Journal of Machine Learning Research 23 (2022) 1-45 Submitted 3/20; Revised 6/22; Published 6/22

Asymptotic Analysis of Sampling Estimators for Randomized Numerical Linear Algebra Algorithms*

Ping Ma PINGMA@UGA.EDU
Department of Statistics
University of Georgia

Yongkai Chen YONGKAICHEN@UGA.EDU
Department of Statistics
University of Georgia

Xinlian Zhang XIZHANG@HEALTH.UCSD.EDU
Department of Family Medicine and Public Health
University of California at San Diego

Xin Xing XINXING@VT.EDU
Department of Statistics
Virginia Tech

Jingyi Ma JYMA@CUFE.EDU.CN
Department of Statistics and Mathematics
Central University of Finance and Economics

Michael W. Mahoney MMAHONEY@STAT.BERKELEY.EDU
International Computer Science Institute and Department of Statistics,
University of California at Berkeley

Editor: Lorenzo Rosasco

Abstract

The statistical analysis of Randomized Numerical Linear Algebra (RandNLA) algorithms within the past few years has mostly focused on their performance as point estimators. However, this is insufficient for conducting statistical inference, e.g., constructing confidence intervals and hypothesis testing, since the distribution of the estimator is lacking. In this article, we develop an asymptotic analysis to derive the distribution of RandNLA sampling estimators for the least-squares problem. In particular, we derive the asymptotic distribution of a general sampling estimator with arbitrary sampling probabilities in a fixed design setting. The analysis is conducted in two complementary settings, i.e., when the objective of interest is to approximate the full sample estimator, and when it is to infer the underlying ground truth model parameters. For each setting, we show that the sampling estimator is asymptotically normally distributed under mild regularity conditions. Moreover, the sampling estimator is asymptotically unbiased in both settings. Based on our asymptotic analysis, we use two criteria, the Asymptotic Mean Squared Error (AMSE) and the Expected Asymptotic Mean Squared Error (EAMSE), to identify optimal sampling probabilities. Several of these optimal sampling probability distributions are new to the literature, e.g., the root leverage sampling estimator and the predictor length sampling

*A short preliminary conference version of this paper has appeared as Ma et al. (2020).

©2022 Ping Ma, Yongkai Chen, Xinlian Zhang, Xin Xing, Jingyi Ma, and Michael W. Mahoney.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v23/20-219.html.

estimator. Our theoretical results clarify the role of leverage in the sampling process, and our empirical results demonstrate improvements over existing methods.

Keywords: least squares, randomized numerical linear algebra, leverage scores, asymptotic distribution, mean squared error, asymptotic mean squared error

1. Introduction

Recent work in Randomized Numerical Linear Algebra (RandNLA) focuses on using random sketches of the input data in order to construct approximate solutions more quickly than with traditional deterministic algorithms. In this article, we consider *statistical* aspects of recently-developed fast RandNLA algorithms for the least-squares (LS) linear regression problem. Given $\mathbf{Y} = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$ and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n \times p}$, we consider the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is the coefficient vector, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$, where ε_is are i.i.d random errors with mean 0 and variance $\sigma^2 < \infty$. This is the so-called fixed design setup since design matrix $\mathbf{X}$ is fixed and given. We assume the sample size n is large and that $\mathbf{X}$ has full column rank. The ordinary least squares (OLS) estimator of $\boldsymbol{\beta}_0$ is

$$\hat{\boldsymbol{\beta}}_{OLS} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{2}$$

where $\|\cdot\|$ is the Euclidean norm. While the OLS estimate is optimal in several senses, the run time for computing it is $O(np^2)$,[1] i.e., it does not grow faster than np^2, up to a constant factor (in fact, with the usual algorithm, it equals np^2, up to a constant factor), as its input size increases, which can be daunting when n and/or p are large.

Motivated by these algorithmic considerations, randomized sketching methods have been developed within RandNLA to achieve improved computational efficiency (Mahoney, 2011; Drineas and Mahoney, 2016; Halko et al., 2011; Woodruff, 2014; Mahoney and Drineas, 2016; Drineas and Mahoney, 2018). With these methods, one takes a (usually nonuniform) random sample of the full data (perhaps after preprocessing or preconditioning with a random projection matrix (Drineas and Mahoney, 2016)), and then the sample is retained as a surrogate for the full data for subsequent computation. Here is an example of this approach for the LS problem.

- **Step 1: Sampling.** Draw a random sample of size $r \ll n$ with replacement from the full sample using sampling probabilities $\{\pi_i\}_{i=1}^n$. Denote the resulting sample as $\mathbf{X}^* = (\mathbf{x}_1^*, \ldots, \mathbf{x}_r^*)^T$, $\mathbf{Y}^* = (Y_1^*, \ldots, Y_r^*)^T$ and the corresponding sampling probability for sampling i^{th} data point $(\mathbf{x}_i^*, Y_i^*)$ as π_i^*, where $i = 1, \ldots, r$. That is, if the j^{th} data point in the resulting sample is the k^{th} point data point in the full sample, we have $\pi_j^* = \pi_k$.

- **Step 2: Estimation.** Calculate the weighted LS solution, using the random sample, by solving

[1] $f(n) = O\big(g(n)\big)$ if there exist positive integer numbers M and n_0 such that $f(n) \leq Mg(n)$ for all $n \geq n_0$.

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\Phi}^*\mathbf{Y}^* - \boldsymbol{\Phi}^*\mathbf{X}^*\boldsymbol{\beta}\|^2 \\ &= (\mathbf{X}^{*T}\boldsymbol{\Phi}^{*2}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\boldsymbol{\Phi}^{*2}\mathbf{Y}^* \end{aligned} \tag{3}$$

where $\boldsymbol{\Phi}^* = \text{diag}\{1/\sqrt{r\pi_i^*}\}_{i=1}^r$.

Popular RandNLA sampling approaches include the uniform sampling estimator (UNIF), the basic leverage-based sampling estimator (BLEV), where $\pi_i^{BLEV} = h_{ii}/\sum_{i=1}^n h_{ii}$, where $h_{ii} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$ are the leverage scores of $\mathbf{X}$, and the shrinkage leverage estimator (SLEV), which involves sampling probabilities $\pi_i^{SLEV} = \lambda h_{ii}/\sum_{i=1}^n h_{ii} + (1-\lambda)/n$, where $\lambda \in (0,1)$ (Drineas et al., 2006, 2008, 2012a; Ma et al., 2014).[2]

In this article, we study the *statistical* properties of these and other estimators. Substantial evidence has shown the practical effectiveness of core RandNLA methods (Ma et al., 2014, 2015; Drineas and Mahoney, 2016) (as well as other randomized approximating methods, including the Hessian sketch (Wang et al., 2017a; Pilanci and Wainwright, 2016) and iterative/divide-and-conquer methods (Avron et al., 2010; Meng et al., 2014)) in providing point estimators. However, this is not sufficient for statistical analysis since the uncertainty of the estimator is lacking. In statistics, uncertainty assessment can be conducted through confidence interval construction and significance testing. It is well-known that the construction of confidence intervals and significance testing are interrelated with each other (Lehmann and Romano, 2006). Performing these two analyses is more difficult than point estimation, since it requires the distributional results of the estimator, rather than just moment conditions or concentration bounds. In the RandNLA literature, distribution results of estimators are still lacking.

There are two main challenges in studying the statistical and distributional properties of RandNLA algorithms. The first challenge is that there are two sources of randomness contributing to the statistical performance of RandNLA sampling estimators: one source is the random errors in the model, i.e., the ε_is, which are typically attributed to measurement error or random noise inherited by $\mathbf{Y}$; and the other source is the randomness in the random sampling procedure within the approximation algorithm. The second challenge is that these two sources of randomness couple together within the estimator in a nontrivial way. More formally, the sampling estimator can be expressed as $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$, where $\mathbf{W}$ is a random diagonal matrix, with the i^{th} diagonal element being related to the probability of choosing the i^{th} sample. The random variable used to denote the random sampling procedure, i.e., $\mathbf{W}$, is involved in the sampling estimator in a nonlinear fashion, and it pre-multiplies $\mathbf{Y}$, which contains randomness from the ε_is.

We address these challenges to studying the asymptotic distribution of general RandNLA sampling estimators for LS problems. Our results are fundamentally different from previous results on the statistical properties of RandNLA algorithms (e.g., Ma et al. (2014, 2015); Raskutti and Mahoney (2015); Chen et al. (2016); Wang et al. (2017b); Clarkson

[2]Importantly, these algorithms are robust to the approximation of leverage scores. This is important since the exact computation of leverage scores takes time of the same order as is needed to solve the OLS problem exactly. Faster approximations to all the leverage scores can be accomplished with the algorithm of Drineas et al. (2012a); and it corresponds to the approximate leverage score (ALEV) method of Ma et al. (2014, 2015). See also Appendix 4 for details.

et al. (2019)), in that we provide asymptotic distribution analysis, rather than finite-sample concentration inequalities. The resulting asymptotic distributions open the possibility of performing statistical inference tasks such as hypothesis testing and constructing confidence intervals, whereas finite sample concentration inequality results may not. It is worth mentioning that the results of asymptotic analysis are usually practically valid as long as the sample size is only moderately large.

1.1 Main Results and Contributions

We study the asymptotic distribution of a general RandNLA sampling estimator for the LS linear regression problem, from both a theoretical and empirical perspective.

Main Theoretical Results. Our main theoretical contribution is to derive the asymptotic distribution of RandNLA estimators in two complementary settings.

Data are a random sample. We first consider the data as a random sample from a population, in which case the goal is to estimate the parameters of the population model (1). In this case, there are two sources of randomness: the noise within the data $\boldsymbol{\varepsilon}$ and the subsampling within the estimator. For this *unconditional inference*, we establish the asymptotic normality, i.e., deriving the asymptotic distribution, of sampling estimators for the linear model under general regularity conditions. The convergence is with respect to the noise ε within the data and the sampling of the full sample. We show that sampling estimators are asymptotically unbiased estimators with respect to the true model coefficients, and we obtain an explicit form for the asymptotic variance, for both fixed number of predictors (Theorem 1) and diverging number of predictors (Theorem 2). **Sampling Estimators.** Using these distributional results, we propose several efficient and asymptotically optimal estimators. Depending on the quantity of interest (e.g., $\boldsymbol{\beta}_0$ versus some linear function of $\boldsymbol{\beta}_0$ such as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_0$ or $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$), we obtain different optimal sampling probabilities (Propositions 1, 2, and 3) that lead to sampling estimators that minimize the Asymptotic Mean Squared Error (AMSE) in the respective context. None of these distributions is proportional to the leverage scores, but one (RL of Proposition 2) is constructed using the square roots of the leverage scores, and another (PL of Proposition 3) is constructed using the row norms of the predictor matrix.

Data are given and fixed. We then consider the data as given/fixed, in which case the goal is to approximate the full sample OLS estimate. There is only one source of randomness: the subsampling within the estimator. In this case, for this *conditional inference*, we establish the asymptotic normality, i.e., deriving the asymptotic distribution, of sampling estimators for the linear model under general regularity conditions. We show that sampling estimators are asymptotically unbiased with respect to the OLS estimate, and we obtain an explicit form of the asymptotic variance and the Expected Asymptotic Mean Squared Error (EAMSE) of sampling estimators (Theorem 3). **Sampling Estimators**. Using these results, we construct sampling probability distributions that lead to sampling estimators that minimize the EAMSE. Depending on the quantity of interest (here, $\hat{\boldsymbol{\beta}}_{OLS}$ versus some linear function of $\hat{\boldsymbol{\beta}}_{OLS}$ such as $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$ or $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$, we obtain different optimal sampling probabilities (Propositions 4, 5, and 6).

Main Empirical Results. We conduct a comprehensive empirical evaluation of the performance of these sampling estimators, on both synthetic and real data sets. This involves both conditional and unconditional inference cases, using predictor matrices generated from various distributions, including heavy-tailed and asymmetric distributions. For all settings under consideration, we calculate the squared bias and variance of the sampling estimators. We demonstrate that the squared bias decreases as sample size increases, and we demonstrate that the squared biases are typically much smaller than the variances. These observations are consistent with our theory stating that the sampling estimators are asymptotically unbiased. The variance of sampling estimators also decreases as sample size increases, indicating the consistency of the sampling estimators. Depending on the specific objective considered, we also demonstrate that the novel estimators we derive have better performance, e.g., smaller variances, than existing ones, confirming the optimality results established in this paper. Another goal of the simulation study is to evaluate the necessity of our regularity conditions for the theorems. In the case of the predictor matrix generated from the t-distribution with 1 degree of freedom, the regularity conditions of our theory are technically not satisfied. The estimators, however, are shown to have performance similar to those in the aforementioned settings. Also, on two real-world data examples, we show that all the observations concerning asymptotic unbiasedness and asymptotic consistency in simulated data sets also appear. In particular, our proposed sampling methods for conditional inference have smaller variances, compared to other leverage-based estimators, such as BLEV/ALEV (Drineas et al., 2006, 2012a) and SLEV (Ma et al., 2014, 2015).

1.2 Related Work

There is a large body of related work in RandNLA (Mahoney, 2011; Drineas and Mahoney, 2016; Halko et al., 2011; Woodruff, 2014; Mahoney and Drineas, 2016; Drineas and Mahoney, 2018). However, very little of this work addresses statistical aspects of the methods. Recently, significant progress has been made in the study of the statistical properties of RandNLA sampling estimators (Ma et al., 2014, 2015; Raskutti and Mahoney, 2015; Chen et al., 2016; Wang et al., 2017b; Clarkson et al., 2019). The work most related to ours is that of Ma et al. (2014, 2015), who employed a Taylor series expansion up to a linear term to study the MSE of RandNLA sampling estimators. Ma et al. (2014, 2015) failed to characterize the detailed convergence performance of the remainder term. They concluded that neither leverage-based sampling (BLEV) nor uniform sampling (UNIF) dominates the other in terms of variance; and they proposed and demonstrated the superiority of the SLEV sampling method. To find the sampling distribution of estimators, leading to statistically-better RandNLA sampling estimators, it is important to examine the convergence properties of the remainder term. To accomplish this, we consider the asymptotic distribution of the sampling estimator. Such asymptotic analysis is common in statistics, and it can substantially simplify the derivation of complicated random variables, leading to simpler analytic expressions (Le Cam, 1986).

Chen et al. (2016) proposed optimal estimators minimizing the variance that account for the randomness of sampling and model error. Our results and those of Chen et al. (2016) have similar goals, but they are different. First, Chen et al. (2016) used bias and variance, while we use AMSE and EAMSE. Second, we consider the asymptotic distribution of the

sampling estimators, going beyond just the bias and variance of Chen et al. (2016). Thus, our results could be used for downstream statistical inferences, e.g., constructing confidence intervals and hypothesis testing, while those of Chen et al. (2016) could not. Third, the exact expression of optimal sampling probabilities in Chen et al. (2016) depends on the unknown true parameter of the model, β_0 and σ^2 (Eqn. (4) in Chen et al. (2016)), while our optimal sampling probabilities (see Section 2) are readily computed from the data. Fourth, Chen et al. (2016) only studied properties of sampling estimators for estimating true model parameters, while we consider both estimating the true parameter and approximating the full sample estimate.

Wang et al. (2017b) proposed an approximated A-optimality criterion, which is based on the conditional variance of the sampling estimator given a subsample. Since the randomness of sampling is not considered in the criterion, they obtained a simple analytic expressions of the optimal results. Clarkson et al. (2019) also consider experimental design from the RandNLA perspective, and they propose a framework for experimental design where the responses are produced by an arbitrary unknown distribution. Their main result yields nearly tight bounds for the classical A-optimality criterion, as well as improved bounds for worst-case responses. In addition, they propose a minimax-optimality criterion (which can be viewed as an extension of both A-optimal design and RandNLA sampling for worst-case regression). Related works on the asymptotic properties of subsampling estimators in logistic regression can be found in Wang et al. (2018) and Wang (2019).

1.3 Outline

The remainder of this article is organized as follows. In Section 2, we introduce technical notations and definitions of MSE, AMSE, and EAMSE, we derive the asymptotic distribution of the sampling estimators, and we propose several criteria which give rise to optimal sampling probability distributions. In Section 3, we present empirical results on simulated data and two real-world data examples. In Section 4, we provide a brief discussion and conclusion. All technical proofs are presented in Appendix 4. Fast approximation methods for approximating the sampling probabilities are presented in the Appendix 4. A short preliminary conference version of this paper has appeared as Ma et al. (2020).

2. Sampling Estimation Methods

In this section, we review the well-known Mean Squared Error (MSE) criterion, and we also define and discuss the standard but less well-known criterion, Asymptotic Mean Squared Error (AMSE) (see section 2.5.2 of Shao (2003)) and its generalization Expected Asymptotic Mean Squared Error (EAMSE). We also derive asymptotic properties of the RandNLA sampling estimator $\tilde{\boldsymbol{\beta}}$ under two scenarios: unconditional inference, which involves estimating the true model parameter $\boldsymbol{\beta}_0$; and conditional inference, which involves approximating the full sample OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$. We use the AMSE and EAMSE to develop two criteria for sampling estimators, and we obtain several optimal estimators.

2.1 MSE and AMSE: Technical Definition

Let $\boldsymbol{T}_n$ be a $p \times 1$ estimator of a $p \times 1$ parameter $\boldsymbol{\nu}$, for every n. One popular quality metric for the estimator $\boldsymbol{T}_n$ is the MSE, which is defined to be

$$\begin{aligned} MSE(\boldsymbol{T}_n; \boldsymbol{\nu}) &= \mathrm{E}[(\boldsymbol{T}_n - \boldsymbol{\nu})^T (\boldsymbol{T}_n - \boldsymbol{\nu})] \\ &= \mathrm{tr}(\mathrm{Var}(\boldsymbol{T}_n)) + (\mathrm{E}(\boldsymbol{T}_n) - \boldsymbol{\nu})^T (\mathrm{E}(\boldsymbol{T}_n) - \boldsymbol{\nu}), \end{aligned}$$

where $\mathrm{Var}(\boldsymbol{T}_n) = E[(T_n - \mathrm{E}(\boldsymbol{T}_n))(T_n - \mathrm{E}(\boldsymbol{T}_n))^T]$. The MSE can be decomposed into two terms: one term, $\mathrm{tr}(\mathrm{Var}(\boldsymbol{T}_n))$, quantifying the *variance* of the estimator; and one term, $(\mathrm{E}(\boldsymbol{T}_n) - \boldsymbol{\nu})^T (\mathrm{E}(\boldsymbol{T}_n) - \boldsymbol{\nu})$, quantifying the *squared bias* of the estimator. To evaluate the RandNLA sampling estimator $\tilde{\boldsymbol{\beta}}$ in estimating the true model parameter $\boldsymbol{\beta}_0$ and the full sample OLS estimate $\hat{\boldsymbol{\beta}}_{OLS}$, we will be interested in the AMSE and EAMSE, respectively. These are the asymptotic counterparts of MSE in large sample theory.

