Information-Based Optimal Subdata Selection for Non-linear Models

Jun Yu[1†], Jiaqi Liu[2†] and HaiYing Wang[2*]

[1]School of Mathematics and Statistics, Beijing Institute of Technology, Beijing, 100811, China.
[2*]Department of Statistics, University of Connecticut, Storrs, 06269, CT, USA.

*Corresponding author(s). E-mail(s): haiying.wang@uconn.edu;
Contributing authors: yujunbeta@bit.edu.cn;
jiaqi.3.liu@uconn.edu;
†These authors contributed equally to this work.

Abstract

Subdata selection methods provide flexible tradeoffs between computational complexity and statistical efficiency in analyzing big data. In this work, we investigate a new algorithm for selecting informative subdata from massive data for a broad class of models, including generalized linear models as special cases. A connection between the proposed method and many widely used optimal design criteria such as A-, D-, and E-optimality criteria is established to provide a comprehensive understanding of the selected subdata. Theoretical justifications are provided for the proposed method, and numerical simulations are conducted to illustrate the superior performance of the selected subdata.

Keywords: Generalized linear models, Information matrix, Massive data, Optimality criteria

1 Introduction

Notably growth of data volumes is ubiquitous in the big data era. Analysis of huge-volume datasets with proven statistical methods meets new challenges due to the limits of the available computational resources. There is an urgent

need to find a balance between computational cost and statistical efficiency for the following two main reasons. Firstly, diminishing marginal utility indicates that analyzing a small but representative data set might be more cost-effective for obtaining the population's information without a loss of too much estimation efficiency. Secondly, a timely responsive analysis is more desired in the big data era. Some intended results may become less useful when the computation takes a long time to run.

Among various big data analysis tactics, subdata selection and subsampling play an important role in achieving a good tradeoff between statistical efficiency and computational cost. The key idea is to utilize a small portion of data to approximate the information contained in the full data. A lot of subsampling and selection methods have been proposed to accommodate different statistical modeling and data analysis. Typical investigations include but are not limited to leverage score based subsampling for linear and vector autoregression models (Ma et al., 2015; Xie et al., 2019; Ma et al., 2020); orthogonal array and Latin hypercube based sampling for linear and Gaussian process regressions (Zhao et al., 2018; Wang et al., 2021; Meng et al., 2020); A-optimality motivated optimal subsampling method for generalized linear models and quantile regressions (Wang et al., 2018; Ai et al., 2021; Wang and Ma, 2021; Ai et al., 2021; Zhang et al., 2021); information-based optimal subdata selection (IBOSS) for linear and logistic regressions (Wang et al., 2019; Cheng et al., 2020); and optimal design subsampling (Deldossi and Tommasi, 2022) for linear regressions. A literature review can be found in Yu et al. (2023).

Subsampling method and subdata selection method have a lot in common, but they are essentially different in the selection scheme. To be precise, subsampling is a randomized algorithm, and the major goal is to approximate the full data estimator using a small subsample. This is very similar to the problem of finite population sampling in this the goal is to estimate finite population parameters (Deville and Särndal, 1992; Särndal et al., 1992). On the other hand, subdata selection is a deterministic algorithm with the major goal to estimate the parameter of the population that generates the full data. This approach is more akin to the problem of optimum experimental design but with substantial differences. Classical optimum experimental design theory concerns theorems and algorithms for choosing optimum designs over a specified region, or from a specified set of candidate points. For large problems, this may be very computational expensive. In the present paper we suggest new ways of selecting the design points using computationally efficient algorithms that are scalable to big data, with the additional feature that the responses are collected too. Please see Drovandi et al. (2017) for more systematic discussions on the connections between experimental design and big data analysis. Compared with existing subsampling methods, subdata selection has its own advantage since the data selection step does not involve additional variation. Consequently, one may expect that the mean squared error decreases as the size of full data increases, even when the subdata size is fixed. Take a linear regression as an example. Wang et al. (2019) showed that the variance of a

random subsampling based estimator converges to zero at a rate proportional to the inverse of the subdata size, while the IBOSS based slope estimator may converge to zero at a rate related to the size of the full data. Thus under some mild conditions, the subdata selection based methods may have "supper efficiency" compared with random subsampling based estimators in terms of the convergence rate of the variance.

Compared with the extensive studies on the subsampling approach, systematic investigations on the subdata selection approach for massive data analysis lag behind. To the best of our knowledge, the most relevant studies are Wang et al. (2019) and Cheng et al. (2020), which focus on linear regression and logistic regression, respectively. These results cannot fulfill the needs of analyzing various massive datasets. In this work, we study an optimal subdata selection procedure for a broad class of models that contains generalized linear models as special cases. The contributions are the following three folds. Firstly, we present a general subdata selection algorithm for a broad class of models which contain the IBOSS algorithm for linear models as a special case. Secondly, we build the connections between the proposed method and some commonly used optimal design criteria, including A-, D-, E-, and T-optimality criteria (Pukelsheim, 2006). It naturally gives a comprehensive understanding of the proposed method and the IBOSS methods in Wang et al. (2019) and Cheng et al. (2020). Thirdly, we show that the information from the new algorithm increases along with the size of full data both theoretically and numerically. These results justify the information-based subdata selection under nonlinear models.

The rest of the paper is organized as follows. Section 2 introduces the problem setups. Section 3 introduces A-, D-, E-, and T-optimality criteria in the context of subdata selection. Section 4 introduces a new algorithm and discusses its rationale under multiple optimality criteria. Section 5 provides some theoretical analysis for the proposed method. Section 6 compares the performance of the proposed algorithm with other commonly used subsampling and subdata selection algorithms numerically. Technical details are postponed to the Appendix.

2 The framework

The regression problem is to determine a statistical relationship between an explanatory variable $\boldsymbol{x}$ and a response variable y. In a parametric setup, the statistical relationship is characterized by a conditional distribution of y given $\boldsymbol{x}$, say $f(y \mid \boldsymbol{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a d dimensional parameter vector of interest. Linear and generalized linear regressions are special cases with different specifications of $f(y \mid \boldsymbol{x}, \boldsymbol{\theta})$. Assume that $(\boldsymbol{x}_1, y_1)$, $\ldots$, $(\boldsymbol{x}_N, y_N)$ are independent data from the distribution of $(\boldsymbol{x}, y)$. With the full data, the maximum likelihood estimator (MLE) is broadly adopted to estimate the unknown parameter $\boldsymbol{\theta}$, which is given by $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} N^{-1} \sum_{i=1}^{N} \log f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$.

The Fisher information plays an important role characterizing estimation efficiency, because it is related to the inverse of the (asymptotic) variance-covariance matrix of the MLE. Herein, the per-observation Fisher information matrix given $\boldsymbol{x}$ is

$$\mathrm{I}(\boldsymbol{x}) = E_{y|\boldsymbol{x}}\left[\left\{\frac{\partial \log f(y \mid \boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\}\left\{\frac{\partial \log f(y \mid \boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\}^{\mathrm{T}}\right], \tag{1}$$

where $E_{y|\boldsymbol{x}}$ means taking the conditional expectation given $\boldsymbol{x}$, and the Fisher information matrix given covariates for $\boldsymbol{\theta}$ based on full data is

$$\mathcal{I}_f = \sum_{i=1}^{N} \mathrm{I}(\boldsymbol{x}_i).$$

To improve the estimation efficiency, optimal experimental design theory (Kiefer, 1959) suggests finding the experimental setting of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ such that $\psi(\mathcal{I}_f)$ attains its maximum. The function ψ is known as the optimality criterion function. Examples of ψ will be introduced in Section 3.

Let $(\boldsymbol{x}_1^*, y_1^*), \ldots, (\boldsymbol{x}_n^*, y_n^*)$ be a subdata of size n chosen deterministically from the full data, in which the rule to determine whether a data point is included or not depend on $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_N)^{\mathrm{T}}$ only. Note that the selection rule only relies on the marginal distribution of $\boldsymbol{x}_i$ which is ancillary to $\boldsymbol{\theta}$. As argued in Efron and Hinkley (1978), statistical inference on parameters should be done based on the observed information when the experimental procedure is ancillary. Thus, the statistical inference procedure based on the selected subdata is the same as using the full data; see also Deldossi and Tommasi (2022). We denote the MLE based on the subdata by $\hat{\boldsymbol{\theta}}^*$. Let δ_i be the indicator that observation $(\boldsymbol{x}_i, y_i)$ is included in the subdata. The Fisher information matrix of a subdata of size k can be written as

$$\mathcal{I}(\boldsymbol{\delta}) = \sum_{i=1}^{N} \delta_i \mathrm{I}(\boldsymbol{x}_i), \tag{2}$$

where $\boldsymbol{\delta} = \{\delta_1, \delta_2, ..., \delta_N\}$ such that $\sum_{i=1}^{N} \delta_i = n$. To have an optimal estimator based on a subdata, one can choose $\boldsymbol{\delta}$ that "maximizes" the above information matrix (2) in the sense of some criterion function $\psi(\cdot)$. The problem is presented as the following optimization problem conditional on the observed full data:

$$\boldsymbol{\delta}^{opt} = \arg\max_{\boldsymbol{\delta}} \psi(\mathcal{I}(\boldsymbol{\delta})), \quad \sum_{i=1}^{N} \delta_i = n. \tag{3}$$

3 Optimality criteria

In this paper, we consider the class of models for which the per-observation information matrix can be expressed as $\mathrm{I}(\boldsymbol{x}_i) = \boldsymbol{z}_i(\boldsymbol{x}_i, \boldsymbol{\theta})\boldsymbol{z}_i^{\mathrm{T}}(\boldsymbol{x}_i, \boldsymbol{\theta})$, where $\boldsymbol{z}_i(\boldsymbol{x}_i, \boldsymbol{\theta})$ is a vector function of $\boldsymbol{x}_i$ that may depend on $\boldsymbol{\theta}$. For simplicity, we write $\boldsymbol{z}_i(\boldsymbol{x}_i, \boldsymbol{\theta})$ as $\boldsymbol{z}_i$ if there is no confusion, and denote $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N)^{\mathrm{T}}$ by $\boldsymbol{Z}$. Similarly, let $\boldsymbol{Z}^* = (\boldsymbol{z}_1^*, \ldots, \boldsymbol{z}_n^*)^{\mathrm{T}}$ be the corresponding quantity for the selected subdata. This kind of information matrix expression occurs in a natural way not only for linear and generalized linear models but also for various other models, like accelerated failure time models and general nonlinear regression models. We present two classes of models with this information expression below to facilitate the later discussion.

