Statistics and Computing (2025) 35:15
https://doi.org/10.1007/s11222-024-10546-x

ORIGINAL PAPER

Optimal subsampling for generalized additive models on large-scale datasets

Lili Li[1] · Bingfan Liu[2] · Xiaodi Liu[1] · Haolun Shi[2] · Jiguo Cao[2]

Received: 17 May 2024 / Accepted: 3 December 2024

Abstract
In the age of big data, the efficient analysis of vast datasets is paramount, yet hindered by computational limitations such as memory constraints and processing duration. To tackle these obstacles, conventional approaches often resort to parallel and distributed computing methodologies. In this study, we present an innovative statistical approach that exploits an optimized subsampling technique tailored for generalized additive models (GAM). Our approach harnesses the versatile modeling capabilities of GAM while alleviating computational burdens and enhancing the precision of parameter estimation. Through simulations and a practical application, we illustrate the efficacy of our method. Furthermore, we provide theoretical support by establishing convergence assurances and elucidating the asymptotic properties of our estimators. Our findings indicate that our approach surpasses uniform sampling in accuracy, while significantly reducing computational time in comparison to utilizing complete large-scale datasets.

Keywords Efficient computation · Big data analysis

1 Introduction

The rapid technological advancements in recent years have made it possible to access an unprecedented amount of data. However, the abundance of data also introduces significant analytical challenges, particularly in terms of computational constraints such as memory and processing time limitations, which often render traditional computational methods impractical. Consequently, subsampling methods for efficient model estimation have become a crucial area of research, particularly in big data scenarios where analyzing the entire dataset is computationally infeasible.

Subsampling techniques have seen significant progress across various models. For instance, Drineas et al. (2006), Mahoney and Drineas (2009), and Drineas et al. (2011) introduce random sampling approaches for linear regression that successfully extract key information from large datasets. Wang et al. (2021) propose an orthogonal subsampling (OSS) method for estimating linear regression models in big data contexts. Further contributions include the entropy-based subsampling method by Sui and Ghosh (2024), which efficiently samples data while measuring information loss, and the algorithmic leveraging methods introduced by Ma et al. (2014) and Zhu et al. (2015), which utilize empirical leverage scores to inform non-uniform sampling. Wang et al. (2019) propose the information-based optimal subsample selection (IBOSS) technique for linear regression, later extended to distributed data by Zhang and Wang (2021).

In the realm of logistic regression, Wang et al. (2018) present a subsampling technique that optimizes the A- and L-optimality criterion. Generalizing this method for the Generalized Linear Model (GLM), Ai et al. (2021) introduce the Optimal Subsampling Method under the A-optimality Cri-

Lili Li and Bingfan Liu have contributed equally to this work.

✉ Jiguo Cao
jiguo_cao@sfu.ca

Lili Li
lili_lee2003@126.com

Bingfan Liu
bingfan_liu@sfu.ca

Xiaodi Liu
xd17616029356@163.com

Haolun Shi
haolun_shi@sfu.ca

[1] Economic Statistic Department, Qingdao University, 308 Ning Xia Road, Qingdao 266071, Shandong, China

[2] Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby V5A 1S6, BC, Canada

Published online: 15 December 2024

terion (OSMAC). Furthermore, Zuo et al. (2021) extended OSMAC to a distributed data scenario.

In the domain of softmax regression, also known as multinomial logistic regression, Yao et al. (2023a) reveal that optimal subsampling strategies can be influenced by model constraints; they propose a subsampling approach invariant to such constraints, minimizing the asymptotic expectation of the mean squared prediction error. Furthermore, Yao et al. (2023b) introduce an optimal Poisson sampling method for softmax regression. For Poisson regression settings, Yu et al. (2022) propose a Poisson subsampling method incorporating A- and L-optimality criteria for maximum quasi-likelihood estimation for distributed data. In a broader context, Yu et al. (2023) unified the problem, introducing a subsampling approach for non-linear models such as accelerated failure time models, and general nonlinear regression models, optimizing design criteria including A-, D-, E-, and T-optimality.

Moreover, subsampling techniques are developed for more complex models. Lee et al. (2022) propose a non-uniform subsampling method tailored to finite mixtures of Gaussian regression models, focusing on minimizing squared errors and reducing computational burden. For a comprehensive overview of optimal subsampling methods in the context of big data scenarios, readers are encouraged to refer to Yao and Wang (2021) and Yu et al. (2024) for a detailed review.

While numerous subsampling methods have been developed for parametric and non-parametric models, there remains a notable gap in research concerning optimal subsampling for semi-parametric models, particularly in big data contexts. GAM stands out as a potent semi-parametric model, as it surpasses linear models in its ability to capture complex nonlinear relationships through non-parametric covariates while maintaining fixed-form parametric variables, making it highly effective for modeling intricate data patterns as explained by Hastie (2017).

Although GAMs provide efficient parameter estimation and favorable convergence rates, traditional estimation methods become computationally impractical for very large datasets. The iterative reweighting process used for estimation leads to complexity and memory usage that scale quadratically with the dataset size. This article addresses this challenge by developing and evaluating optimal subsampling techniques tailored specifically for GAMs. Our proposed method reduces the computational complexity from quadratic to linear in relation to dataset size, while keeping memory usage independent of the full dataset size.