To define the AMSE, let $\boldsymbol{T}_n$ be a $p \times 1$ estimator of a $p \times 1$ parameter $\boldsymbol{\nu}$, for every n, and let $\boldsymbol{\Sigma}_n$ be a sequence of $p \times p$ positive definite matrices. Assume $\boldsymbol{\Sigma}_n^{-1/2}(\boldsymbol{T}_n - \boldsymbol{\nu}) \xrightarrow{d} \boldsymbol{Z}$ as $n \to \infty$, where $\xrightarrow{d}$ denotes convergence in distribution, and assume $\boldsymbol{Z}$ is a $p \times 1$ random vector such that its i^{th} element Z_i satisfies $0 < \mathrm{E}(Z_i^2) < \infty$, for $i = 1, \ldots, p$. Then, the AMSE of $\boldsymbol{T}_n$, denoted $AMSE(\boldsymbol{T}_n; \boldsymbol{\nu})$, is defined to be

$$\begin{aligned} AMSE(\boldsymbol{T}_n; \boldsymbol{\nu}) &= \mathrm{E}(\boldsymbol{Z}^T \boldsymbol{\Sigma}_n \boldsymbol{Z}) \\ &= \mathrm{tr}(\boldsymbol{\Sigma}_n^{1/2} \mathrm{Var}(\boldsymbol{Z}) \boldsymbol{\Sigma}_n^{1/2}) + (\mathrm{E}(\boldsymbol{Z})^T \boldsymbol{\Sigma}_n \mathrm{E}(\boldsymbol{Z})) \\ &= \mathrm{tr}(\mathrm{AVar}(\boldsymbol{T}_n)) + (\mathrm{AE}(\boldsymbol{T}_n) - \boldsymbol{\nu})^T (\mathrm{AE}(\boldsymbol{T}_n) - \boldsymbol{\nu}), \end{aligned} \tag{4}$$

where the second equality was obtained by noticing that the expectation of a quadratic form $\boldsymbol{Z}^T \boldsymbol{\Sigma}_n \boldsymbol{Z}$ follows Theorem 1.5 of Seber and Lee (2003), and $\mathrm{AVar}(\boldsymbol{T}_n) = \boldsymbol{\Sigma}_n^{1/2} \mathrm{Var}(\boldsymbol{Z}) \boldsymbol{\Sigma}_n^{1/2}$ and $\mathrm{AE}(\boldsymbol{T}_n) = \boldsymbol{\nu} + \boldsymbol{\Sigma}_n^{1/2} \mathrm{E}(\boldsymbol{Z})$ denote the *asymptotic variance-covariance matrix* and the *asymptotic expectation* of $\boldsymbol{T}_n$ in estimating $\boldsymbol{\nu}$, respectively. Intuitively, we have that $\boldsymbol{\Sigma}_n^{-1/2}(\mathrm{AE}(\boldsymbol{T}_n) - \boldsymbol{\nu})$ equals $\mathrm{E}(\boldsymbol{Z})$, and $\boldsymbol{\Sigma}_n^{-1/2} \mathrm{AVar}(\boldsymbol{T}_n) \boldsymbol{\Sigma}_n^{-1/2}$ equals $\mathrm{Var}(\boldsymbol{Z})$. Note that the asymptotic variance-covariance is not unique. Different asymptotic variance-covariances could differ by constant multipliers or negligible smaller order terms. Such a difference does not have a significant impact on the resulting asymptotic analysis under very general regularity conditions; see Section 2.5 of Shao (2003). Therefore, we can choose one from the family.

If $\mathrm{E}(\boldsymbol{Z}) = \boldsymbol{0}$, we say $\boldsymbol{T}_n$ is an *asymptotically unbiased estimator* of $\boldsymbol{\nu}$. If $\mathrm{tr}(\mathrm{AVar}(\boldsymbol{T}_n)) \to 0$ as $n \to \infty$, we say $\boldsymbol{T}_n$ is an *asymptotically consistent estimator*.

In this paper, the basic estimator is denoted as $\tilde{\boldsymbol{\beta}}$ (i.e., the counterpart for $\boldsymbol{T}_n$ in the definitions above will be $\tilde{\boldsymbol{\beta}}$, or a linear function of $\tilde{\boldsymbol{\beta}}$). We will obtain the explicit form for both $\mathrm{AVar}(\tilde{\boldsymbol{\beta}})$ and $\mathrm{AE}(\tilde{\boldsymbol{\beta}})$ by deriving the large sample distributions of sampling estimators, when performing unconditional inference in Section 2.2.

2.2 Unconditional Inference: Estimating Model Parameters

For Model (1), from the traditional statistical perspective of using the data to perform inference, one major goal is to estimate the underlying true model parameters, i.e., $\boldsymbol{\beta}_0$. We refer to this as *unconditional inference*. For unconditional inference, both randomness

in the data and randomness in the algorithm contribute to randomness in the RandNLA sampling estimators.

The following theorem states that, in unconditional inference, the asymptotic distribution of the sampling estimator $\tilde{\boldsymbol{\beta}}$ is a normal distribution (with mean $\boldsymbol{\beta}_0$ and variance $\sigma^2\boldsymbol{\Sigma}_0$). The proof of Theorem 1 is provided in Appendix 4.

Theorem 1 (Unconditional inference, fixed p) *Assume the number of predictors p is fixed and the following regularity conditions hold.*

- *(A1)[Data condition]. For sufficiently large n, there exist positive constants b and B such that $b \leq \lambda_{min} \leq \lambda_{max} \leq B$, where λ_{min} and λ_{max} are the minimum and maximum eigenvalues of matrix $\mathbf{X}^T\mathbf{X}/n$, respectively.*
- *(A2)[Sampling condition]. The sample size[3] $r = \Theta(n^{1-\alpha})$, where $0 \leq \alpha < 1$ and the minimum sampling probability $\pi_{min} = \Omega(n^{-\gamma_0})$, where $\gamma_0 \geq 1$. The parameters γ_0 and α satisfy $\gamma_0 + \alpha < 2$.*

Under these assumptions, as the sample size $n \to \infty$, we have

$$(\sigma^2\boldsymbol{\Sigma}_0)^{-\frac{1}{2}}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \quad \xrightarrow{d} \quad \boldsymbol{N}(\mathbf{0}, \mathbf{I}_p), \tag{5}$$

where

$$\boldsymbol{\Sigma}_0 = (\mathbf{X}^T\mathbf{X})^{-1}\left(\mathbf{X}^T(\mathbf{I}_n + \boldsymbol{\Omega})\mathbf{X}\right)(\mathbf{X}^T\mathbf{X})^{-1}, \quad \boldsymbol{\Omega} = diag\{1/r\pi_i\}_{i=1}^n,$$

and $\mathbf{I}_p$ is the $p \times p$ identity. The convergence is with respect to the randomness in the noise $\boldsymbol{\varepsilon}$ and the subsampling of the full sample. Thus, for unconditional inference, the asymptotic mean of $\tilde{\boldsymbol{\beta}}$ is

$$AE(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}_0, \tag{6}$$

i.e., $\tilde{\boldsymbol{\beta}}$ is an asymptotically unbiased estimator of $\boldsymbol{\beta}_0$, and the asymptotic variance of $\tilde{\boldsymbol{\beta}}$ is

$$AVar(\tilde{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{\Sigma}_0. \tag{7}$$

Remark. Theorem 1 considers the case of a fixed parameter dimension p. The case of diverging parameter dimension $p \to \infty$ is considered in Theorem 2 below.

Remark. Theorem 1 shows that, as the number of data points n gets larger, the distribution of $\tilde{\boldsymbol{\beta}}$ is well-approximated by a normal distribution, with mean $\boldsymbol{\beta}_0$ and variance $\sigma^2\boldsymbol{\Sigma}_0$.

Remark. Condition (A1) in Theorem 1 indicates that $\mathbf{X}^T\mathbf{X}/n$ is positive definite (as opposed to being just positive semi-definite). This condition requires the predictor matrix $\mathbf{X}$ to be of full column rank and that the elements in $\mathbf{X}$ are not over-dispersed. This condition ensures the consistency of the full sample OLS estimator (Lai et al., 1978), and it has been used in many regression related problems, e.g., variable selection (Zou, 2006).

Remark. Condition (A2) in Theorem 1 sets restrictions on the subsample size r and minimum sampling probability π_{min}. Since we study subsampling, we require that the subsample size r is of smaller order than the full sample size, i.e, $r = \Theta(n^{1-\alpha})$, where $0 \leq \alpha < 1$.

[3]We say that $f(n) = \Omega(g(n))$ if there exists some positive integer numbers m and n_0, $f_n \geq mg(n)$ for all $n \geq n_0$. Similarly, $f(n) = \Theta(g(n))$ if $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$.

Since the parameters γ_0 and α satisfy $\gamma_0 + \alpha < 2$, as stated in condition (A2), the condition on π_{min}, i.e., $\pi_{min} = \Omega(n^{-\gamma_0})$, can be rewritten as $\pi_{min} \geq \Theta(n^{-(2-\alpha)})$, which provides a lower bound on the *smallest* sampling probability. The smallest sampling probability cannot be too small. Bounding sampling probabilities from below mitigates the inflation of the variance $\mathbf{\Sigma}_0$, which is proportional to the reciprocal sampling probability. The importance of this condition for establishing *statistical* properties of RandNLA algorithms was highlighted by Ma et al. (2014, 2015). Condition (A2) can also be rewritten as $r\pi_{min} \geq \Theta(n^{-1})$, which states that if there is a data point such that its expected number of times being sampled is very small, one compensates by making the sample size large.

Remark. In Theorem 1, the asymptotic variance $AVar(\tilde{\boldsymbol{\beta}})$ can be written as

$$AVar(\tilde{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} + \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Omega}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}, \tag{8}$$

where the first term is the variance of the full sample OLS, and the second term is the variation related to the sampling process. The second term of Eqn. (8) has a "sandwich-type" expression. The center term, $\mathbf{\Omega}$, depends on the reciprocal sampling probabilities, suggesting that extremely small probabilities will result in large asymptotic variance and large AMSE of the corresponding estimator. This was observed previously in the non-asymptotic case by Ma et al. (2015).

Remark. In light of efficient estimation methods such as iterative Hessian sketch and dual random projection, we emphasize that besides estimation, our distribution results can be used for performing additional inference analysis, e.g., constructing a confidence intervals and conducting hypothesis testing. These inference analyses cannot be achieved by other iterative methods as far as we know.

Given Theorem 1, it is natural to ask whether there is an optimal estimator, i.e., one with the smallest AMSE for estimating $\boldsymbol{\beta}_0$. Using the asymptotic results in Theorem 1, we propose the following three estimators.

Estimating $\boldsymbol{\beta}_0$. By Theorem 1, we could express the $AMSE(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)$ as a function of $\{\pi_i\}_{i=1}^n$, as shown, e.g., in Eqn. (9) below. Since this expression is a function of the sampling probabilities, it is straightforward to employ the method of Lagrange multipliers to find the minimizer of the right-hand side of Eqn. (9), subject to the constraint $\sum_{i=1}^n \pi_i = 1$. The minimizer is then the optimal sampling probabilities for estimating $\boldsymbol{\beta}_0$. The proof of Proposition 1 is provided in Appendix 4.

Proposition 1 *For the $AMSE(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)$, we have that*

$$AMSE(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) \;\; = \;\; \sigma^2 tr\{(\mathbf{X}^T\mathbf{X})^{-1}\} + \frac{1}{r}\sum_{i=1}^{n}\frac{\sigma^2}{\pi_i}||(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i||^2. \tag{9}$$

Given (9)*, the sampling estimator with the sampling probabilities*

$$\pi_i = \frac{\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^n \|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}, \quad i = 1, \ldots, n, \tag{10}$$

(which we call the inverse-covariance (IC) sampling estimator) has the smallest $AMSE(\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}_0)$.

Remark. The implication of this optimal estimator is two-fold. On the one hand, as defined, the proposed IC estimator has the smallest AMSE. On the other hand, if given the same tolerance of uncertainty, i.e., to achieve a certain small standard error, the IC estimator requires the smallest sample size.

Remark. The IC sampling probabilities can be computed in $O(np^2)$ time, using standard methods. In Appendix 4, we present an approximation algorithm, similar to that of Drineas et al. (2012a), by which the IC sampling probabilities can be approximated in $O(np\log(n)/\epsilon)$ time, where ϵ is an approximation error parameter.

Estimating linear functions of $\boldsymbol{\beta}_0$. In addition to making inference on $\boldsymbol{\beta}_0$, one may also be interested in linear functions of $\boldsymbol{\beta}_0$, i.e., $\mathbf{L}\boldsymbol{\beta}_0$, where $\mathbf{L}$ is any constant matrix of suitable dimension. Here, we present results for $\mathbf{X}\boldsymbol{\beta}_0$ and $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$ (although clearly similar results hold for other functions of the form $\mathbf{L}\boldsymbol{\beta}_0$).

We start with estimating $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_0$ since, in regression analysis, inference on the true regression line $\mathbf{X}\boldsymbol{\beta}_0$ is crucially important. The proof of Proposition 2 (and other similar propositions below) is similar to that of Proposition 1, and thus it is omitted.

Proposition 2 *For the $AMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}, \mathbf{X}\boldsymbol{\beta}_0)$, we have that*

$$AMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}, \mathbf{X}\boldsymbol{\beta}_0) \;=\; p\sigma^2 + \frac{1}{r}\sum_{i=1}^{n}\frac{\sigma^2}{\pi_i}||\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i||^2. \tag{11}$$

Given (11)*, the sampling estimator with the sampling probabilities*

$$\pi_i \;=\; \frac{\|\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^{n}\|\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|} = \frac{\sqrt{h_{ii}}}{\sum_{i=1}^{n}\sqrt{h_{ii}}}, \quad i = 1, \ldots, n, \tag{12}$$

(which we call the root leverage (RL) sampling estimator) has the smallest $AMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}\boldsymbol{\beta}_0)$.

Remark. Note that

$$\|\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|^2 = (\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i)^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i = h_{ii}.$$

These quantities, the so-called leverage scores (called BLEV, in Ma et al. (2014, 2015)), have been central to RandNLA theory (Mahoney, 2011; Drineas et al., 2012a; Drineas and Mahoney, 2016; Mahoney and Drineas, 2016). Using the main Algorithm 1 in Drineas et al. (2012a), they can be computed in $O(np\log(n)/\epsilon)$ time, where ϵ is an approximation error parameter (and called ALEV by Ma et al. (2014, 2015)).

Remark. The probabilities in RL are a nonlinear transformation of the probabilities in BLEV. Comparing to the BLEV estimator, the RL estimator shrinks the large probabilities and pulls up the small probabilities. Thus, we expect RL to provide an estimator with smaller variances, in a way similar to SLEV.

Remark. Chen et al. (2016) proposed optimal sampling estimators for estimating $\boldsymbol{\beta}_0$ and predicting $\mathbf{Y}$. Their sampling probabilities depend on the unknown parameters, and they proposed the probabilities in (12) as a rough approximation of their proposed probabilities, without demonstration.

We next consider estimating $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$, which is also of interest in regression analysis.

Proposition 3 *For the* $AMSE(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\beta}}, \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0)$*, we have that*

$$AMSE(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\beta}}, \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0) \quad = \quad \sigma^2 tr(\mathbf{X}^T\mathbf{X}) + \frac{\sigma^2}{r}\sum_{i=1}^{n}\frac{1}{\pi_i}||\mathbf{x}_i||^2. \tag{13}$$

Given (13)*, the sampling estimator with the sampling probabilities*

$$\pi_i = \frac{\|\mathbf{x}_i\|}{\sum_{i=1}^{n}\|\mathbf{x}_i\|}, \quad i = 1, \ldots, n, \tag{14}$$

(which we call the predictor-length (PL) sampling estimator) has the smallest value for the $AMSE(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0)$*.*

Remark. The PL probabilities have a connection with the Fisher information of the full sample OLS estimate. The Fisher information measures the "amount of information" about the parameter that is present in the data (see Section 11.10 of Cover and Thomas (2006)). The inverse of the Fisher information matrix gives a lower bound (the Cramer-Rao lower bound) on the variance of any estimator constructed from the data to estimate a parameter (see Section 3.1.3 of Shao (2003)). Since the Fisher information of the full data can be written as the summation of the Fisher information of each data point, i.e., $\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T$, we have that $\text{tr}\{\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\} = \frac{1}{\sigma^2}\sum_{i=1}^{n}||\mathbf{x}_i||^2$. The PL probability is high if the data point has a high contribution to the Fisher information.

Diverging number of predictors, $p \to \infty$**.** Theorem 1 considers the number of predictors/features, p, as fixed. It is also of interest to study the asymptotic properties of RandNLA estimators in the scenario that p diverges with $n \to \infty$ (at a suitable rate relative to n). The following theorem states our results concerning this case. Observe that, in the case of a divergent p, the vector $(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is of divergent dimension. Thus, we characterize its asymptotic distribution via the scalar $\boldsymbol{a}^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, where $\boldsymbol{a}$ is an arbitrary bounded-norm vector. The proof of Theorem 2 is provided in Appendix 4.

Theorem 2 (Unconditional inference, diverging p**)** *In addition to Condition (A1) in Theorem 1, assume the following regularity conditions hold.*

- *(B1)[Data condition]. The number of predictors* p *diverges at a rate* $p = O(n^{1-\kappa})$*,* $2/3 < \kappa < 1$*; and* $\frac{\max_i \|\mathbf{x}_i\|^2}{n} = O(\frac{p}{n})$*, where* $\mathbf{x}_i$ *is the* i^{th} *row of* $\mathbf{X}$*.*
- *(B2)[Sampling condition]: The parameters* α*,* γ_0 *satisfy* $\alpha + \gamma_0 < 3\kappa - 1$*.*

Under these assumptions, as the sample size $n \to \infty$*, we have*

$$(\sigma^2\boldsymbol{a}^T\boldsymbol{\Sigma}_0\boldsymbol{a})^{-\frac{1}{2}}\boldsymbol{a}^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \quad \xrightarrow{d} \quad \boldsymbol{N}(0, 1), \tag{15}$$

$\boldsymbol{a} \in \mathbb{R}^p$ *is a norm one vector, i.e.,* $\|\boldsymbol{a}\|^2 = 1$*. The convergence is with respect to the randomness in the noise* $\boldsymbol{\varepsilon}$ *and the subsampling of the full sample.*

Remark. Condition (B1) indicates the divergence rate of p is lower than the growth rate of the full sample size n. Condition (B2) is more stringent than Condition (A2), and

this is required for accommodating a divergent p. It implies that given the divergence rate $p = n^{1-\kappa}$, $r\pi_{min} > \Theta(n^{2-3\kappa})$, which indicates a lower bound on the expected number of times any data point is sampled. Notice that if we let $\kappa = 1$, which indicates no divergence of p with respect to n, the condition $(B2)$ reduces to the condition $(A2)$ requiring $\alpha+\gamma_0 < 2$.

Remark. It is easy to verify that the sampling estimators in Propositions 1, 2, and 3 are still the optimal sampling estimators for their respective purposes. Thus, we omit restating the results.

2.3 EAMSE: Technical Definition

We shall now define a generalization of the AMSE, called EAMSE. To define the EAMSE, let $\boldsymbol{T}_r$ be an $p\times 1$ estimator of a $p\times 1$ parameter $\boldsymbol{\nu}_{\boldsymbol{\varepsilon}}$, for every sample size r, and let $\boldsymbol{\Sigma}_r$ be a sequence of $p \times p$ positive definite matrices. Assume that $\boldsymbol{\Sigma}_r^{-1/2}(\boldsymbol{T}_r - \boldsymbol{\nu}_{\boldsymbol{\varepsilon}}) \xrightarrow{d} \boldsymbol{Z}_{\boldsymbol{\varepsilon}}$ as $r \to \infty$, and that $\boldsymbol{Z}_{\boldsymbol{\varepsilon}}$ is a $p\times 1$ random vector such that its i^{th} element $Z_{\boldsymbol{\varepsilon}i}$ satisfies $0 < \mathrm{E}(Z_{\boldsymbol{\varepsilon}i}^2) < \infty$, for $i = 1, \ldots, p$. The EAMSE of $\boldsymbol{T}_r$, denoted $EAMSE(\boldsymbol{T}_r; \boldsymbol{\nu}_{\boldsymbol{\varepsilon}})$, is defined to be

$$\begin{aligned} EAMSE(\boldsymbol{T}_r; \boldsymbol{\nu}_{\boldsymbol{\varepsilon}}) &= \mathrm{E}_{\boldsymbol{\varepsilon}}(\mathrm{E}(\boldsymbol{Z}_{\boldsymbol{\varepsilon}}^T \boldsymbol{\Sigma}_r \boldsymbol{Z}_{\boldsymbol{\varepsilon}})) \\ &= \mathrm{E}_{\boldsymbol{\varepsilon}}(\mathrm{tr}(\boldsymbol{\Sigma}_r^{1/2}\mathrm{Var}(\boldsymbol{Z}_{\boldsymbol{\varepsilon}})\boldsymbol{\Sigma}_r^{1/2}) + \mathrm{E}_{\boldsymbol{\varepsilon}}(\mathrm{E}(\boldsymbol{Z}_{\boldsymbol{\varepsilon}})^T \boldsymbol{\Sigma}_r \mathrm{E}(\boldsymbol{Z}_{\boldsymbol{\varepsilon}})) \\ &= \mathrm{E}_{\boldsymbol{\varepsilon}}(\mathrm{tr}(\mathrm{AVar}(\boldsymbol{T}_r))) + \mathrm{E}_{\boldsymbol{\varepsilon}}((\mathrm{AE}(\boldsymbol{T}_r) - \boldsymbol{\nu}_{\boldsymbol{\varepsilon}})^T (\mathrm{AE}(\boldsymbol{T}_r) - \boldsymbol{\nu}_{\boldsymbol{\varepsilon}})), \quad (16) \end{aligned}$$

where $\mathrm{AVar}(\boldsymbol{T}_r) = \boldsymbol{\Sigma}_r^{1/2}\mathrm{Var}(\boldsymbol{Z}_{\boldsymbol{\varepsilon}})\boldsymbol{\Sigma}_r^{1/2}$ and $\mathrm{AE}(\boldsymbol{T}_r) = \boldsymbol{\nu}_{\boldsymbol{\varepsilon}} + \boldsymbol{\Sigma}_r^{1/2}\mathrm{E}(\boldsymbol{Z}_{\boldsymbol{\varepsilon}})$ denote the *asymptotic variance-covariance matrix* and the *asymptotic expectation* of $\boldsymbol{T}_r$ in estimating $\boldsymbol{\nu}_{\boldsymbol{\varepsilon}}$, respectively.