Example 1 Consider a generalized linear regression model

$$g(E(y \mid \boldsymbol{x})) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}, \tag{4}$$

where the distribution of y given $\boldsymbol{x}$ belongs to the exponential family and $g(\cdot)$ is a link function. Let $\mathrm{var}(y \mid \boldsymbol{x})$ and $\mathrm{SD}(y \mid \boldsymbol{x})$ be the variance and standard deviation of y given $\boldsymbol{x}$, respectively. The per-observation information matrix for $\boldsymbol{\theta}$ at $\boldsymbol{x}$ is

$$\mathrm{I}(\boldsymbol{x}) = \frac{\{\dot{g}^{-1}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})\}^2}{\mathrm{var}(y \mid \boldsymbol{x})}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}, \tag{5}$$

where $\dot{g}^{-1}$ is the derivative of the inverse function of g. Thus, $\boldsymbol{z}_i = \{|\dot{g}^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})|/\mathrm{SD}(y_i \mid \boldsymbol{x}_i)\}\boldsymbol{x}_i$ for a generalized linear model.

Example 2 Consider the following general nonlinear regression model

$$y = g(\boldsymbol{x}, \boldsymbol{\theta}) + \varepsilon, \tag{6}$$

where ε follows a normal distribution with mean zero and variance σ^2. The per-observation information matrix for $\boldsymbol{\theta}$ at $\boldsymbol{x}$ is

$$\mathrm{I}(\boldsymbol{x}) = \frac{1}{\sigma^2}\left\{\frac{\partial g(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\}\left\{\frac{\partial g(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\}^{\mathrm{T}}. \tag{7}$$

From (7), $\boldsymbol{z}_i = \sigma^{-1}\partial g(\boldsymbol{x}_i, \boldsymbol{\theta})/\partial \boldsymbol{\theta}$.

In the following, we discuss A-, D-, E-, and T-optimality criteria, which are all defined in Chapter 9 of Pukelsheim (2006). To be precise, A-optimality suggests to maximize $\{\mathrm{tr}(\mathcal{I}^{-1}(\boldsymbol{\delta}))\}^{-1}$, so that the average variance of $\hat{\boldsymbol{\theta}}^*$ is minimized. D-optimality suggests to maximize $\det(\mathcal{I}(\boldsymbol{\delta}))$, so that the volume of the confidence ellipsoid based on $\hat{\boldsymbol{\theta}}^*$ is minimized. E-optimality is to maximize $\lambda_{\min}(\mathcal{I}(\boldsymbol{\delta}))$, so that the maximum variance of $\boldsymbol{l}^T\hat{\boldsymbol{\theta}}^*$ among all unit vector $\boldsymbol{l}$ attains its minimal. Here $\lambda_{\min}(A)$ denotes the smallest eigenvalue of the matrix A. The T-otimality suggests to maximize $\mathrm{tr}(\mathcal{I}(\boldsymbol{\delta}))$ which is different from the criterion with the same name studied in Atkinson and Fedorov (1975). For mathematical rigorousness, $\psi(\mathcal{I}(\boldsymbol{\delta}))$ is defined as zero when $\mathcal{I}(\boldsymbol{\delta})$ is singular.

In the following, we write the mathematical definitions of optimal subdata of size n under A-, D-, E-, and T-optimality criteria. The constant dimension

d is used in the expressions to make different optimality criteria on the same scale.

- A-optimality:

$$\boldsymbol{\delta}_A^{opt} = \arg\max_{\boldsymbol{\delta}} \{\mathrm{tr}(d^{-1}\mathcal{I}^{-1}(\boldsymbol{\delta}))\}^{-1} = \arg\max_{\boldsymbol{\delta}} \left\{ \mathrm{tr}\left(\frac{1}{d} \left(\sum_{i=1}^{N} \delta_i \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}} \right)^{-1} \right) \right\}^{-1}, \sum_{i=1}^{N} \delta_i = n. \tag{8}$$

- D-optimality:

$$\boldsymbol{\delta}_D^{opt} = \arg\max_{\boldsymbol{\delta}} \{\det(\mathcal{I}(\boldsymbol{\delta}))\}^{1/d} = \arg\max_{\boldsymbol{\delta}} \det \left(\sum_{i=1}^{N} \delta_i \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}} \right)^{1/d}, \quad \sum_{i=1}^{N} \delta_i = n. \tag{9}$$

- E-optimality:

$$\boldsymbol{\delta}_E^{opt} = \arg\max_{\boldsymbol{\delta}} \lambda_{\min}(\mathcal{I}(\boldsymbol{\delta})) = \arg\max_{\boldsymbol{\delta}} \lambda_{\min} \left(\sum_{i=1}^{N} \delta_i \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}} \right), \quad \sum_{i=1}^{N} \delta_i = n. \tag{10}$$

- T-optimality:

$$\boldsymbol{\delta}_T^{opt} = \arg\max_{\boldsymbol{\delta}} \mathrm{tr}(d^{-1}\mathcal{I}(\boldsymbol{\delta})) = \arg\max_{\boldsymbol{\delta}} \mathrm{tr} \left(d^{-1} \sum_{i=1}^{N} \delta_i \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}} \right), \quad \sum_{i=1}^{N} \delta_i = n. \tag{11}$$

The A-, D-, and T-optimality criteria correspond to the harmonic mean, geometric mean, and arithmetic mean of the eigenvalues of $\mathcal{I}(\boldsymbol{\delta})$, respectively. For any given subdata set $\boldsymbol{\delta}$, the relationship between the four criteria, namely A-, D-, E, and T-optimality, can be directly derived by the well-known inequality of means. Specifically, it follows that

$$\lambda_{\min}(\mathcal{I}(\boldsymbol{\delta})) \leq \{\mathrm{tr}(d^{-1}\mathcal{I}^{-1}(\boldsymbol{\delta}))\}^{-1} \leq \{\det(\mathcal{I}(\boldsymbol{\delta}))\}^{1/d} \leq \mathrm{tr}(d^{-1}\mathcal{I}(\boldsymbol{\delta})), \tag{12}$$

and these equalities hold when all the eigenvalues are the same. Although there is no strict ordering for the four optimal subdata selection methods, E-optimality subdata selection leads to reasonable efficiencies under A-, D-, and T- optimality criteria, since it maximizes a lower bound of all the other criteria.

4 The proposed method

In this section, we present the subdata selection algorithms under A-, D-, E-, and T-optimality criteria.

The optimal subdata selection under the T-optimality criterion takes a particularly simple form. Thus we first introduce the analytic solution of the T-optimality subdata selection.

Theorem 1 *For a given massive data* $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, N\}$*, the* T*-optimal subdata of size* n *consists of data points with the* n *largest values of* $\{\|\boldsymbol{z}_i\|^2, i = 1, \ldots, N\}$*.*

Finding larger $\|\boldsymbol{z}_i\|$'s to improve the statistical efficiency is common in survey sampling literature. The popular technique that adopts such an idea is the probability proportional to size (PPS) sampling (Särndal et al., 1992). As shown in Hartley and Rao (1962), if the auxiliary information used in the

sampling is correlated with the variables of interest, the PPS sampling reduces sampling variance and improves precision for estimation.

Unlike the T-optimal subdata selection, it is much more difficult to obtain a closed-form solution for A-, D-, or E-optimality criteria in general. A naive exhaustive search for the subdata in the full data satisfying A-, D- or E-optimality criterion is unrealistic because this is often more computationally demanding than finding the full data MLE. In working towards approximate solutions, we first derive some bounds for $\mathcal{I}(\boldsymbol{\delta})$ under A-, D- and E-optimality criteria, which will guide our later algorithms.

Theorem 2 *Let* $\text{var}(\boldsymbol{Z}_j^*) = \sum_{i=1}^n (z_{ij}^* - \bar{z}_j^*)^2/(n-1)$ *be the sample variance for the* j*th column of* $\boldsymbol{Z}^*$ *where* $\bar{z}_j^* = \sum_{i=1}^n z_{ij}^*/n$*, and* $\text{var}_{\min}(\boldsymbol{Z}^*)$ *be the smallest value among* $\{\text{var}(\boldsymbol{Z}_j^*), j = 1, \ldots, d\}$*. For any subdata of size* n*, represented by* $\boldsymbol{\delta}$ *with* $\mathcal{I}(\boldsymbol{\delta}) > 0$*, the following results hold:*

$$\frac{d\lambda_{\min}(\boldsymbol{R}^*)}{\sum_{j=1}^d 1/\text{var}(\boldsymbol{Z}_j^*)} \leq \frac{d}{(n-1)\{\text{tr}(\mathcal{I}^{-1}(\boldsymbol{\delta}))\}} \leq \frac{d\lambda_{\max}(\boldsymbol{R}^*)}{\sum_{j=1}^d 1/\text{var}(\boldsymbol{Z}_j^*)}, \tag{13}$$

$$\lambda_{\min}(\boldsymbol{R}^*)\left(\prod_{j=1}^d \text{var}(\boldsymbol{Z}_j^*)\right)^{1/d} \leq \frac{\{\det(\mathcal{I}(\boldsymbol{\delta}))\}^{1/d}}{n-1} \leq \lambda_{\max}(\boldsymbol{R}^*)\left(\prod_{j=1}^d \text{var}(\boldsymbol{Z}_j^*)\right)^{1/d}, \tag{14}$$

$$\lambda_{\min}(\boldsymbol{R}^*)\text{var}_{\min}(\boldsymbol{Z}^*) \leq \frac{\lambda_{\min}(\mathcal{I}(\boldsymbol{\delta}))}{n-1} \leq \lambda_{\max}(\boldsymbol{R}^*)\text{var}_{\min}(\boldsymbol{Z}^*), \tag{15}$$

where $\boldsymbol{R}^*$ *is the sample correlation matrix of* $\boldsymbol{Z}^*$*, and* $\lambda_{\max}(A)$ *is the maximum eigenvalue of matrix* A*.*

From Theorem 2, one sees that a larger variation in subdata covariates leads to better estimation. This encourages us to select the subdata with large variation in the induced covariates z_j's ($j = 1, \ldots, d$). More importantly, under some mild conditions that $\lambda_{\min}(\boldsymbol{R}^*) > 0$ with fixed d and $\text{var}_{\min}(\boldsymbol{Z}^*) \to \infty$, one can expect that $\lambda_{\min}(\mathcal{I}(\boldsymbol{\delta}))$ with $\boldsymbol{\delta}$ selected by the proposed selection strategy grows much faster than n while $\lambda_{\min}(\mathcal{I}(\boldsymbol{\delta})) = O_P(n)$ under simple random sampling, which implies our methods are more efficient than simple random sampling.