Unlike GLMs, GAMs comprises both parametric and non-parametric covariates, the latter represented by reduced rank smoothing splines under sum-to-zero identifiability constraints. This unique structure renders the derivation of optimal subsampling probability (SSP) for GAMs a challenging task, distinct from that for GLMs. Our approach begins with obtaining initial parameter estimates through uniform random sampling. We then derive the optimal SSP based on A-optimality introduced by Kiefer (1959) and L-optimality criteria introduced by Atkinson et al. (2007). Additionally, we propose a novel estimation technique that combines optimal and random subsamples, offering a refined method for parameter estimation. To support our methodology, we present rigorous mathematical proofs to demonstrate the convergence and asymptotic normality of our estimators.

The subsequent sections of this paper unfold as follows: Sect. 2 introduces the proposed optimal subsampling algorithm for GAMs and its asymptotic properties. Section 3 presents the simulation studies designed to showcase the finite sample performance of the proposed subsampling methods for GAMs, especially in comparison to uniform random sampling. Section 4 demonstrates the proposed method with a real-world application. Section 5 concludes the paper with some discussion.

2 Optimal subsampling method for GAM

2.1 Generalized additive models

Consider a dataset comprising n subjects, where each subject i is characterized by its observed values $(\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}, y_i)$ where $\boldsymbol{x}_i^{(1)} = [x_{i1}^{(1)}, \ldots, x_{id_1}^{(1)}]^\top \in \mathbb{R}^{d_1}$ represents the vector of parametric covariates of dimension d_1 and $\boldsymbol{x}_i^{(2)} = [x_{i1}^{(2)}, \ldots, x_{id_2}^{(2)}]^\top \in \mathbb{R}^{d_2}$ is the vector of non-parametric covariates of dimension d_2 for $i = 1, ..., n$. The scalar-valued response variable of subject i is denoted by y_i. Thus, GAM can be expressed as

$$\begin{aligned} y_i &\sim \mathrm{EF}(\mu_i, \phi), \\ g(\mu_i) &= \boldsymbol{x}_i^{(1)\top}\boldsymbol{\delta}^{(1)} + \sum_j f_j\left(x_{ij}^{(2)}\right), \end{aligned}$$

where EF represents a distribution from exponential family with $\mu_i = \mathrm{E}(y_i)$ and fixed scale parameter ϕ. Furthermore, $g(\cdot)$ denotes the link function. The coefficient vector of the parametric covariates is $\boldsymbol{\delta}^{(1)} \in \mathbb{R}^{d_1}$. Denote $f_j(\cdot)$ to be the smoothing function that relaxes the functional form of $x_{ij}^{(2)}$ for $j = 1, \ldots, d_2$ and be constrained under sum-to-zero identifiability condition, i.e., $\sum_i f_j(x_{ij}^{(2)}) = 0$. Each $f_j(x_{ij}^{(2)})$ can be expressed as a linear expansion using a chosen set of basis functions evaluated at their corresponding values denoted as $\boldsymbol{b}_{ij} = [b_{1_j}(x_{ij}^{(2)}), \ldots, b_{K_j}(x_{ij}^{(2)})]^\top$ where K_j represents the chosen number of basis functions for the j-th covariate. Thus,

$$f_j\left(x_{ij}^{(2)}\right) = \boldsymbol{b}_{ij}^\top \boldsymbol{\delta}_j^{(2)}, \text{ for } \forall j \in \mathbb{R}^{d_2},$$

where $\boldsymbol{\delta}_j^{(2)}$ is the associated coefficient vector for j-th non-parametric covariate.

The objective function of GAM can be formulated as a penalized log-likelihood function as

$$\begin{aligned} l &= \sum_{i=1}^{n} l_i - \sum_{j=1}^{d_2} \lambda_j \int \{\ddot{f}_j(x_{ij}^{(2)})\}^2 \mathrm{d}x_{ij}^{(2)} \\ &= \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - \psi(\theta_i)}{\alpha(\phi)} + c(y_i, \phi) \right\} \\ &\quad - \sum_{j=1}^{d_2} \lambda_j \int \{\ddot{f}_j(x_{ij}^{(2)})\}^2 \mathrm{d}x_{ij}^{(2)}, \end{aligned} \tag{1}$$

where l_i is the log-likelihood of i-th data point. $\theta_i = g(\mu_i)$ represents the unknown canonical parameter. $\alpha(\cdot)$, $\psi(\cdot)$ are functions of θ and ϕ respectively. Denote $c(\cdot)$ to be a function of response y_i and fixed scale parameter ϕ. Additionally,

$$\begin{aligned} \int \{\ddot{f}_j(x_{ij}^{(2)})\}^2 \mathrm{d}x_{ij}^{(2)} &= \int \{\ddot{\boldsymbol{b}}_j^\top \boldsymbol{\delta}_j^{(2)}\}^2 \mathrm{d}x_{ij}^{(2)} \\ &= \int \boldsymbol{\delta}_j^{(2)\top} \ddot{\boldsymbol{b}}_j \ddot{\boldsymbol{b}}_j^\top \boldsymbol{\delta}_j^{(2)} \mathrm{d}x_{ij}^{(2)} \\ &= \boldsymbol{\delta}_j^{(2)\top} \int \ddot{\boldsymbol{b}}_j \ddot{\boldsymbol{b}}_j^\top \mathrm{d}x_{ij}^{(2)} \boldsymbol{\delta}_j^{(2)} \\ &= \boldsymbol{\delta}_j^{(2)\top} \boldsymbol{S}_j \boldsymbol{\delta}_j^{(2)}, \end{aligned}$$