We may think of the EAMSE as the expectation of the AMSE. An important subtlety, however, in the use of the AMSE versus the use of the EAMSE lies in the limiting distribution. In unconditional inference (i.e., where we consider a statistical model, and where we will use the AMSE), the limiting distribution is $\boldsymbol{Z}$, i.e., it does not involve noise within the data $\boldsymbol{\varepsilon}$; whereas, in conditional inference (i.e., where we consider the data set $\mathbf{Y}$ and sample size n as fixed and given, and where we will use the EAMSE), the limiting distribution is $\boldsymbol{Z}_{\boldsymbol{\varepsilon}}$, i.e., it involves noise within the data $\boldsymbol{\varepsilon}$. As in Section 2.2, we will obtain the explicit form for both $\mathrm{AVar}(\tilde{\boldsymbol{\beta}})$ and $\mathrm{AE}(\tilde{\boldsymbol{\beta}})$ by deriving the large sample distributions of sampling estimators, when performing conditional inference (Section 2.4). As we show in Section 2.4, the sequences of $\boldsymbol{\Sigma}_r$ involve statistics based on the full sample. Thus, the motivation for taking the expectation of AMSE to construct EAMSE is to avoid calculating those full sample statistics in proposing the optimal RandNLA sampling estimators in conditional inference.

2.4 Conditional Inference: Approximating the Full Sample OLS Estimate

For Model (1), a second major goal is to approximate the full sample calculations, say the OLS estimate $\hat{\boldsymbol{\beta}}_{OLS}$ in Eqn. (2), regardless of the underlying true model parameter $\boldsymbol{\beta}_0$. We refer to this as *conditional inference*. For conditional inference, we consider the full sample as given, and thus the only source of randomness contributing to the RandNLA sampling estimators is the randomness in the sampling algorithm. The following theorem states that, in conditional inference, the asymptotic distribution of the sampling estimator $\tilde{\boldsymbol{\beta}}$ is a normal distribution (with mean $\boldsymbol{\beta}_{OLS}$ and variance $\sigma^2\boldsymbol{\Sigma}_c$). The proof of Theorem 3 is provided in Appendix 4.

Theorem 3 (Conditional inference) *Assume the following regularity conditions hold.*

- *(C1)[Data condition]. The full sample data $\{\mathbf{X}, \mathbf{Y}\}$, i.e., the full sample size n and the number of predictors p are considered fixed; $\mathbf{X}$ is of full column rank, and $\|\mathbf{x}_i\| < \infty$, for $i = 1, \ldots, n$, where $\mathbf{x}_i$ is the i^{th} row of $\mathbf{X}$.*

- *(C2)[Sampling condition]. The sampling probabilities $\{\pi_i\}_{i=1}^n$ are nonzero.*

Under these assumptions, as the sample size $r \to \infty$, we have

$$(\boldsymbol{\Sigma}_c)^{-\frac{1}{2}}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{OLS}) \quad \overset{d}{\to} \quad \boldsymbol{N}(\mathbf{0}, \mathbf{I}_p)\,, \tag{17}$$

where

$$\boldsymbol{\Sigma}_c = \frac{1}{r}(\mathbf{X}^T\mathbf{X})^{-1}\left(\sum_{i=1}^{n}\frac{e_i^2}{\pi_i}\mathbf{x}_i\mathbf{x}_i^T\right)(\mathbf{X}^T\mathbf{X})^{-1}, \quad e_i = Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{OLS},$$

and $\mathbf{I}_p$ is the $p \times p$ identity. The convergence is with respect to the subsampling of the full sample. Thus, for conditional inference, the asymptotic mean of $\tilde{\boldsymbol{\beta}}$ is

$$AE(\tilde{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}}_{OLS}, \tag{18}$$

i.e., $\tilde{\boldsymbol{\beta}}$ is an asymptotically unbiased estimator of $\boldsymbol{\beta}_{OLS}$, and the asymptotic variance of $\tilde{\boldsymbol{\beta}}$ is

$$AVar(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\Sigma}_c. \tag{19}$$

Remark. Theorem 3 shows that as the sample size r gets larger, the distribution of $\tilde{\boldsymbol{\beta}}$ is well-approximated by a normal distribution, with mean $\hat{\boldsymbol{\beta}}_{OLS}$ and variance $\sigma^2\boldsymbol{\Sigma}_c$. Note that, in theory, r can be bigger than n, i.e., one can perform oversampling, but this is beyond the scope of this paper.

Remark. Similar to unconditional inference, the asymptotic variance $AVar(\tilde{\boldsymbol{\beta}})$ here also has "sandwich-type" expression, where the center term (here, $\left(\sum_{i=1}^n \frac{e_i^2}{\pi_i}\mathbf{x}_i\mathbf{x}_i^T\right)$) depends on the reciprocal sampling probabilities. Thus, we also expect that extremely small probabilities will result in large variances of the corresponding estimators.

Remark. In Theorem 3, $AVar(\tilde{\boldsymbol{\beta}})$ depends on the full sample least square residuals, i.e., the e_is. These are not readily available from the sample. To solve this problem and to obtain meaningful results, we take the expectation of the e_i^2s. The metric we use is thus the EAMSE,

$$EAMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS}) = E_{\boldsymbol{\epsilon}}(AMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS})). \tag{20}$$

The EAMSE is a function of the sampling probabilities $\{\pi_i\}_{i=1}^n$.

It is natural to ask whether there is an optimal estimator, i.e., a sample estimator with the smallest EAMSE for estimating $\boldsymbol{\beta}_{OLS}$. Using the asymptotic results in Theorem 3, we propose the following three estimators for various purposes.

Estimating $\hat{\boldsymbol{\beta}}_{OLS}$. We can use the results of Theorem 3 to obtain expressions of interest for the EAMSE of various quantities. As with the AMSE, these will depend on the sampling probabilities. Thus, we can derive the optimal sampling probabilities for various quantities of interest. We start with $EAMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS})$.

The following proposition gives the minimum $EAMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS})$ sampling estimator. For this result, we denote that $\mathrm{E}_{\boldsymbol{\varepsilon}}(e_i^2) = (1 - h_{ii})\sigma^2$.

Proposition 4 *For the $EAMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS})$, we have that*

$$EAMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS}) = E_{\boldsymbol{\varepsilon}}(tr(AVar(\tilde{\boldsymbol{\beta}}))) = \frac{1}{r}\sum_{i=1}^{n} \frac{(1 - h_{ii})\sigma^2}{\pi_i} ||(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i||^2. \tag{21}$$

Given (21)*, the sample estimator with the sampling probabilities*

$$\pi_i = \frac{\sqrt{1 - h_{ii}}\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^{n}\sqrt{1 - h_{ii}}\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}, i = 1, \ldots, n, \tag{22}$$

(which we call the inverse-covariance negative-leverage (ICNLEV) estimator) has the smallest $EAMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS})$.

Estimating linear functions of $\hat{\boldsymbol{\beta}}_{OLS}$. In addition to approximating $\hat{\boldsymbol{\beta}}_{OLS}$, one may also be interested in linear functions of $\hat{\boldsymbol{\beta}}_{OLS}$. Here, we present results for $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$ and $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$ (although clearly similar results hold for other functions of the form $\mathbf{L}\hat{\boldsymbol{\beta}}_{OLS}$).

We start with estimating $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$.

Proposition 5 *For the $EAMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})$, we have that*

$$EAMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}) = \frac{1}{r}\sum_{i=1}^{n} \frac{(1 - h_{ii})\sigma^2}{\pi_i} ||\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i||^2. \tag{23}$$

Given (23)*, the sample estimator with the sampling probabilities*

$$\pi_i = \frac{\sqrt{1 - h_{ii}}\|\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^{n}\sqrt{1 - h_{ii}}\|\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|} = \frac{\sqrt{(1 - h_{ii})h_{ii}}}{\sum_{i=1}^{n}\sqrt{(1 - h_{ii})h_{ii}}}, i = 1, \ldots, n, \tag{24}$$

(which we call the root leveraging negative-leverage (RLNLEV) estimator) has the smallest value for the $EAMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})$.

We next consider estimating $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$.

Proposition 6 *For the $EAMSE(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})$, we have that*

$$EAMSE(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}) = \frac{1}{r}\sum_{i=1}^{n} \frac{(1 - h_{ii})\sigma^2}{\pi_i} ||\mathbf{x}_i||^2. \tag{25}$$

Given (25)*, the sampling estimator with the sampling probabilities*

$$\pi_i = \frac{\sqrt{1 - h_{ii}}\|\mathbf{x}_i\|}{\sum_{i=1}^{n}\sqrt{1 - h_{ii}}\|\mathbf{x}_i\|}, i = 1, \ldots, n, \tag{26}$$

(which we call the predictor-length negative-leverage (PLNLEV) estimator) has the smallest value for the $EAMSE(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})$.

Estimator	Sampling Probabilities	Criterion	Results
UNIF	$\pi_i = \frac{1}{n}$	$--$	$--$
BLEV/ALEV	$\pi_i = \frac{h_{ii}}{\sum_{i=1}^n h_{ii}}$	$--$	Drineas et al. (2006, 2012a)
SLEV	$\pi_i = \lambda \frac{h_{ii}}{\sum_{i=1}^n h_{ii}} + (1-\lambda)\frac{1}{n}$	$--$	Ma et al. (2014, 2015)
	Data are random	Two sources of randomness	
IC	$\pi_i = \frac{\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^n \|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}$	$AMSE(\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}_0)$	Section 2.2, Eqn. (10)
RL	$\pi_i = \frac{\sqrt{h_{ii}}}{\sum_{i=1}^n \sqrt{h_{ii}}}$	$AMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}\boldsymbol{\beta}_0)$	Section 2.2, Eqn. (12)
PL	$\pi_i = \frac{\|\mathbf{x}_i\|}{\sum_{i=1}^n \|\mathbf{x}_i\|}$	$AMSE(\boldsymbol{X}^T\boldsymbol{X}\tilde{\boldsymbol{\beta}}; \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}_0)$	Section 2.2, Eqn. (14)
	Data are given and fixed	One source of randomness	
ICNLEV	$\pi_i = \frac{\sqrt{1-h_{ii}}\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^n \sqrt{1-h_{ii}}\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}$	$EAMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS})$	Section 2.4, Eqn. (22)
RLNLEV	$\pi_i = \frac{\sqrt{(1-h_{ii})h_{ii}}}{\sum_{i=1}^n \sqrt{(1-h_{ii})h_{ii}}}$	$EAMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})$	Section 2.4, Eqn. (24)
PLNLEV	$\pi_i = \frac{\sqrt{1-h_{ii}}\|\mathbf{x}_i\|}{\sum_{i=1}^n \sqrt{1-h_{ii}}\|\mathbf{x}_i\|}$	$EAMSE(\boldsymbol{X}^T\boldsymbol{X}\tilde{\boldsymbol{\beta}}; \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS})$	Section 2.4, Eqn. (26)

Table 1: Summary of three existing sampling estimators (UNIF, BLEV, SLEV) and the six sampling estimators (IC, RL, PL, ICNLEV, RLNLEV, PLNLEV) presented in this paper.

Remark. All these proposed metrics can be approximated in the time it takes to approximate leverage scores, i.e., to implement a random projection, using the algorithm of Drineas et al. (2012a), since they are essentially strongly related to leverage scores.

As a summary, the six proposed estimators (IC, RL, PL, ICNLEV, RLNLEV, PLNLEV), along with three existing estimators (UNIF, BLEV/ALEV, SLEV) are presented in Table 1.

2.5 Relationship of the Sampling Estimators

Here, we study the relationships between the probability distributions given by IC, RL, PL, ICNLEV, RLNLEV, PLNLEV, and those given by SLEV and BLEV.

2.5.1 "Shrinkage" Properties of Proposed Estimators

We illustrate the "shrinkage" property of the proposed optimal sampling probabilities, compared to the BLEV sampling probabilities. For convenience, we refer to the numerators of the sampling probabilities in a sampling estimators as the scores, e.g., the RL score is $\sqrt{h_{ii}}$ and the RLNLEV score is $\sqrt{(1-h_{ii})h_{ii}}$. In Figure 1, we plot the RL score, RLNLEV score, and SLEV score ($0.9h_{ii} + 0.1p/n$ with $p/n = 0.2$) as functions of the leverage score h_{ii} (i.e., the BLEV score in Figure 1). Observe that the RLNLEV scores amplify small h_{ii}s but shrink large h_{ii}s. RL scoress provide nonlinear amplification of h_{ii}s. The probability counterparts of RL scores, i.e., RL scores divided by their summation, shrink large h_{ii}s and amplify small h_{ii}s. The SLEV scores also shrink large h_{ii}s and amplify small h_{ii}s, but in a linear fashion.

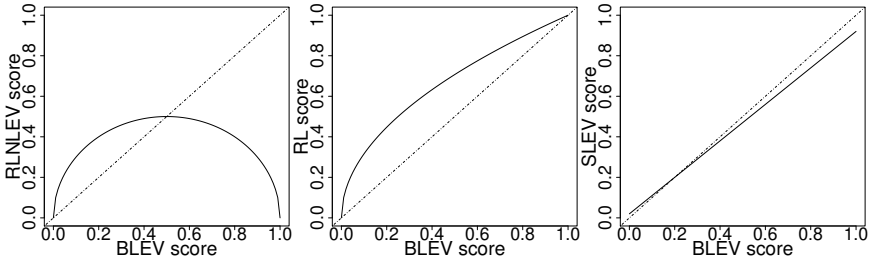


Figure 1: Relationship between different sampling methods. Left panel: RLNLEV score ($\sqrt{(1-h_{ii})h_{ii}}$) versus BLEV score (h_{ii}). Middle panel: RL score ($\sqrt{h_{ii}}$) versus BLEV score (h_{ii}). Right panel: SLEV score ($0.9h_{ii}+0.1p/n$, where $p/n=0.2$) versus BLEV score (h_{ii}).

Sampling methods (IC,RL,PL) are derived under the AMSE criterion, whereas sampling methods (ICNLEV, RLNLEV, PLNLEV) are deveried under the EAMSE criterion. Thus, they are not comparable. However, they do have interesting connections. Comparing the sampling probabilities of (IC,RL,PL) to those of (ICNLEV, RLNLEV,PLNLEV), we can see that the only difference is the inclusion of the shrinkage $\sqrt{1-h_{ii}}$ in the latter probabilities. This is due to the following fact: (IC,RL,PL) aim for the ground truth, whereas (ICNLEV, RLNLEV, PLNLEV) aim for OLS estimators. When aiming for the ground truth, we have a regression problem with constant variance noise, i.e., $\mathrm{E}(\epsilon_i^2)=\sigma^2$. When aiming for OLS estimators, we have a regression problem with heteroscedastic residuals. That is, treating the OLS estimator $\hat{\beta}_{OLS}$ as the "ground truth," we have that the variance of residual $e_i=Y_i-x_i^T\hat{\beta}_{OLS}$ is $\mathrm{E}_{\boldsymbol{\varepsilon}}(e_i^2)=(1-h_{ii})\sigma^2$. Thus, our sampling methods naturally take such a heterogeneity into account, leading to sampling probabilities proportional to standard deviation of residuals $\sqrt{1-h_{ii}}$.

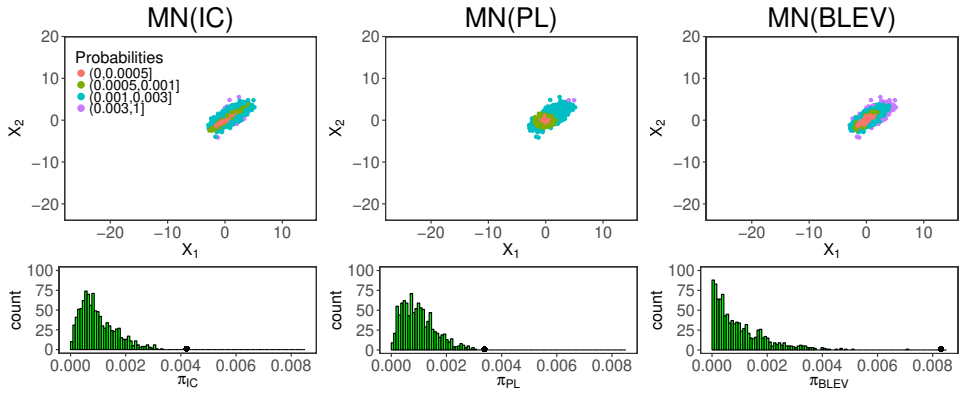
2.5.2 The Role of h_{ii}s.

On the one hand, if the h_{ii}s are homogeneous, then the sampling probabilities of the ICNLEV estimator ($\frac{\sqrt{1-h_{ii}}\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^n\sqrt{1-h_{ii}}\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}$) and those of the IC estimator ($\frac{\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^n\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}$) will be similar to each other. On the other hand, since $\sum_{i=1}^n h_{ii}=p$, given a fixed value of p, we expect that h_{ii}s are small when sample size n is large. When $h_{ii}=o(1)$ for all $i=1,\ldots,n$, i.e., h_{ii}s are extremely small compared to 1, the sampling probabilities of the ICNLEV estimator and those of the IC estimator will also be similar. Analogous arguments also apply to PLNLEV and PL.

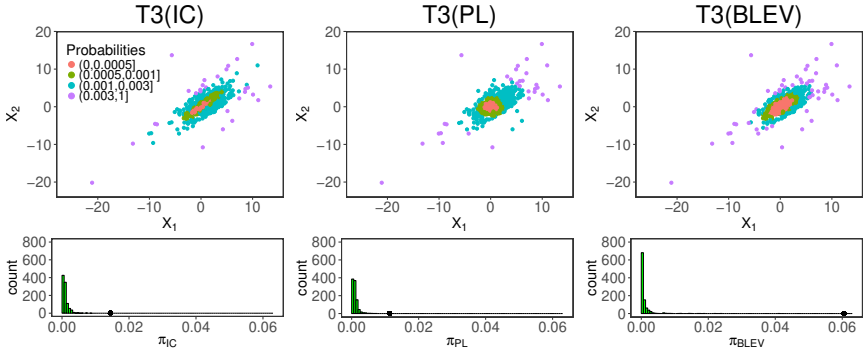
2.5.3 Two Examples.

We now use two examples to illustrate the relationship between the sampling probabilities in various sampling estimators.

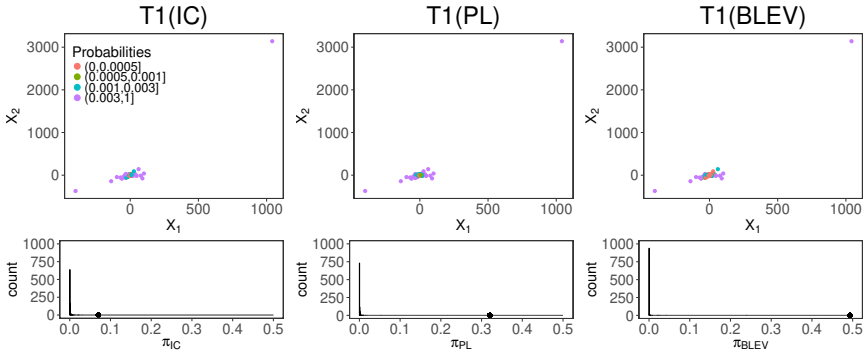
Example 1: Orthogonal predictor matrix, i.e., $\mathbf{X}^T\mathbf{X}=\mathbf{I}$. Consider a linear regression model with an orthogonal predictor matrix, i.e., $\mathbf{X}^T\mathbf{X}=\mathbf{I}$. In this case, we have $h_{ii}=\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i=\|\mathbf{x}_i\|^2$. Further, the ICNLEV score, RLNLEV score, and PLNLEV score are the same and equal $\sqrt{(1-h_{ii})h_{ii}}$. Analogously, the IC score coincides with the RL score and the PL score, and all equal $\|\mathbf{x}_i\|=\sqrt{h_{ii}}$.



(a) Scatter plots (first row) of data points generated from a bivariate normal distribution with colors coding the sampling probability in the IC (left panel), PL (middle panel), and BLEV (right panel). Below each scatter plot is the histogram of the corresponding sampling probabilities, with the dot representing the maximum probability.