Improving the subdata induced covariate variances for all dimensions simultaneously may still exceed the limits of computational budgets. To further accelerate the selection procedure, we suggest enlarging the subsample's range of each dimension separately. This is because a sample range usually serves as a proxy for a sample standard deviation in the statistical literature. Such idea was investigated back in Tippett (1925), and it has been widely adopted in hypothesis testing (David et al., 1954) and statistical process control (Montgomery, 2019).

Along with this thinking, we only need to collect subdata with extreme values of $\boldsymbol{Z}_j$'s ($j = 1, \ldots, d$), both small and large, occurring with the same frequency. This agrees with the common statistical knowledge that selecting

the edge points of the data region improves statistical efficiency. For example, the optimal design points for the generalized linear model constructed in Yang et al. (2011) and Schmidt and Schwabe (2017) lie on the edge of some covariates. The information-based optimal subdata selection methods for linear and logistic regression also encourage selecting both small and large values of $\boldsymbol{Z}_j$'s $(j = 1, \ldots, d)$, occurring with the same frequency. Readers may refer to Wang et al. (2019); Cheng et al. (2020) for more details.

When $\boldsymbol{Z}$ is given, we propose to apply the IBOSS algorithm of Wang et al. (2019), which consists of selecting $2r$ data points with the r data points with the smallest $\boldsymbol{z}_{ij}$ values and r data points with the r largest $\boldsymbol{z}_{ij}$ values for $j = 1, \ldots, d$. The subdata is formed by collecting these $n = 2rd$ points together. However, care is necessary as $\boldsymbol{z}_i$'s are often functions of $\boldsymbol{\theta}$. To select a subdata based on the information matrix, $\boldsymbol{\theta}$ has to be replaced by a pilot estimate, say $\tilde{\boldsymbol{\theta}}$. Denote $\tilde{\boldsymbol{z}}_i = \boldsymbol{z}_i(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}})$. We will perform subdata selection by applying the IBOSS algorithm to $\tilde{\boldsymbol{Z}} = (\tilde{\boldsymbol{z}}_1, \ldots, \tilde{\boldsymbol{z}}_N)^{\mathrm{T}}$, which is summarized in the following:

Algorithm 1 Suppose that $r = n/(2d)$ is an integer.

(1) Take a random subsample of size n_0 from the full sample and use it to obtain a pilot estimate of $\boldsymbol{\theta}$, $\tilde{\boldsymbol{\theta}}$. Calculate $\tilde{\boldsymbol{z}}_i = \boldsymbol{z}_i(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}})$, for $i = 1, ..., N$.

(2) Using a partition-based selection algorithm, perform the following steps:
   (a) For $\tilde{z}_{i1}$, $1 \leq i \leq N$, include $2r$ data points with the r smallest $\tilde{\boldsymbol{z}}_{i1}$ values and r data points with the r largest $\tilde{\boldsymbol{z}}_{i1}$ values;
   (b) For $j = 2, ..., d$, exclude data points that were previously selected, and from the remainder select r data points with the smallest $\tilde{\boldsymbol{z}}_{ij}$ values and r data points with the largest $\tilde{\boldsymbol{z}}_{ij}$ values for the subdata.

(3) Calculate and return the MLE and associated statistics using the selected subdata.

Let us note that this approach has already been followed by Deldossi and Tommasi (2022) for comparing the IBOSS algorithm to other selection methods in a logistic regression model.

Remark 1 Given a pilot estimate $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, for any subdata of size k, represented by $\boldsymbol{\delta}$, the results in Theorem 2 still hold when $\boldsymbol{\theta}$ is replaced by $\tilde{\boldsymbol{\theta}}$. Under some mild conditions that $\mathcal{I}(\boldsymbol{\delta})$ is continuous with respect to $\boldsymbol{\theta}$, and $\tilde{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$. The proposed algorithm is still a viable approach to enhance the estimation efficiency based on the selected subdata.

Remark 2 For Algorithm 1, the time to obtain $\tilde{\boldsymbol{\theta}}$ is often $O(n_0 d^2 \xi_0)$ where ξ_0 is the number of iterations in the optimization procedure. The time to calculate $\tilde{\boldsymbol{Z}}$ is $O(Nd)$. By partition-based selection algorithm (Musser, 1997; Martínez, 2004), the time to select the r largest and smallest elements of $\boldsymbol{Z}_j$ has an average time complexity $O(N)$. The time to calculate $\hat{\boldsymbol{\theta}}^*$ is often $O((n_0 + n)d^2\xi)$ where ξ is the number of iterations in the optimization procedure. If n_0, n and d are all much smaller than N, the time complicity of Algorithm 1 is $O(Nd)$.

Remark 3 The proposed method provides an economic solution for semi-supervised learning problems, in which the response variable is expensive to measure. Specifically, although the computational cost may be higher compared with the uniform subsampling method, it requires fewer data to be labeled in order to achieve the same statistical efficiency. This is because we only need the response values for the selected subdata to estimate $\boldsymbol{\theta}$, and $\boldsymbol{Z}$ does not depend on the response.

It is worth mentioning that both T-optimal subdata selection and Algorithm 1 are sensitive to correlations of $\boldsymbol{z}_i$'s (or $\tilde{\boldsymbol{z}}_i$'s). Moreover, the T-optimal subdata selection is also influenced by scales of $\boldsymbol{Z}$ or $\tilde{\boldsymbol{Z}}$. For simplicity, we only consider $\tilde{\boldsymbol{Z}}$ here. If $\tilde{\boldsymbol{Z}}$ has strong co-linearity, some extreme values of $\tilde{\boldsymbol{Z}}_j$'s may cluster together. Consequently, it may lead to the corresponding information matrix $\mathcal{I}(\boldsymbol{\delta})$ being close to singular. To solve this problem, we use the idea of principal components. Let the singular value decomposition (SVD) of $\tilde{\boldsymbol{Z}}$ be $\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{V}}$, where $\tilde{\boldsymbol{U}}$ is a matrix whose columns contain the left singular vectors of $\tilde{\boldsymbol{Z}}$, $\tilde{\boldsymbol{V}}$ is a matrix whose columns contain the right singular vectors of $\tilde{\boldsymbol{Z}}$, and $\tilde{\boldsymbol{\Lambda}}$ is a diagonal matrix with the (non-negative) singular values in decreasing order. We suggest to apply Algorithm 1 using $\tilde{\boldsymbol{U}}$ instead of $\tilde{\boldsymbol{Z}}$. The advantages of working with $\tilde{\boldsymbol{u}}_i$'s are two folds. Firstly, the columns of $\tilde{\boldsymbol{U}}$ are orthogonal. Secondly, the data have been normalized with $\|\tilde{\boldsymbol{U}}_j\|^2 = 1$ for all columns. We summarize this procedure in the following algorithm:

Algorithm 2 Suppose that $r = n/(2d)$ is an integer.

(1) Take a random subsample of size n_0 from the full sample and use it to obtain a pilot estimate of $\boldsymbol{\theta}$, $\tilde{\boldsymbol{\theta}}$. Calculate $\tilde{\boldsymbol{Z}} = (\tilde{\boldsymbol{z}}_1, ..., \tilde{\boldsymbol{z}}_N)^{\mathrm{T}}$.

(2) Perform a SVD $\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{V}}$ to obtain the left-singular-vector matrix $\tilde{\boldsymbol{U}}$ of $\tilde{\boldsymbol{Z}}$.

(3) Implement steps (3) and (4) of Algorithm 1 using columns of $\tilde{\boldsymbol{U}}$ instead of columns of $\tilde{\boldsymbol{Z}}$.

Remark 4 Performing subdata data selection on $\tilde{\boldsymbol{U}}$ and $\tilde{\boldsymbol{Z}}$ are different in general. Selecting subdata based on $\tilde{\boldsymbol{U}}$ can be regarded as a subdata selection on "normalized" and "orthogonalized" covariates, which often makes the algorithm more stable. However, the additional step of SVD to obtain $\tilde{\boldsymbol{U}}$ requires additional $O(Nd^2)$ time.

Similarly, the T-optimality motivated algorithm based on SVD can be done. We omit the detail due to its simplicity. For linear regression, $\|\tilde{\boldsymbol{u}}_i\|^2$ corresponds to the leverage score of the ith data point. Some theoretical results for this kind of selection methods for linear regression are available in Xie et al. (2019); Yu and Wang (2022). Moreover, selecting high leverage score data points is beneficial not only in deterministic subdata selection but also in the random subsampling approach. A typical example is leverage score subsampling. Readers may refer to Ma et al. (2015, 2020) for more details.

5 Asymptotic Analysis

In this section, we derive the asymptotic properties of the proposed modifications of the IBOSS algorithm. Note that the A-, D-, and E-optimality criteria in (8), (9), and (10) are equivalent to minimizing $\mathrm{tr}(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta}))$, $\{\det(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta}))\}^{1/d}$, and $\lambda_{\max}(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta}))$, respectively. For simplicity, we focus on the case that the pilot is given in the first step and consider the behavior of $\mathrm{tr}(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta}))$, $\{\det(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta}))\}^{1/d}$, $\lambda_{\max}(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta}))$, in the next Theorem 3. The results for the T-optimality based selection, presented as the behavior of $\mathrm{tr}(\tilde{\mathcal{I}}(\boldsymbol{\delta}))$, are in the next Theorem 5. We assume that $N \to \infty$, d is fixed, and the pilot $\tilde{\boldsymbol{\theta}}$ is a constant or given independently of the full data in this section.

Let $|\tilde{z}|_{(N)j} = \max(|\tilde{z}_{1j}|, |\tilde{z}_{2j}|, \ldots, |\tilde{z}_{Nj}|)$. The following theorem shows the goodness of the subdata selected by Algorithm 1 in terms of A-,D-, and E-optimality.

Theorem 3 *Let $\tilde{\mathcal{I}}(\boldsymbol{\delta})$ be the information matrix based on subdata of size $n = 2dr$ selected using Algorithm 1. If $\tilde{z}_{(r)j} - \tilde{z}_{(1)j} = o_P(\tilde{z}_{(N)j} - \tilde{z}_{(1)j})$ and $\tilde{z}_{(N)j} - \tilde{z}_{(N-r+1)j} = o_P(\tilde{z}_{(N)j} - \tilde{z}_{(1)j})$, then the following results hold.*

$$\mathrm{tr}(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta})) \leq \sum_{j=1}^{d} \frac{4d^2}{2\lambda_{\min}(\tilde{\boldsymbol{R}})r|\tilde{\boldsymbol{z}}|^2_{(N)j}} \left(1 + o_P(1)\right), \tag{16}$$

$$\{\det(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta}))\}^{1/d} \leq \frac{4d^2}{2\lambda_{\min}(\tilde{\boldsymbol{R}})r} \left(\prod_{j=1}^{d} |\tilde{\boldsymbol{z}}|^2_{(N)j}\right)^{-1/d} \left(1 + o_P(1)\right), \tag{17}$$

$$\lambda_{\max}(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta})) \leq \frac{4d^2}{2\lambda_{\min}(\tilde{\mathbf{R}})r \min(|\tilde{\boldsymbol{z}}|^2_{(N)j}, j = 1, \ldots, d)} \left(1 + o_P(1)\right), \tag{18}$$

where $\tilde{\mathbf{R}}$ is the sample correlation matrix of $\tilde{\boldsymbol{z}}_i$'s ($i = 1, ..., n$) in the subdata.