where $\ddot{f}_j(\cdot)$ denotes the second order derivative of the function $f_j(\cdot)$. The vector $\ddot{\boldsymbol{b}}_j = [\ddot{b}_{1_j}(x_{ij}^{(2)}), \ldots, \ddot{b}_{K_j}(x_{ij}^{(2)})]^\top$ represents the second order derivative of basis functions of $x_{ij}^{(2)}$. Let λ_j be the smoothing parameter which is treated as the tuning parameter in the model for $j = 1, \ldots, d_2$. Matrix $\boldsymbol{S}_j$ is the integration of the covariance matrix of the vector of basis functions for j-th covariate, which does not depend on $x_{ij}^{(2)}$ after the integration.

For notation simplicity, we employ an n-by-p matrix $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]^\top$, where $\boldsymbol{X}_i = [\boldsymbol{x}_i^{(1)\top}, \boldsymbol{b}_{i1}^\top, \ldots, \boldsymbol{b}_{id_2}^\top]^\top$ for $i = 1, \ldots, n$, to represent the design matrix. The corresponding coefficient vector in the model is denoted as $\boldsymbol{\beta} = [\boldsymbol{\delta}^{(1)\top}, \boldsymbol{\delta}_1^{(2)\top}, \ldots, \boldsymbol{\delta}_{d_2}^{(2)\top}]^\top$. Consequently, (1) can be further expressed as

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} l_i(\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{S}_\lambda \boldsymbol{\beta} \\ &= \sum_{i=1}^{n} \left\{ \frac{y_i\theta - \psi(\theta)}{\alpha(\phi)} + c(y_i, \phi) \right\} - \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{S}_\lambda \boldsymbol{\beta}, \end{aligned} \tag{2}$$

where $\boldsymbol{S}_\lambda$ is a zero-padded block-wise matrix with non-zero diagonal blocks given by $\lambda_j \boldsymbol{S}_j$ for $j = 1, \ldots, d_2$.

Denote the score and Fisher information of $l(\boldsymbol{\beta})$ to be $s(\boldsymbol{\beta})$ and $I(\boldsymbol{\beta})$.

$$s(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} w_i a_i (y_i - \mu_i) \boldsymbol{X}_i - \boldsymbol{S}_\lambda \boldsymbol{\beta}, \tag{3}$$

$$I(\boldsymbol{\beta}) = E\left\{ -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \right\} = \sum_{i=1}^{n} w_i \boldsymbol{X}_i \boldsymbol{X}_i^\top - \boldsymbol{S}_\lambda, \tag{4}$$

where $w_i = [\mathrm{var}(y_i)(\frac{\partial g}{\partial \mu_i})^2]^{-1}$ and $a_i = \frac{\partial g}{\partial \mu_i}$. Given the values of λ_j for $j = 1, \ldots, d_2$, the parameter vector $\boldsymbol{\beta}$ can be estimated by using the penalized iteratively re-weighted log-likelihood method on the objective function (2) as

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = \widehat{\boldsymbol{\beta}}^{(t)} + I\left(\widehat{\boldsymbol{\beta}}^{(t)}\right)^{-1} s\left(\widehat{\boldsymbol{\beta}}^{(t)}\right), \tag{5}$$

where t is the iteration index.

2.2 A general subsampling method for GAM

In situations where the size of the dataset n is exceptionally large, computing the penalized log-likelihood estimate from the full dataset, $\widehat{\boldsymbol{\beta}}$, becomes challenging. To address this issue, subsampling approaches emerge as viable solutions. In this section, we delve into a general subsampling method for GAM which is showcased in Algorithm 1. The algorithm generates a general subsampling estimate, $\widetilde{\boldsymbol{\beta}}$, computed on a subsample of size $r(\ll n)$ to approximate $\widehat{\boldsymbol{\beta}}$. Without loss of generality, we choose the uniform subsampling method as the general subsampling method used in this paper.

Algorithm 1: General subsampling algorithm for GAM

General Subsampling:

Denote the SSP of each subject in the full dataset to be π_i for $i = 1, \ldots, n$.

Draw a subsample of size r based on SSP π_i from uniform random sampling.

Estimation under General Subsampling:

Denote the size, log-likelihood, re-calculated SSP of each subject and $\boldsymbol{S}_\lambda$ in the subsample to be as r, l_i^*, π_i^* and $\boldsymbol{S}_\lambda^*$ for $i = 1, \ldots, r$. It is worth noting that, under uniform random sampling, subsampling probabilities $\pi_i = \pi_i^* = \frac{1}{n}$ are assumed for each unique subject i.

Given the predefined maximum number of iteration T, if the iteration index $t < T$, iteratively estimate $\widetilde{\boldsymbol{\beta}}$ by minimizing the penalized re-weighted log-likelihood defined on the subsample

$$l^*(\boldsymbol{\beta}) = \frac{1}{r}\sum_{i=1}^{r} \frac{l_i^*(\boldsymbol{\beta})}{\pi_i^*} - \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{S}_\lambda^* \boldsymbol{\beta} \tag{6}$$

using Equation (5).