(b) Same as in (a), except that the data points are generated from a bivariate noncentral t distribution with three degrees of freedom.



(c) Same as in (a), except that the data points are generated from a bivariate noncentral t distribution with one degree of freedom.

Figure 2: Scatter plots of 1000 data points generated from three distributions in Example 2 in Section 2.5.3 and the histograms of sampling probabilities.

Example 2: A two dimensional example. Consider also a toy example of a linear regression model with $p = 2$ correlated predictors. We generated 1000 data points for two predictors from a multivariate normal distribution, a multivariate noncentral t distribution

with three degrees of freedom, and a multivariate noncentral t distribution with one degree of freedom.

In Figure 2, we present scatter plots of these data points. In each scatter plot, the color of points indicates the magnitude of sampling probabilities in IC, PL and BLEV methods. Below each scatter plot, we also present histograms of the corresponding sampling probabilities. Examination of Figure 2 reveals one pattern shared by all sampling distributions, i.e., the sampling probabilities of data points in the center are smaller than those of data points at the boundary. In addition, note that, compared to $\pi_i^{PL} \propto \|\mathbf{x}_i\|$, both $\pi_i^{IC} \propto \|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|$ and $\pi_i^{BLEV} \propto \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$ depend on $(\mathbf{X}^T\mathbf{X})^{-1}$, which normalizes the scale of the predictors. Such normalization causes the directions (principle components) with the smallest eigenvalue have higher sampling probabilities. This is evident in Figure 2. Specifically, the colored clouds associated to IC and BLEV are more skewed, and the data points with high probabilities in IC and BLEV form contours toward the exterior of the data clouds. For PL, the colors simply depend upon the distance (norm) from $(0, 0)$.

The histograms in each row also show the key difference between the sampling probabilities of BLEV and those of IC and PL, i.e., the sampling probability distribution of BLEV is more dispersed than others. In other words, there are a significant number of data points with either extremely large or extremely small probabilities in BLEV. This phenomenon is also observed in Figure 3 in Section 3.

3. Empirical Results

In this section, we present a summary of the main results of our empirical analysis, which consisted of an extensive analyses on simulated and real data sets.

3.1 Simulation Setting

We generated synthetic data from Model (1) with $p = 10$, $n = 5000$,[4] and random error $\varepsilon_i \overset{iid}{\sim} N(0, 1)$. We set the first and last two entries of $\boldsymbol{\beta}_0$ to be 1 and the rest to be 0.1. We generated the predictors from the following distributions.

- Multivariate normal distribution $\mathbf{N}(\mathbf{1}, \mathbf{D})$, where $\mathbf{1}$ is a $p \times 1$ column vector of 1s, and the $(i, j)^{th}$ element of $\mathbf{D}$ is set to $1 \times 0.7^{|i-j|}$, for $i, j = 1, \ldots, p$. We refer to this as MN data.
- Multivariate noncentral t-distribution with 3 degrees of freedom, noncentrality parameter $\mathbf{1}$, and scale matrix $\mathbf{D}$, i.e., $t_3(\mathbf{1}, \mathbf{D})$. We refer to this as T3 data.
- Log-normal distribution $\mathbf{LN}(\mathbf{1}, \mathbf{D})$. We refer to this as LN data.
- Multivariate noncentral t-distribution with 1 degree of freedom, noncentrality parameter $\mathbf{1}$, and scale matrix $\mathbf{D}$, i.e., $t_1(\mathbf{1}, \mathbf{D})$. We refer to this as T1 data.

Note that for $t_1(\mathbf{1}, \mathbf{D})$, its expectation and variance do not exist. This violates Condition (A1) in Theorem 1. We include this distribution in the simulation to explore the estimators' performance, in a situation which has no guarantees by our theoretical analysis.

[4] We have also done simulation with p up to 100 and n up to 1×10^6. The observations discussed below in this section are generally applicable

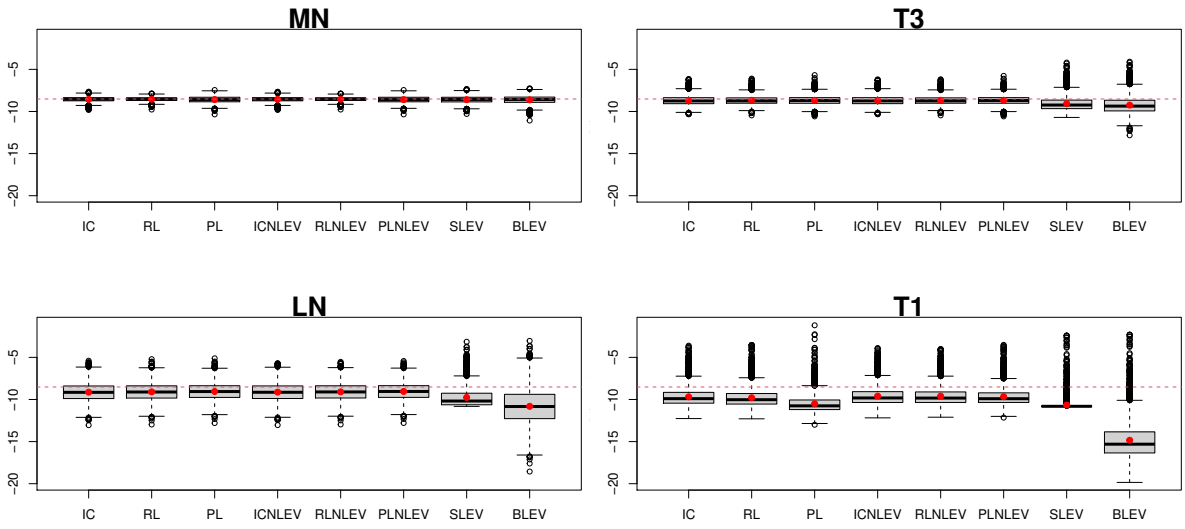


Figure 3: Box plots of the sampling probabilities (in log scale) of all data points for IC, RL, PL, ICNLEV, RLNLEV, PLNLEV, SLEV, and BLEV (from left to right in each panel) for MN, T3, LN, and T1 data, for p=10 and n =5000. In each box plot, the dot inside the box indicates the mean of corresponding sampling probabilities (in log scale). The dashed red lines represent the uniform sampling probabilities.

In Figure 3, we present box plots of the sampling probabilities (in log scale) of all the data points in IC, RL, PL, ICNLEV, RLNLEV, PLNLEV, SLEV, and BLEV (from left to right) for MN, T3, LN, and T1. The sampling probability distributions of BLEV are more dispersive than those of other estimators. There exist a significant number of extremely small sampling probabilities in BLEV, especially when the data distribution has heavier tails, such as is the case for LN and T1. These extremely small sampling probabilities in BLEV are effectively mitigated in SLEV. However, the medians of the sampling probabilities in SLEV are still smaller than the first quartiles of the sampling probabilities in ICNLEV, IC, PLNLEV, and PL in T3, LN, and T1. The relatively small sampling probabilities in BLEV and SLEV will inflate the variance of the sampling estimators (recall the expression for the asymptotic variances in Theorems 1 and 3). Thus, it is expected that BLEV and SLEV will give rise to estimates with relatively large variances, especially when data were generated from more heavy-tailed distributions, e.g., LN and T1. It is also observed that sampling probabilities in PL in T1 have large variance. The observation is consistent with the fact that PL sampling probabilities are constructed directly from the more heavy-tailed T1 data.

3.2 Sampling Estimators for Estimating Model Parameters

Here, we evaluate the performance of the proposed sampling estimators in estimating $\boldsymbol{\beta}_0$, $\boldsymbol{X}\boldsymbol{\beta}_0$, and $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}_0$. Under the simulation settings of Section 3.1, we generated 100 replicates of MN, T3, LN, and T1 data. We applied IC, RL, PL, SLEV (with $\lambda = 0.9$ here

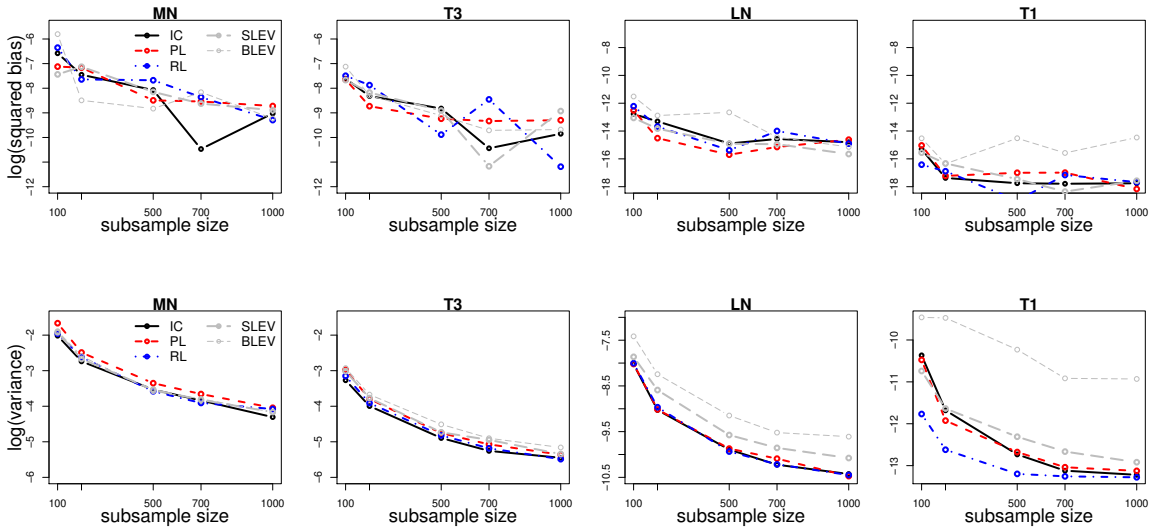


Figure 4: Squared biases (first row) and variances (second row) of IC, RL, PL, SLEV, and BLEV estimates in estimating $\boldsymbol{\beta}_0$ (in log scale) at different sample sizes.

and after), and BLEV to each replicated data set to obtain sampling estimates at sample sizes $r = 100, 200, 500, 700, 1000$. Then, we calculated the squared bias and variance for each method. In particular, let $\tilde{\boldsymbol{\beta}}_b$ be the subsampling estimator in the b^{th} replicated sample, and $\bar{\tilde{\boldsymbol{\beta}}} = \frac{1}{100}\sum_{b=1}^{100}\tilde{\boldsymbol{\beta}}_i$ be the sample mean of $\tilde{\boldsymbol{\beta}}_b$ across the replicated samples. The sample squared bias and variance of $\tilde{\boldsymbol{\beta}}$ are calculated as below,

$$\begin{aligned} \text{Sample Squared Bias}(\tilde{\boldsymbol{\beta}}) &= \|\bar{\tilde{\boldsymbol{\beta}}} - \boldsymbol{\beta}_0\|^2, \\ \text{Sample Variance}(\tilde{\boldsymbol{\beta}}) &= \frac{1}{100}\sum_{b=1}^{100}\|\tilde{\boldsymbol{\beta}}_b - \bar{\tilde{\boldsymbol{\beta}}}\|^2. \end{aligned} \tag{27}$$

The sample squared biases and variances of $\boldsymbol{X}\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{X}^T\boldsymbol{X}\tilde{\boldsymbol{\beta}}$ are analogously calculated.

In Figure 4, we plot the squared biases (first row) and the variances (second row) (in log scale) for IC, RL, PL, SLEV, and BLEV estimates in estimating $\boldsymbol{\beta}_0$ in MN, T3, LN, and T1. First, both the squared biases and the variances show decreasing patterns as r increases. The squared biases of different methods are similar to each other and are much smaller than the corresponding variances. These observations are expected, since Theorem 1 states that the RandNLA estimators are asymptotically unbiased and consistent estimators of $\boldsymbol{\beta}_0$. Second, the variances of estimates using IC, whose sampling probabilities minimize $AMSE(\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}_0)$, are slightly smaller than the variances of estimates using other methods in MN and T3, at most sample sizes. The variances of estimates using IC, RL, and PL are all smaller than those of BLEV and SLEV estimates in T3. As mentioned in the discussion of Figure 3, the larger variances of BLEV estimates are caused by the extremely small sampling probabilities in BLEV. Taking a weighted average of the sampling probability distribution of BLEV and that of UNIF shows a beneficial effect on the variances for SLEV

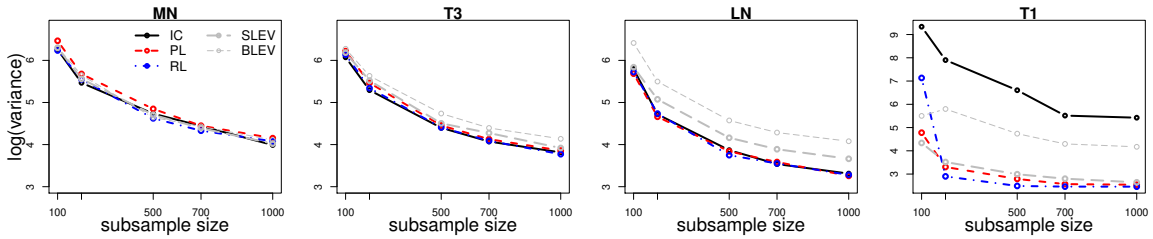


Figure 5: The variances of IC, RL, PL, SLEV, and BLEV estimates in predicting $\boldsymbol{X\beta_0}$ (in log scale) at different sample sizes.

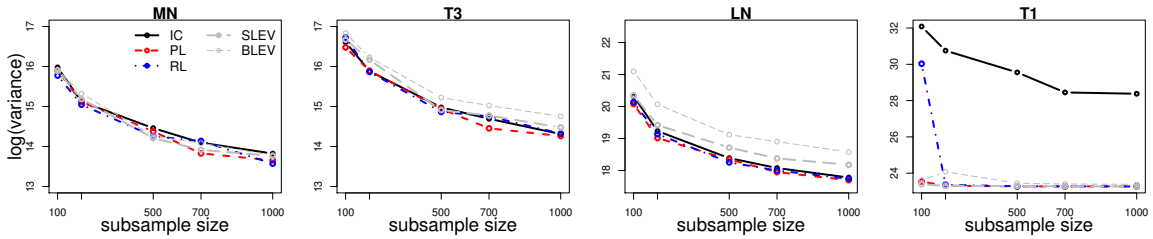


Figure 6: The variances of IC, RL, PL, SLEV, and BLEV estimates in estimating $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$ (in log scale) at different sample sizes.

estimators. However, the variances of SLEV estimators are still larger than those of IC in T3, LN, and T1 at larger sample sizes. Third, for T1, despite the violation of the regularity condition in Theorem 1, our proposed estimators IC, RL, and PL still outperform BLEV and SLEV in terms of variances, when sample size is greater than 200. Fourth, the squared biases and variances of all estimates get smaller from left panels to right panels.

For estimating $\boldsymbol{X\beta_0}$ and $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$, the biases of all sampling estimators are very similar to each other and are much smaller than the corresponding variances. This observation is consistent with what we observed in estimating $\boldsymbol{\beta}_0$ in Figure 4. We thus only present the variances of IC, RL, PL, SLEV, and BLEV estimates in estimating $\boldsymbol{X\beta_0}$ and $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$ at different sample sizes in Figure 5 and Figure 6. As shown, the variances of the estimates for estimating both $\boldsymbol{X\beta_0}$ and $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$, using PL, IC, and RL, are smaller than the variances of estimates using BLEV and SLEV in T3 and LN, at most sample sizes. We observe that IC does perform as well as BLEV and SLEV for T1 data, which has a very large variation with many outliers as suggested in Figure 2. This observation is consistent with the fact that the IC probability relies on the computation of $(\mathbf{X}^T\mathbf{X})^{-1}$, which is very unstable for T1 distributed $\mathbf{X}$, resulting in larger variances for the resulting estimates.

3.3 Sampling Estimators for Approximating the Full Sample OLS Estimate

Here, we evaluate the performance of the proposed sampling estimators for approximating $\hat{\boldsymbol{\beta}}_{OLS}$, $\boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS}$, and $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS}$. Under the simulation settings of Section 3.1, we generated four data sets without replicates from MN, T3, LN, and T1, respectively. For

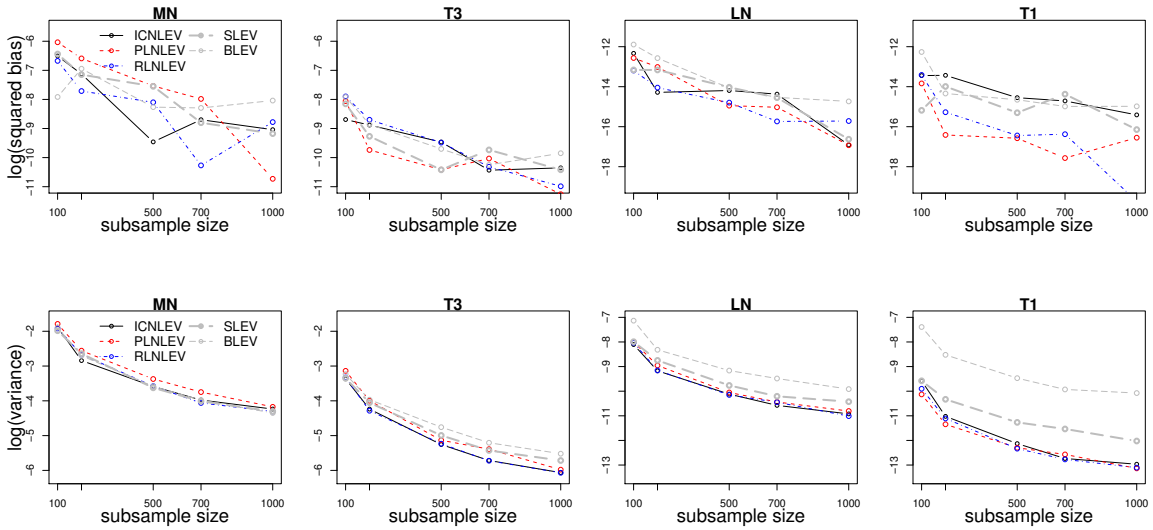


Figure 7: Squared biases (first row) and variances (second row) of ICNLEV, RLNLEV, PLNLEV, SLEV, and BLEV estimates in approximating $\hat{\boldsymbol{\beta}}_{OLS}$ (in log scale) at different sample sizes.

each data set, the full sample OLS estimate was calculated. We set samples sizes at $r = 100, 200, 500, 700, 1000$. We repeatedly applied ICNLEV, RLNLEV, PLNLEV, SLEV, and BLEV methods 100 times at each sample size to get sampling estimates $\tilde{\boldsymbol{\beta}}_b$, where $b = 1, \ldots, 100$. Using these estimates, we calculated the squared bias and variance for each method for approximating $\hat{\boldsymbol{\beta}}_{OLS}$.

In Figure 7, we plot the squared biases and variances (in log scale) for ICNLEV, RLNLEV, PLNLEV, SLEV, and BLEV estimates for approximating $\hat{\boldsymbol{\beta}}_{OLS}$ at different sample sizes in all data sets. Several observations are worth noting in Figure 7. First, the squared biases are negligible compared to the corresponding variances. For all sampling methods, both the squared biases and the variances decrease as sample size increases. These observations are in agreement with Theorem 3, which states that the sampling estimators are asymptotically unbiased estimators of $\hat{\boldsymbol{\beta}}_{OLS}$, provided that the regularity conditions are satisfied. Second, the variances of estimates using ICNLEV and RLNLEV are slightly smaller than the variances of estimates using other methods in T3 and LN at most sample sizes. The variances of estimates using ICNLEV, RLNLEV, and PLNLEV are consistently smaller than those of SLEV and BLEV in LN and T1. Third, all sampling estimators perform better in LN and T1 than in T3 and MN, i.e., the squared biases and variances of all estimates in LN and T1 are smaller than those in T3 and MN.

To examine the performance of the RandNLA sampling estimators for approximating $\hat{\mathbf{Y}}_{OLS} (= \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})$, we plot the variances (in log scale) of $\mathbf{X}\tilde{\boldsymbol{\beta}}_b$, at different sample sizes, for all sampling estimators in Figure 8. The variances of estimates using RLNLEV, whose sampling probabilities minimize $EAMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})$, are slightly smaller than those of estimates using other methods at all sample sizes in T3 and at most sample sizes in LN.