As pointed out in Theorem 5 of Wang et al. (2019), the assumptions $\tilde{z}_{(r)j} - \tilde{z}_{(1)j} = o_P(\tilde{z}_{(N)j} - \tilde{z}_{(1)j})$ and $\tilde{z}_{(N)j} - \tilde{z}_{(N-r+1)j} = o_P(\tilde{z}_{(N)j} - \tilde{z}_{(1)j})$ hold under some mild conditions on the tail probability of $\tilde{\boldsymbol{z}}$. The proposed method enjoys a super-efficiency compared with simple random sampling in terms of $\mathrm{tr}(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta}))$ and $\det(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta}))$ if $|\tilde{z}|_{(N)j} \to \infty$ for some j. In addition, it also enjoys a super-efficiency in terms of $\lambda_{\max}(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta}))$ when $|\tilde{z}|_{(N)j} \to \infty$ for all j. Note that the sample maximum $|\tilde{z}|_{(N)j}$ goes to infinity as N goes to infinity if the distribution of $\tilde{\boldsymbol{z}}$ is not bounded, regardless of the subdata size n. This implies that when all components of $\tilde{\boldsymbol{z}}$ come from unbounded distributions, then even if n is a fixed constant, the information for the selected subdata will grow with the full data size N. We discuss this aspect in the following two specific examples.

Example 3 Consider a logistic regression with

$$P(y = 1 \mid \boldsymbol{x}, \boldsymbol{\theta}) = p_{\boldsymbol{x},\boldsymbol{\theta}} = g^{-1}(\boldsymbol{x}^T\boldsymbol{\theta}) = \frac{\exp(\boldsymbol{x}^T\boldsymbol{\theta})}{1 + \exp(\boldsymbol{x}^T\boldsymbol{\theta})}. \tag{19}$$

The $\tilde{\boldsymbol{z}}_i = w(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}})\boldsymbol{x}_i$, $i = 1, ..., N$, with $w(\boldsymbol{x}, \boldsymbol{\theta}) = \{p_{\boldsymbol{x},\boldsymbol{\theta}}(1 - p_{\boldsymbol{x},\boldsymbol{\theta}})\}^{1/2}$. The following proposition states that $\tilde{\boldsymbol{z}}$ has an unbounded support with normally distributed covariates under mild conditions.

Proposition 4 *Suppose that $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \Sigma)$ with $\Sigma > 0$, and the nonrandom $\tilde{\boldsymbol{\theta}}$ is finite with more than one nonzero element. The support of any component of $\tilde{\boldsymbol{z}}$ for the logistic regression in Example 3 is unbounded; and it is $(-\infty, \infty)$.*

Proposition 4 indicates that $|\tilde{\boldsymbol{z}}|_{(N)j}$ goes to infinity as $N \to \infty$ for $j = 1, \ldots, d$ under the required conditions. Thus it is clear to see that $\mathrm{tr}(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta}))$, and $\det(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta}))$ go to zero even for a fixed k. As for $\lambda_{\max}(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta}))$, when there is no intercept in the model, it also goes to zero for a fixed n. If an intercept is included (i.e., $\boldsymbol{x}_{i1} = 1$), then $\tilde{\boldsymbol{z}}_{i1} = w(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}) \leq 0.5$, which is a bounded random variable. Thus for a fixed k, $\lambda_{\max}(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta}))$ does not go to zero. However, we always have $\lambda_{\max}(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta})) = O(n^{-1})$.

Example 4 Consider the following Poisson regression

$$E(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = g^{-1}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}) = \exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}). \tag{20}$$

We have $\boldsymbol{z}_i = \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}/2)\boldsymbol{x}_i$ $(i = 1, ..., N)$, for a full data of size N. Similarly to Example 3, suppose that $\boldsymbol{x}$ comes from a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance $\Sigma > 0$, and assume that $\tilde{\boldsymbol{\theta}}$ has more than one nonzero elements. It is easy to show that all components of $\tilde{\boldsymbol{z}}$ come from unbounded distributions. Thus the quantity $\lambda_{\max}(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta})), \mathrm{tr}(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta}))$, and $\det(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta}))$ will go to zero even for a fixed n. When an intercept term is included in the model, the result still holds.

Remark 5 According to Wiely's theorem (Horn and Johnson, 2013), we have $\lambda_{\max}(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta})) \leq \lambda_{\max}(\tilde{\boldsymbol{\Lambda}}^{-1})\lambda_{\max}((\sum_{i=1}^{N} \delta_i \tilde{\boldsymbol{u}}_i \tilde{\boldsymbol{u}}_i^{\mathrm{T}})^{-1})$. Thus the result in Theorem 3 can be easily extended to Algorithm 2 by replacing $\tilde{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{Z}}$ by $\tilde{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{U}}$. Similarly, note the fact that $\mathrm{tr}(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta})) \leq \lambda_{\max}(\tilde{\boldsymbol{\Lambda}}^{-1})\mathrm{tr}((\sum_{i=1}^{N} \delta_i \tilde{\boldsymbol{u}}_i \tilde{\boldsymbol{u}}_i^{\mathrm{T}})^{-1})$, and $\det(\tilde{\mathcal{I}}^{-1}(\boldsymbol{\delta})) = \det(\tilde{\boldsymbol{\Lambda}})^{-2}\det((\sum_{i=1}^{N} \delta_i \tilde{\boldsymbol{u}}_i \tilde{\boldsymbol{u}}_i^{\mathrm{T}})^{-1})$. We can obtain similar results for Algorithm 2 as Theorem 3. Thus, we omit the details.

At the end of this section, we consider the behavior of $\mathrm{tr}(\tilde{\mathcal{I}}(\boldsymbol{\delta}))$ with $\boldsymbol{\delta}$ obtaining by the n largest values of $\{\|\tilde{\boldsymbol{u}}_i\|^2, i = 1, \ldots, N\}$. It is easy to see the following result by noting the fact that $\|\tilde{\boldsymbol{z}}_i\|^2 = \|\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{u}}_i\|^2 \geq \lambda_{\min}(\tilde{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{Z}})\|\tilde{\boldsymbol{u}}_i\|^2$.

Theorem 5 *Let $\tilde{\mathcal{I}}(\boldsymbol{\delta})$ be the Fisher information for the subdata selected by the n largest values of $\{\|\tilde{\boldsymbol{u}}_i\|^2, i = 1, \ldots, N\}$. Then the following result holds.*

$$\mathrm{tr}(\tilde{\mathcal{I}}(\boldsymbol{\delta})) \geq \lambda_{\min}(\tilde{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{Z}})\sum_{i=1}^{n} \|\tilde{\boldsymbol{u}}_{(N-i+1)}\|^2, \tag{21}$$

where $\|\tilde{\boldsymbol{u}}_{(N-i+1)}\|^2$ is the $(N-i+1)$th order statistics of $\{\|\tilde{\boldsymbol{u}}_i\|^2, i = 1, \ldots, N\}$ with $\|\tilde{\boldsymbol{u}}_{(1)}\|^2 \leq \|\tilde{\boldsymbol{u}}_{(2)}\|^2 \leq \ldots \leq \|\tilde{\boldsymbol{u}}_{(N)}\|^2$.

6 Numerical Examples

In this section, we use numerical experiments to evaluate the practical performance of the proposed methods.

Since a pilot subdata of size n_0 is required to obtain the pilot estimator $\tilde{\boldsymbol{\theta}}$, we combine it with the estimator $\hat{\boldsymbol{\theta}}$ from a proposed algorithm to obtain the final subdata estimator

$$\check{\boldsymbol{\theta}} = (\tilde{\boldsymbol{M}}_X + \hat{\boldsymbol{M}}_X)^{-1}(\tilde{\boldsymbol{M}}_X\tilde{\boldsymbol{\theta}} + \hat{\boldsymbol{M}}_X\hat{\boldsymbol{\theta}}), \tag{22}$$

where $\tilde{\boldsymbol{M}}_X$ and $\hat{\boldsymbol{M}}_X$ are Hessian matrices for the objective functions on the pilot subdata and IBOSS subdata, respectively. The estimation performance of the final estimators are evaluated with the empirical mean squared error

$$\text{MSE} = \frac{1}{T}\sum_{t=1}^{T}\|\check{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\|^2, \tag{23}$$

where T is the number of times the simulation was repeated.

6.1 Logistic regression

In this subsection, we consider the logistic regression illustrated in Example 3. The pilot estimator is obtained using the case-control sampling, which is more stable than the uniform sampling for imbalanced data. We follow the settings in Wang et al. (2018) to generate full data. To be precise, full data are generated from the logistic regression model in (19), where we set $\boldsymbol{\theta}$ as a 7 dimension vector with all elements being 0.5. Let $\boldsymbol{\Sigma}$ be a matrix with entries $\boldsymbol{\Sigma}_{ij} = 0.5^{I(i\neq j)}$, where $I(\cdot)$ is the indicator function. The covariates $\boldsymbol{x}_i$'s are generated from the following six distributions.

(a) **mzNormal**, $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$. The generated responses were balanced with around half being 0's and half being 1's.
(b) **nzNormal**, $\boldsymbol{x} \sim N(\boldsymbol{1.5}, \boldsymbol{\Sigma})$. The generated responses were imbalanced with around 5% of 0's and 95% of 1's.
(c) **ueNormal**, $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{W\Sigma W})$, where $\boldsymbol{W}$ is a diagonal matrix with the j-th diagonal element being j^{-1}. This setting yields balanced responses.
(d) **mixNormal**. $\boldsymbol{x} \sim 0.5N(\boldsymbol{1}, \boldsymbol{\Sigma}) + 0.5N(-\boldsymbol{1}, \boldsymbol{\Sigma})$. The covariate distribution was bimodal and the responses were balanced.
(e) **Mvt$_3$**. $\boldsymbol{x} \sim t_3(\boldsymbol{0}, \boldsymbol{\Sigma})/10$. The covariate distribution had heavy tails and the responses were balanced.
(f) **EXP**, components of $\boldsymbol{x}$ are independently generated from an exponential distribution with rate parameter 2. The covariate distribution was skewed, and the responses were imbalanced with around 17% of 0's and 83% of 1's.

