standard normal $\mathscr{N}(0,1)$. Alternatively, one can get an estimate of $q_\alpha(p)$ by bootstrapping or pre-subsampling in the Scoring Algorithm 2 in the case of Big Data.*

Remark 2. *In nonuniform subsampling for data of massive size, a sampling distribution $\boldsymbol{\pi}$ must be computed before actually sampling. An optimal distribution $\boldsymbol{\pi}$ typically has the same computational complexity as the original problem. To tackle this problem, one may take a uniform pre-subsample of small size and compute an approximation $\tilde{\boldsymbol{\pi}}_0$ to $\boldsymbol{\pi}$ as described in the Scoring Algorithm, choosing suitable values of the ROE ϵ, α and $q_\alpha(p)(= \chi_\alpha^2(p))$. To determine the pre-subsample size, one may take* $\det(\Sigma) = 1$ *and get*

$$n_{p,0}(\epsilon, \alpha) = \frac{\pi q_\alpha(p)}{\Gamma^{2/p}(p/2+1)} \frac{1}{\epsilon^2}. \tag{6}$$

For large p, one may use $\tilde{n}_{p,0}(\epsilon, \alpha)$ in 4 with $\det(\Sigma) = 1$. *Noting the fact that p and n must satisfy $p = o(\sqrt{n})$ (Portnoy, 1987), one may take the sample size to be*

$$n_0 = \max(n_{p,0}(\epsilon, \alpha),\ \tilde{n}_{p,0}(\epsilon, \alpha),\ \sqrt{n}/(c_0 \log(n)),\ p),$$

where c_0 is a constant ($c_0 = 1$ in our study). More generally, this can be used for SSD in the uniform sampling and data splitting in machine learning.

Remark 3. *Given α and sample size r, one obtains the* observed ROE *from solving 3 for ϵ,*

$$\epsilon(\alpha, r) = \frac{\sqrt{\pi q_\alpha(p)}}{\Gamma^{1/p}(p/2+1)} \frac{\sqrt[2p]{\det(\Sigma)}}{\sqrt{r}}. \tag{7}$$

We shall use it to compare the efficiency of sampling distributions, together with the criterion of MSE, see our extensive simulations and real data application below.

3 ASN and the A-optimal Distributions

In this section, we prove an almost sure ASN result, derive the A-optimal distributions and discuss its relationship to the leverage-scores-based distribution, construct the Scoring Algorithm, and introduce truncation.

Figure 1: Algorithm 1 (Computing the subsampling estimator $\hat{\boldsymbol{\beta}}_r^*$)

1. Construct a distribution $\boldsymbol{\pi}$ on the data points $(\mathbf{x}_i, y_i)$'s, use it to draw a subsample $(\mathbf{X}^*, \mathbf{y}^*)$ of size $r << n$ from $(\mathbf{X}, \mathbf{y})$, and formulate the diagonal matrix $\mathbf{W}^* = \text{diag}(1/r\boldsymbol{\pi}^*)$ with $\boldsymbol{\pi}^*$ the corresponding probability vector.
2. Calculate the weighted least squares estimator $\hat{\boldsymbol{\beta}}_r^* = (\mathbf{X}^{*\top}\mathbf{W}^*\mathbf{X}^*)^{-1}\mathbf{X}^{*\top}\mathbf{W}^*\mathbf{y}^*$.

3.1 Asymptotic Normality

We give a set of conditions for the almost sure asymptotic normality of $\hat{\boldsymbol{\beta}}_r^*$ for an arbitrary sampling distribution. Occasionally, we write $\boldsymbol{\pi} = \boldsymbol{\pi}_n$ and $\pi_i = \pi_{n,i}$ to stress their dependene on the sample size n.

(M1)
$$\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top(\varepsilon_i^2 - \sigma^2)}{n\pi_{n,i}} = O(1), \quad a.s.$$

(M2) There is a $p \times p$ symmetric matrix Γ whose smallest eigenvalue is bounded away from zero, i.e., $\lambda_{\min}(\Gamma) \geq b_0 > 0$ for some constant b_0, such that
$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top = \Gamma + o(1).$$

(M3)
$$\frac{1}{n}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i\|^4}{n\pi_{n,i}} = O(1) \quad a.s.$$

(M4) $\mathbb{L}_n(\boldsymbol{\pi}) =: n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top/(n\pi_{n,i})$ satisfies $0 < b \leq \lambda_{\min}\mathbb{L}_n(\boldsymbol{\pi}) \leq \lambda_{\max}\mathbb{L}_n(\boldsymbol{\pi}) \leq B < \infty$ a.s. for constants b, B, where $\lambda_{\min}$ and $\lambda_{\max}$ denote the maximum and minimum eigenvalues, respectively.

(M5) Lindeberg condition: the double array $\boldsymbol{\eta}_{n,i} := \mathbf{x}_i\varepsilon_i/(n\pi_{n,i})$, $i = 1, 2, \ldots, n$, $n \geq 1$ satisfies that for any $t > 0$,
$$\frac{1}{n}\sum_{j=1}^{n}\frac{\|\mathbf{x}_i\|^2\varepsilon_i^2}{n\pi_{n,i}}\mathbf{1}\left[\frac{\|\mathbf{x}_i\||\varepsilon_i|}{n\pi_{n,i}} \geq \sqrt{r}t\right] = o(1), \quad a.s. \quad r \to \infty.$$

(D1) Condition (M1) can be verified using the result on the SLLN for weighted i.i.d. rv's of Baxter, *et al.* (2004)[1]. Specifically, for a sequence $\{a_i\}$, $\frac{1}{n}\sum_{j=1}^{n}|a_i|^q = O(1)$ for some $q > 1$ implies $\frac{1}{n}\sum_{j=1}^{n}a_i\xi_i \to 0$ a.s. for an i.i.d. $\{\xi_n\}$ with $\text{E}(\xi_1) = 0$ and $E(|\xi_1|) < \infty$.

(D2) Condition (M2) was used in Lemma 3.1 of Portnoy (1984)[15].

Theorem 1. *Assume (M1)–(M5). Suppose that for every $\varrho > 0$,*
$$\max_{1\leq i\leq n}\|\mathbf{x}_i\| = o(n^{1/2}\log^{-\varrho}(n)), \quad a.s. \tag{8}$$
Suppose that there exists some $\rho > 2$ such that
$$E(|\varepsilon_1|^\rho) < \infty. \tag{9}$$
Then $\hat{\boldsymbol{\beta}}_r^*$ is asymptotically normal along almost all the sample paths of the sequence $\{(\mathbf{x}_i, y_i)\}$, i.e.,
$$\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\pi})\sqrt{r}(\hat{\boldsymbol{\beta}}_r^* - \hat{\boldsymbol{\beta}}_{\text{ols}}) \implies \mathcal{N}(0, \mathbf{I}_p), \quad a.s. \quad r \to \infty, \tag{10}$$
where $\boldsymbol{\Sigma}(\boldsymbol{\pi}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\text{Diag}(\hat{\boldsymbol{\varepsilon}}^2/\boldsymbol{\pi})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}$ with $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$.

Remark 4. *For the uniform distribution, $\pi_i = 1/n$, (M1) – (M5) are usual assumptions, which are independent of the sampling distribution $\boldsymbol{\pi}$. This is true in general if $n\pi_i \geq l_0$ for some positive constant l_0. For later use, we shall denote the usual assumptions by (M1') – (M5').*

Remark 5. *The leverage scores are widely used in the development of stochastic algorithms, see e.g. Ma,* et al.*(2015) [13]. The scores induce a distribution given by $\boldsymbol{\ell} = (h_{i,i}/p) =: (\ell_i)$, where $h_{i,i}$ are the diagonal entries of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. Assume that there exist positive constants c_1, c_2 such that uniformly in n,*
$$\lambda_{\max}(n^{-1}\mathbf{X}^\top\mathbf{X}) \leq c_1, \quad \|\mathbf{x}_i\| \geq c_2, \quad i = 1, 2, \ldots, n. \tag{11}$$
From $h_{i,i} = \mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i$ it follows $\ell_i \geq c_2^2/(pc_1n)$. Thus (M1') – (M5') are sufficient conditions for (M1) – (M5).

3.2 The A-optimal Distributions

For a $q \times p$ matrix $\mathbf{A}$, we minimize the trace norm $\operatorname{Tr}(\boldsymbol{\Sigma}_{\mathbf{A}})$ over distributions on the data points, where

$$\boldsymbol{\Sigma}_{\mathbf{A}}(\boldsymbol{\pi}) = \mathbf{A}\Sigma(\boldsymbol{\pi})\mathbf{A}^\top = \mathbf{A}(\mathbf{X}^\top\mathbf{X})^{-1}\Sigma_c(\boldsymbol{\pi})(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top, \tag{12}$$

with $\Sigma_c(\boldsymbol{\pi}) = \mathbf{X}^\top \mathrm{Diag}(\hat{\varepsilon}^2/r\boldsymbol{\pi})\mathbf{X}$. Let $\hat{\boldsymbol{\theta}} = \mathbf{A}\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$. The plug-in estimate $\hat{\boldsymbol{\theta}}^* = \mathbf{A}\hat{\boldsymbol{\beta}}^*_r$ of $\hat{\boldsymbol{\theta}}$ has $\mathrm{Var}^*(\hat{\boldsymbol{\theta}}^*) = \boldsymbol{\Sigma}_{\mathbf{A}}(\boldsymbol{\pi})$. Consider $\boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\beta})$, where $\mathbf{g}$ has the continuous partial derivative $\dot{\mathbf{g}}$. Then $\hat{\boldsymbol{\theta}}^* = \mathbf{g}(\hat{\boldsymbol{\beta}}^*_r)$ is a subsampling estimator to approximate $\hat{\boldsymbol{\theta}} = \mathbf{g}(\hat{\boldsymbol{\beta}}_{\mathrm{ols}})$, and an A-optimal distribution for $\hat{\boldsymbol{\theta}}^*$ to approximate $\hat{\boldsymbol{\theta}}$ is given by taking $\mathbf{A} = \dot{\mathbf{g}}(\bar{\boldsymbol{\beta}})$ for some pilot estimator $\bar{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.

An A-optimal distribution depends on data, parameters, and the estimation method. With these in mind and for convenience, we introduce the following definition.

Definition 2. *Given a σ-field $\mathcal{F}$, a distribution $\boldsymbol{\pi}$ supported on the data points is said to be A-optimal for the subsampling estimate $\hat{\boldsymbol{\theta}}^*$ to approximate an estimate $\hat{\boldsymbol{\theta}}$ of parameter $\boldsymbol{\theta}$ if $\boldsymbol{\pi}$ asymptotically minimizes the trace norm of the conditional variance-covariance matrix $\mathrm{Var}(\hat{\boldsymbol{\theta}}^*|\mathcal{F})$ of $\hat{\boldsymbol{\theta}}^*$ given $\mathcal{F}$.*

If $\mathcal{F}$ is the σ-field generated by $\{(\mathbf{x}_i, y_i)\}$ $(\{\mathbf{x}_i\})$, then $\boldsymbol{\pi}$ is referrred to as $\hat{A}$ $(\bar{A})$-optimal.