We establish the consistency and asymptotic normality for this general subsampling estimator, $\widetilde{\boldsymbol{\beta}}$, leading to an enhanced

sampling procedure for obtaining optimal subsample. To derive these results, we make the following assumptions:

Assumption 1 $\boldsymbol{D}_x = \frac{1}{n}\left(\sum_{i=1}^{n} w_i \boldsymbol{X}_i \boldsymbol{X}_i^\top - \boldsymbol{S}_\lambda\right)$ is a positive-definite matrix.

Assumption 2 $\frac{1}{n^2}\sum_{i=1}^{n} \frac{\|\boldsymbol{X}_i\|^k}{\pi_i} = O_P(1)$ for $k = 2, 4$ where $||\cdot||$ is the Euclidean norm.

Assumption 3 $\frac{1}{n^{2+\tau}}\sum_{i=1}^{n} \frac{\|\boldsymbol{X}_i\|^{2+\tau}}{\pi_i^{1+\tau}} = O_P(1)$ when $\tau > 0$.

Assumption 4 $\frac{1}{n^2 r^2}\left(\sum_{i=1}^{n} w_i \boldsymbol{X}_i \boldsymbol{X}_i^\top\right)^2 = O_P(1)$.

The aforementioned assumptions establish general moment conditions on covariates distributions, essential for deriving the estimator's properties. These conditions can be easily satisfied when covariates follow a sub-Gaussian distribution, as noted in Buldygin and Kozachenko (1980) and Wang et al. (2018). Given Assumptions 1 and 2, we present the following lemma to aid in deriving the properties of $\widetilde{\boldsymbol{\beta}}$.

Lemma 1 *If Assumptions 1 and 2 hold, conditional on σ-field $\mathcal{F}_n$,*

$$\widetilde{\boldsymbol{D}}_x - \boldsymbol{D}_x = O_{P|\mathcal{F}_n}\left(r^{-1/2}\right), \tag{7}$$

$$\frac{1}{n}\frac{\partial l^*(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = O_{P|\mathcal{F}_n}\left(r^{-1/2}\right), \tag{8}$$

where

$$\begin{aligned}\widetilde{\boldsymbol{D}}_x &= \mathrm{E}\left[\frac{1}{n}\frac{\partial^2 l^*(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top}\right] \\ &= \frac{1}{n}\left(\frac{1}{r}\sum_{i=1}^{r}\frac{1}{\pi_i^*} w^{*i} \boldsymbol{X}_i^* \boldsymbol{X}_i^{*\top} - \boldsymbol{S}_\lambda\right), \\ \boldsymbol{D}_x &= \mathrm{E}\left[\frac{1}{n}\frac{\partial^2 l(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top}\right] \\ &= \frac{1}{n}\left(\sum_{i=1}^{n} w_i \boldsymbol{X}_i \boldsymbol{X}_i^\top - \boldsymbol{S}_\lambda\right).\end{aligned} \tag{9}$$

where $\boldsymbol{X}_i^*$ and w_i^* represent the covariate vector and weight parameter of the subsample respectively.

Lemma 1 provides insights into the first and second-order condition of the penalized re-weighted log-likelihood function evaluated at $\widehat{\boldsymbol{\beta}}$. This lemma sets the stage for the following Theorem 1 which establishes the consistency of the general estimator, $\widetilde{\boldsymbol{\beta}}$ with respect to the full-data estimator, $\widehat{\boldsymbol{\beta}}$. All proofs are deferred to the Appendix section at the end of this paper.

Theorem 1 *If Assumptions 1 and 2 hold, as $n \to \infty$, $\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}} = O_{P|\mathcal{F}_n}(r^{-1/2})$.*

Theorem 1 asserts that, under the conditions of Assumptions 1 and 2, the general subsample estimator $\widetilde{\boldsymbol{\beta}}$ is consistent with $\widehat{\boldsymbol{\beta}}$ in probability conditional on σ-field $\mathcal{F}_n$. As subsample size r increases, the difference between $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$ diminishes at the rate of $r^{-1/2}$. Furthermore, we present the asymptotic distribution of this difference in Theorem 2.

Theorem 2 *If Assumption 3 hold, as $n \to \infty$ and $r \to \infty$, conditional on σ-field $\mathcal{F}_n$,*

$$\boldsymbol{V}^{-1/2}(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{I}),$$

where

$$\boldsymbol{V} = \boldsymbol{D}_x^{-1}\boldsymbol{V}_c\boldsymbol{D}_x^{-1} = O_P(r^{-1}), \tag{10}$$

$$\begin{aligned}\boldsymbol{V}_c &= \mathrm{Var}\left\{\frac{1}{n} l^*(\widehat{\boldsymbol{\beta}})\right\} \\ &= \frac{1}{rn^2}\sum_{i=1}^{n}\frac{w_i^2(y_i - \mu_i)^2 \boldsymbol{X}_i \boldsymbol{X}_i^\top}{\pi_i} = O_{P|\mathcal{F}_n}(r^{-1}).\end{aligned} \tag{11}$$

Theorem 2 asserts that, under Assumption 3, as $n \to \infty$ and $r \to \infty$, conditional on σ-field $\mathcal{F}_n$, the asymptotic distribution of $\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}$ is a multivariate normal distribution with mean $\boldsymbol{0}$ and variance $\boldsymbol{V}$, i.e., $N(\boldsymbol{0}, \boldsymbol{V})$.