We implement the proposed Algorithm 1 (denoted oD), Algorithm 2 (denoted oDsvd), and the T-optimal subdata selection as described in Theorem 5 (denoted oTsvd). For comparison, we also perform the OSMAC method in (Wang et al., 2018) with mMSE and mVc probabilities, IBOSS

based subdata selection method proposed in Cheng et al. (2020) (denoted as CWY's method), and the uniform subsampling without replacement. The tuning parameter in CWY's method is set to be 0.5 for balanced and 2.5 for unbalanced responses as suggested in Cheng et al. (2020).

We first consider the scenario when the subdata size are fixed and the full data size increases. Specifically, we set $N = 5 \times 10^4$, 10^5, 5×10^5, 10^6, 5×10^6 and 10^7 with fixed $n_0 = 1500$ and $n = 2500$ for the proposed algorithms and the OSMAC method. For fair comparison the uniform sampling use a sample size of $n_0 + n = 4000$. The simulations are repeated for $T = 1000$ times. The estimation results are shown in Figure 1.

Algorithm 1 (oD) and Algorithm 2 (oDsvd) outperform OSMAC based procedures and CWY's method under all covariate distributions considered in Figure 1. The oTsvd is better than Algorithms 1 and 2 in cases (a)-(e) with symmetric covariate distributions, but it does not perform well in case (f) in which the covariate distribution is skewed. The three random sampling based methods (uniform, mMSE, and mVc) do not have significant improvement as the dataset size n grows. The decrease of the MSE in mMSE and mVc methods as N increases for small N is because sampling with replacement was used in the sampling step. There were some replicates in the resulting subsample for smaller N which made the resulting estimator not that efficient. The MSE of uniform sampling based estimator is very stable. In contrary, the proposed algorithms have decreasing MSEs as N increases.

To evaluate the impact of the pilot estimator, we fix full sample size $N = 10^6$, optimal subdata size $n = 2500$, and varied pilot subdata size n_0. The empirical MSEs of pre-combined estimators $\hat{\boldsymbol{\theta}}$ are reported in Figure 2 for Cases (a) and (b).

In Figure 2, the MSE decreases as the pilot size n_0 increases from 200 to 1500, and the improvement is very limited when n_0 is greater than 1000 for both two cases. Therefore the pilot estimator is accurate enough when $n_0 = 1500$, and simply fixed $n_0 = 1500$ as we used in this example yielded satisfactory performance.

As discussed in Example 3, the intercept parameter has a different behavior compared with the slope parameters since its corresponding covariate is bounded. To further evaluate our methods, we also consider adding an intercept term $\theta_0 = 0.5$ to the aforementioned logistic regression model. To be precise, we replace $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}$ in (19) in the previous setup by $\theta_0 + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}_1$ and estimate the 8-dimension parameter, where θ_0 is the intercept parameter and $\boldsymbol{\theta}_1$ is the slope parameter. With the added intercept, the percentages of 1's in the generated responses become around 56%, 96%, 60%, 54%, 62%, and 89% for the cases (a)-(f), respectively. To exam the performance of the proposed method on the intercept and the slope separately, we present the MSEs for the intercept and slope parameters in Figure 3 and Figure 4, respectively.

In Figure 3, we see that the proposed procedures fall short compared to OSMAC methods and perform similarly to CWY's method in intercept estimation, except in case (b). In Figure 4, the relative performance between random

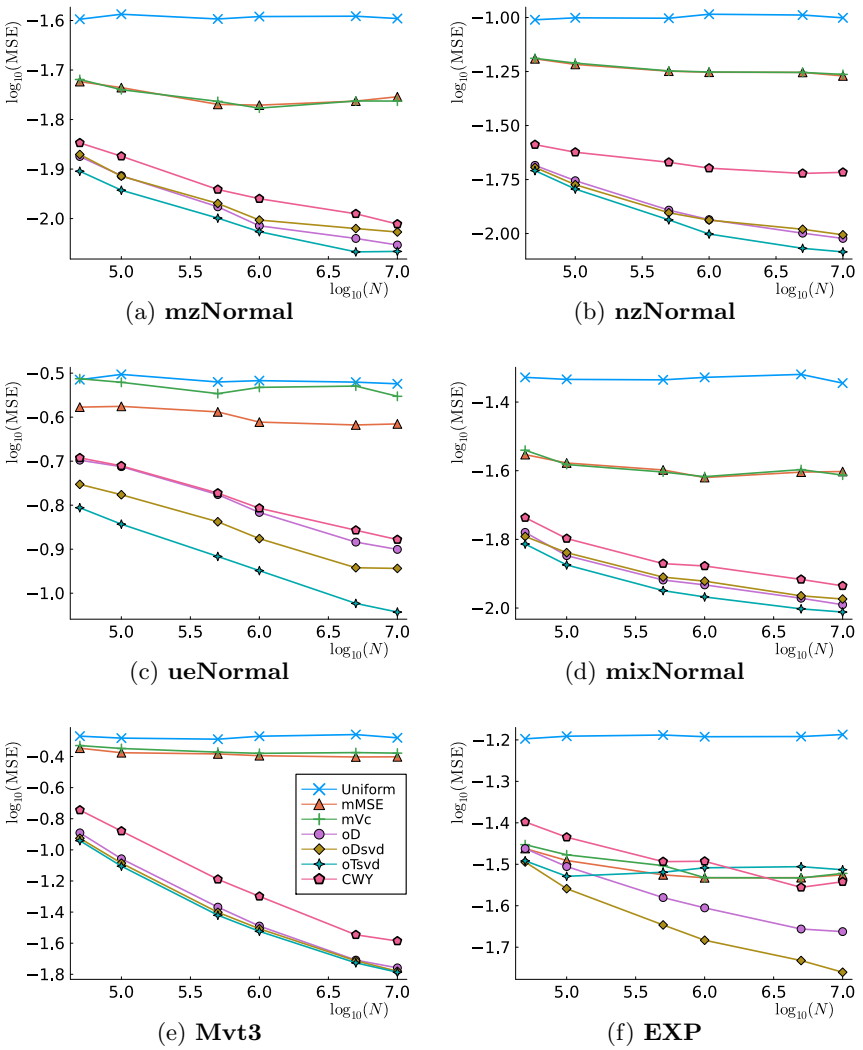
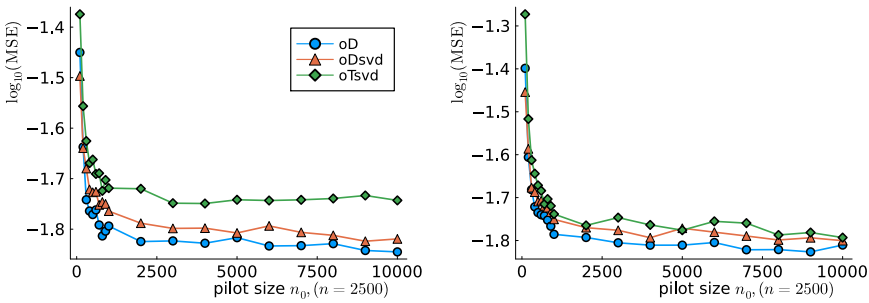


Fig. 1: Empirical MSE of $\check{\boldsymbol{\theta}}$ for logistic regression without an intercept when the full sample size increases.

sampling based methods and the proposed methods in estimating the slope parameter are resembling to the scenario without an intercept, except that a larger full sample size is required for the proposed method to outperform the OSMAC methods in case (f).

Next, we evaluate the impact of the subdata size. The full sample size is fixed at $N = 10^6$ and the subsample size n vary from $n = 1000, 2000, 3000, 4000,$ to 5000. The pilot subsample size $n_0 = 1000$. The results for the logistic regression model with an intercept are given in Figure 5. We



(a) **mzNormal** (b) **nzNormal**

Fig. 2: Empirical MSE of $\hat{\boldsymbol{\theta}}$ in logistic regression without intercept with varying pilot sample size.

Table 1: MSEs and the corresponding computing times using R.

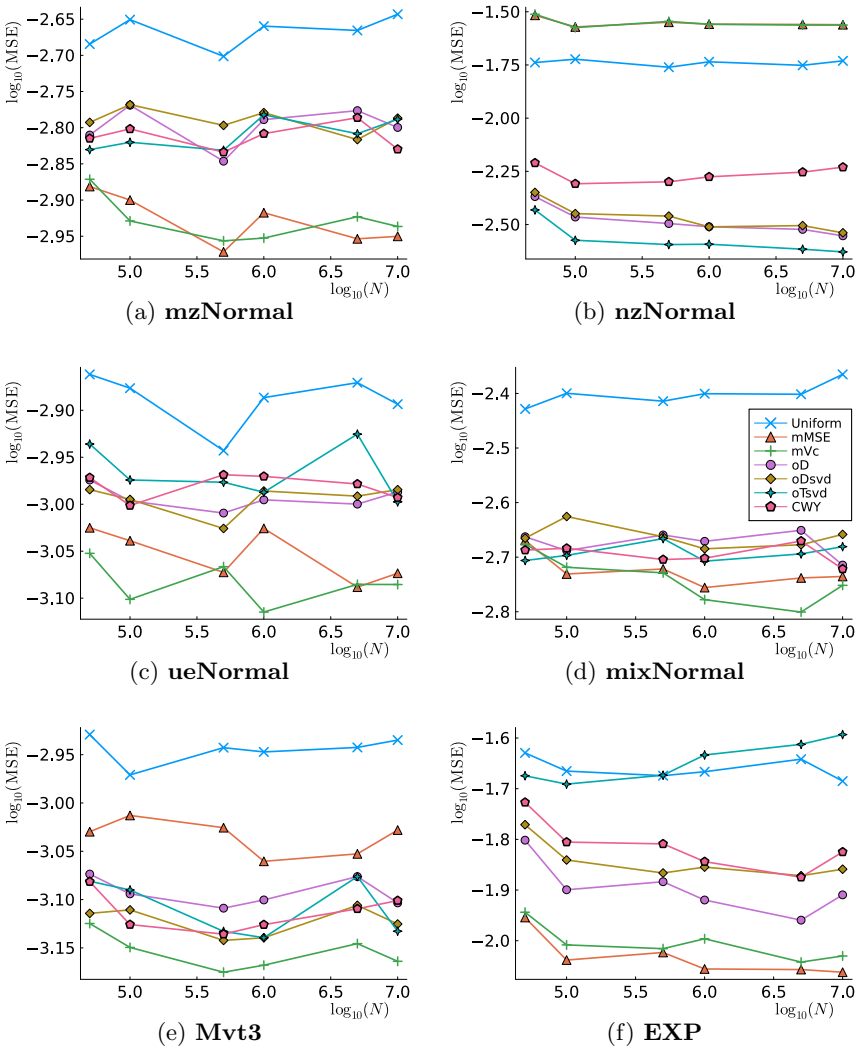
Method	d=7			d=50		
	n	MSE	Time(ms)	n	MSE	Time(ms)
Uniform	5000	0.02038	4.05	52000	0.0901	1132
oD	1000	0.02189	145	10000	0.0856	1056
oDsvd	1000	0.02233	246	10000	0.0831	5061
oTsvd	1000	0.01967	165	10000	0.1161	4657
Uniform	13000	0.00787	8.66	250000	0.0183	5752
oD	4000	0.00763	149	40000	0.0177	1405
oDsvd	4000	0.00804	244	40000	0.0170	5613
oTsvd	4000	0.00781	166	40000	0.0184	5055
Uniform	16000	0.00648	11.1	330000	0.0140	7544
oD	5000	0.00657	147	50000	0.0139	1596
oDsvd	5000	0.00682	249	50000	0.0135	5633
oTsvd	5000	0.00666	178	50000	0.0145	5152

present the results only for Case (a) as an example because results for other cases are similar and thus omitted.