The $\hat{A}$-optimal Distribution $\hat{\boldsymbol{\pi}}_2$. Minimizing the trace norm of the variance-covariance matrix $\boldsymbol{\Sigma}_{\mathbf{A}}$ in 12, we obtain the $\hat{A}$-optimalizer $\hat{\boldsymbol{\pi}}_{\mathbf{A}}$. We now invoke the Lagrange multipliers to get

Proposition 1. *Let $\mathbf{A}$ be a $q \times p$ matrix which is independent of $\boldsymbol{\pi}$. Assume that $\mathbf{A}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i \neq 0$ and $h_{i,i} \neq 1$ for all i. Then the square roots of the diagonal entries of $\hat{\mathbf{H}}_{2,\mathbf{A}}$ induce the unique $\hat{A}$-optimal distribution $\hat{\boldsymbol{\pi}}_{\mathbf{A}}$ for $\mathbf{A}\hat{\boldsymbol{\beta}}^*_r$ to approximate $\mathbf{A}\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$, where $\hat{\mathbf{H}}_{2,\mathbf{A}} = \mathrm{Diag}(\hat{\varepsilon})\mathbf{H}_{2,\mathbf{A}}\mathrm{Diag}(\hat{\varepsilon})$ with*

$$\mathbf{H}_{2,\mathbf{A}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top\mathbf{A}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top. \tag{13}$$

We shall refer to $\hat{\mathbf{H}}_{2,\mathbf{A}}$ as the $\hat{A}$-*optimal score matrix.* Write $p_i \propto b_i$ if $p_i = b_i/\sum_j b_j$ for all i. Then $\hat{\boldsymbol{\pi}}_{\mathbf{A}}$ is given by

$$\hat{\pi}_{\mathbf{A},i} \propto \|\mathbf{A}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i\|\,|\hat{\varepsilon}_i|. \tag{14}$$

For $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{1-\alpha/2}$, set $\mathbf{H}_\alpha = \mathbf{H}_{2,\mathbf{A}}$ and $\hat{\mathbf{H}}_\alpha = \hat{\mathbf{H}}_{2,\mathbf{A}}$, so that

$$\mathbf{H}_\alpha = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\alpha}\mathbf{X}^\top, \quad \hat{\mathbf{H}}_\alpha = \mathrm{Diag}(\hat{\varepsilon})\mathbf{H}_\alpha\mathrm{Diag}(\hat{\varepsilon}), \quad \alpha \in \mathbb{R}.$$

It then follows $\hat{\mathbf{H}}_\alpha$ is the $\hat{A}$-optimal score matrix for $\hat{\boldsymbol{\theta}}^*_\alpha = (\mathbf{X}^\top\mathbf{X})^{1-\alpha/2}\hat{\boldsymbol{\beta}}^*_r$ to approximate $\hat{\boldsymbol{\theta}}_\alpha = (\mathbf{X}^\top\mathbf{X})^{1-\alpha/2}\hat{\boldsymbol{\beta}}_{\mathrm{ols}} = (\mathbf{X}^\top\mathbf{X})^{-\alpha/2}\mathbf{X}^\top\mathbf{y}$, with the unique $\hat{A}$-optimal distribution $\hat{\boldsymbol{\pi}}_\alpha$ given by

$$\hat{\pi}_{\alpha,i} \propto \sqrt{h_{\alpha,i,i}}|\hat{\varepsilon}_i|, \quad \text{where} \quad h_{\alpha,i,i} = \mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-\alpha}\mathbf{x}_i. \tag{15}$$

As a result, $\hat{\boldsymbol{\pi}}_2 = (\sqrt{h_{2,i,i}}|\hat{\varepsilon}_i|)$ is the unique $\hat{A}$-optimal distribution for $\hat{\boldsymbol{\beta}}^*_r$ to approximate $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$.

Remark 6. *While $\hat{\pi}_{0,i} \propto \|\mathbf{x}_i\||\hat{\varepsilon}_i|$ has less computational cost than $\hat{\boldsymbol{\pi}}_\alpha$ ($\alpha \neq 0$) (as only $\|\mathbf{x}_i\|$ and $|\hat{\varepsilon}_i|$ must be computed), $\hat{\pi}_{1,i} \propto \sqrt{h_{i,i}}|\hat{\varepsilon}_i|$ can be computed using the fast algorithm given in Drineas, et al. (2006)[6].*

Remark 7. *Notice that the unique $\hat{A}$-optimal distribution for $\hat{\boldsymbol{\theta}}^*_0 = (\mathbf{X}^\top\mathbf{X})\hat{\boldsymbol{\beta}}^*_r$ to approximate $\hat{\boldsymbol{\theta}}_0 = (\mathbf{X}^\top\mathbf{X})\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$ is $\hat{\boldsymbol{\pi}}_0$, neither $\hat{\boldsymbol{\pi}}_2$ nor any other distribution. This note applies in general.*

The $\bar{A}$-optimal $\bar{\boldsymbol{\pi}}_2$ and its Approximation $\tilde{\boldsymbol{\pi}}_2$. Consider minimizing the trace norm of the conditional variance-covariance matrix given $\mathbf{X}$. Since $\hat{\tau}_{\mathbf{A}}(\boldsymbol{\pi}) = \mathrm{Tr}(\boldsymbol{\Sigma}_{\mathbf{A}}(\boldsymbol{\pi})) = r^{-1}\sum_{i=1}^n \|\mathbf{a}_i\|^2\hat{\varepsilon}_i^2/\pi_i$ and $\mathrm{Var}(\hat{\varepsilon}|\mathbf{X}) = (\mathbf{I}_n - \mathbf{H})\sigma^2$, we integrate out the squared residuals in the trace $\hat{\tau}_{\mathbf{A}}(\boldsymbol{\pi})$ to get

$$\bar{\tau}_{\mathbf{A}}(\boldsymbol{\pi}) = \mathrm{E}(\tau_{\mathbf{A}}(\boldsymbol{\pi})|\mathbf{X}) = \frac{\sigma^2}{r}\sum_{i=1}^n \frac{\|\mathbf{a}_i\|^2(1-h_{i,i})}{\pi_i}, \quad \mathbf{a}_i = \mathbf{A}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i. \tag{16}$$

Suppose that $h_{i,i}$'s satisfy $\max_{i=1,\ldots,n} h_{i,i} = o(1)$. One then obtains an approximation to the trace as follows:

$$\tilde{\tau}_{\mathbf{A}}(\boldsymbol{\pi}) = \frac{\sigma^2}{r}\sum_{i=1}^n \frac{\|\mathbf{a}_i\|^2}{\pi_i}.$$

Minimizing $\bar{\tau}_{\mathbf{A}}(\boldsymbol{\pi})$ and $\tilde{\tau}_{\mathbf{A}}(\boldsymbol{\pi})$ yields the sampling distributions $\bar{\boldsymbol{\pi}}_{\mathbf{A}}$ and $\tilde{\boldsymbol{\pi}}_{\mathbf{A}}$, respectively. Note the conditional version of $\hat{\mathbf{H}}_{2,\mathbf{A}}$ in 13 takes the form,

$$\bar{\mathbf{H}}_{2,\mathbf{A}} = \mathrm{Diag}((1-h_{i,i})^{1/2})\mathbf{H}_{2,\mathbf{A}}\mathrm{Diag}((1-h_{i,i})^{1/2}).$$

Thus $\bar{\boldsymbol{\pi}}_{\mathbf{A}}$ is given by $\bar{\pi}_{\mathbf{A},i} \propto \|\mathbf{a}_i\|\sqrt{1-h_{i,i}}$. For $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{1-\alpha/2}$, let $\bar{\mathbf{H}}_\alpha = \bar{\mathbf{H}}_{2,\mathbf{A}}$. The $\bar{A}$-optimal $\bar{\boldsymbol{\pi}}_\alpha$ is

$$\bar{\pi}_{\alpha,i} \propto \sqrt{h_{\alpha,i,i}}\sqrt{1-h_{i,i}}. \tag{17}$$

Hence $\bar{\boldsymbol{\pi}}_2$ is the unique $\bar{A}$-optimal distribution for $\hat{\boldsymbol{\beta}}_r^*$ to approximate $\hat{\boldsymbol{\beta}}_{\text{ols}}$. Likewise, $\tilde{\boldsymbol{\pi}}_\alpha$ is given by

$$\tilde{\pi}_{\alpha,i} \propto \sqrt{h_{\alpha,i,i}}. \tag{18}$$

Remark 8. *As in Remark 6, while $\bar{\boldsymbol{\pi}}_1, \tilde{\boldsymbol{\pi}}_1$ can be fast computed, $\bar{\boldsymbol{\pi}}_0, \tilde{\boldsymbol{\pi}}_0$ enjoy computational ease. The latter are, respectively, the optimal sampling (OPT) and predictor-length (PL) sampling given in Zhu,* et al. *(2015).*

Comparison and Truncation. Since $\hat{\boldsymbol{\pi}}_{\mathbf{A}}$ minimizes $\hat{\tau}_{\mathbf{A}}(\boldsymbol{\pi})$, it follows from Proposition 1 that $\hat{\tau}_{\mathbf{A}}(\hat{\boldsymbol{\pi}}_{\mathbf{A}}) \le \hat{\tau}_{\mathbf{A}}(\bar{\boldsymbol{\pi}}_{\mathbf{A}})$. Hence, by 16, we obtain

$$\mathrm{E}(\hat{\tau}_{\mathbf{A}}(\hat{\boldsymbol{\pi}}_{\mathbf{A}})) \le \mathrm{E}(\hat{\tau}_{\mathbf{A}}(\bar{\boldsymbol{\pi}}_{\mathbf{A}})) = \bar{\tau}_{\mathbf{A}}(\bar{\boldsymbol{\pi}}_{\mathbf{A}}).$$

This shows that $\hat{\boldsymbol{\pi}}_{\mathbf{A}}$ is, on average, better A-optimizing than $\bar{\boldsymbol{\pi}}_{\mathbf{A}}$. Our extensive simulations and real data applications exhibited that $\hat{\boldsymbol{\pi}}_{\mathbf{A}}$ was much better optimizing than both $\bar{\boldsymbol{\pi}}_{\mathbf{A}}$ and $\tilde{\boldsymbol{\pi}}_{\mathbf{A}}$.

Truncation. Observe that 14 implies that $(\mathbf{x}_i, y_i)$ must be drawn with probability $\hat{\pi}_{\mathbf{A},i}$ proportional to $|\hat{\varepsilon}_i|$. Since each probability is inversely used in constructing $\hat{\boldsymbol{\beta}}_r^*$, $\hat{\pi}_{\mathbf{A}}$ must be truncated from below in order to guarantee appropriate statistical properties for $\hat{\boldsymbol{\beta}}_r^*$. In fact, similar to Remark 5, we have

Remark 9. *Assume 11. Then $h_{\alpha,i,i} = \mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-\alpha}\mathbf{x}_i \ge c_2^2 c_1^{-\alpha} n^{-\alpha}$ for $\alpha \ge 0$ and all i. Assume, furthermore, that there exist positive constants c_0, c_3 such that*

$$\lambda_{\min}(n^{-1}\mathbf{X}^\top\mathbf{X}) \ge c_0, \quad \|\mathbf{x}_i\| \le c_3, \quad i = 1, 2, \ldots, n,\ n \ge 1.$$

Then $\sum_{i=1}^n \sqrt{h_{\alpha,i,i}}|\hat{\varepsilon}_i| \le c_3 c_0^{-\alpha/2} n^{-\alpha/2} S$, where $S = \sum_{i=1}^n |\hat{\varepsilon}_i|$. Therefore, by 15, 17 – 18,

$$\hat{\pi}_{\alpha,i} \ge c|\hat{\varepsilon}_i|/S, \quad \bar{\pi}_{\alpha,i} \ge c\sqrt{1-h_{i,i}}/\bar{S}, \quad \tilde{\pi}_{\alpha,i} \ge c, \quad i = 1, 2, \ldots, n,$$

where $\bar{S} = \sum_{i=1}^n \sqrt{1-h_{i,i}}$ and $c = (c_0/c_1)^{\alpha/2} c_2/c_3$. As in Remark 5, (M1') – (M5'), which are used for the uniform distribution in Remark 4, are sufficient conditions for (M1) – (M5) for $\tilde{\boldsymbol{\pi}}_\alpha$; for $\bar{\boldsymbol{\pi}}_\alpha$ if, additionally, $h_{i,i}$ are uniformly bounded away (by a constant) from one; but not for $\hat{\boldsymbol{\pi}}_\alpha$. This exhibits that the above conditions are not enough for $\hat{\pi}_{\alpha,i}$ to be bounded away from zero. In fact, it is necessary to truncate $\hat{\boldsymbol{\pi}}_\alpha$ from below for $\hat{\boldsymbol{\pi}}_\alpha$ to satisfy (M1)–(M5).