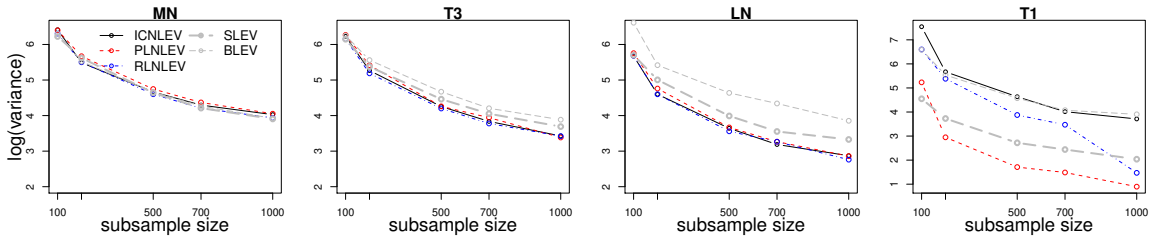


Figure 8: The variances of ICNLEV, RLNLEV, PLNLEV, SLEV, and BLEV estimates in approximating $\hat{\mathbf{Y}}_{OLS}(= \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})$ (in log scale) at different sample sizes.

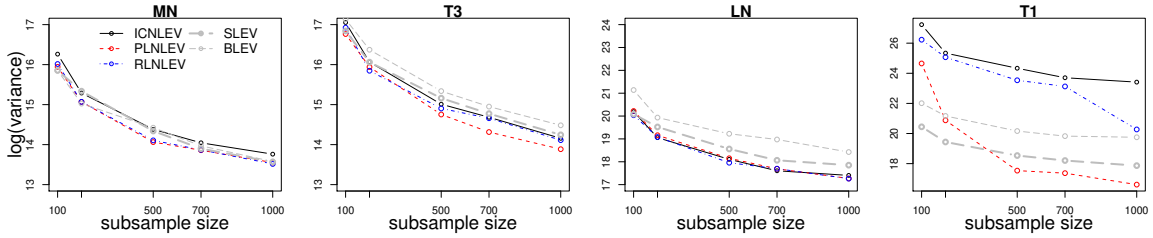


Figure 9: The variances of ICNLEV, RLNLEV, PLNLEV, SLEV, and BLEV estimates in approximating $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$ (in log scale) at different sample sizes.

To assess the performance of the RandNLA sampling estimators for approximating $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$, we plot the variances (in log scale) of $\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\beta}}_b$, at different sample sizes, for all sampling estimators in Figure 9. For all estimators, the variances decrease as the sample size increases. Also, in T3, the variances of estimates using PLNLEV, whose sampling probabilities minimize $EAMSE(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})$ are smaller than the variances of estimates using other methods at most sample sizes. In this case, despite the violation of the conditions for the proper definition of EAMSE in T1, the variances of PLNLEV estimates are still the smallest, when sample sizes are greater than 200.

3.4 Flight Delay Data set

Here, we evaluate the performance of the sampling estimators on a flight delay data set we compiled from the website of the US Department of Transportation.[5] The data set contains records of $3,274,894$ US domestic flights during weekdays from Mondays to Thursdays in 2017. There are five variables for each flight record: arrival delay (difference in minutes between scheduled and actual arrival time, and early arrivals show negative numbers), arrival taxi in time (in minutes), departure taxi out time (in minutes), departure delays (difference in minutes between scheduled and actual departure time, and early depar-

[5]U. S. Bureau of Transportation Statistics. Rita airline delay data was downloaded from: `https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236`.

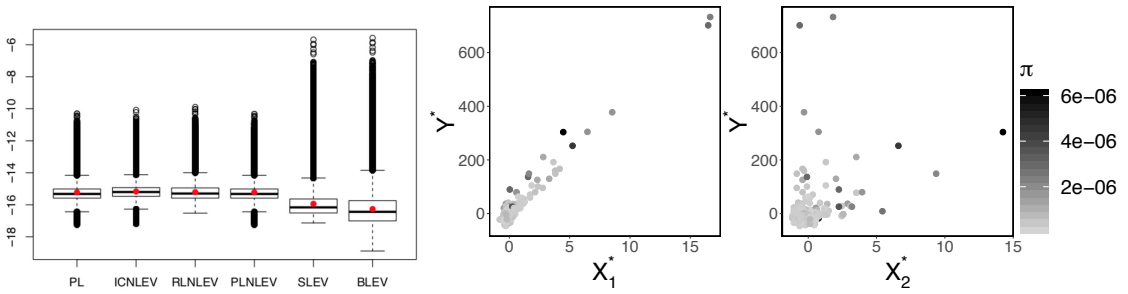


Figure 10: Flight delay data set. Left: the box plots of sampling probabilities (in log scale) of all data points in PL, ICNLEV, RLNLEV, PLNLEV, and BLEV. Middle and Right: the scatter plots of the 200 sampled response vector (ARRIVAL_DELAY) and two predictors (DEPARTURE_DELAY and TAXI_OUT) using the ICNLEV sampling probability distribution.

tures show negative numbers), and computer reservation system based elapsed time of the flight (in minutes; a measure for the distance of the flight). We are interested in predicting the arrival delay of each flight using the rest of the variables. We fitted Model (1), with the response being flight arrival delay. In addition to using the four variables (other than arrival delay) in our data set as linear predictors, we also included their quadratic and all pairwise interaction terms. We thus have 14 predictors. Considering the large number of flights, we use the sampling methods to approximate the full sample OLS estimate. Note that the computation of the OLS estimates in Sections 3.4 and 3.5 was conducted in a supercomputer in the Pittsburgh Supercomputing Center.

In the left panel of Figure 10, we present the box plots of sampling probabilities (in log scale) of all data points in PL, ICNLEV, RLNLEV, PLNLEV, SLEV, and BLEV. Observe that the sampling probability distributions are right-skewed, similar to those in Figure 3 in the simulation study. Using the sampling probability distribution in ICNLEV, we took a sample of size 200 from the full data. The middle and right panels in Figure 10 are the scatter plots of the sampled response and the first two predictors, respectively. These scatter plots provide a visual sketch of the full sample data.

We repeatedly applied the PL, ICNLEV, IC, PLNLEV, SLEV, and BLEV methods to this data set for 100 times at sample size $r = 20p, 50p, 70p, 100p, 200p$, where $p = 14$. We calculated the squared bias and variance of the resulting estimates in approximating $\hat{\boldsymbol{\beta}}_{OLS}$, $\hat{\boldsymbol{Y}}_{OLS}$ and $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$, for each method. The results are summarized in Figure 11. Observe that the squared biases of all methods are all much smaller than the corresponding variances for all methods at all sample sizes. For approximating $\hat{\boldsymbol{\beta}}_{OLS}$, the ICNLEV estimates have the smallest variance consistently at all sample sizes among all estimators. For approximating $\hat{\boldsymbol{Y}}_{OLS}$ and $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$, the estimates using PLNLEV, PL, and RLNLEV are very similar to each other, and they have better performance in terms of variances at all sample sizes than those using BLEV and SLEV.

3.5 "YearPredictionMSD" Data set

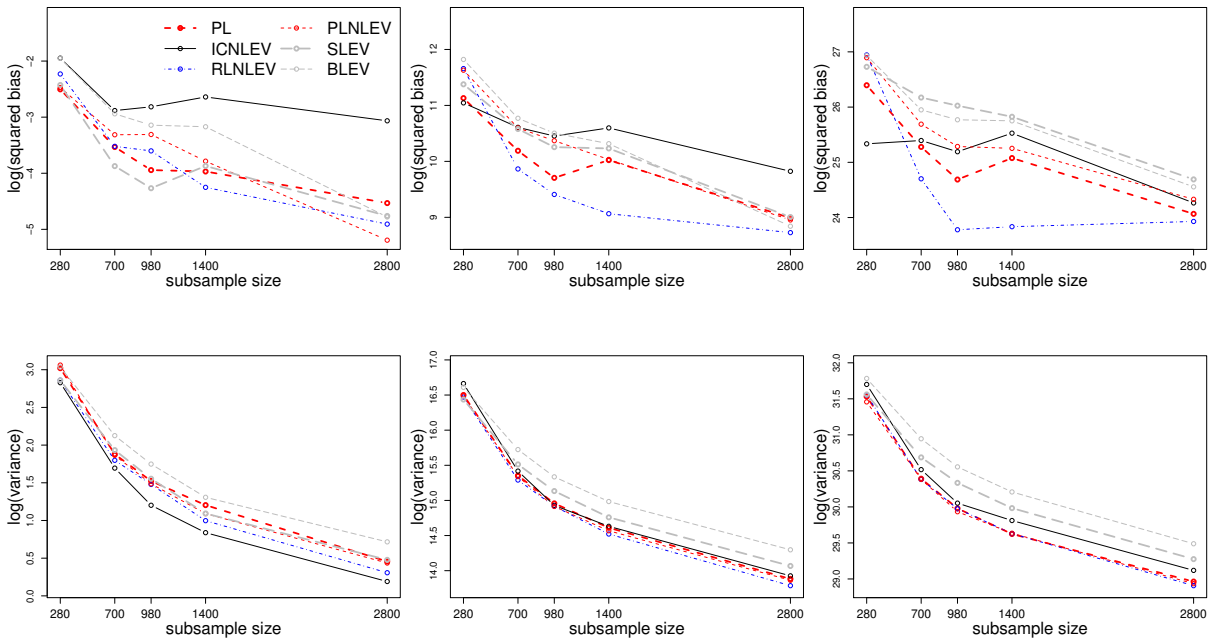


Figure 11: Squared biases (first row) and variances (second row) of PL, ICNLEV, RLNLEV, PLNLEV, SLEV, and BLEV estimates for approximating $\hat{\boldsymbol{\beta}}_{OLS}$ (first column), $\hat{\boldsymbol{Y}}_{OLS}$ (second column) and $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$ (third column) (in log scale) at different sample sizes for Airline Delay data.

Here, we evaluate the performance of the sampling estimators on the "YearPrediction-MSD" data set (Bertin-Mahieux et al., 2011), which we downloaded from the UCI machine learning repository.[6] The data set consists of records of 515,345 songs released between the year 1922 and 2011. For each song, multiple segments are taken, and each segment is characterized by 12 timbre features. These timbre features capture timbral characteristics, such as brightness and flatness, of each segment. The mean and variance of each timbre feature, as well as the covariances between every two timbre features, are calculated. Our primary interest for our analysis is to use all timbre feature information to predict the year of release. We fitted Model (1), where the response is the year (in log scale) of releasing of the song, and the predictors include all timbre features.

In the left panel of Figure 12, we present the box plots of sampling probabilities (in log scale) of all data points in PL, ICNLEV, RLNLEV, PLNLEV, SLEV, and BLEV. Inspecting the box plots reveals that all sampling distributions are right-skewed and that the sampling distributions of SLEV and BLEV are much more dispersed than those of other estimators. Using the sampling probability distribution in ICNLEV, we took a sample of size 200 from the full data. The middle and right panels of Figure 12 are the scatter plots of the sampled response and two timbre features, respectively.

[6] See `http://archive.ics.uci.edu/ml/`.

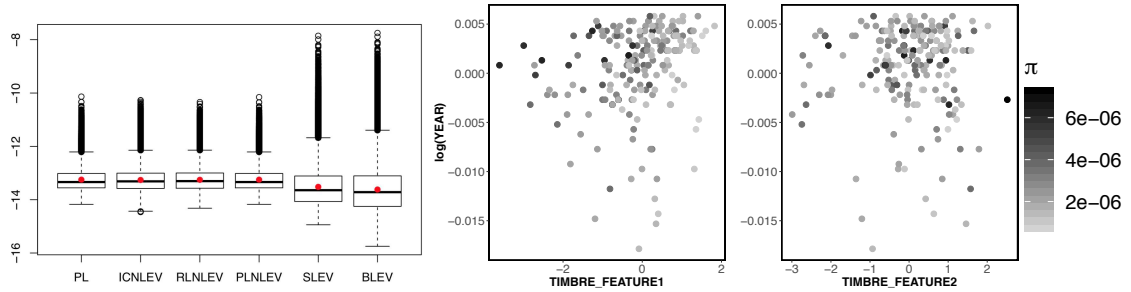


Figure 12: "YearPredictionMSD" data. Left: the box plots of sampling probabilities (in log scale) for all data points for PL, ICNLEV, RLNLEV, PLNLEV, SLEV, and BLEV. A sample of size 200 is taken from the full data using the sampling probabilities of ICNLEV. Middle and Right: the scatter plots of sampled response and two timbre feature predictors.

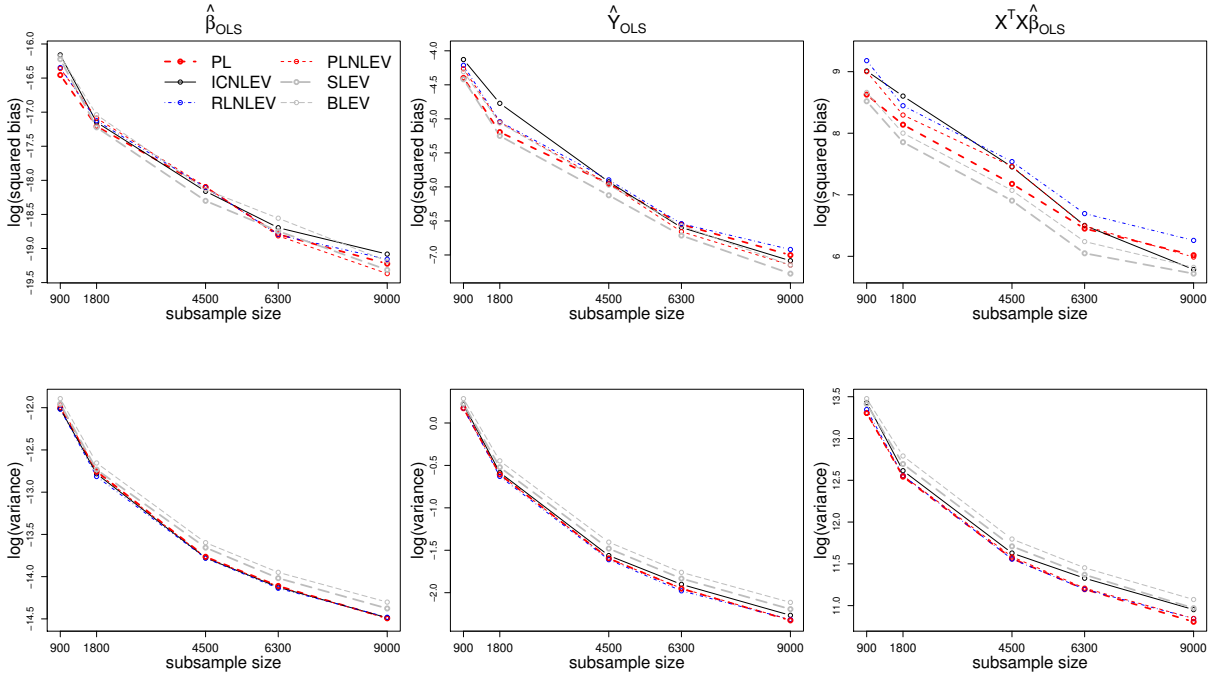


Figure 13: Squared biases (first row) and variances (second row) of PL, ICNLEV, RLNLEV, PLNLEV, SLEV, and BLEV estimates for approximating $\hat{\boldsymbol{\beta}}_{OLS}$ (first column), $\hat{\boldsymbol{Y}}_{OLS}$ (second column), and $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$ (third column) (in log scale) at different sample sizes for "YearPredictionMSD" data.

We repeatedly applied the ICNLEV, RLNLEV, PLNLEV, SLEV, and BLEV methods to the data set for 100 times at sample sizes $r = 10p, 20p, 50p, 70p, 100p$, where $p = 90$. In Figure 13, we plot the squared biases and the variances (in log scale) of the estimates for all weighted sampling methods for approximating $\hat{\boldsymbol{\beta}}_{OLS}$, $\hat{\boldsymbol{Y}}_{OLS}$, and $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$. For all three scenarios, the squared biases are much smaller than the corresponding variances, for all methods at all sample sizes. For approximating $\hat{\boldsymbol{\beta}}_{OLS}$, the variances of ICNLEV,

RLNLEV, and PLNLEV estimates are comparable to each other and consistently smaller than those of SLEV and BLEV estimates at all sample sizes. For approximating $\hat{\boldsymbol{Y}}_{OLS}$ and $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$, the variances of PLNLEV and PL estimates are consistently smaller than those of other estimates.

4. Conclusion

We have studied the asymptotic properties of RandNLA sampling estimators in LS linear regression models. We showed that under certain regularity conditions on the data distributions and sampling probability distributions, the sampling estimators are asymptotically normally distributed. Moreover, the sampling estimators are asymptotically unbiased for approximating the full sample OLS estimate and for estimating true coefficients. Based on these asymptotic results, we proposed optimality criteria to assess the performance of the sampling estimators, based on AMSE and EAMSE. In particular, we developed six sampling estimators, i.e., IC, RLEV, PL, ICNLEV, RLNLEV, and PLNLEV, for minimizing AMSE and EAMSE, under a variety of settings. These empirical results demonstrate that these new sampling estimators outperform the conventional ones in the literature. For generalization, depending on the application, one may consider criteria other than AMSE and EAMSE. For example, when hypothesis testing problems are of primary interest, the power of the test is a more reasonable choice to serve as a criterion. Developing scalable sampling methods to optimize criteria such as this are of interest.

Acknowledgment

We would like to thank Shusen Wang for providing constructive comments on an earlier version of this paper and Bin Yu for helpful discussions. PM, XZ, and XX acknowledge NSF and NIH for providing partial support of this work. MWM acknowledges ARO, DARPA, NSF, and ONR for providing partial support of this work.

Appendix A. Proofs of Our Main Results

In this Appendix, we collect the proofs of our main results.

A.1 Notation and Technical Preliminaries

Let K_i represent the number of times the i^{th} observation is sampled. It is easy to see that $(K_1, \ldots, K_n)$ follows a multinomial distribution, $\text{Mult}(r, \{\pi_i\}_{i=1}^n)$, with sample size r as the total number of trials. Define $\mathbf{K} = \text{diag}\{K_i\}_{i=1}^n$, $\boldsymbol{\Omega} = \text{diag}\{1/r\pi_i\}_{i=1}^n$, and $\mathbf{W} = \boldsymbol{\Omega}\mathbf{K}$. For the i^{th} diagonal element of matrix $\mathbf{W}$, denoted as W_i, we have

$$\text{E}(W_i) = 1, \quad \text{Var}(W_i) = \frac{(1-\pi_i)}{r\pi_i}, \quad \text{Cov}(W_i, W_j) = -\frac{1}{r}, \quad i \neq j, \quad i, j = 1, \ldots, n. \tag{28}$$

Simple algebra yields that the sampling estimator of Eqn. (3) can be written as

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^{*T}\boldsymbol{\Phi}^{*2}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\boldsymbol{\Phi}^{*2}\mathbf{Y}^* = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}. \tag{29}$$

O_p Notation. *The O_p notation is the stochastic counterpart of the nonstochastic big-O notation, i.e., it describes the limiting behavior of (or the order of) a sequence of random variables, rather than that of sequence of fixed or deterministic variables.*

For a sequence of random variables, $\{A_n\}$, and a sequence of constants, $\{a_n\}$, the notation $A_n = O_p(a_n)$, means that $\{A_n/a_n\}$ is stochastically bounded (or bounded in probability). That is, for any $\tau > 0$, there exit a constant $K(\tau)$ and an integer $n(\tau)$ such that if $n \geq n(\tau)$, then

$$P(|A_n/a_n| \leq K(\tau)) \geq 1 - \tau.$$

More details and examples of this can be found in Section 14.4 of Bishop et al. (1975) and Section 1.2 of Serfling (2001).

Remark. If $Var(A_n) = O(n^{2\delta})$ and $E(A_n) = 0$, where δ is a real number, then we have that $\{A_n/n^{\delta}\}$ is bounded in probability by Chebyshev's inequality. We write $A_n = O_p(n^{\delta})$.

Remark. Throughout this paper, for a sequence of matrices $\mathbf{A}_n$, we write $\mathbf{A}_n = O(n^{\delta})$ to denote that $\|\mathbf{A}_n\|_{\infty} = O(n^{\delta})$, where $\|\mathbf{A}_n\|_{\infty} = \max_i \sum_j |\mathbf{A}_n[i,j]|$, and $\mathbf{A}_n[i,j]$ is the $(i,j)^{th}$ entry of $\mathbf{A}_n$. For a sequence of matrices $\mathbf{A}_n$, we write $\mathbf{A}_n = O_p(n^{\delta})$ to denote that $\|\mathbf{A}_n\|_{\infty} = O_p(n^{\delta})$.