As expected, all methods perform better as subdata size n increases. The relative performance between the proposed methods and the random subsampling based methods are in general similar to those in Figure 3 and Figure 4 for the intercept and slope parameters, respectively.

In the following, we record the computing times together with the MSEs for Case (a) with $d = 7$ and $d = 50$. All the other settings are the same as in Figure 5 except that the subdata size of the uniform sampling method is enlarged so that its MSEs' are comparable with other methods and the pilot sample size is increased to $n_0 = 2000$ for $d = 50$. All the computations were carried out on a Desktop computer with AMD 3950x processor using the R programming language. The results are reported in Table 1.

From Table 1, one sees that the uniform sampling has its own advantage compared with other methods, especially when $d = 7$. This is because the



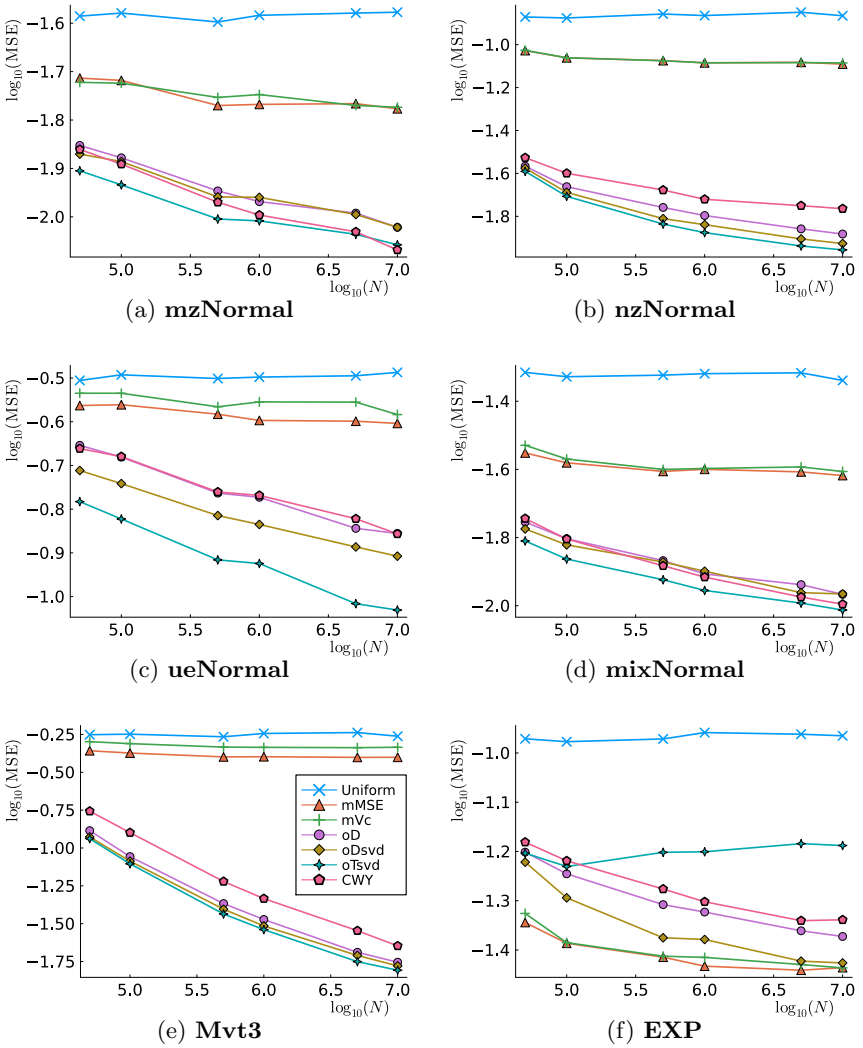
(a) **mzNormal** (b) **nzNormal**

(c) **ueNormal** (d) **mixNormal**

(e) **Mvt3** (f) **EXP**

Fig. 3: Empirical MSE of the intercept estimator $\check{\theta}_0$ in logistic regression with varying full data sample size.

estimation is relatively easier when d is small. However, this does not mean that uniform sampling is always better than the proposed method, because it has higher costs in other aspects. If the available memory is very limited while the computational time is relatively cheap or if the responses are expensive to measure (e.g., data need human annotation as we discussed in Remark 3), the proposed method is more economic. When $d = 50$, the proposed methods may outperform the uniform subsampling methods in terms of computation time, especially when n is large. This is because calculating the parameter



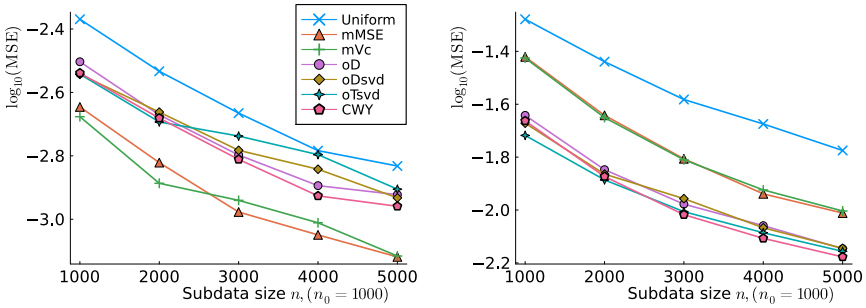
(a) **mzNormal** (b) **nzNormal**

(c) **ueNormal** (d) **mixNormal**

(e) **Mvt3** (f) **EXP**

Fig. 4: Empirical MSE of the slope estimator $\check{\boldsymbol{\theta}}_1$ in logistic regression model with an intercept when the full data sample size increases.

estimates based on the selected subdata becomes harder and requires more time. Note that Newton's method requires $O(\zeta nd^2)$ time, where ζ are the number of iterations. The proposed oD method only needs $O(Nd)$ times for data selection, and it require much more smaller n than the uniform sampling approach in achieving similar MSEs and thus requires much shorter times in parameter estimation.



Empirical MSE of $\check{\theta}_0$ Empirical MSE $\check{\boldsymbol{\theta}}_1$

Fig. 5: Empirical MSEs of intercept estimator $\check{\theta}_0$ and slope estimator $\check{\boldsymbol{\theta}}_1$ in logistic regression with an intercept when subdata size n increases for Case (a) **mzNormal**.

6.2 Poisson regression

In this subsection, we evaluate the proposed methods with the Poisson regression in Example 4. The pilot estimator is obtained from uniform sampling without replacement. Full data is generated from the Poisson regression model given by (20) with an intercept term, i.e., $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}$ is replaced by $\theta_0 + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}_1$ where θ_0 is the intercept parameter and $\boldsymbol{\theta}_1$ is the slope parameters. Here $\boldsymbol{\theta}_1$ is a 6-dimension vector with all elements being 0.25, and θ_0 is also set to be 0.25. The following two distributions are used to generate the covariates.

(a) **Normal** $\boldsymbol{x} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\Sigma_{uv} = 0.5^{|u-v|}$ is the (u, v)-th entry of $\boldsymbol{\Sigma}$. Components of $\boldsymbol{x}$ are dependent with a symmetric distribution whose support are unbounded to both positive and negative infinity.

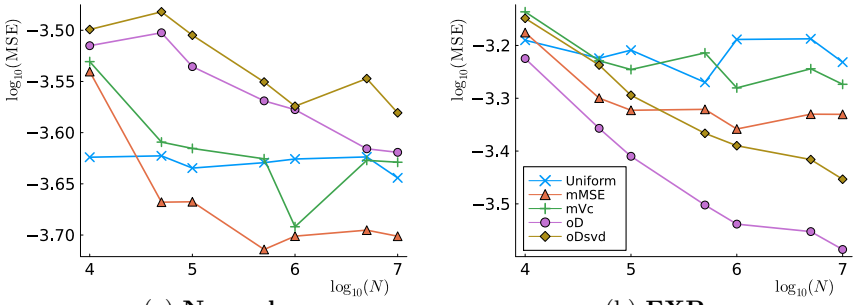
(b) **EXP**, components of $\boldsymbol{x}$ are independent Exponential (2) with rate parameter 2. This distribution is skewed, bounded from below away from negative infinity but unbounded from infinity.

Similarly to the example of logistic regression, we implement the proposed Algorithm 1 (denoted oD) and Algorithm 2 (denoted oDsvd), together with the results from OSMAC with mVc and mMSE probabilities for Poisson regression (Ai et al., 2021), and uniform sampling without replacement for comparisons.

As in the previous subsection, the estimation accuracy for the intercept and slope parameters are reported in Figure 6 and Figure 7, respectively. The full data size N varies from 10^4 to 10^7, and the subdata sizes are fixed at $n_0 = 1500$ for pilot estimates and $n = 2500$ for the proposed methods. Each procedure was repeated for $T = 1000$ times to obtain reliable empirical MSEs.