Truncation was used in constructing the generalized bootstrap estimator by Chatterjee and Bose (2002)[4]. Specifically, we truncate $\hat{\boldsymbol{\pi}}_{\mathbf{A}}$ from below by L/n and define $\hat{\boldsymbol{\pi}}_{\mathbf{A}}(l)$ by

$$\hat{\pi}_{\mathbf{A},i}(l) \propto \hat{\pi}_{\mathbf{A},i}\mathbf{1}[\hat{\pi}_{\mathbf{A},i} \ge L/n] + (l/n)\mathbf{1}[\hat{\pi}_{\mathbf{A},i} < L/n], \quad i = 1, 2, \ldots, n,$$

where L is a threshold value. Typically $0 < L \le 1$. This is, in fact, a mixture distribution of the optimal and the uniform distributions. For fast computing, we may drop "unimportant" observations by taking $l = 0$, otherwise $l = L$. See p. 18 (Tropp, 2019)[17] for further discussion. As $\bar{\pi}_{\mathbf{A},i} = 0$ at $h_{i,i} = 1$, we truncate $\bar{\pi}_{\mathbf{A},i}$ similarly from below by $\bar{\pi}_{\mathbf{A},i}(l)$. Although $\tilde{\boldsymbol{\pi}}_{\mathbf{A}}$ is positive, we also truncate it and define the likewise $\tilde{\boldsymbol{\pi}}_{\mathbf{A}}(l)$.

To determine the value of L, we must take it into consideration the desired running time and the accuracy. Our extensive numerical results exhibited that even high percentages of truncation led to only slight loss of efficiency.

The Scoring Algorithm. Like a typical optimal sampling, the A-optimal sampling $\hat{\boldsymbol{\pi}}_2$, $\bar{\boldsymbol{\pi}}_2$ and $\tilde{\boldsymbol{\pi}}_2$ have the same running time $O(np^2)$ as the full data LSE $\hat{\boldsymbol{\beta}}_{\text{ols}}$. We provide a fast algorithm in Fig. 2.

Since the computational bottleneck is to invert $\mathbf{X}^\top\mathbf{X}$, we shall approximate it by the subsampling $(\mathbf{X}_0^{*\top}\mathbf{X}_0^*)^{-1}$ based on a computationally easy pre-subsample $(\mathbf{X}_0^*, \mathbf{y}_0^*)$ from the data $(\mathbf{X}, \mathbf{y})$. Let the resulting estimator and residuals be

$$\hat{\boldsymbol{\beta}}_0^* = (\mathbf{X}_0^{*\top}\mathbf{X}_0^*)^{-1}\mathbf{X}_0^{*\top}\mathbf{y}_0^*, \quad \hat{\boldsymbol{\varepsilon}}_0^* = \mathbf{y}_1 - \mathbf{X}_1\hat{\boldsymbol{\beta}}_0^*,$$

where $(\mathbf{X}_1, \mathbf{y}_1)$ is the remaining observations in $(\mathbf{X}, \mathbf{y})$. Compute one of

$$\mathbf{H}_{0,\alpha}^* = \mathbf{X}_1(\mathbf{X}_0^{*\top}\mathbf{X}_0^*)^{-\alpha}\mathbf{X}_1^\top, \quad \hat{\mathbf{H}}_{0,\alpha}^*, \quad \text{and} \quad \bar{\mathbf{H}}_{0,\alpha}^*, \quad \alpha = 1, 2. \tag{19}$$

Our simulations in Section 4 exhibited that the Scoring Algorithm performed paticularly well.

Figure 2: The Scoring Algorithm

1. Take a uniform pre-subsample $(\mathbf{X}^*_0, \mathbf{y}^*_0)$ of size r_0 from $(\mathbf{X}, \mathbf{y})$, and use it to compute $\mathbf{H}^*_{0,\alpha}$ ($\bar{\mathbf{H}}^*_{0,\alpha}$ or $\hat{\mathbf{H}}^*_{0,\alpha}$) given in 19.
2. Call Algorithm 1 in Fig. 1 with the subsample size r and the A-optimal distribution $\boldsymbol{\pi}$.

Remark 10. *The Algorithm in Fig. 2 can be implemented in $O(\max(r_0, r)\ p^2)$ much faster than the original running time $O(np^2)$ as $\max(r_0, r) << n$.*

The Leverage Scores and its Relationship with the A-optimal Distributions.

The formula $\ell_i = \mathbf{u}_i^\top \mathbf{u}_i/p$ indicates that ℓ_i depends *only* on the singular vector $\mathbf{u}_i$ of $\mathbf{X}$. Meanwhile, since the $\hat{A}$-optimal $\hat{\pi}_{2,i}$ depends on $h_{2,i,i}$, which can be written as

$$h_{2,i,i} = \mathbf{u}_i^\top \mathrm{Diag}(1/\sigma_1^2, \ldots, 1/\sigma_p^2)\mathbf{u}_i,$$

it follows that $\hat{\pi}_{2,i}$ depends on not only $\mathbf{u}_i$ but all the singular values σ_i's of $\mathbf{X}$. These suggest that $\boldsymbol{\ell}$ is not efficient in extracting information as it ignores the information in the singular values.

Suppose that $\mathbf{X}$ is column-orthonormal. Then $h_{i,i} = \|\mathbf{x}_i\|^2$ and

$$\bar{\pi}_{2,i} \propto \begin{cases} \sqrt{h_{i,i}} + o(1), & h_{i,i} = o(1), \\ \sqrt{1 - h_{i,i}} + o(1), & h_{i,i} = 1 - o(1). \end{cases}$$

When sampling according to $\boldsymbol{\ell}$, the ith observation is drawn with probability proportional to $h_{i,i}$, especially in the vicinity of $h_{i,i} = 1$. The $\bar{A}$-optimality, however, dictates that in this vicinity the ith observation must be drawn with the probability proportional to $\sqrt{1 - h_{i,i}}$ — decreasing with $h_{i,i}$. In fact, the increasing relationship occurs in the vicinity of $h_{i,i} = 0$ with the probability proportional to $\sqrt{h_{i,i}}$. Similarly, $\hat{\pi}_{2,i} \propto h^{1/2}_{2,i,i}|\hat{\varepsilon}_i|$, suggesting data points closer to the regression hyperplane is less informative than those farther away.

4 Simulations and Real Data Applications

In this Section, we report simulations and real data application about the numerical behaviors of the A-optimal distributions and their comparison with the uniform and the leverage scores (lev) based distributions.

Simulated MSE. As in Zhu, *et al.* (2015)[22], we chose the coefficient $\boldsymbol{\beta} = (\mathbf{1}_{30}^\top,\ 0.1 \cdot \mathbf{1}_{20}^\top)^\top$, generated $p = 50$-dimensional covariate vector $\mathbf{x}$ (treated as non-random) from Gaussian $\mathrm{N}(\mathbf{0}, \Sigma)$ (GA), Log-normal $\exp(\mathrm{N}(\mathbf{0}, \Sigma))$(LN), and Mixing Gaussian $0.5\mathrm{N}(\mathbf{0}, \Sigma) + 0.5\mathrm{N}(\mathbf{0}, 25\Sigma)$(MG) with $\Sigma_{ij} = 2 * 0.8^{|i-j|}$. The random error ε was generated from the normal ($\mathscr{N}$) and the logistic ($\mathscr{L}$), both with zero mean and unit standard deviation. For sample size $n = 10^5$ and a few subsample sizes r, we calculated the empirical mean squared errors of $\hat{\boldsymbol{\beta}}^*_r$ as follows:

$$\mathrm{EMSE}(\hat{\boldsymbol{\beta}}^*_r) = \frac{1}{M} \sum_{m=1}^{M} \|\hat{\boldsymbol{\beta}}^*_m - \hat{\boldsymbol{\beta}}_{\mathrm{ols}}\|^2, \quad M = 500.$$

Reported on Tables 6–10 are the ratios of the EMSE of $\hat{\boldsymbol{\beta}}^*_r$ to that of the uniform subsampling estimator, where the sampling distributions are untruncated in Table 6 and truncated in Tables 7-10; the residual $\hat{\boldsymbol{\varepsilon}}$ was computed based on the full sample $(\mathbf{X}, \mathbf{y})$ in Tables 6-7 and on a uniform pre-subsample $(\mathbf{X}^*_0, \mathbf{y}^*_0)$ of size $0.1n$ in Tables 9-10. In addition, the Scoring Algorithm in Fig. 2 was used in Table 10.

Observe first that the ratios in all the tables are almost all less than one, indicating that the uniform sampling is ineffective in extracting information. This is most noticeable for $\hat{A}$-optimal sampling, and for the LN covariate in which some of the ratios were as low as 25%. Note that the LN is skewed, whereas both GA and MG are symmetric in which the uniform sampling had better performance. Second, the small differences of the ratios in all the tables indicated that the uniform pre-subsampling of a small size resulted in small loss of efficiency, and that the Scoring Algorithm worked well. Third, the $\hat{A}$-optimal sampling performed the best, and gave substantially smaller EMSE ratios than $\bar{A}$-, $\tilde{A}$- and the leverage scores based sampling. In particular, $\hat{\boldsymbol{\pi}}_2$ gave the smallest EMSE ratios in Table 6, when the subsample size reached half the full sample size, which was mostly kept for the truncated sampling distributions in Tables 7-10.

The Running Time. Reported on Table 12 are the running times of the Scoring Algorithm and the LSE. They were measured on a computing cluster with 16 processors running at 2.60GHz with 250GB of memory. The *R* package (ver

Table 1: PUMS: the empirical MSE ratios of the subsampling LSE $\hat{\boldsymbol{\beta}}_r^*$ using the A-optimal Subsampling to using the Uniform. The Scoring Algorithm was used with 1% truncation for sample size $n = 6,688,524$.

$r : n$	$\hat{\pi}_0$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\bar{\pi}_0$	$\bar{\pi}_1$	$\bar{\pi}_2$	$\tilde{\pi}_0$	$\tilde{\pi}_1$	$\tilde{\pi}_2$
.05%	.599	.304	.208	.981	.628	.578	.935	.661	.597
.10%	.556	.279	.185	.969	.619	.677	.968	.700	.615
.50%	.542	.270	.174	.923	.719	.610	.923	.675	.611
1.0%	.524	.291	.178	.879	.688	.586	.868	.729	.592
5.0%	.449	.268	.178	.935	.640	.608	.949	.684	.549

3.3.1) was used to carry out the numerical computations. Since $\mathbf{X}^\top\mathbf{X}$ was approximated by the subsampling $\mathbf{X}_0^{*\top}\mathbf{X}_0^*$, the time-consuming part is the matrix multiplications in $\hat{\mathbf{H}}_2^*$. Instead of using *solve* to find the inverse, we called *svd* to obtain a singular value decomposition of $\mathbf{X}_0^*$ to compute the sampling distribution $\hat{\boldsymbol{\pi}}_2$, and called *lm* to compute both the subsampling estimator $\hat{\boldsymbol{\beta}}_r^*$ and the full data $\hat{\boldsymbol{\beta}}_{\text{ols}}$. The Scoring Algorithm saved time in comparison with the LSE. The times spent on the matrix multiplications were found to be about 30% of the total running times, which can be improved by fast matrix multiplication. Here $\mathbf{x}$ was generated from GA and ε from $\mathscr{N}(0,1)$. The results for the other distributions of $\mathbf{x}$ and ε considered in Table 6 are similar (not reported here).