2.3 An optimal subsampling method for GAM

In this subsection, we introduce an optimal subsampling method designed to improve the general subsampling method, which relies on uniform random sampling. The optimal subsampling method aims to decrease the variance of the error $\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}$ while minimizing the asymptotic mean squared error between $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$, under the condition that $\widetilde{\boldsymbol{\beta}}$ is consistent with $\widehat{\boldsymbol{\beta}}$.

In this context, $\boldsymbol{V}$, the variance of error $\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}$, is a matrix. To diminish the variance of $\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}$, we draw inspiration from the concept of A-optimality from experimental design and define the trace of $\boldsymbol{V}$, $\mathrm{tr}(\boldsymbol{V})$, as the measure of the variance matrix of $\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}$. Notably, when minimizing $\mathrm{tr}(\boldsymbol{V})$, one is exactly minimizing the asymptotic mean squared error between $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$. Consequently, we present the following theorem, outlining the derivation of the optimal SSP.

Theorem 3 *If SSP satisfy*

$$\pi_i^{\mathrm{mmse}} = \frac{|w_i||y_i - \mu_i|\|\boldsymbol{D}_x^{-1}\boldsymbol{X}_i\|}{\sum_{j=1}^{n}|w_j||y_j - \mu_j|\|\boldsymbol{D}_x^{-1}\boldsymbol{X}_j\|}, \tag{12}$$

for i=1, …, n, tr($\boldsymbol{V}$) achieves minimum.

The significance of Theorem 3 lies in its revelation that the optimal SSP, denoted as π_i^{mmse}, is contingent upon $||\boldsymbol{D}_x^{-1}\boldsymbol{X}_i||$ and $|y_i - \mu_i|$. This dependency manifests through the interplay of covariates and the absolute error between y_i and μ_i. Specifically, with regard to the effect of the response, the theorem implies that a higher prediction error corresponds to a greater SSP. In essence, Theorem 3 suggests that, to achieve the minimum asymptotic mean squared error, one should prioritize selecting data points associated with a substantial loss.

To calculate π_i^{mmse}, it is necessary to evaluate $\|\boldsymbol{D}_x^{-1}\boldsymbol{X}_i\|$ for $i = 1, \ldots, n$ which takes $O(np^2)$ operations. Note that π_i^{mmse} minimizes the trace of $\boldsymbol{V} = \boldsymbol{D}_x^{-1}\boldsymbol{V}_c\boldsymbol{D}_x^{-1}$ through $\boldsymbol{V}_c$. Recognizing that computing π_i^{mmse} involves significant computational complexity, we shift our focus to the variance matrix $\boldsymbol{V}_c$ and propose an alternative optimal SSP, denoted as $\pi_i^{\text{mV}_c}$. As outlined in Theorem 4, $\pi_i^{\text{mV}_c}$ minimizes $\text{tr}(\boldsymbol{V}_c)$ while demanding less computation for its calculation.

Theorem 4 If the SSP is chosen such that

$$\pi_i^{\text{mV}_c} = \frac{|w_i||y_i - \mu_i|\|\boldsymbol{X}_i\|}{\sum_{j=1}^n |w_j||y_j - \mu_j|\|\boldsymbol{X}_j\|},$$

$\text{tr}(\boldsymbol{V}_c)$ *achieves minimum.*

By leveraging Theorem 2, we can deduce that

$$\boldsymbol{D}_x(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{V}_c).$$

This indicates that $\text{tr}(\boldsymbol{V}_c)$ represents the asymptotic mean squared error of $\boldsymbol{D}_x\widetilde{\boldsymbol{\beta}}$, denoted as $\text{tr}(\boldsymbol{V}_c) = \text{AMSE}(\boldsymbol{D}_x\widetilde{\boldsymbol{\beta}})$. Therefore, the SSP provided by Theorem 4 effectively minimizes $\text{AMSE}(\boldsymbol{D}_x\widetilde{\boldsymbol{\beta}})$. Calculating $\pi_i^{\text{mV}_c}$ requires $O(np)$ operations to compute $\|\boldsymbol{X}_i\|$, significantly reducing the computational burden compared to $O(np^2)$ operations for evaluating $\|\boldsymbol{D}_x^{-1}\boldsymbol{X}_i\|$ in the case of π_i^{mmse}, especially when dealing with high-dimensional design matrices. Moreover, $\pi_i^{\text{mV}_c}$ exhibits similar behavior to π_i^{mmse} in subsampling, as it also prioritizes the selection of units associated with large absolute estimation errors, $|y_i - \mu_i|$. Notably, the criterion to minimize $\text{tr}(\boldsymbol{V}_c)$ aligns with the concept of L-optimality in experimental design, which seeks to minimize the trace of a linear transformation of the asymptotic covariance matrix.

However, computing either π_i^{mmse} or $\pi_i^{\text{mV}_c}$ necessitates the estimation of coefficient vector, $\widehat{\boldsymbol{\beta}}$, on the entire dataset which is impractical to compute in the context of large data scenarios. Consequently, directly applying these optimal SSP with Algorithm 1 to estimate model coefficients is not feasible. To address this, we introduce a two-stage algorithm, outlined in Algorithm 2. Throughout the remainder of the article, we refer to the optimal coefficient estimator obtained from this two-stage algorithm as $\breve{\boldsymbol{\beta}}$.