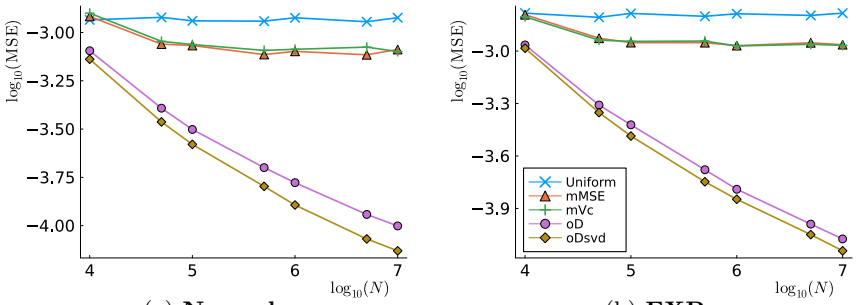
In Figure 7, the proposed methods are significantly better than other methods in terms of estimating the slope parameter, and the overall relative performance is similar to the results for logistic regression reported in Subsection 6.1. It is worth mentioning that the MSE of the intercept term for the proposed methods also decreases under Poisson regression, which is different

from what we have seen under logistic regression. This is because, unlike the case of logistic regression, the sample maximum $|\tilde{\boldsymbol{z}}|_{(N)1}$ goes to infinity as N goes to infinity for the intercept term.



(a) **Normal** (b) **EXP**

Fig. 6: Empirical MSEs of intercept estimator $\check{\theta}_0$ in Poisson regression with varying full sample size.



(a) **Normal** (b) **EXP**

Fig. 7: Empirical MSEs of slope estimator $\check{\boldsymbol{\theta}}_1$ (bottom panel) in Poisson regression with varying full sample size.

Acknowledgments. The authors sincerely thank the editors and referees for their valuable comments and insightful suggestions, which led to further improvement of this article. Yu's work was supported by NSFC grants 12001042, Beijing Institute of Technology Research Fund Program for Young Scholars, and Beijing Municipal Natural Science Foundation No. 1232019. Liu and Wang's research was supported by NSF grant CCF 2105571 and UConn CLAS Research in Academic Themes funding.

Appendix A Proofs

A.1 Proof of Theorem 1

Proof Note that

$$\boldsymbol{\delta}_T^{opt} = \arg\max_{\boldsymbol{\delta}} \operatorname{tr}\left(\mathcal{I}(\boldsymbol{\delta})\right) = \arg\max_{\boldsymbol{\delta}} \delta_i \|\boldsymbol{x}_i\|^2.$$

Thus, the T-optimal subdata is the subdata with the k largest values of $\|\boldsymbol{x}_i\|^2$. □

A.2 Proof of Theorem 2

Before proving Theorem 2, we need the following lemma.

Lemma A.1 *Let $\lambda_1, \ldots, \lambda_d$ be the d eigenvalues of $\mathcal{I}(\boldsymbol{\delta})$ with $\lambda_{\min} = \lambda_1 \leq \ldots \leq \lambda_d = \lambda_{\max}$. Assume that $\operatorname{var}(\boldsymbol{Z}_1^*) \leq \ldots \leq \operatorname{var}(\boldsymbol{Z}_d^*)$ with $\operatorname{var}(\boldsymbol{Z}_j^*)$ being the sample variance for the jth column of $\boldsymbol{Z}^*$. For any subdata of size n, represented by $\boldsymbol{\delta}$, it holds that*

$$(n-1)\lambda_{\min}(\boldsymbol{R}^*)\operatorname{var}(\boldsymbol{Z}_j^*) \leq \lambda_j(\mathcal{I}(\boldsymbol{\delta})) \leq (n-1)\lambda_{\max}(\boldsymbol{R}^*)\operatorname{var}(\boldsymbol{Z}_j^*), \tag{A.1}$$

where $\boldsymbol{R}^*$ is the sample correlation matrix of $\boldsymbol{Z}^*$.

Proof Recall $\boldsymbol{Z}^* = (\boldsymbol{z}_1^*, ..., \boldsymbol{z}_n^*)^{\mathrm{T}}$. Let $\mathbf{1}$ be a $n \times 1$ vector of ones, and $\operatorname{var}(\boldsymbol{Z}_j^*)$ be the sample variance for the jth column of $\boldsymbol{Z}^*$, $j = 1, ..., d$. It follows that

$$\begin{aligned}
\mathcal{I}(\boldsymbol{\delta}, \boldsymbol{\theta}) =& \boldsymbol{Z}^{*\mathrm{T}} \boldsymbol{Z}^* \\
\geq& \boldsymbol{Z}^{*\mathrm{T}} \left(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}\right) \boldsymbol{Z}^* \\
=& (n-1) \begin{bmatrix} \sqrt{\operatorname{var}(\boldsymbol{Z}_1^*)} & & \\ & \ddots & \\ & & \sqrt{\operatorname{var}(\boldsymbol{Z}_d^*)} \end{bmatrix} \mathbf{R}^* \begin{bmatrix} \sqrt{\operatorname{var}(\boldsymbol{Z}_1^*)} & & \\ & \ddots & \\ & & \sqrt{\operatorname{var}(\boldsymbol{Z}_d^*)} \end{bmatrix}.
\end{aligned}$$

Note the fact that for any matrices $A \geq 0$ and $B \geq 0$, it holds that $\lambda_{\min}(B)\lambda_j(A^2) \leq \lambda_j(ABA) \leq \lambda_{\max}(B)\lambda_j(A^2)$. The desired result follows immediately by letting $B = \boldsymbol{R}^*$ and $A = \operatorname{diag}(\sqrt{\operatorname{var}(\boldsymbol{Z}_1^*)}, \ldots, \sqrt{\operatorname{var}(\boldsymbol{Z}_d^*)})$. □

Proof of Theorem 2 Recall that $\lambda_1, \ldots, \lambda_d$ are d eigenvalues of $\mathcal{I}(\boldsymbol{\delta})$ with $0 < \lambda_{\min} = \lambda_1 \leq \ldots \leq \lambda_d = \lambda_{\max}$. We have

$$\{\operatorname{tr}(d^{-1}\mathcal{I}^{-1}(\boldsymbol{\delta}))\}^{-1} = \left(d^{-1}\sum_{j=1}^{d} \lambda_j^{-1}\right)^{-1}, \tag{A.2}$$

$$\{\det(\mathcal{I}(\boldsymbol{\delta}))\}^{1/d} = \left(\prod_{j=1}^{d} \lambda_j\right)^{1/d}. \tag{A.3}$$

Thus (13)–(15) can be easily obtained by Lemma A.1. □

A.3 Proof of Theorem 3

Proof For each sample variance,

$$\begin{aligned}\mathrm{var}(\tilde{\boldsymbol{Z}}_j^*) =& \frac{1}{n-1}\sum_{i=1}^{n}(\tilde{z}_{ij}^* - \bar{\tilde{z}}_j^*)^2 \\ =& \frac{(\tilde{z}_{(N)j} - \tilde{z}_{(1)j})^2}{k-1}\sum_{i=1}^{n}\left(\frac{\tilde{z}_{ij}^* - \bar{\tilde{z}}_j^*}{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}}\right)^2 \\ \geq& \frac{(\tilde{z}_{(N)j} - \tilde{z}_{(1)j})^2}{n-1}\left(\sum_{i=1}^{r} + \sum_{i=N-r+1}^{N}\right)\left(\frac{\tilde{z}_{(i)j} - \bar{\tilde{z}}_j^*}{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}}\right)^2. \end{aligned} \tag{A.4}$$

For the first summation in (A.4),

$$\sum_{i=1}^{r}\left(\frac{\tilde{z}_{(i)j} - \bar{\tilde{z}}_j^*}{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}}\right)^2 = \frac{1}{n^2}\sum_{i=1}^{r}\left(\frac{\sum_{s=1}^{n}\tilde{z}_{sj}^* - n\tilde{z}_{(i)j}}{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}}\right)^2. \tag{A.5}$$

Each term in the summation of (A.5) can be written as

$$\begin{aligned}&\frac{\sum_{s=1}^{n}\tilde{z}_{sj}^* - n\tilde{z}_{(i)j}}{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}} \\ =& \sum_{s=N-r+1}^{N}\frac{\tilde{z}_{(s)j} - \tilde{z}_{(i)j}}{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}} + \sum_{s=1}^{r}\frac{\tilde{z}_{(s)j} - \tilde{z}_{(i)j}}{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}} \\ &+ \sum_{l\neq j}\left(\sum_{s=1}^{r} + \sum_{s=N-r+1}^{N}\right)\frac{\tilde{z}_j^{(s)l} - \tilde{z}_{(i)j}}{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}}, \end{aligned} \tag{A.6}$$

where $\tilde{z}_j^{(s)l}$ is the jth dimension of the subdata point selected according to $\{\tilde{z}_{il}, i = 1, \ldots, N\}$ in the second step of Algorithm 1. From the Assumption 1, we have that for $s, i \leq r$, $(\tilde{z}_{(s)j} - \tilde{z}_{(i)j})/(\tilde{z}_{(N)j} - \tilde{z}_{(1)j}) = o_P(1)$ and $(\tilde{z}_j^{(s)l} - \tilde{z}_{(i)j})/(\tilde{z}_{(N)j} - \tilde{z}_{(1)j})$ is either positive or $o_P(1)$. Thus (A.6) implies

$$\frac{\sum_{s=1}^{n}\tilde{z}_{sj}^* - n\tilde{z}_{(i)j}}{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}} \geq \sum_{s=N-r+1}^{N}\frac{\tilde{z}_{(s)j} - \tilde{z}_{(i)j}}{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}}. \tag{A.7}$$

From assumptions 1 and 2, for $s \geq N - r + 1$ and $i \leq r$, as $N \to \infty$,

$$\frac{\tilde{z}_{(s)j} - \tilde{z}_{(i)j}}{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}} = \frac{\tilde{z}_{(s)j} - \tilde{z}_{(N)j}}{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}} + \frac{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}}{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}} + \frac{\tilde{z}_{(1)j} - \tilde{z}_{(i)j}}{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}} = 1 + o_P(1) \tag{A.8}$$

From (A.5), (A.7) and (A.8),

$$\sum_{i=1}^{r}\left(\frac{\tilde{z}_{(i)j} - \bar{\tilde{z}}_j^*}{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}}\right)^2 \geq \frac{1}{n^2}\sum_{i=1}^{r}\left(\sum_{s=N-r+1}^{N}\frac{\tilde{z}_{(s)j} - \tilde{z}_{(i)j}}{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}} + o_P(1)\right)^2 = \frac{r^3}{n^2} + o_P(1). \tag{A.9}$$

Similarly,

$$\sum_{i=N-r+1}^{N}\left(\frac{\tilde{z}_{(i)j} - \bar{\tilde{z}}_j^*}{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}}\right)^2 \geq \frac{r^3}{n^2} + o_P(1). \tag{A.10}$$

Combining (A.4), (A.9) and (A.10),

$$\mathrm{var}(\tilde{\boldsymbol{Z}}_j^*) \geq \frac{2r^3(\tilde{z}_{(N)j} - \tilde{z}_{(1)j})^2}{n^2(n-1)}(1 + o_P(1)). \tag{A.11}$$

If $\tilde{z}_{(1)j}/\tilde{z}_{(N)j} \xrightarrow{P} 0$ or $\pm\infty$, then

$$\frac{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}}{|\tilde{z}|_{(N)j}} = 1 + o_P(1). \tag{A.12}$$

Combining (A.11) and (A.12) shows that

$$\mathrm{var}(\tilde{\boldsymbol{Z}}_j^*) \geq \frac{2r^3|\tilde{z}|_{(N)j}^2}{n^2(n-1)}(1 + o_P(1)). \tag{A.13}$$

Thus, the desired results come from Theorem 2 and Slutsky's theorem.