SSD Reported on Table 13 are the sample sizes for the Lev and the $\hat{\boldsymbol{\pi}}_2$ sampling. We chose the values of $\hat{\delta}$ in $\mathrm{P}(\|\hat{\boldsymbol{\beta}}_r^* - \hat{\boldsymbol{\beta}}\| > \hat{\delta}) < \alpha$ with $\alpha = 0.01$ so that the sample sizes using the Unif sampling were $k \cdot 10^3$ for $k = 1, 2, 5, 10, 20, 50$. The data of sample size $n = 10^5$ were generated from the normal ($\mathscr{N}$) and the logistic ($\mathscr{L}$) distributions for the error ε, and the Gauss mixing (GA), the logarithmic normal (LN) and the mixing Gauss (MG) distributions for the covariate $\mathbf{x}$. The data were truncated at 30%. One observes that the sample sizes using the Lev were almost the same as the Unif, and the sizes using the $\hat{\boldsymbol{\pi}}_2$ were mostly only half the sizes of the Unif (hence also the Lev). Similar results were also obtained (not reported here) for the larger values of $\alpha = 0.05, 0.10$ and the smaller value 10% of truncation.

Income Census Data. The Public Use Microdata Sample (PUMS) contains a sample of actual responses to the American Community Survey. The PUMS dataset includes variables for nearly every question on the survey, and new variables that were derived from multiple survey responses. Each record in the file represents a single person, or – in the household-level– a single housing unit. In the person-level file, individuals are organized into households, making possible the study of people within the contexts of their families and other household members. The PUMS files for an individual year, such as 2016, contain data on approximately one percent of the United States population. The files, covering a five-year period such as 2012-2016, contain data on approximately five percent of the United States population.

We downloaded the 5-year (2012-2016) PUMS data from the US census website[2]. After cleaning, the sample size was reduced to $n = 6,688,524$. We fit the data with the linear regression model to study the influence of the covariates on the response PINCP(total personal income). We used 16 covariates including AGEP(age), COW(class of work), ENG(ability to speak English), GCL(Grandparents living with grandchildren), MAR(Marital status), SCHL(Educational attainment), SCIENGP(Fields of Degrees in Science and Engineering), SEX(Sex), DIS(Disability), NATIVITY, etc.

Reported in Tables 1 and 3 are the ratios of the EMSE of the subsampling LSE $\hat{\boldsymbol{\beta}}_r^*$ using the A-optimal distributions to using the uniform distribution, and in Tables 2 and 4 are the running times in second, based on 500 repetitions. Here the Scoring Algorithm in Fig. 2 was used with truncation rates at 1% and 10%.

The MSE ratios in Table 5 were obtained using 25 covariates in which the variables with multiple levels were converted to indicator variables, thus the data-cleaning led to $n = 6,103,746$ observations. The results in this table indicated that (1) the leverage-scores based sampling was very efficient although it was still less efficient than the A-optimal distributions and far less efficient in our simulations, and (2) only 0.16% of the full data retained almost all efficiency. All the MSE ratios are significantly less than one, suggesting that the A-optimal distributions substantially outperformed the uniform distribution. $\hat{\boldsymbol{\pi}}_2$ gave the smallest MSE ratios, about 0.18, a tremendous improvement over the uniform. The running times were much faster than 24.20 seconds for the full-data LSE $\hat{\boldsymbol{\beta}}_{\text{ols}}$. The Scoring Algorithm and the truncation only resulted in slight loss of efficiency.

[2] https://www.census.gov/programs-surveys/acs/data/pums.html

Table 2: PUMS: The average running times in second for computing Table 1

$r:n$	$\hat{\pi}_0$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\bar{\pi}_0$	$\bar{\pi}_1$	$\bar{\pi}_2$	$\tilde{\pi}_0$	$\tilde{\pi}_1$	$\tilde{\pi}_2$
.05%	.209	.208	.203	.213	.212	.205	.215	.212	.205
.10%	.216	.215	.212	.222	.221	.217	.224	.22	.217
.50%	.334	.333	.327	.346	.340	.337	.338	.342	.334
1.0%	.521	.515	.515	.531	.532	.527	.534	.534	.524
5.0%	1.89	1.89	1.88	1.96	1.95	1.95	1.97	1.97	1.95

Table 3: PUMS: Same as Table 1 but with 10% truncation: the empirical MSE ratios of the subsampling LSE $\hat{\boldsymbol{\beta}}_r^*$ using the A-optimal Subsampling to using the Uniform. The Scoring Algorithm was used with 1% truncation for sample size $n = 6,688,524$.

$r:n$	$\hat{\pi}_0$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\bar{\pi}_0$	$\bar{\pi}_1$	$\bar{\pi}_2$	$\tilde{\pi}_0$	$\tilde{\pi}_1$	$\tilde{\pi}_2$
.05%	.575	.276	.184	.952	.684	.622	.893	.688	.596
.10%	.528	.268	.175	.930	.650	.580	.920	.680	.623
.50%	.490	.254	.181	.958	.666	.620	.882	.655	.645
1.0%	.493	.243	.177	.941	.667	.593	.958	.691	.605
5.0%	.495	.244	.176	.933	.684	.610	.929	.658	.601

5 Proof for Asymptotic Normality

A rv $\mathbf{w} = (w_1, \cdots, w_n)^\top \sim \mathrm{sMult}(\boldsymbol{\pi}, r)$ (the scaled multinomial distribution) for $\boldsymbol{\pi} \in [0,1]^n$ with $\sum_{i=1}^n \pi_i = 1$ if

$$\mathrm{P}\Big(w_1 = \frac{k_1}{r\pi_1}, \ldots, w_n = \frac{k_n}{r\pi_n}\Big) = \frac{r!}{\prod_{i=1}^n k_i!} \prod_{i=1}^n \pi_i^{k_i}, \quad k_i \geq 0, \sum_{i=1}^n k_i = r. \tag{20}$$

It is customary to express $\hat{\boldsymbol{\beta}}_r^*$ in the full data using $\mathbf{w}$, decoupling the resampling scheme from the data. Stochastically equivalently,

$$\hat{\boldsymbol{\beta}}_r^* \overset{d}{=} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y}, \quad \mathbf{W} = \mathrm{Diag}(\mathbf{w}), \tag{21}$$

where $\mathbf{x} \overset{d}{=} \mathbf{y}$ denotes $\mathbf{x}$ and $\mathbf{y}$ have the same distribution. Note that the laws $\mathrm{P}_{\mathbf{w}}$ and P^* governed by $\mathrm{sMult}(\boldsymbol{\pi}, r)$ and $\boldsymbol{\pi}$, respectively, are stochastically equivalent, see, e.g., page 2055, Præstgaard and Wellner (1993)[16] and Zhu, *et al.* (2015)[22]. Such equivalence is commonly used in the bootstrap theory, see Sections 3.5–3.6, Van de Vaart and Wellner (1996)[18]. We shall use P^* also for $\mathrm{P}_{\mathbf{w}}$, and write E^*, Var^*, etc. for the expected value, variance, etc. It is easy to check

$$\mathrm{E}^*(\mathbf{w}) = \mathbf{1}, \quad \mathrm{Cov}^*(\mathbf{w}) = (1/r)(\mathrm{Diag}(1/\boldsymbol{\pi}) - \mathbf{1}\mathbf{1}^\top). \tag{22}$$

Lemma 1. *Assume (M2). Suppose 8 holds for all $\varrho > 0$ and 9 holds for some $\rho > 2$. Then*

$$\|\hat{\boldsymbol{\beta}}_{\mathrm{ols}} - \boldsymbol{\beta}_0\| = O(n^{-1/2} \log_2^{1/2}(n)), \quad a.s. \tag{23}$$

Hence,

$$\max_{1 \leq i \leq n} |\mathbf{x}_i^\top (\hat{\boldsymbol{\beta}}_{\mathrm{ols}} - \boldsymbol{\beta}_0)| = o(1), \quad a.s. \tag{24}$$

PROOF. We show without loss of generality that 23 holds for the first component $\hat{\beta}_1$ of $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$. To do so, we shall apply Theorem 2 of Lai and Wei (1982)[10], for which we need to verify

$$\lim_{n \to \infty} A_n = \infty, \quad \limsup_{n \to \infty} A_{n+1}/A_n < \infty, \quad \text{and} \tag{25}$$

$$\max_{1 \leq i \leq n} |x_{i,1} - \mathbf{k}_n^\top \mathbf{H}_n^{-1} \mathbf{t}_i| = o(n^{1/2} \log^{-\varrho}(n)) \tag{26}$$

for all $\varrho > 0$, where $\mathbf{x}_i = (x_{i,1}, \mathbf{t}_i^\top)^\top$, $\mathbf{k}_n = \sum_{i=1}^n x_{i,1} \mathbf{t}_i$, $\mathbf{H}_n = \sum_{i=1}^n \mathbf{t}_i \mathbf{t}_i^\top$, and $A_n = \sum_{i=1}^n (x_{i,1} - \mathbf{k}_n^\top \mathbf{H}_n^{-1} \mathbf{t}_i)^2$. Partition $\mathbf{M}_0$ as follows:

$$\mathbf{M}_0 = \begin{pmatrix} m_{1,1} & \mathbf{m}_1^\top \\ \mathbf{m}_1 & \mathbf{M}_{1,1} \end{pmatrix}.$$

It follows from (M2) that

$$\frac{1}{n} \sum_{i=1}^n x_{i,1}^2 = m_{1,1} + o(1), \quad \frac{\mathbf{k}_n}{n} = \mathbf{m}_1 + o(1), \quad \frac{\mathbf{H}_n}{n} = \mathbf{M}_{1,1} + o(1). \tag{27}$$

Table 4: PUMS: The average running times in second for computing Table 3

$r:n$	$\hat{\pi}_0$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\bar{\pi}_0$	$\bar{\pi}_1$	$\bar{\pi}_2$	$\tilde{\pi}_0$	$\tilde{\pi}_1$	$\tilde{\pi}_2$
.05%	.202	.204	.205	.211	.209	.205	.214	.210	.208
.10%	.214	.214	.213	.222	.222	.210	.223	.221	.217
.50%	.329	.331	.330	.347	.336	.333	.347	.336	.340
1.0%	.518	.517	.521	.536	.527	.527	.537	.526	.528
5.0%	1.91	1.89	1.88	1.95	1.95	1.95	1.96	1.96	1.95

Table 5: PUMS: the empirical MSE ratios of the subsampling LSE $\hat{\boldsymbol{\beta}}_r^*$ using the A-optimal Subsampling to using the Uniform.

				The Scoring Algorithm r_0						The Scoring Algorthm r_0		
Truncation	$r:n$	Lev	$\boldsymbol{\pi}_2$	$5*10^3$	10^4	$2*10^4$	$r:n$	Lev	$\boldsymbol{\pi}_2$	$5*10^3$	10^4	$2*10^4$
.0	.001	.372	.114	.131	.125	.120	.005	.360	.109	.128	.119	.119
.1	.001	.385	.115	.127	.122	.121	.005	.367	.110	.124	.118	.118
.3	.001	.350	.115	.125	.121	.122	.005	.364	.114	.128	.122	.118

The last two equalities imply $\mathbf{k}_n^\top \mathbf{H}_n^{-1} = \mathbf{m}_1^\top \mathbf{M}_{1,1} + o(1)$. Hence,

$$n^{-1} A_n = m_{1,1} - \mathbf{m}_1^\top \mathbf{M}_{1,1}^{-1} \mathbf{m}_1 + o(1).$$

Since the above difference is positive as it is the inverse of the positive definite matrix $\mathbf{M}_0$, it follows that 25 holds, while 26 follows from the triangle inequality, $\|\mathbf{t}_i\| \le \max_{1\le i\le n} \|\mathbf{x}_i\|$ and 8. Apply now Theorem 2 of Lai and Wei (1982)[10] to finish the proof. □

PROOF (of Theorem 1). Let

$$\bar{\mathbf{w}} = \mathbf{w} - \mathbf{1}, \quad \bar{\mathbf{W}} = \mathbf{W} - \mathbf{I}, \quad \boldsymbol{\Delta}^* = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} - (\mathbf{X}^\top \mathbf{X})^{-1}. \tag{28}$$