Remark. The notation $A_n = \Omega(a_n)$ means that for some constant c and n_0, $A_n \geq ca_n$ for all $n \geq n_0$. The notation $A_n = \Theta(a_n)$ means that $A_n = O(a_n)$ and $A_n = \Omega(a_n)$.

Other than in the statement and proof of Theorem 2, we assume that the dimension p is fixed in all lemmas and theorems. The Cramer-Wold Device and Lemma 1 below govern the proofs for Theorem 1, Theorem 2, and Theorem 3.

Cramer-Wold Device. *For random vectors $\mathbf{Z}_n = (Z_{n1}, \ldots, Z_{np})^T$ and $\mathbf{Z} = (Z_1, \ldots, Z_p)^T$, a necessary and sufficient condition for $\mathbf{Z}_n \xrightarrow{d} \mathbf{Z}$ is that $\boldsymbol{b}^T\mathbf{Z}_n \xrightarrow{d} \boldsymbol{b}^T\mathbf{Z}$ as $n \to \infty$, for each $\boldsymbol{b} \in \mathbb{R}^p$.*

Remark. To derive the asymptotic distribution for the sampling estimator $\tilde{\boldsymbol{\beta}}$ in (29), which is a vector of random variables, we use the Cramer-Wold device to reduce the derivation of the asymptotic distribution for *vectors* to the usual *scalar* case. For more details about the Cramer-Wold device, see Section 29 of Billingsley (1995).

Convergence of Geometric Series of Matrices. *Let $\mathbf{A}$ be an $n \times n$ square matrix. We use $\rho(\mathbf{A})$ to denote the spectral radius of matrix $\mathbf{A}$, i.e., $\rho(\mathbf{A}) = \max\{|\lambda_1|, \ldots, |\lambda_n|\}$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of matrix $\mathbf{A}$. If $\rho(\mathbf{A}) < 1$, then $(\mathbf{I} - \mathbf{A})$ is invertible, and the series*

$$\mathbf{S} \;=\; \mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \ldots$$

converges to $(\mathbf{I} - \mathbf{A})^{-1}$.

Remark. The convergence of geometric series of matrices will be used in the proof of Lemma 1 below. For more details and a proof of this result, see Section 1.5 of Hubbard and Hubbard (1999).

Uniform Equicontinuity. The family of functions $\{f_n\}$ defined on $[a,b]$ is said uniformly equicontinuous if and only if for any $\epsilon > 0$, there exits a constant $\tau > 0$, such that for any $n \in N$, and any $s, t \in [a,b], |s-t| < \tau$,then

$$|f_n(t) - f_n(s)| < \epsilon.$$

Remark. We use uniform equicontinuity in the proof of Lemma 3.

Lemma 1 *Assume that* $0 < \pi_i < 1$*, for* $i = 1, \ldots, n$*. If*

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X} \;=\; O_p\left(\frac{1}{r^{\frac{\delta}{2}}}\right), \tag{30}$$

where δ *is a positive constant, then the weighted sample estimator in (29) can be written as*

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{e} + O_p(1/r^{\delta}), \tag{31}$$

where $\boldsymbol{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$*.*

Proof

Eqn (30) implies $\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X}\|_\infty = O_p\left(r^{-\frac{\delta}{2}}\right)$. Noting that $||\cdot||_\infty$ is sub-multiplicative (Theorem 1.3.8.6 in van de Geijn and Myers (2019)), we have

$$((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X})^j \;=\; O_p\left(r^{-\frac{j}{2}\delta}\right), \quad j \geq 1. \tag{32}$$

Therefore, by the convergence of geometric series of matrices,

$$[\mathbf{I} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X}]^{-1} = \mathbf{I} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X} + O_p\left(\frac{1}{r^{\delta}}\right). \tag{33}$$

Once again, by the submultiplicaive property of the matrix norm, $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X}$, $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{Y}$, and $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\boldsymbol{e}$ are of the same order, since the variances of $\mathbf{Y}$ and $\boldsymbol{e}$ are both bounded. It follows that

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{W}\mathbf{Y}) \\
&= [\mathbf{I} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X}]^{-1}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{W}\mathbf{Y}) \\
&= [\mathbf{I} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X} + O_p(1/r^{\delta})](\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y} + \mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{Y}),
\end{aligned} \tag{34}$$

where the expansion in (34) is by the convergence of geometric series of matrices and the assumption that $\delta > 0$. Note that $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X}$ and $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{Y}$ are of the same order $O_p(1/r^{\delta})$ since the variance of $\mathbf{Y}$ is bounded. Therefore, $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{Y}$ is of the order $O(1/r^{2\delta})$. It follows that

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} &= [\mathbf{I} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X} + O_p(1/r^{\delta})](\hat{\boldsymbol{\beta}}_{OLS} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{Y}) \\
&= \hat{\boldsymbol{\beta}}_{OLS} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\boldsymbol{e} + O_p(1/r^{\delta}) \\
&= \hat{\boldsymbol{\beta}}_{OLS} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{e} + O_p(1/r^{\delta}),
\end{aligned} \tag{35}$$

where the equality in (35) holds since $\mathbf{X}^T\boldsymbol{e} = 0$. This completes the proof.

Remark. Lemma 1 relates the sampling estimator $\tilde{\boldsymbol{\beta}}$ to the quantity $\hat{\boldsymbol{\beta}}_{OLS}$, with an order constraint on the residual term, i.e., $O_p(1/r^{\delta/2})$. In the application of Lemma 1 to the proof of Theorem 1 (asymptotic normality of $\tilde{\boldsymbol{\beta}}$ in estimating $\boldsymbol{\beta}_0$), we subtract $\boldsymbol{\beta}_0$ from

both sides of (35) to relate $\tilde{\boldsymbol{\beta}}$ to $\boldsymbol{\beta}_0$. In the proof of Theorem 3, Lemma 1 is directly applied (asymptotic normality of $\tilde{\boldsymbol{\beta}}$ in approximating $\hat{\boldsymbol{\beta}}_{OLS}$).

Remark. The assumption that $\delta > 0$ implies that $\rho((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X}) \to 0$ as $r \to \infty$. By the convergence of geometric series of matrices, the inverse of $[\mathbf{I}+(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X}] = \mathbf{X}^T\mathbf{W}\mathbf{X}$ exists, and the expansion in (34) is valid asymptotically. In the proof of Theorem 1, Theorem 2, and Theorem 3, we will verify the condition in Lemma 1, i.e., that $\delta > 0$. The exact magnitude of δ depends on $(\mathbf{W}-\mathbf{I})$, and it is different in Theorem 1 and Theorem 3.

In Appendix 4 and Appendix 4, we present the proofs of Theorem 1 and Theorem 3, respectively. The proof of Theorem 1 is much more complicated than that of Theorem 3. In conditional inference of Theorem 3, the data are given and the only randomness comes from sampling. However, in unconditional inference of Theorem 1, we consider both unobserved hypothetical data sampled from the underlying population as well as the sample sampled from observations. Thus, one more layer of randomness needs to take into account.

A.2 Proof of Theorem 1

We start by establishing several preliminary technical lemmas, and then we will present the main proof of Theorem 1.

A.2.1 Preliminary Material for the Proof of Theorem 1

To facilitate the proof of Theorem 1, we first present the Hajek-Sidak central limit theorem (CLT), as well as Lemma 2 and Lemma 3, as follows.

Theorem 4 (Hajek-Sidak CLT) *Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables such that $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ are both finite. Define $T_n = d_1X_1 + \ldots + d_nX_n$, then*

$$\frac{T_n - \mu\sum_{i=1}^n d_i}{\sigma\sqrt{\sum_{i=1}^n d_i^2}} \xrightarrow{d} N(0,1)\,, \tag{36}$$

whenever the Noether condition,

$$\frac{\max_{1\le i\le n} d_i^2}{\sum_{i=1}^n d_i^2} \to 0, \quad \text{as } n \to \infty, \tag{37}$$

is satisfied.

Remark. The Hajek-Sidak CLT is used in the proof of Lemma 2.

Lemma 2 *Define $\mathbf{U} = diag(U_1, \ldots, U_n)$, where for $i = 1, \ldots, n$ the independent random variables $U_i \sim Poisson(r\pi_i)$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$, where ε_is are i.i.d. with mean 0 and variance σ^2. If conditions (A1) and (A2) in Theorem 1 hold, then as $n \to \infty$,*

$$(\sigma^2\boldsymbol{\Sigma}_0)^{-\frac{1}{2}}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{U}\boldsymbol{\varepsilon} \quad \xrightarrow{d} \quad \boldsymbol{N}(\mathbf{0}, \mathbf{I}_p), \tag{38}$$

where $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Omega}$ are defined in Theorem 1.

Proof

We derive the asymptotic normality of the random vector $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Omega}\mathbf{U}\boldsymbol{\varepsilon}$ using the Cramer-Wold device. For any nonzero constant vector $\boldsymbol{b} \in \mathbb{R}^p$, we write

$$\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Omega}\mathbf{U}\boldsymbol{\varepsilon} = \sum_{i=1}^{n} d_i\zeta_i, \tag{39}$$

where $d_i = \boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\frac{\sqrt{r\pi_i+r^2\pi_i^2}}{r\pi_i}$ and $\zeta_i = U_i\varepsilon_i/\sqrt{r\pi_i + r^2\pi_i^2}$, $\mathrm{E}(\zeta_i) = 0$, and where $\mathrm{Var}(\zeta_i) = \sigma^2$.

Since Eqn. (39) is a weighted average of independent random variables ζ_i, it suffices to verify the Noether condition (37) of the Hajek-Sidak CLT to show the asymptotic normality of $\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Omega}\mathbf{U}\boldsymbol{\varepsilon}$. For d_i^2, we have

$$d_i^2 \leq \left(1 + \frac{1}{r\pi_{min}}\right)(\boldsymbol{a}^T\mathbf{x}_i)^2 \leq \left(1 + \frac{1}{r\pi_{min}}\right)\boldsymbol{a}^T\boldsymbol{a}M_x, \tag{40}$$

where $\boldsymbol{a} = (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{b}$, $M_x = \max\{\mathbf{x}_i^T\mathbf{x}_i\}_{i=1}^n$, and the last inequality is derived using the Cauchy-Schwarz inequality. Thus, $\max_{1\leq i\leq n} d_i^2 \leq (1+\frac{1}{r\pi_{min}})\boldsymbol{a}^T\boldsymbol{a}M_x$. For $\sum_{i=1}^n d_i^2$, we have

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}(1 + \frac{1}{r\pi_i})\boldsymbol{a}^T\mathbf{x}_i\boldsymbol{a}^T\mathbf{x}_i \geq (1 + \frac{1}{r\pi_{max}})\boldsymbol{a}^T\mathbf{X}^T\mathbf{X}\boldsymbol{a} \geq (n + \frac{n}{r\pi_{max}})\lambda_{min}\boldsymbol{a}^T\boldsymbol{a}, \tag{41}$$

where λ_{min} is the minimum eigenvalue of $\mathbf{X}^T\mathbf{X}/n$. Combining (40) and (41), we have

$$\lim_{n\to\infty}\frac{\max_{1\leq i\leq n} d_i^2}{\sum_{i=1}^n d_i^2} \leq \lim_{n\to\infty}\frac{(1+\frac{1}{r\pi_{min}})M_x}{(n+\frac{n}{r\pi_{max}})\lambda_{min}} \leq \frac{M_x}{\lambda_{min}}\lim_{n\to\infty}\frac{1+r\pi_{min}}{(nr\pi_{min}+\frac{n\pi_{min}}{\pi_{max}})} = 0, \tag{42}$$

where the last equality is obtained since condition (A2) implies $nr\pi_{min} \to \infty$ as $n \to \infty$. Since

$$\sum_{i=1}^{n}\mathrm{Var}(d_i\zeta_i) = \sigma^2\sum_{i=1}^{n}(\boldsymbol{a}^T\mathbf{x}_i)^2(1+\frac{1}{r\pi_i}) = \sigma^2\boldsymbol{a}^T\mathbf{X}^T(\mathbf{I}_p + \mathbf{\Omega})\mathbf{X}\boldsymbol{a},$$

then by Theorem 4, we have,

$$\begin{aligned}\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Omega}\mathbf{U}\boldsymbol{\varepsilon} &\xrightarrow{d} N(0, \sigma^2\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I}_p + \mathbf{\Omega})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{b}).\\ &= N(0, \sigma^2\boldsymbol{b}^T\mathbf{\Sigma}_0\boldsymbol{b})\end{aligned} \tag{43}$$

By applying the Cramer-Wold device, we have $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Omega}\mathbf{U}\boldsymbol{\varepsilon} \xrightarrow{d} N(0, \sigma^2\mathbf{\Sigma}_0)$. The proof is thus complete.

In the following statement and proof of Lemma 3, as well as in the proof of Theorem 1 below, we use $A|B$ to denote random variable A given random variable B.

Lemma 3 *Given any nonzero constant vector $\boldsymbol{b} \in \mathbb{R}^p$, as $n \to \infty$ we have*

$$(\sigma^2\boldsymbol{b}^T\mathbf{\Sigma}_0\boldsymbol{b})^{-\frac{1}{2}}\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Omega}\mathbf{U}\boldsymbol{\varepsilon}|\sum_{i=1}^{n}U_i = r \quad \xrightarrow{d} \quad N(0,1), \tag{44}$$

where $\mathbf{\Omega}$, $\mathbf{U}$, and $\mathbf{\Sigma}_0$ are defined in Lemma 2, and where r is the subsample size.

Proof

For $i = 1, \ldots, n$, we have

$$\text{Cov}(\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}U_i\varepsilon_i, \sum_{i=1}^{n} U_i) = \sum_{i=1}^{n} \boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\text{Cov}(U_i\varepsilon_i, U_i) = 0, \tag{45}$$

and thus we have

$$\text{Cov}(\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{U}\boldsymbol{\varepsilon}, \sum_{i=1}^{n} U_i) = 0.$$

By Eqn. (43), we have

$$(\sigma^2\boldsymbol{b}^T\boldsymbol{\Sigma}_0\boldsymbol{b})^{-\frac{1}{2}}\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{U}\boldsymbol{\varepsilon} \quad \xrightarrow{d} \quad N(0,1).$$

By the Lyapunov CLT (Section 27 in Billingsley (1995)), we have

$$\frac{1}{\sqrt{r}}(\sum_{i=1}^{n} U_i - r) \xrightarrow{d} N(0,1).$$

Combining this with the fact that U_i is independent of $\boldsymbol{\varepsilon}$, we have

$$\begin{pmatrix} (\sigma^2\boldsymbol{b}^T\boldsymbol{\Sigma}_0\boldsymbol{b})^{-\frac{1}{2}}\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{U}\boldsymbol{\varepsilon} \\ \frac{1}{\sqrt{r}}(\sum_{i=1}^{n} U_i - r) \end{pmatrix} \quad \xrightarrow{d} \quad \mathbf{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right). \tag{46}$$

Furthermore, with

$$(\sigma^2\boldsymbol{b}^T\boldsymbol{\Sigma}_0\boldsymbol{b})^{-\frac{1}{2}}\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{U}\boldsymbol{\varepsilon}|\sum_{i=1}^{n} U_i = r \quad \xrightarrow{d} \quad N(0,1) \tag{47}$$

provided, we can show the convergence of conditional distributions is the uniform equicontinuity of conditional characteristic functions (Steck, 1957), as we do below.

Here, for the ease of notation, we define $Q_n = \boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{U}\boldsymbol{\varepsilon}$, $L_n = \frac{1}{\sqrt{r}}(\sum_{i=1}^{n} U_i - r)$, and $s_n^2 = \sigma^2\boldsymbol{b}^T\boldsymbol{\Sigma}_0\boldsymbol{b}$. Let

$$\psi_n(t_n; t) = \text{E}\left(\exp\left(\text{i}t\frac{Q_n}{s_n} \mid \sum_{i=1}^{n} U_i = t_n\right)\right), \tag{48}$$

where i denotes the imaginary unit. Hence, we aim to show the uniform equicontinuity of $\psi_n(t_n; t)$. When $L_n = l_n$, $\sum_{i=1}^{n} U_i = r + \sqrt{r}l_n$; when $L_n = l_n + h$, $\sum_{i=1}^{n} U_i = r + \sqrt{r}l_n + \sqrt{r}h$. Note that

$$(Q_n | L_n = l_n + h) \overset{d}{=} \boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}(\mathbf{M} + \mathbf{R})\boldsymbol{\varepsilon},$$

where $\overset{d}{=}$ denotes two random variables have the same ditribution, $\mathbf{M} = \text{diag}\{M_i\}_{i=1}^{n}$, $(M_1, \ldots, M_n) \sim \text{Mult}(h\sqrt{r}, (\pi_1, \ldots, \pi_n))$, $\mathbf{R} = \text{diag}\{R_i\}_{i=1}^{n}$, and $(R_1, \ldots, R_n) \sim \text{Mult}(r +$

$\sqrt{r}l_n, (\pi_1, \ldots, \pi_n))$. Thus, we have that

$$\begin{aligned}
&|\psi_n(l_n + h; t) - \psi_n(l_n; t)| \\
&\quad = \left| \mathrm{E}\left(\exp\left(\mathrm{i}\frac{t}{s_n} \boldsymbol{b}^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}(\mathbf{M}+\mathbf{R})\boldsymbol{\varepsilon} \right) \right) - \mathrm{E}\left(\exp\left(\mathrm{i}\frac{t}{s_n} \boldsymbol{b}^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{R}\boldsymbol{\varepsilon} \right) \right) \right| \\
&\quad \leq \mathrm{E}\left(\left| \exp\left(\mathrm{i}\frac{t}{s_n} \boldsymbol{b}^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}(\mathbf{M}+\mathbf{R})\boldsymbol{\varepsilon} \right) - \exp\left(\mathrm{i}\frac{t}{s_n} \boldsymbol{b}^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{R}\boldsymbol{\varepsilon} \right) \right| \right) \quad (49) \\
&\quad \leq \frac{t}{s_n} \mathrm{E}\left(\left| \boldsymbol{b}^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}(\mathbf{M}+\mathbf{R})\boldsymbol{\varepsilon} - \boldsymbol{b}^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{R}\boldsymbol{\varepsilon} \right| \right) \quad (50) \\
&\quad = \frac{t}{s_n} \mathrm{E}\left(\left| \boldsymbol{b}^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{M}\boldsymbol{\varepsilon} \right| \right) \\
&\quad \to 0 \qquad \text{as } h \to 0,
\end{aligned}$$

where (49) is by Jensen's inequality, and (50) is by the fact that

$$|e^{\mathrm{i}a} - e^{\mathrm{i}b}| = \sqrt{2(1 - \cos(\frac{a-b}{2}))} = 2|\sin(\frac{a-b}{2})| \leq |a - b|,$$

for any a, b. Thus, the uniform equicontinuity of conditional characteristic function is verified, and the proof is complete.

Remark. The proof of Lemma 3 is a simplified version of the proof of Theorem 2.1 in Morris (1975).

Remark. The key difference between Lemma 2 and Lemma 3 is that we consider a conditional distribution in Lemma 3, whereas we consider an unconditional distribution in Lemma 2.

We will also need the following lemma, the proof of which can be found in Section 6.4 of Sheldon (2006).