If $\tilde{z}_{(N)j} \to \infty$ and $\tilde{z}_{(1)j}$ is bounded below, or $\tilde{z}_{(1)j} \to -\infty$ and $\tilde{z}_{(N)j}$ is bounded above, then

$$\frac{\tilde{z}_{(N)j} - \tilde{z}_{(1)j}}{|\tilde{z}|_{(N)j}} = 1 + o_P(1). \tag{A.14}$$

From (A.11) and (A.14), it follows that

$$\mathrm{var}(\tilde{\boldsymbol{Z}}_j^*) \geq \frac{2r^3|\tilde{z}|_{(N)j}^2}{n^2(n-1)}(1 + o_P(1)). \tag{A.15}$$

Thus, the desired results come from Theorem 2 and Slutsky's theorem.

□

A.4 Some details of Example 3 and Proposition 4

Lemma A.2 *Suppose $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \Sigma)$ with $\Sigma > 0$ and $\boldsymbol{\theta}$ has more than one nonzero elements. Then it holds that $(x_j, \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x})^{\mathrm{T}}$ is still a nondegenerate normal distribution for all j.*

Lemma A.3 *Suppose $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \Sigma)$ with $\Sigma > 0$, and $\boldsymbol{\theta}$ lies in a compact ball with more than one nonzero element. Then for any given M, and $C > 0$, it holds that*

$$\mathbb{P}\Big(\frac{x_j}{e^{-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}/2} + e^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}/2}} > M,\ |\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}| \leq 2C\Big) > 0,$$

$$\mathbb{P}\Big(\frac{x_j}{e^{-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}/2} + e^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}/2}} < -M,\ |\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}| \leq 2C\Big) > 0,$$

for all j.

Proof of Proposition 4 Note that the jth dimension of $\tilde{\boldsymbol{z}}$ can be written as

$$\tilde{z}_j = \frac{x_j}{e^{-\boldsymbol{x}^{\mathrm{T}}\tilde{\boldsymbol{\theta}}/2} + e^{\boldsymbol{x}^{\mathrm{T}}\tilde{\boldsymbol{\theta}}/2}}.$$

For any j, one can see that

$$\mathbb{P}\Big(\frac{x_j}{e^{-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}/2} + e^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}/2}} > M\Big) \geq \mathbb{P}\Big(\frac{x_j}{e^{-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}/2} + e^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}/2}} > M,\ |\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}| \leq 2C\Big) > 0,$$

where M and C are some constant independent of $\boldsymbol{x}$.

Similarly, one can show that

$$\mathbb{P}\Big(\frac{x_j}{e^{-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}/2} + e^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}/2}} < -M\Big) \geq \mathbb{P}\Big(\frac{x_j}{e^{-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}/2} + e^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}/2}} < -M,\ |\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}| \leq 2C\Big) > 0.$$

From Lemma A.3, the desired result follows. □

Details of Example 3 Grounded on the two lemmas, we can see that for any $M > 0$ and $j = 1, \ldots, d$,

$$\mathbb{P}(\tilde{z}_{j,(N)} > M) = 1 - \mathbb{P}(\tilde{z}_{j,(N)} \leq M) \tag{A.16}$$

$$= 1 - \mathbb{P}^N(\tilde{z}_j \leq M) \tag{A.17}$$

$$= 1 - \mathbb{P}^N\Big(\frac{x_j}{e^{-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}/2} + e^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}/2}} \leq M\Big). \tag{A.18}$$

From Lemma A.3, it is clear to see that

$$\mathbb{P}^N\Big(\frac{x_j}{e^{-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}/2} + e^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}/2}} \leq M\Big) \to 0. \tag{A.19}$$

Thus the result follows. □

Proof of Lemma A.2 Without loss of generality, we only consider the case $j = 1$ here. Note that $\boldsymbol{x}$ is a multivariate normal distribution. Let $\boldsymbol{e}_j$ be a unit vector with jth element being one and $C = (\boldsymbol{e}_j, \boldsymbol{\theta})$. Note that $\boldsymbol{\theta}$ has more than one nonzero element by assumption. It is clear to see the rank of C equals two. Since $\Sigma > 0$, the rank of $C^{\mathrm{T}}\Sigma C$ is also two. The $\check{\boldsymbol{x}} = C^{\mathrm{T}}\boldsymbol{x} = (x_j, \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x})^{\mathrm{T}}$ is also a non degenerate normal distribution with mean $C^{\mathrm{T}}\boldsymbol{\mu} = (\mu_j, \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\theta})^{\mathrm{T}}$, and variance

$$C^{\mathrm{T}}\Sigma C = \check{\Sigma} = \begin{pmatrix} \sigma_{jj} & \check{\sigma}_{12} \\ \check{\sigma}_{12}^{\mathrm{T}} & \check{\sigma}_{22} \end{pmatrix},$$

where μ_j is the jth elements in $\boldsymbol{\mu}$, σ_{jj} is the (j,j)th element of Σ, $\check{\sigma}_{12}$ equals to $(\Sigma_{\cdot j}^{\mathrm{T}}\boldsymbol{\theta})$ with $\Sigma_{\cdot j}$ being the jth column of Σ, and $\check{\sigma}_{22} = \boldsymbol{\theta}^{\mathrm{T}}\Sigma\boldsymbol{\theta}$.

□

Proof of Lemma A.3 For the first result, simple calculation yields that

$$\mathbb{P}\Big(\frac{x_j}{e^{-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}/2} + e^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}/2}} > M,\ |\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}| \leq 2C\Big) \tag{A.20}$$

$$\geq \mathbb{P}\Big(x_j > 2e^C M,\ |\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}| \leq 2C\Big) > 0, \tag{A.21}$$

since $(x_j, \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})^{\mathrm{T}}$ are non degenerate normal distribution by the fact proved in Lemma A.2. The second result is quite similar. Thus we omit it.

□

References

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

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

Atkinson, A.C. and V.V. Fedorov. 1975. The design of experiments for discriminating between two rival models. *Biometrika* 62: 57–70 .

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

David, H.A., H.O. Hartley, and E.S. Pearson. 1954. The distribution of the ratio, in a single normal sample, of range to standard deviation. *Biometrika* 41: 482–493 .

Deldossi, L. and C. Tommasi. 2022. Optimal design subsampling from big datasets. *Journal of Quality Technology* 54: 93–101 .

Deville, J.C. and C.E. Särndal. 1992. Calibration estimators in survey sampling. *Journal of the American statistical Association 87*(418): 376–382 .

Drovandi, C.C., C. Holmes, J.M. McGree, K. Mengersen, S. Richardson, and E.G. Ryan. 2017. Principles of experimental design for big data analysis. *Statistical science: a review journal of the Institute of Mathematical Statistics 32*(3): 385 .

Efron, B. and D.V. Hinkley. 1978. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika 65*(3): 457–483 .

Hartley, H.O. and J.N.K. Rao. 1962. Sampling with unequal probabilities and without replacement. *The Annals of Mathematical Statistics* 33: 350 – 374 .

Horn, R.A. and C.R. Johnson. 2013. *Matrix Analysis* (2nd ed.). Cambridge University Press.

Kiefer, J. 1959. Optimum experimental designs. *Journal of the Royal Statistical Society. Series B (Methodological) 21*(2): 272–319 .

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

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

Martínez, C. 2004. On partial sorting. Technical report, 10th Seminar on the Analysis of Algorithms.

Meng, C., R. Xie, A. Mandal, X. Zhang, W. Zhong, and P. Ma. 2020. Lowcon: A design-based subsampling approach in a misspecified linear model. *Journal of Computational and Graphical Statistics*: in press. https://doi.org/10.1080/10618600.2020.1844215 .

Montgomery, D.C. 2019. *Introduction to Statistical Quality Control (8th Edition)*. Introduction to Statistical Quality Control.

Musser, D.R. 1997. Introspective sorting and selection algorithms. *Software: Practice and Experience 27*(8): 983–993 .

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

Särndal, C.E., B. Swensson, and J. Wretman. 1992. *Model assisted survey sampling*. Springer.

Schmidt, D. and R. Schwabe. 2017. Optimal design for multiple regression with information driven by the linear predictor. *Statistica Sinica* 27: 1371–1384 .

Tippett, L.H.C. 1925. On the extreme individuals and the range of samples taken from a normal population. *Biometrika* 17: 364–387 .

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

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

Wang, H., R. Zhu, and P. Ma. 2018. Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association* 113: 829–844. https://doi.org/10.1080/01621459.2017.1292914 .

Wang, L., J. Elmstedt, W.K. Wong, and H. Xu. 2021. Orthogonal subsampling for big data linear regression. *The Annals of Applied Statistics* 15: 1273–1290 .

Xie, R., Z. Wang, S. Bai, P. Ma, and W. Zhong 2019. Online decentralized leverage score sampling for streaming multidimensional time series. In K. Chaudhuri and M. Sugiyama (Eds.), *Proceedings of Machine Learning Research*, Volume 89 of *Proceedings of Machine Learning Research*, pp. 2301–2311. PMLR.

Yang, M., B. Zhang, and S. Huang. 2011. Optimal designs for generalized linear models with multiple design variables. *Statistica Sinica* 21: 1415–1430 .

Yu, J., M. Ai, and Z. Ye. 2023. A review on design inspired subsampling for big data. *Statistical Papers*. https://doi.org/10.1007/s00362-022-01386-w .

Yu, J. and H. Wang. 2022. Subdata selection algorithm for linear model discrimination. *Statistical Papers*: in press. https://doi.org/10.1007/s00362-022-01299-8 .

Zhang, T., Y. Ning, and D. Ruppert. 2021. Optimal sampling for generalized linear models under measurement constraints. *Journal of Computational and Graphical Statistics* 30: 106–114 .

Zhao, Y., Y. Amemiya, and Y. Hung. 2018. Efficient Gaussian process modeling using experimental design-based subagging. *Statistica Sinica* 28: 1459–1479 .