Then $\mathrm{E}_{\mathbf{w}}(\bar{\mathbf{w}}) = 0$, $\mathrm{E}_{\mathbf{w}}(\bar{\mathbf{W}}) = 0$, and stochastically equivalently,

$$\boldsymbol{\Delta}^* \overset{d}{=} (\mathbf{X}^{*\top} \mathbf{W}^* \mathbf{X}^*)^{-1} - (\mathbf{X}^\top \mathbf{X})^{-1}, \ \mathbf{X}^\top \bar{\mathbf{W}} \mathbf{y} \overset{d}{=} \mathbf{X}^{*\top} \mathbf{W}^* \mathbf{y}^* - \mathbf{X}^\top \mathbf{y}. \tag{29}$$

Let $\boldsymbol{\Delta}_1^* = -(\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \bar{\mathbf{W}} \mathbf{X})$. Stochastically equivalently,

$$\bar{\boldsymbol{\Delta}}_1^* =: \mathbf{I} - \boldsymbol{\Delta}_1^* = (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{W} \mathbf{X}). \tag{30}$$

Recall $\bar{\mathbf{W}}$ and $\boldsymbol{\Delta}^*$ in 28 and write

$$(\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} = (\mathbf{X}^\top \mathbf{X})^{-1} + \boldsymbol{\Delta}^*, \quad \mathbf{W}\mathbf{y} = \mathbf{y} + \bar{\mathbf{W}}\mathbf{y}.$$

Substitution of them in the full-data formula 21 of $\hat{\boldsymbol{\beta}}_r^*$ yields

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_r^* &= (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y} \\
&= \big((\mathbf{X}^\top \mathbf{X})^{-1} + \boldsymbol{\Delta}^*\big) \mathbf{X}^\top \big(\mathbf{y} + \bar{\mathbf{W}}\mathbf{y}\big) \\
&= \hat{\boldsymbol{\beta}}_{\text{ols}} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \bar{\mathbf{W}} \mathbf{y} + \boldsymbol{\Delta}^* \mathbf{X}^\top \mathbf{y} + \boldsymbol{\Delta}^* \mathbf{X}^\top \bar{\mathbf{W}} \mathbf{y} \\
&= \hat{\boldsymbol{\beta}}_{\text{ols}} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \bar{\mathbf{W}} \hat{\boldsymbol{\varepsilon}} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \bar{\mathbf{W}} \hat{\mathbf{y}} + \boldsymbol{\Delta}^* \mathbf{X}^\top \mathbf{y} + \boldsymbol{\Delta}^* \mathbf{X}^\top \bar{\mathbf{W}} \mathbf{y} \\
&= \hat{\boldsymbol{\beta}}_{\text{ols}} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \bar{\mathbf{W}} \hat{\boldsymbol{\varepsilon}} + \boldsymbol{\Delta}^* \mathbf{X}^\top \bar{\mathbf{W}} \hat{\boldsymbol{\varepsilon}} + [(\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \bar{\mathbf{W}} \hat{\mathbf{y}} + \boldsymbol{\Delta}^* \mathbf{X}^\top \mathbf{y}].
\end{aligned}$$

Substituting $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ in the square bracket, the sum inside it is identically zero. Since all the preceding statements hold on the subspace in which $\mathbf{X}^\top \mathbf{W} \mathbf{X}$ is invertible, we show 31-32,

$$\hat{\boldsymbol{\beta}}_r^* = \hat{\boldsymbol{\beta}}_{\text{ols}} + \frac{1}{r} \sum_{j=1}^{r} (\mathbf{X}^\top \mathbf{X})^{-1} \frac{\mathbf{x}_j^* \hat{\varepsilon}_j^*}{\pi_j^*} + \mathbf{r}^*, \tag{31}$$

valid on the subspace in which $\mathbf{X}^{*\top} \mathbf{W}^* \mathbf{X}^*$ is invertible, where $\mathbf{r}^*$ is given by

$$\mathbf{r}^* = \big((\mathbf{X}^{*\top} \mathbf{W}^* \mathbf{X}^*)^{-1} - (\mathbf{X}^\top \mathbf{X})^{-1}\big)(\mathbf{X}^{*\top} \mathbf{W}^* \hat{\boldsymbol{\varepsilon}}^*). \tag{32}$$

Let A_n^* be the event on which $\bar{\boldsymbol{\Delta}}_1^*$ is nonsingular. Using $\boldsymbol{\Delta}_1^* (\bar{\boldsymbol{\Delta}}_1^*)^{-1} = (\bar{\boldsymbol{\Delta}}_1^*)^{-1} \boldsymbol{\Delta}_1^*$, we express

$$\boldsymbol{\Delta}^* = \boldsymbol{\Delta}_1^* (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} = \boldsymbol{\Delta}_1^* (\bar{\boldsymbol{\Delta}}_1^*)^{-1} (\mathbf{X}^\top \mathbf{X})^{-1} = (\bar{\boldsymbol{\Delta}}_1^*)^{-1} \boldsymbol{\Delta}_1^* (\mathbf{X}^\top \mathbf{X})^{-1},$$

valid on A_n^*. Recalling $\boldsymbol{\delta}^* = (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\bar{\mathbf{W}}\hat{\boldsymbol{\varepsilon}})$, we thus obtain

$$\mathbf{r}^* = \boldsymbol{\Delta}^*\mathbf{X}^\top\bar{\mathbf{W}}\hat{\boldsymbol{\varepsilon}} = (\bar{\boldsymbol{\Delta}}_1^*)^{-1}\boldsymbol{\Delta}_1^*\boldsymbol{\delta}^* \quad \text{valid on } A_n^*. \tag{33}$$

By the second equality in 22, one gets

$$\mathrm{E}^*(\|\boldsymbol{\Delta}_1^*\|^2) \le \frac{1}{r}\sum_{i=1}^n \frac{h_{2,i,i}}{\pi_i}\|\mathbf{x}_i\|^2.$$

Using $\hat{\varepsilon}_i^2 \le 2\varepsilon_i^2 + 2\|\boldsymbol{\beta}_0\|^2\|\mathbf{x}_i\|^2$, one has

$$\mathrm{E}^*(\|\boldsymbol{\delta}^*\|^2) \le \frac{1}{r}\sum_{i=1}^n \frac{h_{2,i,i}}{\pi_i}\hat{\varepsilon}_i^2 \le \frac{2}{r}\sum_{i=1}^n \frac{h_{2,i,i}}{\pi_i}\varepsilon_i^2 + \frac{2\|\boldsymbol{\beta}_0\|^2}{r}\sum_{i=1}^n \frac{h_{2,i,i}}{\pi_i}\|\mathbf{x}_i\|^2.$$

It thus follows from (M1) and (M3) that

$$r[\mathrm{E}^*(\|\boldsymbol{\Delta}_1^*\boldsymbol{\delta}^*\|)]^2 \le r\mathrm{E}^*(\|\boldsymbol{\Delta}_1^*\|^2)\mathrm{E}^*(\|\boldsymbol{\delta}^*\|^2) = o(1), \quad a.s.$$

This, $\bar{\boldsymbol{\Delta}}^* = \mathbf{I} + o_{P^*}(1)$ a.s. and the expression 33 for the remainder $\mathbf{r}^*$ prove $\sqrt{r}\mathbf{r}^* = o_{P^*}(1)$ a.s. Consequently, by 31, it suffices to show for any $\mathbf{t} \in \mathbb{R}^p$ with $\|\mathbf{t}\| = 1$,

$$\frac{\sigma_n^{-1}(\mathbf{t})}{\sqrt{r}}\sum_{j=1}^r \mathbf{t}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\frac{\mathbf{x}_j^*\hat{\varepsilon}_j^*}{\pi_{nj}^*} \Longrightarrow \mathcal{N}(0,1), \quad a.s. \quad r \to \infty, \tag{34}$$

where $\sigma_n^2(\mathbf{t}) = \mathbf{t}^\top\boldsymbol{\Sigma}(\boldsymbol{\pi})\mathbf{t}$. As $\mathbf{X}^\top\hat{\boldsymbol{\varepsilon}} = 0$, we have

$$\mathrm{E}^*\big(\mathbf{x}_j^*\hat{\varepsilon}_j^*/\pi_{nj}^*\big) = \mathbf{X}^\top\hat{\boldsymbol{\varepsilon}} = 0, \quad \mathrm{Var}^*(\mathbf{x}_j^*\hat{\varepsilon}_j^*/\pi_{nj}^*) = \mathbf{X}^\top\mathrm{Diag}(\hat{\boldsymbol{\varepsilon}}^2/\boldsymbol{\pi})\mathbf{X}. \tag{35}$$

Let $\xi_j^* = \mathbf{t}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_j^*\hat{\varepsilon}_j^*/\pi_{nj}^*$. It is shown below for every $\eta > 0$,

$$\sigma_n^{-2}(\mathbf{t})\mathrm{E}^*(|\xi_1^*|^2\mathbf{1}[|\xi_1^*| > \sqrt{r}\sigma_n(\mathbf{t})\eta]) \to 0, \quad a.s. \quad r \to \infty. \tag{36}$$

We now apply the Lindeberg-Feller theorem (e.g. Theorem 7.2.1. of Chung (2001)[5]) to claim 34. To show 36, we prove below

$$\frac{1}{n^2}\sum_{i=1}^n \frac{\mathbf{x}_i\mathbf{x}_i^\top}{\pi_i}(\hat{\varepsilon}_i^2 - \varepsilon_i^2) = o(1), \quad a.s. \tag{37}$$

Let $\boldsymbol{\Sigma}_c = n^{-2}\mathbf{X}^\top\mathrm{Diag}(\hat{\boldsymbol{\varepsilon}}^2/\boldsymbol{\pi})\mathbf{X}$. Then $\boldsymbol{\Sigma}(\boldsymbol{\pi}) = (n^{-1}\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{\Sigma}_c(n^{-1}\mathbf{X}^\top\mathbf{X})^{-1}$. It follows from 37 and (M1) that

$$\begin{aligned}\boldsymbol{\Sigma}_c &= \frac{1}{n^2}\sum_{i=1}^n \frac{\mathbf{x}_i\mathbf{x}_i^\top}{\pi_i}\sigma^2 + \frac{1}{n^2}\sum_{i=1}^n \frac{\mathbf{x}_i\mathbf{x}_i^\top}{\pi_i}(\varepsilon_i^2 - \sigma^2) + \frac{1}{n^2}\sum_{i=1}^n \frac{\mathbf{x}_i\mathbf{x}_i^\top}{\pi_i}(\hat{\varepsilon}_i^2 - \varepsilon_i^2) \\ &= \frac{1}{n^2}\sum_{i=1}^n \frac{\mathbf{x}_i\mathbf{x}_i^\top}{\pi_i}\sigma^2 + o(1), \quad a.s.\end{aligned}$$

We now use (M2) to get

$$\boldsymbol{\Sigma}(\boldsymbol{\pi}) = \sigma^2\Gamma_n^{-1}\frac{1}{n^2}\sum_{i=1}^n \frac{\mathbf{x}_i\mathbf{x}_i^\top}{\pi_i}\Gamma_n^{-1} + o(1), \quad a.s. \tag{38}$$

This immediately yields for any unit vector $\mathbf{t}$,

$$\sigma_n^2(\mathbf{t}) = \sigma^2\mathbf{t}^\top\Gamma_n^{-1}\frac{1}{n^2}\sum_{i=1}^n \frac{\mathbf{x}_i\mathbf{x}_i^\top}{\pi_i}\Gamma_n^{-1}\mathbf{t} + o(1), \quad a.s. \tag{39}$$

By (M2)–(M4), there are constants b_0, B_0 such that

$$0 < b_0 \le \sup \sup_{\|\mathbf{t}\|=1} \sigma_n^2(\mathbf{t}) \le B_0 < \infty, \quad a.s.$$

This shows that 36 is implied by the following (shown below)

$$L(r,n) := \mathrm{E}^*(|\xi_1^*|^2\mathbf{1}[|\xi_1^*| > \sqrt{r}b_0\eta]) \to 0, \quad a.s. \quad r \to \infty. \tag{40}$$

To prove 37, we use (M1) and (M3) to get

$$\frac{1}{n^2}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i\|^2}{\pi_i}\varepsilon_i^2=\frac{1}{n^2}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i\|^2}{\pi_i}(\varepsilon_i^2-\sigma^2)+O(1)=O(1),\ a.s. \tag{41}$$