Lemma 4 *If independent random variables $U_i \sim Poisson(\lambda_i)$, $i = 1, \ldots, n$, then*

$$(U_1, \ldots, U_n) | \sum_{i=1}^n U_i = r \sim Mult\left(r, \left\{ \frac{\lambda_i}{\sum_{i=1}^n \lambda_i} \right\}_{i=1}^n \right).$$

A.2.2 Main Part of the Proof of Theorem 1

We first verify that the condition in Lemma 1 holds. To do this, we derive the magnitude of δ in Eqn. (30). Note that

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W} - \mathbf{I})\mathbf{X} = (\mathbf{X}^T\mathbf{X}/n)^{-1}\mathbf{X}^T(\mathbf{W} - \mathbf{I})\mathbf{X}/n.$$

By Condition (A1), we have

$$\|(\mathbf{X}^T\mathbf{X}/n)^{-1}\|_\infty \leq \sqrt{p}\|(\mathbf{X}^T\mathbf{X}/n)^{-1}\|_2 \leq \sqrt{p}/\lambda_{min}. \quad (51)$$

Since the dimension p we considered in Theorem 1 is fixed, we further have $\|(\mathbf{X}^T\mathbf{X}/n)^{-1}\|_\infty = O(1)$. Thus, the order of $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W} - \mathbf{I})\mathbf{X}$ depends on that of $\mathbf{X}^T(\mathbf{W} - \mathbf{I})\mathbf{X}/n$. We

next derive the order of the $(s,t)^{th}$ element of $\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X}/n$, i.e., of $\frac{1}{n}\sum_{i=1}^n x_{is}x_{it}(W_i-1)$. To do so, we have

$$\mathrm{E}\left(\frac{\sum_{i=1}^n x_{is}x_{it}(W_i-1)}{n}\right) = \frac{\sum_{i=1}^n x_{is}x_{it}\mathrm{E}(W_i-1)}{n} = 0, \tag{52}$$

and

$$\begin{aligned}
\mathrm{Var}\left(\frac{\sum_{i=1}^n x_{is}x_{it}(W_i-1)}{n}\right) &= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^n x_{is}x_{it}(W_i-1)\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^n (x_{is}x_{it})^2\frac{1-\pi_i}{r\pi_i} - 2\sum_{i<j} x_{is}x_{it}x_{js}x_{jt}\frac{1}{r}\right) \\
&= \frac{1}{rn^2}\left[\sum_{i=1}^n (x_{is}x_{it})^2\frac{1-\pi_i}{\pi_i} - \left(\left(\sum_{i=1}^n x_{is}x_{it}\right)^2 - \sum_{i=1}^n (x_{is}x_{it})^2\right)\right] \\
&= \frac{1}{r}\left[\sum_{i=1}^n \frac{(x_{is}x_{it})^2}{n^2\pi_i} - \left(\sum_{i=1}^n \frac{x_{is}x_{it}}{n}\right)^2\right] \\
&= O\left(\frac{1}{rn^2}\sum_{i=1}^n \frac{1}{\pi_i}\right).
\end{aligned} \tag{53}$$

Combining the fact that $\sum_{i=1}^n \frac{1}{\pi_i} \le \frac{n}{\pi_{min}}$ and Condition (A2), we have that the order of $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X}$ is not larger than $O_p(n^{-\frac{2-\gamma_0-\alpha}{2}})$. Thus $0 < 2-\gamma_0-\alpha \le \delta$ in Eqn. (30), and we verify that the condition in Lemma 1 holds.

Subtracting $\boldsymbol{\beta}_0$ from both sides of Eqn. (31) in Lemma 1, we get

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{e} + \hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0 + O_p\left(\frac{1}{r^\delta}\right), \tag{54}$$

where $\boldsymbol{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$. Since $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0) = O(1/n)$, we have $\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0 = O_p(1/\sqrt{n})$. Thus, both $\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0$ and the residual term in the right hand side of (54) are negligible. Hence, the asymptotic distribution of $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ is equivalent to that of $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{e}$.

Thus, for the rest of the proof, we derive the asymptotic normality of $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{e}$. Note that

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{e} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\varepsilon} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\boldsymbol{e}-\boldsymbol{\varepsilon}), \tag{55}$$

where $\boldsymbol{\varepsilon}$ is the random noise in Model (1). We will show that the order of $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\boldsymbol{e}-\boldsymbol{\varepsilon})$ is bounded by calculating the variances of s^{th} element of $\mathbf{X}^T\mathbf{W}(\boldsymbol{e}-\boldsymbol{\varepsilon})/n$. We have

$$\begin{aligned}
&\mathrm{Var}\left(\frac{\sum_{i=1}^n x_{is}W_i(e_i-\varepsilon_i)}{n}\right) \\
&\quad= \frac{1}{n^2}\left(\sum_{i=1}^n x_{is}^2\mathrm{Var}(W_i(e_i-\varepsilon_i)) + 2\sum_{i<j} x_{is}x_{js}\mathrm{Cov}\left[W_i(e_i-\varepsilon_i), W_j(e_j-\varepsilon_j)\right]\right).
\end{aligned} \tag{56}$$

Now, we analyze the two terms on the right hand side of Eqn. (56). For the first term, we have that

$$\begin{aligned}\sum_{i=1}^{n}\mathrm{Var}(W_i(e_i-\varepsilon_i)) &= \sum_{i=1}^{n}\mathrm{E}(W_i^2(e_i-\varepsilon_i)^2)\\ &= \sum_{i=1}^{n}(\mathrm{Var}(W_i)\mathrm{Var}(e_i-\varepsilon_i)+(\mathrm{E}W_i)^2\mathrm{Var}(e_i-\varepsilon_i))\\ &= \sum_{i=1}^{n}(\frac{1-\pi_i}{r\pi_i}h_{ii}\sigma^2+h_{ii}\sigma^2)=O\left(\frac{1}{r\pi_{min}}\right), \end{aligned} \tag{57}$$

where the last equality holds since $\sum_{i=1}^{n}h_{ii}=p$.

For the second term, we have that

$$\begin{aligned}\sum_{i<j}\mathrm{Cov}\left(W_i(e_i-\varepsilon_i),W_j(e_j-\varepsilon_j)\right) &= \sum_{i<j}\mathrm{E}(W_iW_j(e_i-\varepsilon_i)(e_j-\varepsilon_j))\\ &= \sum_{i<j}\mathrm{E}(W_iW_j)\mathrm{E}((e_i-\varepsilon_i)(e_j-\varepsilon_j))\\ &= \sum_{i<j}(1-\frac{1}{r})h_{ij}\sigma^2=O\left(n\right), \end{aligned} \tag{58}$$

where the facts that $\mathrm{E}[(e_i-\epsilon_i)(e_j-\epsilon_j)]=h_{ij}\sigma^2$ and $\mathrm{E}(W_iW_j)=1-\frac{1}{r}$ are used in the third equality. Substituting (57) and (58) into (56), we have that

$$\mathrm{Var}\left(\frac{\sum_{i=1}^{n}x_{is}W_i(e_i-\varepsilon_i)}{n}\right)=O\left(\frac{1}{n}\right) \tag{59}$$

Combining (55) and (59), we aim to show that $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\boldsymbol{e}-\boldsymbol{\epsilon})$ is of higher order than $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\varepsilon}$. Thus, if we establish the asymptotic normality of $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\varepsilon}$, then the asymptotic normality of $\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0$ in Eqn. (54) will follow directly.

Note that $\mathbf{W}$ can be written as $\mathbf{W}=\boldsymbol{\Omega}\mathbf{K}$. By Lemma 4, it follows that $(K_1,\ldots,K_n)$ and $[(U_1,\ldots,U_n)|\sum_{i=1}^{n}U_i=r]$ are identically distributed. Hence,

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\varepsilon} \quad \text{and} \quad (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{U}\boldsymbol{\varepsilon}|\sum_{i=1}^{n}U_i$$

are identically distributed. Thus, Lemma 3 can be applied, and the asymptotic normality is obtained using the Cramer-Wold device.

Finally, combining Eqn. (54), Lemma 2, and Lemma 3, we have that

$$(\sigma^2\boldsymbol{\Sigma}_0)^{-\frac{1}{2}}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \quad \xrightarrow{d} \quad \mathbf{N}(\mathbf{0},\mathbf{I}_p), \quad \text{as } n\to\infty, \tag{60}$$

where $\boldsymbol{\Sigma}_0=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I}_p+\boldsymbol{\Omega})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$. This completes the proof.

A.3 Proof of Proposition 1

By Theorem 1, the asymptotic squared bias for $\tilde{\boldsymbol{\beta}}$ is 0. By the definition of AMSE in Eqn. (4), $AMSE(\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}_0) = tr(Avar(\tilde{\boldsymbol{\beta}}))$, i.e., the expression given in Eqn. (9). We consider minimizing $AMSE(\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}_0)$ as a function of $\{\pi_i\}_{i=1}^n$. It is straightforward to employ the method of Lagrange multipliers to find the minimizer of the right-hand side of Eqn. (9), subject to the constraint $\sum_{i=1}^n \pi_i = 1$. If we do this, then we let

$$L(\pi_1, \ldots, \pi_n) = tr(Avar(\tilde{\boldsymbol{\beta}})) + \lambda(\sum_{i=1}^n \pi_i - 1).$$

Then, we can solve $\partial L / \partial \pi_i = 0$, $i = 1, \ldots, n$, for the optimal sampling probabilities.

The proofs of Propositions 2–6 all follow in a manner similar to that of Proposition 1, and thus they will be omitted.

A.4 Proof of Theorem 2

We first verify the condition that $\delta > 0$ in Lemma 1; and the rest of the proof of Theorem 2, in which we allow the number of predictors p to diverge, is readily derived from that of Theorem 1.

By (51), we have $\|(\mathbf{X}^T\mathbf{X}/n)^{-1}\|_\infty = O(\sqrt{p})$. We next derive the order of $\|\mathbf{X}^T(\mathbf{W} - \mathbf{I})\mathbf{X}/n\|_\infty$. To do so, we derive the magnitude of the absolute sum of the s^{th} row of $\mathbf{X}^T(\mathbf{W} - \mathbf{I})\mathbf{X}/n$, i.e., of $\frac{1}{n}\sum_{t=1}^p |\sum_{i=1}^n x_{is}x_{it}(W_i - 1)|$. We have

$$\begin{aligned}
\mathrm{E}\left(\frac{\sum_{t=1}^p |\sum_{i=1}^n x_{is}x_{it}(W_i - 1)|}{n}\right)^2 &\leq p\sum_{t=1}^p \mathrm{E}\left(\frac{\sum_{i=1}^n x_{is}x_{it}(W_i - 1)}{n}\right)^2 \qquad (61)\\
&\leq \frac{p}{r}\sum_{t=1}^p \left[\sum_{i=1}^n \frac{(x_{is}x_{it})^2}{n^2\pi_i} - \left(\sum_{i=1}^n \frac{x_{is}x_{it}}{n}\right)^2\right]\\
&\leq \frac{p}{r\pi_{min}}\sum_{i=1}^n \frac{x_{is}^2}{n}\left[\sum_{t=1}^p \frac{x_{it}^2}{n}\right]\\
&= O\left(\frac{p^2}{rn\pi_{min}}\right), \qquad (62)
\end{aligned}$$

where (61) is by Cauchy-Schwarz inequality and Eqn. (62) is by the results of Eqn. (53) and Condition (B1). Notice the above equations holds for any row of $\mathbf{X}^T(\mathbf{W} - \mathbf{I})\mathbf{X}/n$, whose bound is $O_p(p/n^{\frac{2-\gamma_0-\alpha}{2}})$ by Markov's inequality. Therefore, by the submultiplicative property of the matrix norm, we have that the order of $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W} - \mathbf{I})\mathbf{X}$ is no larger than $O_p(p^{\frac{3}{2}}n^{\frac{-2+\gamma_0+\alpha}{2}})$. Recall that $p = O(n^{1-\kappa})$. Under Condition (B1), we have $3(1-\kappa) - 2 + \gamma_0 + \alpha < 0$. We thus can find some δ such that $0 < \delta \leq 3\kappa - \gamma_0 - \alpha_0 - 1$ and verify that the assumption in Lemma 1 holds.

By combining Eqns. (54) and (55), it follows that

$$\boldsymbol{a}^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \boldsymbol{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\varepsilon} + \boldsymbol{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\boldsymbol{e} - \boldsymbol{\varepsilon}) + \boldsymbol{a}^T(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0) + O_p\left(\frac{1}{r^\delta}\right).$$

By results in (Huber, 1973; Yohai and Maronna, 1979; Portnoy, 1984, 1985), we note that $\|\boldsymbol{a}\|^2 = 1$, and $\boldsymbol{a}^T(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0) = O_p(1/\sqrt{n})$, which is of the highest order. Further, by a similar argument in Theorem 1 (from (55) to (59)) we have that $\boldsymbol{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\boldsymbol{e}-\epsilon)$ is of higher order than $\boldsymbol{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\varepsilon}$. To prove Theorem 2, it suffices to establish the asymptotic normality of $\boldsymbol{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\varepsilon}$. This follows from applying Condition (B2) to Lemma 2 and by noting that $M_x = n\max_i \|\mathbf{x}_i\|^2 = O(p)$ in (42).

A.5 Proof of Theorem 3

Given the data $\{\mathbf{X}, \mathbf{Y}\}$, we first determine the value of δ in Eqn. (30) in order to use Lemma 1. Since $\|\mathbf{x}_i\| < \infty$, where $\mathbf{x}_i$ is the i^{th} row of $\mathbf{X}$, each element of $\mathbf{X}^T\mathbf{X}$ is a fixed matrix and is finite in norm. Since the $(s,t)^{th}$ element of $\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X}$ is equal to $\sum_{i=1}^n x_{is}x_{it}(W_i - 1)$, it follows that

$$\text{Var}\left(\sum_{i=1}^n x_{is}x_{it}(W_i-1)\right) = \frac{1}{r}\left(\sum_{i=1}^n (x_{is}x_{it})^2\frac{1-\pi_i}{\pi_i} - 2\sum_{i<j} x_{is}x_{it}x_{js}x_{jt}\right) = O_p\left(\frac{1}{r}\right), \quad (63)$$

i.e., $\delta = 1$ in Eqn. (30).

Next, note that $\mathbf{K}$ can be written as $\mathbf{K} = \sum_{j=1}^r \mathbf{K}^{(j)}$, where $\mathbf{K}^{(j)} = \text{Diag}\{K_i^{(j)}\}_{i=1}^n$, and where $(K_1^{(j)}, \ldots, K_n^{(j)}) \overset{iid}{\sim} \text{Mult}(1, \{\pi_i\})_{i=1}^n$, for $j = 1, \ldots, n$. Combining Eqn. (31) in Lemma 1 and Eqn. (63), we can show that

$$\begin{aligned}\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{OLS} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{e} + O_p(1/r) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\sum_{j=1}^r \mathbf{X}^T\boldsymbol{\Omega}\mathbf{K}^{(j)}\boldsymbol{e} + O_p(1/r).\end{aligned}$$

Given this, we can use the Cramer-Wold device to establish the asymptotic normality of

$$(\mathbf{X}^T\mathbf{X})^{-1}\sum_{j=1}^r \mathbf{X}^T\boldsymbol{\Omega}\mathbf{K}^{(j)}\boldsymbol{e}.$$

To do this, for any constant vector $\boldsymbol{b} \in \mathbb{R}^p$ such that $\boldsymbol{b} \neq \mathbf{0}$, we will consider the quantity $\sum_{j=1}^r \boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{K}^{(j)}\boldsymbol{e}$. This is a summation of r independent random variables. Since the elements in $\boldsymbol{X}$ and $\boldsymbol{e}$ are fixed numbers, finite in norm, and $\pi_i > 0$, the Noether condition in Hajek-Sidek CLT is satisfied.

Without loss of generality, we have

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{K}^{(1)}\boldsymbol{e}) &= \mathrm{Var}\left(\sum_{i=1}^{n}\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\frac{1}{r\pi_i}K_i^{(1)}e_i\right) \\
&= \sum_{i=1}^{n}(\boldsymbol{a}^T\mathbf{x}_ie_i\frac{1-\pi_i}{r\pi_i}e_i\mathbf{x}_i^T\boldsymbol{a}) - 2\sum_{i\le j}\boldsymbol{a}^T\mathbf{x}_ie_i\frac{1}{r}e_j\mathbf{x}_j^T\boldsymbol{a} \\
&= \frac{1}{r}\boldsymbol{a}^T\left(\sum_{i=1}^{n}\frac{e_i^2}{\pi_i}\mathbf{x}_i\mathbf{x}_i^T\right)\boldsymbol{a} - \frac{1}{r}\boldsymbol{a}^T\left(\sum_{i=1}^{n}\mathbf{x}_ie_i^2\mathbf{x}_i^T + 2\sum_{i<j}\mathbf{x}_ie_ie_j\mathbf{x}_j^T\right)\boldsymbol{a} \\
&= \frac{1}{r}\boldsymbol{a}^T\left(\sum_{i=1}^{n}\frac{e_i^2}{\pi_i}\mathbf{x}_i\mathbf{x}_i^T\right)\boldsymbol{a} - \frac{1}{r}\boldsymbol{a}^T\mathbf{X}^T\boldsymbol{e}\boldsymbol{e}^T\mathbf{X}\boldsymbol{a} \\
&= \frac{1}{r}\boldsymbol{a}^T\left(\sum_{i=1}^{n}\frac{e_i^2}{\pi_i}\mathbf{x}_i\mathbf{x}_i^T\right)\boldsymbol{a}, \qquad (64)
\end{aligned}$$

where $\boldsymbol{a} = (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{b}$, and where Eqn. (64) follows since $\mathbf{X}^T\boldsymbol{e} = \mathbf{0}$. By the Lindeberg-Lévy CLT, we have that

$$\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\sum_{j=1}^{r}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{K}^{(j)}\boldsymbol{e} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{b}^T\boldsymbol{\Sigma}_c\boldsymbol{b}),$$

where $\boldsymbol{\Sigma}_c = (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\Sigma}_e(\mathbf{X}^T\mathbf{X})^{-1}$ and $\boldsymbol{\Sigma}_e = \frac{1}{r}\sum_{i=1}^{n}\frac{e_i^2}{\pi_i}\mathbf{x}_i\mathbf{x}_i^T$. Thus, by the Cramer-Wold device, Theorem 3 follows.

Appendix B. Approximation Analysis

In this Appendix, we consider several related algorithmic questions. First, what is the effect (on MSE/AMSE/EAMSE) of computing approximately-optimal sampling probabilities? Second, can approximate sampling probabilities be computed more quickly than computing them exactly (which often takes time of the same order as solving the original problem exactly)? We will show that approximately optimal sampling probabilities incur negligible error, relative to exactly optimal sampling probabilities; and we show that approximate sampling probabilities can be computed quickly, either using the main algorithm of Drineas et al. (2012a), or using a variant of the main algorithm of Drineas et al. (2012a).

B.1 Approximation of RL estimator and its relative-error for $AMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}, \mathbf{X}\boldsymbol{\beta}_0)$

Recall that the RL sampling estimator with sampling probabilities $\pi_i = \sqrt{h_{ii}}/\sum_j\sqrt{h_{jj}}$ has the smallest $AMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}, \mathbf{X}\boldsymbol{\beta}_0)$. These depend on the leverage scores, which can be expensive to compute exactly. We can use Theorem 1 in Drineas et al. (2012a) to approximate these sampling probabilities. Here is Theorem 1 of Drineas et al. (2012a).

Theorem B.1 *Let* $\mathbf{X}$ *be a full-rank* $n \times p$ *matrix, with* $n \gg p$*; let* $\epsilon \in (0, 1/2]$ *be an error parameter; and define the statistical leverage scores* $l_i = h_{ii}$*. Then, using Algorithm 1 in*

Drineas et al. (2012a), which returns values $\tilde{l}_i$, for all $i \in \{i, ..., n\}$, we have that with probability at least 0.8,

$$|l_i - \tilde{l}_i| \leq \epsilon l_i$$

holds for all $i \in \{1, ..., n\}$. Assuming $p \leq n \leq e^p$, the running time of the algorithm is

$$O(np \ln (p\epsilon^{-1}) + np\epsilon^{-2} \ln n + p^3\epsilon^{-2}(\ln n)(\ln p\epsilon^{-1})).$$

We now examine the impact of using the approximated sampling probabilities on the resulting AMSE. Let C denote the value of the smallest $AMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}, \mathbf{X}\boldsymbol{\beta}_0)$ and $\tilde{C}$ denote the value of $AMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}, \mathbf{X}\boldsymbol{\beta}_0)$ using the approximately-optimal leverage scores from Theorem B.1. The following theorem states that $\tilde{C}$ is a relative-error approximation of C.

Theorem B.2 *Assume the conditions in Theorem B.1 hold. Then, using the approximations from Algorithm 1 in Drineas et al. (2012b), we have*

$$C \leq \tilde{C} \leq \sqrt{\frac{1+\epsilon}{1-\epsilon}} C$$

with probability at least 0.8.