Algorithm 2: Two-stage optimal subsampling algorithm for estimating (5) in GAM

Stage 1:

Draw a subsample of size r_0 using uniform random sampling with SSP $\pi_i = 1/n$.

Given the pre-defined maximum number of iteration T, obtain the estimated model coefficients $\widetilde{\boldsymbol{\beta}}_0$ as an approximation of the value of $\widehat{\boldsymbol{\beta}}$.

Replace $\widehat{\boldsymbol{\beta}}$ by $\widetilde{\boldsymbol{\beta}}_0$, and compute optimal SSP π_i^{mmse} and $\pi_i^{\text{mV}_c}$.

Stage 2:

Draw a second subsample of size r with replacement using optimal subsampling with the optimal SSPs, π_i^{mmse} and $\pi_i^{\text{mV}_c}$.

Given T, obtain the estimate $\breve{\boldsymbol{\beta}}$ using the combined subsamples of size $r_0 + r$.

Algorithm 2 outlines a two-stage optimal subsampling procedure for estimating the optimal coefficients $\breve{\boldsymbol{\beta}}$. In the first stage, the algorithm approximates the value of $\widehat{\boldsymbol{\beta}}$ and computes an estimate of the optimal SSP. The approximation estimator, $\widetilde{\boldsymbol{\beta}}_0$, obtained in this stage is consistent, as demonstrated in Theorem 1. Additionally, its asymptotic variance is bounded by r_0^{-1} in probability. While $\widetilde{\boldsymbol{\beta}}_0$ can provide a relatively accurate approximation of $\widehat{\boldsymbol{\beta}}$ as r_0 approaches closer to n, using a large r_0 in large data scenarios is impractical and can make the algorithm inefficient. The second stage is designed for a more thorough estimation, and hence, r_0 is typically chosen to be small. For efficiency considerations, Wang et al. (2018) suggests choosing r_0 to be as small as $o(r^{-1/2})$. This ensures computational efficiency while maintaining the algorithm's effectiveness.

The computational complexity of estimating the model using the proposed optimal subsampling method is analyzed. Let T be the maximum iterations for the GAM to converge in Stage 1, and $\text{d} = \text{d}_1 + \text{d}_2$ be the total number of coefficients in the model. Then the estimation for the coefficient, $\widetilde{\boldsymbol{\beta}}_0$, in Stage 1 requires $O(T(r_0^2 d + r_0 d^2))$ computation. In Stage 2, calculating the optimal subsampling probabilities, π_i^{mmse} and $\pi_i^{\text{mV}_c}$, has complexities of $O(nd^2)$ and $O(nd)$ respectively, for very large n. The final estimation of the refined $\breve{\boldsymbol{\beta}}$ requires $O(T\{d(r_0+r)^2+(r_0+r)d^2\})$. Thus, when n is large, the overall complexity under the A-optimality condition is $O(nd^2)$, while under the L-optimality criterion, it reduces to $O(nd)$. In contrast, uniform random subsampling that selects a subset of size $r_0 + r$ requires $O(T\{(r_0 + r)^2 d + (r_0 + r)d^2\})$ computation. When using the full dataset, model estimation has a quadratic complexity in n, requiring $O(T(n^2 d))$ computation. Therefore, the proposed method substantially reduces computational costs compared to using the entire dataset, although the uniform subsampling method requires the least computation.

Regarding memory requirements, in Stage 1, the proposed optimal subsampling method requires $O(r_0^2 + r_0 d + d^2)$ memory. In Stage 2, computing the optimal subsampling

probabilities, π_i^{mmse} and π_i^{mVc}, requires $O(d^2)$ and $O(d)$ memory respectively. The final estimation of $\breve{\beta}$ demands $O((r_0+r)^2+(r_0+r)d+d^2)$ memory. Overall, the proposed method requires $O(d^2+(r_0r)d+(r_0+r))$ memory. In contrast, the uniform subsampling method needs $O((r_0+r)^2+(r_0+r)d+d^2)$ memory. Without subsampling, fitting GAM on the full dataset demands $O(n^2)$ memory. The subsampling methods significantly reduce the memory requirement for model estimation.

We further provide the readers with theoretical guarantees on the proposed estimator $\breve{\boldsymbol{\beta}}$ obtained through Algorithm 2 under the previously introduced assumptions. To facilitate the derivation of the theorems, we first present Lemma 2.

Lemma 2 *If Assumptions 1 and 2 hold, then conditional on σ-field $\mathcal{F}_n$,*

$$\widetilde{\boldsymbol{D}}_x^{\widetilde{\boldsymbol{\beta}}_0} - \boldsymbol{D}_x = O_{P|\mathcal{F}_n}\left(r^{-1/2}\right), \tag{13}$$

$$\frac{1}{n}\frac{\partial l^*_{\widetilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = O_{P|\mathcal{F}_n}\left(r^{-1/2}\right), \tag{14}$$

where

$$\begin{aligned}\widetilde{\boldsymbol{D}}_x^{\widetilde{\boldsymbol{\beta}}_0} &= \mathrm{E}\left\{\frac{1}{n}\frac{\partial^2 l^*_{\widetilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\right\}\\ &= \frac{1}{n}\left\{\frac{1}{r}\sum_{i=1}^{r}\frac{1}{\pi_i^*(\widetilde{\boldsymbol{\beta}}_0)}w_i^*\boldsymbol{X}_i^*\boldsymbol{X}_i^{*\top} - \boldsymbol{S}_\lambda\right\}.\end{aligned} \tag{15}$$

Lemma 2 characterizes the behavior of the first and second-order condition of the penalized log-likelihood function which are re-weighted by the SSP computed based on $\widetilde{\boldsymbol{\beta}}_0$. Lemma 2 prepares the theoretical support for the subsequent theorems which establish the consistency and asymptotic distribution of the optimal coefficient estimators estimated based on the SSP computed from $\widetilde{\boldsymbol{\beta}}_0$.