By 24, we have uniformly in $i=1,\dots,n$,

$$\hat{\varepsilon}_i-\varepsilon_i=\mathbf{x}_i^\top(\hat{\boldsymbol{\beta}}_{\text{ols}}-\boldsymbol{\beta}_0)=o(1),\quad \hat{\varepsilon}_i+\varepsilon_i=2\varepsilon_i+o(1),\quad a.s. \tag{42}$$

Thus $\hat{\varepsilon}_i^2-\varepsilon_i^2=o(1)\varepsilon_i$ a.s. uniformly in i. This yields 37 in view of

$$\Big\|\frac{1}{n^2}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\pi_i}\varepsilon_i\Big\|^2\le\frac{1}{n^2}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i\|^2}{\pi_i}\frac{1}{n^2}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i\|^2}{\pi_i}\varepsilon_i^2=O(1),\quad a.s.$$

where 41 and (M3) were used. To finish, it remains to prove 40. This follows from (M2), (M5), the first equality in 42, and

$$\begin{aligned}L(r,n)&=\sum_{i=1}^{n}\frac{|\mathbf{t}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i|^2}{\pi_i}\hat{\varepsilon}_i^2\mathbf{1}\big[\frac{|\mathbf{t}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i|}{\pi_i}|\hat{\varepsilon}_i|\ge\sqrt{r}b_0\eta\big]\\&\le 2\|\Gamma_n^{-1}\|_o^2\frac{1}{n^2}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i\|^2\hat{\varepsilon}_i^2}{\pi_i}\mathbf{1}\big[\frac{\|\mathbf{x}_i\||\hat{\varepsilon}_i|}{n\pi_i}\ge\frac{\sqrt{r}b_0\eta}{\|\Gamma_n^{-1}\|_o}\big]\\&\le 4\|\Gamma_n^{-1}\|_o^2\frac{1}{n^2}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i\|^2\varepsilon_i^2}{\pi_i}\mathbf{1}\big[\frac{\|\mathbf{x}_i\||\varepsilon_i|}{n\pi_i}\ge\frac{\sqrt{r}b_0\eta}{2\|\Gamma_n^{-1}\|_o}\big]\\&\longrightarrow 0,\quad a.s.\quad r\to\infty.\qquad\qquad \square\end{aligned}$$

References

[1] Baxter, J., Jones, R., Lin, M. and Olsen, J. (2004). SLLN for Weighted Independent Identically Distributed Random Variables. *J. Theoret. Probab.*, **17**: 165–181. doi:10.1023/B:JOTP.0000020480.84425.8d.

[2] Barbe, P. and Bertail, P. (1995). *Weighted bootstrap*. Lecture Notes in Statist. Vol. 98, Springer, New York.

[3] Candés, E.J. and Tao, T. (2009). Exact Matrix Completion via Convex Optimization. *Found Comput Math* **9**: 717. doi:10.1007/s10208-009-9045-5.

[4] Chatterjee, S. and Bose, A. (2002). Dimension asymptotics for generalized bootstrap in linear regression. *Ann. Inst. Statist. Math.* **54** (2): 367–381.

[5] Chung, K.L. (2001). *A Course in Probability Theory*. Academic Press, San Diego, CA.

[6] Drineas P., Kannan R. and Mahoney M.W. (2006). Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM Journal on Computing*, **36**: 132–157.

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

[8] Gordon, R., Litvak, A., Schütt, C. and Werner, E. (2002). Orlicz norms of sequences of random variables. *Ann. Probab.* **30**(4): 1833-1853.

[9] Kleiner, A., Talwalkar, A., Sarkar, P. and Jordan, M. I. (2014). A scalable bootstrap for massive data. *J. R. Stat. Soc. Series B Stat. Methodol.* **76**(4), 795–816.

[10] Lai, T. L. and C. Z. Wei (1982). A Law of the Iterated Logarithm for Double Arrays of Independent Random Variables with Applications to Regression and Time Series Models. *Ann. Probab.* **10**(2): 320–335.

[11] Liang, F., Cheng, Y., Song, Q., Park, J., and Yang, P. (2013). stochastic approximation method for analysis of large geostatistical data. *J. Am. Stat. Assoc.* **108**(501): 325–339.

[12] Ma, P. and Sun, X. (2014). Leveraging for big data regression. *Computational Statistics*. **7** (1): 70-76.

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

[14] Mahoney, M. W. (2011). Randomized algorithms for matrices and data. *arXiv:1104.5557v3* [cs.DS]

[15] Portnoy, S. (1984). Asymptotic Behavior of M-Estimators of p Regression Parameters when p^2/n is Large. I. Consistency. *Ann. Statist.* **12** (4): 1298–1309.

Table 6: Simulated empirical MSE ratios of the subsampling LSE $\hat{\boldsymbol{\beta}}_r^*$ using the A-optimal Subsampling to using the Uniform for sample size $n = 10^5$ and subsample sizes r. The residual $\hat{\varepsilon}$ was computed using the full sample.

$\mathbf{x}$	ε	$r : n$	$\hat{\boldsymbol{\pi}}_2$	$\hat{\boldsymbol{\pi}}_1$	$\hat{\boldsymbol{\pi}}_0$	$\bar{\boldsymbol{\pi}}_2$	$\bar{\boldsymbol{\pi}}_1$	$\bar{\boldsymbol{\pi}}_0$	$\tilde{\boldsymbol{\pi}}_2$	$\tilde{\boldsymbol{\pi}}_1$	$\tilde{\boldsymbol{\pi}}_0$	Lev
GA	$\mathscr{N}$	.5%	.823	.832	.884	.992	.975	.999	.968	.960	1.03	.979
		1%	.784	.783	.813	1.01	.995	1.07	.994	1.01	1.06	1.02
		10%	.649	.658	.685	.981	.983	1.03	.985	.997	1.02	1.00
		45%	.653	.638	.651	.983	.991	1.03	.994	.982	1.06	1.01
		50%	.620	.629	.660	.961	.971	1.04	.965	.964	1.02	.995
	$\mathscr{L}$	.5%	.795	.813	.873	.990	1.02	1.03	.988	.998	1.04	1.02
		1%	.728	.716	.752	1.00	.985	1.02	.990	.987	1.02	.999
		10%	.618	.615	.661	1.03	1.03	1.06	1.01	1.02	1.07	1.04
		45%	.565	.588	.610	.980	.975	1.01	.987	.989	1.04	.998
		50%	.586	.599	.613	1.00	1.00	1.03	.990	.988	1.04	.983
LN	$\mathscr{N}$	.5%	.302	.303	.322	.333	.327	.352	.328	.332	.360	.493
		1%	.281	.278	.306	.338	.334	.366	.338	.331	.360	.599
		10%	.262	.267	.282	.381	.387	.401	.379	.389	.404	.851
		45%	.276	.278	.286	.419	.425	.447	.415	.425	.453	.952
		50%	.280	.280	.293	.430	.428	.450	.431	.435	.441	.977
	$\mathscr{L}$	.5%	.283	.284	.315	.324	.333	.361	.330	.335	.361	.486
		1%	.256	.253	.279	.331	.330	.361	.332	.331	.361	.576
		10%	.238	.238	.254	.382	.388	.404	.382	.385	.402	.848
		45%	.253	.253	.266	.412	.422	.450	.428	.426	.444	.942
		50%	.253	.253	.268	.420	.425	.450	.418	.427	.446	.959
MG	$\mathscr{N}$	.5%	.558	.551	.593	.644	.651	.675	.633	.636	.687	.900
		1%	.515	.506	.542	.655	.662	.709	.649	.651	.690	.948
		10%	.451	.454	.476	.682	.695	.723	.685	.683	.714	1.02
		45%	.438	.446	.458	.684	.692	.719	.694	.682	.698	1.01
		50%	.433	.438	.459	.671	.680	.721	.667	.697	.710	1.00
	$\mathscr{L}$	.5%	.554	.555	.562	.664	.658	.696	.648	.670	.690	.933
		1%	.500	.500	.509	.662	.685	.706	.672	.659	.698	.953
		10%	.399	.408	.428	.658	.654	.713	.673	.660	.690	.971
		45%	.395	.397	.417	.666	.684	.699	.673	.692	.712	.974
		50%	.407	.410	.428	.710	.685	.712	.690	.683	.722	.995

[16] PRÆSTGAARD, J. AND WELLNER, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, **21** (4): 2053–2086.

[17] TROPP, J.A. (2019). Matrix Concentration & Computational Linear Algebra. https://resolver.caltech.edu/CaltechAUTHORS:20190715-125341188.

[18] VAN DE VAART AND WELLNER (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag New York,Inc. 1996.

[19] WANG, H., YANG, M. and STUFKEN, J. (2019). Information-Based Optimal Subdata Selection for Big Data Linear Regression. *J. Am. Stat. Assoc.* **114** (52): 393–405.

[20] WANG, H., ZHU, R., AND MA, P. (2015). Optimal subsampling for large sample logistic regression. *J. Am. Stat. Assoc.* **113** (522): 829–844.

[21] XU, P., YANG, J., ROOSTA-KHORASANI, F., RÉ, C. AND MAHONEY, M.W. (2016). Subsampled Newton Methods with Non-uniform Sampling. *arXiv:1607.00559.v2* [math.OC].

[22] ZHU, R., MA, P., MAHONEY, M.W. AND YU, B. (2015). Optimal subsampling Approaches for Large Sample Linear Regression. *arXiv:1509.0511.v1* [stat.ME].

Table 7: Same as Table 6 except that the sampling distributions are truncated: Simulated empirical MSE ratios of the subsampling LSE $\hat{\boldsymbol{\beta}}_r^*$ using the A-optimal Subsampling to using the Uniform for sample size $n = 10^5$ and subsample sizes r. The residual $\hat{\varepsilon}$ was computed using the full sample.

$\mathbf{x}$	ε	$r : n$	$\hat{\boldsymbol{\pi}}_2$	$\hat{\boldsymbol{\pi}}_1$	$\hat{\boldsymbol{\pi}}_0$	$\bar{\boldsymbol{\pi}}_2$	$\bar{\boldsymbol{\pi}}_1$	$\bar{\boldsymbol{\pi}}_0$	$\tilde{\boldsymbol{\pi}}_2$	$\tilde{\boldsymbol{\pi}}_1$	$\tilde{\boldsymbol{\pi}}_0$	Lev
Truncation 10%												
GA	$\mathscr{N}$	.5%	.800	.812	.852	.985	.964	1.03	1.00	.994	1.02	.976
		1%	.718	.740	.755	.961	1.00	1.03	.985	1.01	1.04	.994
		10%	.646	.646	.680	.982	1.01	1.04	.989	1.00	1.04	1.00
	$\mathscr{L}$	.5%	.744	.775	.788	.977	.981	1.02	.952	.985	1.02	.973
		1%	.668	.686	.714	.964	.999	1.03	.960	.983	1.02	.996
		10%	.595	.588	.625	.995	.999	1.04	1.01	.995	1.02	.998
LN	$\mathscr{N}$	.5%	.305	.302	.322	.337	.330	.353	.323	.320	.361	.458
		1%	.269	.275	.286	.339	.329	.362	.331	.336	.359	.569
		10%	.260	.263	.278	.384	.392	.402	.386	.390	.409	.813
	$\mathscr{L}$	.5%	.287	.279	.303	.328	.327	.370	.324	.331	.358	.462
		1%	.253	.258	.277	.335	.340	.369	.331	.334	.364	.580
		10%	.247	.247	.258	.396	.406	.425	.398	.391	.424	.840
MG	$\mathscr{N}$	.5%	.545	.559	.572	.637	.632	.656	.633	.656	.670	.888
		1%	.522	.514	.556	.656	.674	.708	.661	.688	.705	.964
		10%	.455	.449	.477	.692	.685	.715	.695	.681	.724	.985
	$\mathscr{L}$	.5%	.527	.534	.558	.653	.638	.676	.636	.650	.684	.905
		1%	.478	.476	.504	.664	.663	.697	.649	.665	.687	.955
		10%	.412	.416	.430	.680	.671	.718	.676	.664	.695	.959
Truncation 30%												
GA	$\mathscr{N}$	.5%	.753	.749	.802	.980	.983	1.02	.971	1.01	1.02	1.01
		1%	.705	.689	.726	.970	.974	1.01	.967	.980	1.01	.971
		10%	.664	.684	.708	.995	1.01	1.06	.991	1.03	1.03	1.00
	$\mathscr{L}$	.5%	.701	.712	.730	.978	.983	1.01	.980	.989	.999	.990
		1%	.658	.673	.694	1.00	1.01	1.02	.991	1.00	1.02	1.01
		10%	.612	.619	.638	.989	.983	1.00	.987	.994	1.03	.998
LN	$\mathscr{N}$	.5%	.295	.301	.330	.340	.334	.373	.344	.342	.384	.422
		1%	.269	.266	.295	.326	.332	.356	.323	.341	.356	.500
		10%	.263	.264	.280	.391	.394	.416	.393	.391	.414	.741
	$\mathscr{L}$	.5%	.290	.291	.309	.348	.344	.390	.350	.344	.376	.434
		1%	.258	.257	.276	.331	.342	.364	.337	.340	.362	.515
		10%	.247	.251	.263	.403	.399	.426	.395	.398	.423	.747
MG	$\mathscr{N}$	.5%	.560	.546	.580	.646	.645	.659	.652	.651	.666	.866
		1%	.504	.510	.532	.659	.657	.698	.667	.656	.692	.886
		10%	.456	.466	.481	.685	.681	.733	.685	.677	.709	.945
	$\mathscr{L}$	.5%	.524	.535	.546	.655	.650	.642	.652	.658	.670	.861
		1%	.468	.481	.500	.660	.663	.688	.663	.670	.689	.926
		10%	.421	.420	.439	.693	.692	.710	.675	.693	.724	.939