Proof (of Theorem B.2)

By Proposition 2 in our main text, we get the smallest value of $AMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}, \mathbf{X}\boldsymbol{\beta_0})$, denoted by C, by setting the sampling probabilities $\pi_i = \frac{\sqrt{h_{ii}}}{\sum_{j=1}^n \sqrt{h_{jj}}} = \frac{\sqrt{l_i}}{\sum_{j=1}^n \sqrt{l_j}}$, $i = 1, \ldots, n$,

$$\begin{aligned} C &= \min_{\pi_1, \cdots, \pi_n} AMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}, \mathbf{X}\boldsymbol{\beta_0}) \\ &= \min_{\pi_1, \cdots, \pi_n} (p\sigma^2 + \frac{1}{r}\sum_{i=1}^n \frac{\sigma^2}{\pi_i} ||\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i||^2) \\ &= p\sigma^2 + \frac{\sigma^2}{r}\sum_{i=1}^n \frac{\sum_{j=1}^n \sqrt{l_j}}{\sqrt{l_i}} ||\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i||^2 \end{aligned}$$

Instead of exact sampling probabilities, we calculate approximated sampling probabilities $\tilde{\pi}_i = \frac{\sqrt{\tilde{l}_i}}{\sum_{j=1}^n \sqrt{\tilde{l}_j}}$, $i = 1, \ldots, n$. Now the corresponding value of $AMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}, \mathbf{X}\boldsymbol{\beta_0})$, denoted by $\tilde{C}$, is

$$\begin{aligned} \tilde{C} &= p\sigma^2 + \frac{\sigma^2}{r}\sum_{i=1}^n \frac{\sum_{j=1}^n \sqrt{\tilde{l}_j}}{\sqrt{\tilde{l}_i}} ||\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i||^2 \\ &= p\sigma^2 + \frac{\sigma^2}{r}\sum_{i=1}^n \frac{\sum_{j=1}^n \sqrt{\tilde{l}_j}}{\sqrt{\tilde{l}_i}} l_i, \end{aligned}$$

where the last equality holds since we have $||\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i||^2 = l_i$. Since C is the smallest value of $AMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}, \mathbf{X}\boldsymbol{\beta_0})$ for all possible π_is, we have $C \leq \tilde{C}$.

By Theorem B.1, with probability at least 0.8, we have

$$\frac{\sum_{j=1}^{n}\sqrt{\tilde{l}_j}}{\sqrt{\tilde{l}_i}} \leq \frac{\sum_{j=1}^{n}\sqrt{(1+\epsilon)l_j}}{\sqrt{(1-\epsilon)l_i}}.$$

Combining all these facts, we have

$$\begin{aligned} C \leq \tilde{C} &\leq p\sigma^2 + \sqrt{\frac{1+\epsilon}{1-\epsilon}}\frac{\sigma^2}{r}(\sum_{i=1}^{n}\sqrt{l_i})^2 \\ &\leq \sqrt{\frac{1+\epsilon}{1-\epsilon}}C. \end{aligned}$$

The proof is thus completed.

B.2 Approximation of IC estimator and its relative-error for $AMSE(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)$

Recall that the IC sampling estimator with sampling probabilities

$$\pi_i = \|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\| / \sum_j \|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_j\|$$

has minimal $AMSE(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)$. These are related to but different than leverage scores. However, we can approximate these IC sampling probabilities quickly. Our algorithm to approximate $s_i = \|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|^2$ is a modified version of Algorithm 1 in Drineas et al. (2012a).

Algorithm 1 Modified Algorithm 1 in Drineas et al. (2012b)

Input: $\mathbf{X} \in \mathbb{R}^{n\times p}$ (with SVD $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$), error parameter $\epsilon \in (0, 1/2]$.
Output: $\tilde{s}_i, \quad i = 1, \cdots, n$.

- 1. Let $\boldsymbol{\Pi}_1 \in \mathbb{R}^{r_1\times n}$ be an $\epsilon - FJLT$ for $\mathbf{U}$, using Lemma 3 in Drineas et al. (2012b) with $r_1 = \Omega(\frac{q\ln n}{\epsilon^2}\ln\frac{d\ln n}{\epsilon^2})$.
- 2. Compute $\boldsymbol{\Pi}_1\mathbf{X} \in \mathbb{R}^{r_1\times p}$ and its SVD, $\boldsymbol{\Pi}_1\mathbf{X} = \mathbf{U}_{\boldsymbol{\Pi}_1\mathbf{X}}\boldsymbol{\Sigma}_{\boldsymbol{\Pi}_1\mathbf{X}}\mathbf{V}_{\boldsymbol{\Pi}_1\mathbf{X}}^T$. Let $\tilde{\mathbf{R}}^{-1} = (\boldsymbol{\Pi}_1\mathbf{X})^\dagger((\boldsymbol{\Pi}_1\mathbf{X})^\dagger)^T$. (Here, $(\boldsymbol{\Pi}_1\mathbf{X})^\dagger = \mathbf{V}_{\boldsymbol{\Pi}_1\mathbf{X}}(\boldsymbol{\Sigma}_{\boldsymbol{\Pi}_1\mathbf{X}})^{-1}\mathbf{U}_{\boldsymbol{\Pi}_1\mathbf{X}}^T$ is the Moore-Penrose pseudoinverse of $\boldsymbol{\Pi}_1\mathbf{X}$.)
- 3. View the normalized rows of $\mathbf{X}\tilde{\mathbf{R}}^{-1} \in \mathbb{R}^{n\times p}$ as n vectors in $\mathbb{R}^p$, and construct $\boldsymbol{\Pi}_2 \in \mathbb{R}^{p\times r_2}$ to be an $\epsilon - JLT$ for n^2 vectors, using Lemma 1 in Drineas et al. (2012b) with $r_2 = O(\epsilon^{-2}\ln n)$
- 4. Construct the matrix product $\boldsymbol{\Omega} = \mathbf{X}\tilde{\mathbf{R}}^{-1}\boldsymbol{\Pi}_2$.
- 5. For all $i = 1, \cdots, n$ compute and return $\tilde{s}_i = \|\boldsymbol{\Omega}_{(i)}\|_2^2$.

Remark. The major difference between this algorithm and Algorithm 1 in Drineas et al. (2012a) is step 2; while the running times of the two algorithms are the same, i.e., the IC scores can be computed more quickly than solving the original problem.

The following theorem provides our main quality-of-approximation and running time result for Algorithm 1.

Theorem B.3 *Let* $\mathbf{X}$ *be a full-rank* $n \times p$ *matrix, with* $n \gg p$*; let* $\epsilon \in (0, 1/2]$ *be an error parameter. Then, using Algorithm 1, we have that with probability at least 0.8,*

$$|s_i - \tilde{s}_i| \leq \epsilon s_i$$

holds for all $i \in \{1, ..., n\}$*. Assuming* $p \leq n \leq e^p$*, the running time of the algorithm is*

$$O(np \ln (p\epsilon^{-1}) + np\epsilon^{-2} \ln n + p^3\epsilon^{-2}(\ln n)(\ln p\epsilon^{-1})).$$

To prove Theorem B.3, we introduce some notation. We define

$$u_i = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i, \quad \hat{u}_i = (\tilde{\mathbf{R}}^{-1})^T\mathbf{X}^T e_i, \quad \text{and} \quad \tilde{u}_i = \mathbf{\Pi}_2^T(\tilde{\mathbf{R}}^{-1})^T\mathbf{X}^T\mathbf{e}_i.$$

Then, $s_i = \|u_i\|_2^2$, $\hat{s}_i = \|\hat{u}_i\|_2^2$, and $\tilde{s}_i = \|\tilde{u}_i\|_2^2$. The proof of Theorem B.3 will follow from the following two lemmas.

Lemma B.4 *We assume that the same conditions of Theorem B.3 hold. Then, using Algorithm 1, we have that with probability at least 0.8,*

$$|s_i - \hat{s}_i| \leq 2\left(\frac{B}{b}\right)^2\left(\frac{\epsilon}{1-\epsilon}\right)s_i,$$

holds for all $i \in \{1, ..., n\}$*, where* B *and* b *are the upper and lower bounds of eigenvalues of the of matrix* $\mathbf{X}^T\mathbf{X}/n$*, respectively as stated in Condition (A1) in Theorem 1.*

Lemma B.5 *We assume that the same conditions of Theorem B.3 hold. Then, using Algorithm 1, we have that with probability at least 0.8,*

$$|\hat{s}_i - \tilde{s}_i| \leq 2\epsilon\hat{s}_i.$$

holds for all $i \in \{1, ..., n\}$*.*

The proof of Lemma B.5 can be found in Section 4.2 Drineas et al. (2012a).

Proof (of Lemma B.4)

Combining the SVD of $\mathbf{X}$, i.e., $\mathbf{X} = \mathbf{U\Sigma V}^T$, and Eqn. (10) in Lemma 2 in Drineas et al. (2012a), we have

$$\begin{aligned}\hat{s}_i &= \mathbf{e}_i^T\mathbf{X}(\tilde{\mathbf{R}}^{-1})(\tilde{\mathbf{R}}^{-1})^T\mathbf{X}^T\mathbf{e}_i \\ &= \mathbf{e}_i^T\mathbf{U}(\mathbf{\Pi}_1\mathbf{U})^\dagger(\mathbf{\Pi}_1\mathbf{U})^{\dagger T}\mathbf{\Sigma}^{-2}(\mathbf{\Pi}_1\mathbf{U})^\dagger(\mathbf{\Pi}_1\mathbf{U})^{\dagger T}\mathbf{U}^T\mathbf{e}_i.\end{aligned}$$

Let the SVD of $\mathbf{\Psi}$ be $\mathbf{\Psi} = \mathbf{U}_\mathbf{\Psi}\mathbf{\Sigma}_\mathbf{\Psi}\mathbf{V}_\mathbf{\Psi}^T$, where $\mathbf{V}_\mathbf{\Psi}$ is a full rotation in p dimensions. Then, $\mathbf{\Psi}^\dagger\mathbf{\Psi}^{\dagger T} = \mathbf{V}_\mathbf{\Psi}\mathbf{\Sigma}_\mathbf{\Psi}^{-2}\mathbf{V}_\mathbf{\Psi}^T$. We denote $\mathbf{I}_d - \mathbf{V}_\mathbf{\Psi}\mathbf{\Sigma}_\mathbf{\Psi}^{-2}\mathbf{V}_\mathbf{\Psi}^T$ by $\mathbf{M}_\Psi$. By Eqn. (9) of Lemma 2 in

Drineas et al. (2012a), we have $\|\mathbf{M}_{\mathbf{\Psi}}\|_2 < \frac{\epsilon}{1-\epsilon}$. Hence,

$$\begin{aligned}
|s_i - \hat{s}_i| &= \mathbf{e}_i^T\mathbf{U}\mathbf{\Sigma}^{-2}\mathbf{U}^T\mathbf{e}_i - \mathbf{e}_i^T\mathbf{U}(\mathbf{\Pi}_1\mathbf{U})^{\dagger}(\mathbf{\Pi}_1\mathbf{U})^{\dagger T}\mathbf{\Sigma}^{-2}(\mathbf{\Pi}_1\mathbf{U})^{\dagger}(\mathbf{\Pi}_1\mathbf{U})^{\dagger T}\mathbf{U}^T\mathbf{e}_i \\
&= \mathbf{e}_i^T\mathbf{U}(\mathbf{\Sigma}^{-2} - (\mathbf{\Pi}_1\mathbf{U})^{\dagger}(\mathbf{\Pi}_1\mathbf{U})^{\dagger T}\mathbf{\Sigma}^{-2}(\mathbf{\Pi}_1\mathbf{U})^{\dagger}(\mathbf{\Pi}_1\mathbf{U})^{\dagger T})\mathbf{U}^T\mathbf{e}_i \\
&\leq \|\mathbf{I}_p - \mathbf{\Sigma}(\mathbf{\Pi}_1\mathbf{U})^{\dagger}(\mathbf{\Pi}_1\mathbf{U})^{\dagger T}\mathbf{\Sigma}^{-2}(\mathbf{\Pi}_1\mathbf{U})^{\dagger}(\mathbf{\Pi}_1\mathbf{U})^{\dagger T}\mathbf{\Sigma}\|_2 s_i \\
&= \|\mathbf{I}_p - \mathbf{\Sigma}\mathbf{V}_{\mathbf{\Psi}}\mathbf{\Sigma}_{\mathbf{\Psi}}^{-2}\mathbf{V}_{\mathbf{\Psi}}^T\mathbf{\Sigma}^{-2}\mathbf{V}_{\mathbf{\Psi}}\mathbf{\Sigma}_{\mathbf{\Psi}}^{-2}\mathbf{V}_{\mathbf{\Psi}}\mathbf{\Sigma}\|_2 s_i \\
&= \|\mathbf{\Sigma}\mathbf{M}_{\mathbf{\Psi}}\mathbf{\Sigma}^{-2}\mathbf{M}_{\mathbf{\Psi}}\mathbf{\Sigma} - \mathbf{\Sigma}^{-1}\mathbf{M}_{\mathbf{\Psi}}\mathbf{\Sigma} - \mathbf{\Sigma}\mathbf{M}_{\mathbf{\Psi}}\mathbf{\Sigma}^{-1}\|_2 s_i \\
&= (\|\mathbf{\Sigma}\|_2^2\|\mathbf{M}_{\mathbf{\Psi}}\|_2^2/\|\mathbf{\Sigma}^{-1}\|_2^2 + 2\|\mathbf{\Sigma}\|_2\|\mathbf{M}_{\mathbf{\Psi}}\|_2/\|\mathbf{\Sigma}^{-1}\|_2)s_i \\
&\leq \left(\left(\frac{B}{b}\right)^2\left(\frac{\epsilon}{1-\epsilon}\right)^2 + \frac{B}{b}\frac{\epsilon}{1-\epsilon}\right)s_i \\
&\leq 2\left(\frac{B}{b}\right)^2\left(\frac{\epsilon}{1-\epsilon}\right)s_i.
\end{aligned}$$

This completes the proof of the lemma.

Proof (of Theorem B.3)

Combining Lemma B.4 and Lemma B.5, we have

$$\begin{aligned}
|s_i - \tilde{s}_i| &\leq |s_i - \hat{s}_i + \hat{s}_i - \tilde{s}_i| \leq |s_i - \hat{s}_i| + |\hat{s}_i - \tilde{s}_i| \\
&\leq 2\left(\frac{B}{b}\right)^2\left(\frac{\epsilon}{1-\epsilon}\right)s_i + 2\epsilon\hat{s}_i \\
&\leq 2\left(\frac{B}{b}\right)^2\left(\frac{\epsilon}{1-\epsilon}\right)s_i + 2\epsilon\left(1 + 2\left(\frac{B}{b}\right)^2\left(\frac{\epsilon}{1-\epsilon}\right)\right)s_i \\
&\leq \left(2 + 4\left(\frac{B}{b}\right)^2\right)\epsilon s_i.
\end{aligned}$$

The theorem follows after rescaling ϵ, thus completing the proof.

Let D be the smallest value of $AMSE(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)$ and $\tilde{D}$ be the value of $AMSE(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)$ using the proposed approximation method. The following theorem states that $\tilde{D}$ is a relative-error approximation of D.

Theorem B.6 *Assume the conditions in Theorem B.3 hold. Then, by using the Algorithm 1, we have*

$$D \leq \tilde{D} \leq \sqrt{\frac{1+\epsilon}{1-\epsilon}}D$$

with probability at least 0.8.

The proof is almost identical to the proof of Theorem B.2, with only changes of notations, and it is thus omitted.

B.1 Approximations of other estimators and their relative-errors

The sampling probabilities of ICNLEV, RLNLEV, and PLNLEV estimators are proportional to $\sqrt{1-l_i}\|s_i\|$, $\sqrt{1-l_i}\sqrt{l_i}$ and $\sqrt{1-l_i}\|\mathbf{x}_i\|$, respectively. By replacing l_i with $\tilde{l}_i$ and s_i with $\tilde{s}_i$ respectively, we can use Theorem B.1 and Theorem B.3 to prove that the proposed method provides relative-error approximation for these numerators. Thus, similar theoretical results for the quality-of-approximation can be proven for the rest of the sampling estimators using the proposed approximation methods. We omit the details.

References

Haim Avron, Petar Maymounkov, and Sivan Toledo. Blendenpik: Supercharging lapack's least-squares solver. *SIAM Journal on Scientific Computing*, 32(3):1217–1236, 2010.

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

Patrick Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. Wiley, 1995.

Yvonne M Bishop, Stephen E Fienberg, and Paul W Holland. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, 1975.

Siheng Chen, Rohan Varma, Aarti Singh, and Jelena Kovačcević. A statistical perspective of sampling scores for linear regression. In *Information Theory (ISIT), 2016 IEEE International Symposium*, pages 1556–1560. IEEE, 2016.

Kenneth L Clarkson, Manfred K Warmuth, Michael Mahoney, and Michal Derezinski. Minimax experimental design: Bridging the gap between statistical and worst-case approaches to least-squares regression. *Proceedings of Machine Learning Research vol*, 99:1–20, 2019.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006. ISBN 0471241954.

Petros Drineas and Michael W. Mahoney. RandNLA: Randomized Numerical Linear Algebra. *Communications of the ACM*, 59(6):80–90, 2016. ISSN 0001-0782. doi: 10.1145/2842602. URL `http://doi.acm.org/10.1145/2842602`.

Petros Drineas and Michael W. Mahoney. Lectures on randomized numerical linear algebra. In M. W. Mahoney, J. C. Duchi, and A. C. Gilbert, editors, *The Mathematics of Data*, IAS/Park City Mathematics Series, pages 1–48. AMS/IAS/SIAM, 2018.

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Sampling algorithms for l_2 regression and applications. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1127–1136, 2006.

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decomposition. *SIAM Journal on Matrix Analysis and Applications*, 30:844–881, 2008.

Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506, 2012a.

Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506, 2012b.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

Barbara Hubbard and John H. Hubbard. *Vector Calculus, Linear Algebra, and Differential Forms: A Unified Approach*. Prentice Hall, 1999. ISBN 9780136574460.

Peter J Huber. Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, pages 799–821, 1973.

T.L. Lai, Herbert Robbins, and C.Z. Wei. Strong consistency of least squares estimates in multiple regression. *Proceedings of the National Academy of Sciences*, 75(7):3034–3036, 1978.

Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, 1986.

Erich L Lehmann and Joseph P Romano. *Testing Statistical Hypotheses*. Springer Science & Business Media, 2006.

Ping Ma, Mahoney. W. Mahoney, and B. Yu. A statistical perspective on algorithmic leveraging. In *Proceedings of the 31th ICML Conference*, pages 91–99, 2014.

Ping. Ma, Mahoney. W. Mahoney, and B Yu. A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, 16:861–911, 2015.

Ping Ma, Xinlian Zhang, Xin Xing, Jingyi Ma, and Michael Mahoney. Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. In *Proceedings of the 23rd International Workshop on Artificial Intelligence and Statistics*, 2020.

Michael Mahoney. *Randomized Algorithms for Matrices and Data*. Foundations and Trends in Machine Learning. NOW Publishers, Boston, 2011. Also available at: arXiv:1104.5557.

Michael W. Mahoney and Petros Drineas. Structural properties underlying high-quality randomized numerical linear algebra algorithms. In P. Bühlmann, P. Drineas, M. Kane, and M. van de Laan, editors, *Handbook of Big Data*, pages 137–154. CRC Press, 2016.

Xiangrui Meng, Michael A Saunders, and Michael W Mahoney. LSRN: A parallel iterative solver for strongly over- or under-determined systems. *SIAM Journal on Scientific Computing*, 36(2):C95–C118, 2014.

Carl Morris. Central limit theorems for multinomial sums. *The Annals of Statistics*, 3(1): 165–188, 1975.

Mert Pilanci and Martin J. Wainwright. Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares. *Journal of Machine Learning Research*, 17 (53):1–38, 2016.

Stephen Portnoy. Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *The Annals of Statistics*, pages 1298–1309, 1984.

Stephen Portnoy. Asymptotic behavior of M estimators of p regression parameters when p^2/n is large; II. Normal approximation. *The Annals of Statistics*, pages 1403–1417, 1985.

Garvesh Raskutti and Michael W. Mahoney. A statistical perspective on randomized sketching for ordinary least-squares. In *Proceedings of the 32th ICML Conference*, pages 617–625, 2015.

George A.F. Seber and Alan J. Lee. *Linear Regression Analysis*. Wiley, 2nd edition, 2003.

Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 2001.

Jun Shao. *Mathematical Statistics*. Springer Texts in Statistics. Springer Verlag, 2003. ISBN 9780387953823.

Ross Sheldon. *A First Course in Probability*. Pearson Education India, 7th edition, 2006.

George P. Steck. *Limit Theorems for Conditional Distributions*. University of California Press, Berkeley, 1957.

Robert van de Geijn and Margaret Myers. *Advanced Linear Algebra: Foundations to Frontiers*. ulaff.net, 2019.

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

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

Jialei Wang, Jason D Lee, Mehrdad Mahdavi, Mladen Kolar, and Nathan Srebro. Sketching meets random projection in the dual: A provable recovery algorithm for big and high-dimensional data. *Electronic Journal of Statistics*, 11(2):4896–4944, 2017a.

Yining Wang, Adams Wei Yu, and Aarti Singh. On computationally tractable selection of experiments in measurement-constrained regression models. *Journal of Machine Learning Research*, 18(143):1–41, 2017b.

David P Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Victor J Yohai and Ricardo A Maronna. Asymptotic behavior of M-estimators for the linear model. *The Annals of Statistics*, pages 258–268, 1979.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.