Theorem 5 *If Assumptions 2 and 3 hold, when $r_0r^{-1/2}\to 0$, $n\to\infty$ and $r_0\to\infty$, $\breve{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}} = O_{P|\mathcal{F}_n}(r^{-1/2})$.*

Theorem 5 establishes the consistency of the proposed two-stage optimal estimator, $\breve{\boldsymbol{\beta}}$, to the full-data estimator, $\widehat{\boldsymbol{\beta}}$. Their difference is bounded by $r^{-1/2}$ in probability and diminishes as r increases. Further insights into the asymptotic distribution of the proposed estimator are provided in Theorem 6 below.

Theorem 6 *If $r_0r^{-1/2}\to 0$ and Assumption 3 holds, as $n\to\infty$, given σ-field $\mathcal{F}_n$ and $\widetilde{\boldsymbol{\beta}}_0$,*

$$(\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2}\left(\breve{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\right)\xrightarrow{d} N(\boldsymbol{0},\boldsymbol{I}),$$

where

$$\begin{aligned}\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0} &= \boldsymbol{D}_x^{-1}\boldsymbol{V}_c^{\widetilde{\boldsymbol{\beta}}_0}\boldsymbol{D}_x^{-1},\\ \boldsymbol{V}_c^{\widetilde{\boldsymbol{\beta}}_0} &= \mathrm{Var}\left\{\frac{1}{n}\dot{l}^*_{\widetilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}})\right\}\\ &= \frac{1}{rn^2}\sum_{i=1}^{n}\frac{w_i^2(y_i-\mu_i)^2\boldsymbol{X}_i\boldsymbol{X}_i^\top}{\pi_i(\widetilde{\boldsymbol{\beta}}_0)} = O_{P|\mathcal{F}_n}(r^{-1}).\end{aligned}$$

Theorem 6 establishes that the proposed two-stage optimal estimator, $\breve{\boldsymbol{\beta}}$, is asymptotically normal and consistent with $\widehat{\boldsymbol{\beta}}$ with variance $\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_o}$. Compared to the general subsampling estimator, although both are consistent, the proposed estimator is associated with lower asymptotic variance, benefiting from the two-stage subsampling with an optimized variance criterion. In other words, under the condition where both estimators have the same value of variance, the proposed method significantly reduces the number of data points required for computation, fulfilling the purpose of this proposed sampling method for big data analysis.

3 Simulation studies

In this section, we conduct simulation studies to evaluate the finite-sample performance of the proposed optimal subsampling method. Our primary interest lies in assessing how well the proposed estimator, $\breve{\boldsymbol{\beta}}$, approximates the full-data estimator, $\widehat{\boldsymbol{\beta}}$, in large-scale data scenarios. Additionally, we explore the computational time differences when compared to using uniform random sampling.

Each data point i is associated with two parametric covariate, $x_{i2}^{(1)}, x_{i3}^{(1)}$, and three nonparametric covariates, $x_{i1}^{(2)}, x_{i2}^{(2)}$ and $x_{i3}^{(2)}$. For notation convenience, we represent the intercept of the model to be $x_{i1}^{(1)}=1$ such that the parametric part of the model input $\boldsymbol{x}_i^{(1)} = [x_{i1}^{(1)}, x_{i2}^{(1)}, x_{i3}^{(1)}]^\top = [1, x_{i2}^{(1)}, x_{i3}^{(1)}]^\top$. We simulate six scenarios to generate various covariate distributions, and in each scenario, the full data size is simulated to be $n=50000$:

Scenario 1 (Uniform): Covariates follow uniform distribution Unif$(-1,1)$, i.e., $x_{i2}^{(1)}, x_{i3}^{(1)}, x_{i1}^{(2)}, x_{i2}^{(2)}, x_{i3}^{(2)} \sim$ Unif$(-1,1)$.

Scenario 2 (Normal): Covariates follow normal distribution N(0, 1), i.e., $x_{i2}^{(1)}, x_{i3}^{(1)}, x_{i1}^{(2)}, x_{i2}^{(2)}, x_{i3}^{(2)} \sim$ N(0, 1).

Scenario 3 (Exponential): Covariates follow exponential distribution N(0, 1), i.e., $x_{i2}^{(1)}, x_{i3}^{(1)}, x_{i1}^{(2)}, x_{i2}^{(2)}, x_{i3}^{(2)} \sim$ Exp(0.5).

Scenario 4 (Gamma): Covariates follow gamma distribution Gamma(5, 0.5), i.e., $x_{i2}^{(1)}, x_{i3}^{(1)}, x_{i1}^{(2)}, x_{i2}^{(2)}, x_{i3}^{(2)} \sim$ Gamma(5, 0.5).