Table 8: The ROE of the subsampling LSE $\hat{\boldsymbol{\beta}}_r^*$ for sample size $n = 10^5$ and subsample sizes r. The residual $\hat{\boldsymbol{\varepsilon}}$ was computed using the full sample.

$\mathbf{x}$	ε	$r : n$	Unif	Lev	$\hat{\boldsymbol{\pi}}_2$	Unif	Lev	$\hat{\boldsymbol{\pi}}_2$	Unif	Lev	$\hat{\boldsymbol{\pi}}_2$
			$\alpha = 0.01$			$\alpha = 0.05$			$\alpha = 0.10$		
Truncation 10%											
GA	$\mathscr{N}$	.5%	.2524	.2522	.2013	.2376	.2374	.1896	.2298	.2296	.1834
		1%	.1784	.1783	.1424	.1680	.1679	.1340	.1625	.1624	.1297
		10%	.0564	.0564	.0450	.0531	.0531	.0424	.0514	.0514	.0410
	$\mathscr{L}$	.5%	.2534	.2531	.1938	.2386	.2383	.1824	.2308	.2305	.1765
		1%	.1792	.1789	.1370	.1687	.1685	.1290	.1632	.1630	.1248
		10%	.0567	.0566	.0433	.0534	.0533	.0408	.0516	.0515	.0395
LN	$\mathscr{N}$	.5%	.0394	.0380	.0204	.0371	.0358	.0192	.0359	.0346	.0186
		1%	.0279	.0269	.0145	.0262	.0253	.0136	.0254	.0245	.0132
		10%	.0088	.0085	.0046	.0083	.0080	.0043	.0080	.0077	.0042
	$\mathscr{L}$	.5%	.0395	.0394	.0207	.0372	.0371	.0195	.0360	.0358	.0189
		1%	.0279	.0278	.0147	.0263	.0262	.0138	.0255	.0253	.0134
		10%	.0088	.0088	.0046	.0083	.0083	.0044	.0080	.0080	.0042
MG	$\mathscr{N}$	.5%	.0699	.0696	.0464	.0659	.0655	.0437	.0637	.0634	.0423
		1%	.0495	.0492	.0328	.0466	.0463	.0309	.0450	.0448	.0299
		10%	.0156	.0156	.0104	.0147	.0146	.0098	.0142	.0142	.0095
	$\mathscr{L}$	.5%	.0701	.0696	.0445	.0660	.0656	.0419	.0638	.0634	.0405
		1%	.0495	.0492	.0315	.0467	.0464	.0296	.0451	.0448	.0287
		10%	.0157	.0156	.0100	.0148	.0147	.0094	.0143	.0142	.0091
Truncation 30%											
GA	$\mathscr{N}$	.5%	.2524	.2518	.2054	.2376	.2370	.1934	.2298	.2293	.1871
		1%	.1784	.1780	.1452	.1680	.1676	.1368	.1625	.1621	.1323
		10%	.0564	.0563	.0459	.0531	.0530	.0432	.0514	.0513	.0418
	$\mathscr{L}$	.5%	.2534	.2527	.1974	.2386	.2379	.1859	.2308	.2301	.1798
		1%	.1792	.1787	.1396	.1687	.1682	.1314	.1632	.1627	.1271
		10%	.0567	.0565	.0441	.0534	.0532	.0416	.0516	.0515	.0402
LN	$\mathscr{N}$	.5%	.0394	.0365	.0207	.0371	.0344	.0195	.0359	.0333	.0188
		1%	.0279	.0258	.0146	.0262	.0243	.0138	.0254	.0235	.0133
		10%	.0088	.0082	.0046	.0083	.0077	.0044	.0080	.0074	.0042
	$\mathscr{L}$	.5%	.0395	.0378	.0210	.0372	.0356	.0198	.0360	.0344	.0191
		1%	.0279	.0267	.0148	.0263	.0252	.0140	.0255	.0243	.0135
		10%	.0088	.0084	.0047	.0083	.0080	.0044	.0080	.0077	.0043
MG	$\mathscr{N}$	.5%	.0699	.0684	.0470	.0659	.0644	.0442	.0637	.0623	.0428
		1%	.0495	.0483	.0332	.0466	.0455	.0313	.0450	.0440	.0303
		10%	.0156	.0153	.0105	.0147	.0144	.0099	.0142	.0139	.0096
	$\mathscr{L}$	.5%	.0701	.0684	.0450	.0660	.0644	.0424	.0638	.0623	.0410
		1%	.0495	.0484	.0318	.0467	.0456	.0300	.0451	.0441	.0290
		10%	.0157	.0153	.0101	.0148	.0144	.0095	.0143	.0139	.0092

Table 9: The ROE of the subsampling LSE $\hat{\boldsymbol{\beta}}_r^*$ for sample size $n = 10^5$ and subsample sizes r. The residual $\hat{\varepsilon}$ was approximated using a uniform pre-subsample $\mathbf{X}_0^*$ of size $r_0 : n = 10\%$.

$\mathbf{x}$	ε	$r : n$	$\hat{\boldsymbol{\pi}}_2$	$\hat{\boldsymbol{\pi}}_1$	$\hat{\boldsymbol{\pi}}_0$	$\bar{\boldsymbol{\pi}}_2$	$\bar{\boldsymbol{\pi}}_1$	$\bar{\boldsymbol{\pi}}_0$	$\tilde{\boldsymbol{\pi}}_2$	$\tilde{\boldsymbol{\pi}}_1$	$\tilde{\boldsymbol{\pi}}_0$	Lev
Truncation 10%												
GA	$\mathscr{N}$	.5%	.843	.814	.868	.985	1.01	1.06	1.00	1.01	1.03	1.02
		1%	.730	.724	.781	1.00	.984	1.05	.975	.996	1.04	.994
		10%	.649	.651	.699	1.02	.986	1.04	.996	1.01	1.05	1.00
	$\mathscr{L}$	.5%	.808	.798	.846	.988	1.01	1.04	1.02	1.02	1.05	1.02
		1%	.690	.694	.715	.982	.992	1.02	.977	.990	1.03	1.00
		10%	.593	.603	.631	.991	1.01	1.05	.984	1.02	1.03	.986
LN	$\mathscr{N}$	.5%	.288	.287	.311	.320	.324	.365	.322	.325	.353	.474
		1%	.268	.269	.298	.336	.335	.365	.338	.342	.364	.585
		10%	.279	.281	.295	.391	.387	.415	.401	.399	.417	.834
	$\mathscr{L}$	.5%	.256	.266	.286	.318	.316	.349	.312	.320	.341	.455
		1%	.249	.255	.277	.339	.340	.355	.335	.326	.349	.572
		10%	.253	.258	.270	.382	.391	.407	.383	.390	.409	.828
MG	$\mathscr{N}$	.5%	.555	.554	.587	.636	.643	.665	.628	.638	.673	.879
		1%	.527	.533	.547	.660	.669	.708	.676	.681	.690	.944
		10%	.460	.466	.491	.714	.701	.743	.697	.707	.745	1.02
	$\mathscr{L}$	.5%	.531	.516	.549	.630	.633	.652	.632	.640	.661	.888
		1%	.471	.483	.501	.650	.652	.674	.643	.639	.679	.922
		10%	.422	.420	.442	.670	.673	.691	.666	.676	.709	.972
Truncation 30%												
GA	$\mathscr{N}$	.5%	.772	.781	.828	1.01	.989	1.01	1.00	1.01	1.04	1.00
		1%	.694	.718	.743	.978	1.01	1.02	.972	.978	1.04	1.00
		10%	.683	.686	.706	1.02	1.01	1.03	1.03	1.00	1.04	1.01
	$\mathscr{L}$	.5%	.722	.714	.739	.983	.996	1.03	.985	1.01	1.03	.996
		1%	.652	.666	.691	.996	.988	1.04	.993	.990	1.02	1.01
		10%	.614	.613	.623	.987	.998	1.02	.982	.995	1.03	.998
LN	$\mathscr{N}$	.5%	.288	.294	.309	.328	.333	.368	.329	.337	.366	.413
		1%	.263	.264	.283	.321	.346	.364	.336	.338	.368	.510
		10%	.279	.275	.297	.396	.398	.419	.383	.395	.416	.732
	$\mathscr{L}$	.5%	.258	.270	.294	.332	.336	.365	.332	.331	.362	.408
		1%	.240	.236	.260	.325	.328	.376	.328	.326	.355	.493
		10%	.258	.258	.275	.394	.406	.429	.395	.400	.420	.753
MG	$\mathscr{N}$	.5%	.555	.557	.576	.644	.656	.677	.656	.645	.675	.862
		1%	.517	.510	.548	.677	.675	.689	.688	.675	.691	.925
		10%	.469	.477	.491	.703	.706	.716	.694	.696	.711	.977
	$\mathscr{L}$	.5%	.515	.522	.546	.627	.640	.665	.655	.633	.667	.846
		1%	.486	.483	.508	.664	.663	.703	.663	.671	.684	.943
		10%	.438	.438	.458	.717	.701	.746	.701	.713	.729	.972

Table 10: Same as Table 9 except that the Scoring Algorithm was used and the sampling distributions were approximated resulted from approximating $(\mathbf{X}^\top\mathbf{X})^{-1}$ by $(\mathbf{X}_0^{*\top}\mathbf{X}_0^*)^{-1}$.

$\mathbf{x}$	ε	$r:n$	$\hat{\boldsymbol{\pi}}_2$	$\hat{\boldsymbol{\pi}}_1$	$\hat{\boldsymbol{\pi}}_0$	$\bar{\boldsymbol{\pi}}_2$	$\bar{\boldsymbol{\pi}}_1$	$\bar{\boldsymbol{\pi}}_0$	$\tilde{\boldsymbol{\pi}}_2$	$\tilde{\boldsymbol{\pi}}_1$	$\tilde{\boldsymbol{\pi}}_0$	Lev
Truncation 10%												
GA	$\mathscr{N}$	.5%	.834	.844	.869	.999	1.01	1.06	1.01	1.01	1.05	.986
		1%	.740	.749	.777	.965	1.01	1.03	.977	1.03	1.05	1.01
		10%	.663	.670	.709	1.03	1.02	1.04	1.02	1.01	1.04	1.02
	$\mathscr{L}$	.5%	.764	.779	.813	.976	1.01	1.02	.970	.995	1.02	.981
		1%	.680	.688	.734	.987	.978	1.02	.994	.969	1.01	.986
		10%	.604	.605	.641	.987	.982	1.05	1.01	1.02	1.03	1.02
LN	$\mathscr{N}$	.5%	.288	.280	.313	.340	.335	.364	.338	.331	.357	.485
		1%	.274	.262	.280	.358	.344	.360	.338	.337	.351	.612
		10%	.292	.285	.301	.409	.404	.424	.394	.403	.413	.858
	$\mathscr{L}$	.5%	.284	.275	.308	.357	.348	.369	.341	.335	.370	.486
		1%	.242	.242	.258	.338	.334	.359	.332	.319	.343	.588
		10%	.259	.254	.262	.394	.385	.402	.381	.387	.394	.852
MG	$\mathscr{N}$	.5%	.555	.550	.591	.619	.639	.670	.632	.630	.667	.889
		1%	.534	.539	.556	.687	.686	.712	.682	.676	.713	.979
		10%	.458	.454	.475	.668	.686	.705	.664	.688	.713	.973
	$\mathscr{L}$	.5%	.528	.537	.560	.635	.641	.702	.639	.643	.669	.919
		1%	.481	.497	.509	.657	.662	.681	.653	.659	.690	.930
		10%	.419	.420	.442	.684	.672	.713	.676	.687	.712	.980
Truncation 30%												
GA	$\mathscr{N}$	.5%	.782	.776	.800	.999	1.02	1.06	1.00	1.01	1.03	.990
		1%	.726	.735	.747	.971	1.01	1.03	.992	1.02	1.03	1.00
		10%	.676	.691	.715	1.03	1.03	1.03	1.01	1.03	1.03	1.01
	$\mathscr{L}$	.5%	.719	.716	.746	.978	1.00	1.01	.976	.996	1.01	.975
		1%	.655	.670	.711	.992	.982	1.01	.992	.978	1.01	.987
		10%	.615	.617	.642	.991	.985	1.04	1.00	1.01	1.01	1.01
LN	$\mathscr{N}$	.5%	.302	.285	.303	.357	.349	.379	.342	.337	.358	.438
		1%	.274	.276	.291	.355	.357	.382	.342	.345	.358	.544
		10%	.288	.289	.303	.414	.412	.431	.410	.404	.420	.780
	$\mathscr{L}$	.5%	.287	.279	.302	.368	.352	.383	.359	.343	.377	.438
		1%	.244	.244	.253	.346	.342	.364	.335	.331	.343	.528
		10%	.260	.250	.268	.395	.399	.407	.392	.394	.401	.773
MG	$\mathscr{N}$	.5%	.547	.547	.574	.629	.643	.676	.652	.632	.665	.866
		1%	.528	.529	.536	.677	.683	.713	.681	.677	.704	.940
		10%	.462	.455	.483	.679	.695	.689	.674	.676	.731	.936
	$\mathscr{L}$	.5%	.528	.523	.556	.634	.642	.674	.647	.639	.662	.879
		1%	.474	.471	.497	.656	.676	.672	.657	.664	.682	.887
		10%	.422	.429	.446	.675	.665	.712	.679	.683	.722	.939

Table 11: Same as Table 10 except that the MSE ratios are replaced with the ROE.

$\mathbf{x}$	ε	$r:n$	Unif	Lev	$\hat{\boldsymbol{\pi}}_2$	Unif	Lev	$\hat{\boldsymbol{\pi}}_2$	Unif	Lev	$\hat{\boldsymbol{\pi}}_2$
			$\alpha = 0.01$			$\alpha = 0.05$			$\alpha = 0.10$		
Truncation 10%											
GA	$\mathscr{N}$	.5%	.2524	.2522	.2024	.2376	.2374	.1906	.2298	.2297	.1844
		1%	.1784	.1783	.1432	.1680	.1679	.1348	.1625	.1624	.1304
		10%	.0564	.0564	.0453	.0531	.0531	.0426	.0514	.0514	.0412
	$\mathscr{L}$	.5%	.2534	.2531	.1951	.2386	.2383	.1836	.2308	.2305	.1776
		1%	.1792	.1790	.1379	.1687	.1685	.1299	.1632	.1630	.1256
		10%	.0567	.0566	.0436	.0534	.0533	.0411	.0516	.0515	.0397
LN	$\mathscr{N}$	.5%	.0394	.0388	.0218	.0371	.0366	.0205	.0359	.0354	.0198
		1%	.0279	.0275	.0154	.0262	.0259	.0145	.0254	.0250	.0140
		10%	.0088	.0087	.0049	.0083	.0082	.0046	.0080	.0079	.0044
	$\mathscr{L}$	.5%	.0395	.0399	.0220	.0372	.0376	.0207	.0360	.0363	.0200
		1%	.0279	.0282	.0155	.0263	.0266	.0146	.0255	.0257	.0141
		10%	.0088	.0089	.0049	.0083	.0084	.0046	.0080	.0081	.0045
MG	$\mathscr{N}$	.5%	.0699	.0696	.0470	.0659	.0655	.0443	.0637	.0634	.0428
		1%	.0495	.0492	.0333	.0466	.0463	.0313	.0450	.0448	.0303
		10%	.0156	.0156	.0105	.0147	.0146	.0099	.0142	.0142	.0096
	$\mathscr{L}$	.5%	.0701	.0696	.0452	.0660	.0656	.0426	.0638	.0634	.0412
		1%	.0495	.0492	.0320	.0467	.0464	.0301	.0451	.0449	.0291
		10%	.0157	.0156	.0101	.0148	.0147	.0095	.0143	.0142	.0092
Truncation 30%											
GA	$\mathscr{N}$	.5%	.2524	.2518	.2060	.2376	.2370	.1940	.2298	.2293	.1876
		1%	.1784	.1780	.1457	.1680	.1676	.1371	.1625	.1621	.1327
		10%	.0564	.0563	.0461	.0531	.0530	.0434	.0514	.0513	.0420
	$\mathscr{L}$	.5%	.2534	.2527	.1981	.2386	.2379	.1865	.2308	.2301	.1804
		1%	.1792	.1787	.1401	.1687	.1682	.1319	.1632	.1627	.1276
		10%	.0567	.0565	.0443	.0534	.0532	.0417	.0516	.0515	.0403
LN	$\mathscr{N}$	.5%	.0394	.0373	.0218	.0371	.0351	.0205	.0359	.0339	.0199
		1%	.0279	.0264	.0154	.0262	.0248	.0145	.0254	.0240	.0140
		10%	.0088	.0083	.0049	.0083	.0078	.0046	.0080	.0076	.0044
	$\mathscr{L}$	.5%	.0395	.0383	.0220	.0372	.0361	.0207	.0360	.0349	.0200
		1%	.0279	.0271	.0156	.0263	.0255	.0146	.0255	.0247	.0142
		10%	.0088	.0086	.0049	.0083	.0081	.0046	.0080	.0078	.0045
MG	$\mathscr{N}$	.5%	.0699	.0684	.0474	.0659	.0644	.0446	.0637	.0623	.0432
		1%	.0495	.0483	.0335	.0466	.0455	.0316	.0450	.0440	.0305
		10%	.0156	.0153	.0106	.0147	.0144	.0100	.0142	.0139	.0097
	$\mathscr{L}$	.5%	.0701	.0684	.0455	.0660	.0644	.0428	.0638	.0623	.0414
		1%	.0495	.0484	.0322	.0467	.0456	.0303	.0451	.0441	.0293
		10%	.0157	.0153	.0102	.0148	.0144	.0096	.0143	.0139	.0093

Table 12: The running times (in seconds) of the Scoring Algorithm in Fig. 2 and the LSE for sample size n and subsample sizes r with $\mathbf{x} \sim$ GA and $\varepsilon \sim \mathscr{N}(0, 1)$.

	The Scoring Algorithm						LSE
$n \backslash r$	$.05n$	$.10n$	$.20n$	$.30n$	$.40n$	$.50n$	n
$6 * 10^6$	11.807	12.576	18.671	23.276	29.296	30.050	36.344
$6 * 10^5$	0.882	0.981	1.502	1.896	2.266	2.784	3.809
$6 * 10^4$	0.116	0.134	0.161	0.175	0.173	0.201	0.234
$6 * 10^3$	0.012	0.013	0.017	0.018	0.030	0.029	0.027

Table 13: Sample sizes using the leverage-scores-based (Lev) and the $\hat{\pi}_2$ sampling for sample size $n = 10^5$, truncation at 30%, ε and $\mathbf{x}$ generated respectively from two and three distributions, and $\hat{\hat{\delta}}$ in $\mathrm{P}(\|\hat{\boldsymbol{\beta}}_r^* - \hat{\boldsymbol{\beta}}\| > \hat{\delta}) < .01$ chosen such that the sample sizes using the uniform sampling are 10^3, $2 \cdot 10^3$, $5 \cdot 10^3$, 10^4, $2 \cdot 10^4$, $5 \cdot 10^4$.

$\mathbf{x}$	GA			LN			MG		
ε	$\hat{\hat{\delta}}$	Lev	$\hat{\boldsymbol{\pi}}_2$	$\hat{\hat{\delta}}$	Lev	$\hat{\boldsymbol{\pi}}_2$	$\hat{\hat{\delta}}$	Lev	$\hat{\boldsymbol{\pi}}_2$
$\mathscr{N}$	.432	986	652	.062	841	262	.119	959	445
	.304	2003	1318	.044	1665	516	.084	1915	900
	.192	5021	3298	.028	4173	1307	.054	4728	2217
	.137	9957	6521	.020	8276	2594	.038	9399	4397
	.096	19904	13113	.014	16575	5196	.027	19294	9037
	.061	49766	32657	.009	42005	13041	.017	47872	22397
$\mathscr{L}$	.433	997	600	.062	906	271	.120	945	406
	.306	2001	1203	.044	1802	548	.085	1924	818
	.194	4986	3026	.027	4545	1358	.054	4738	2036
	.137	9886	5966	.019	9010	2741	.038	9478	4052
	.097	20043	11993	.014	18211	5482	.027	18986	8093
	.061	49805	29906	.009	45844	13675	.017	47412	20141

Table 14: Sample sizes determined by 3 and $n(\mathrm{MSE}) = \mathrm{trace}(\Sigma)/\mathrm{MSE}$ using the leverage-scores based (Lev) and the $\hat{\boldsymbol{\pi}}_2$ sampling for $n = 10^5$, $p = 50$, $\alpha = 0.05$, truncation at 10%, ε and $\mathbf{x}$ generated respectively from two and three distributions, and the ROE ϵ was chosen such that the sample sizes using the uniform sampling are $10^3, 10^4$.

ε		$\mathscr{N}$						$\mathscr{L}$					
$\mathbf{x}$	Unif	ϵ	Lev	$\hat{\boldsymbol{\pi}}_2$	MSE	Lev	$\hat{\boldsymbol{\pi}}_2$	ϵ	Lev	$\hat{\boldsymbol{\pi}}_2$	MSE	Lev	$\hat{\boldsymbol{\pi}}_2$
GA	10^3	.9583	1997	1273	.0558	1997	1261	.9584	1995	1170	.0563	1995	1158
GA	10^4	.9430	9985	6365	.0112	9985	6304	.9431	9972	5848	.0113	9975	5789
LN	10^3	.9234	1862	539	.0012	1807	514	.9234	1984	551	.0012	1950	533
LN	10^4	.9086	9309	2692	.0002	9034	2566	.9087	9917	2755	.0002	9748	2664
MG	10^3	.9340	1979	881	.0043	1979	872	.9340	1976	808	.0043	1975	7799
MG	10^4	.9191	9893	4403	.0009	9895	4360	.9191	9876	4037	.0009	9871	3992