Scenario 5 (Poisson): Covariates follow Poisson distribution Pois(10), i.e., $x_{i2}^{(1)}, x_{i3}^{(1)}, x_{i1}^{(2)}, x_{i2}^{(2)}, x_{i3}^{(2)} \sim$ Pois(10).

Scenario 6 (Weibull): Covariates follow Weibull distribution Weibull(3, 1), i.e., $x_{i2}^{(1)}, x_{i3}^{(1)}, x_{i1}^{(2)}, x_{i2}^{(2)}, x_{i3}^{(2)} \sim$ Weibull(3, 1).

Define the inverse of the link function $\mu_i = g^{-1}(\cdot)$ using the following equation for all i:

$$\begin{aligned} g^{-1}(\cdot) &= x_{i1}^{(1)} + x_{i2}^{(1)} + x_{i3}^{(1)} + f_1\left(x_{i1}^{(2)}\right) + f_2\left(x_{i2}^{(2)}\right) \\ &\quad + f_3\left(x_{i3}^{(2)}\right) + \epsilon_i \\ &= 1 + x_{i2}^{(1)} + x_{i3}^{(1)} + x_{i1}^{(2)} + \left(x_{i1}^{(2)}\right)^2 \\ &\quad + x_{i2}^{(2)} + \left(x_{i2}^{(2)}\right)^2 + x_{i3}^{(2)} + \left(x_{i3}^{(2)}\right)^2 + \epsilon_i, \end{aligned}$$

where $x_{i2}^{(1)}, x_{i3}^{(1)}, x_{i1}^{(2)}, x_{i2}^{(2)}$ and $x_{i3}^{(2)}$ are standardized, respectively, before input into the link function for numerical stability considerations in computation. The smooth functions are defined as $f_j(\cdot) = x_{ij}^{(2)} + (x_{ij}^{(2)})^2$ for $j = 1, 2, 3$. The true parameter $\boldsymbol{\beta} = [1, 1, 1, 1, 1, 1, 1, 1, 1]^\top$. Some measurement error $\epsilon_i \sim N(0, 0.001)$ is added for all i. The response variable y_i follows a Bernoulli distribution, i.e., $y_i \sim$ Bernoulli(μ_i) where μ_i is modeled by the inverse of the logit function, i.e., $\mu_i = g^{-1}(\boldsymbol{X}_i\boldsymbol{\beta}) = \exp\left(\boldsymbol{X}_i\boldsymbol{\beta}\right)/\{1 + \exp\left(\boldsymbol{X}_i\boldsymbol{\beta}\right)\}$.

The subsample sizes for the first and second stages are initially fixed at $r_0 = 250$ and $r = 250$, respectively, to emulate a scenario where a maximum of $(r_0 + r)/n \times 100\% = 1\%$ of data is feasible for computation. Let $\widehat{\boldsymbol{\beta}}^{\text{uni}}$ represent the subsampling estimator from uniform random sampling (uni), and $\breve{\boldsymbol{\beta}}^{\text{mmse}}$ and $\breve{\boldsymbol{\beta}}^{\text{mV}_c}$ represent optimal subsampling estimators under A-optimality (mmse) and L-optimality criteria (mV_c), respectively. To account for randomness and uncertainty, the simulation process is repeated for $K = 1000$ runs in each scenario. The computing hardware used in this study is equipped with Intel i7-8750H CPU and 16 GB RAM. We utilize the average 2-norm error to measure the estimation accuracy in each of the six scenarios and present them in Table 1. The average 2-norm error is defined as:

$$\frac{1}{K}\sum_{k=1}^{K} \|\widehat{\boldsymbol{\beta}}_{(k)}^{\text{uni}} - \widehat{\boldsymbol{\beta}}\|_2,$$

$$\frac{1}{K}\sum_{k=1}^{K} \|\breve{\boldsymbol{\beta}}_{(k)}^{\text{mmse}} - \widehat{\boldsymbol{\beta}}\|_2,$$

$$\frac{1}{K}\sum_{k=1}^{K} \|\breve{\boldsymbol{\beta}}_{(k)}^{\text{mV}_c} - \widehat{\boldsymbol{\beta}}\|_2,$$

where $\boldsymbol{\beta}_{(k)}$ is the coefficient estimator from the k-th simulation run.

Table 1 illustrates that $\breve{\boldsymbol{\beta}}^{\text{mmse}}$ from optimal subsampling under A-optimality criterion exhibits overall smaller average errors in all six scenarios. Although slightly larger, $\breve{\boldsymbol{\beta}}^{\text{mV}_c}$ shows similar average errors to those of $\breve{\boldsymbol{\beta}}^{\text{mmse}}$. In contrast, $\widehat{\boldsymbol{\beta}}^{\text{uni}}$ from uniform random sampling reveals larger errors and variance in all six scenarios. It's worth noting that for covariate distributions that are skewed, such as exponential, gamma, Poisson, and Weibull distributions, the errors and variance of all three estimators become larger than those of Uniform or Normal distributions. Moreover, $\breve{\boldsymbol{\beta}}^{\text{mmse}}$ exhibits significantly smaller variance compared to all the other estimators.

Table 1 Average 2-norm errors of subsampling estimators from uniform random sampling (uni), optimal subsampling under A-optimality (mmse) and optimal subsampling under L-optimality (mV_c) computed at subsample size $(r_0 + r) = 500$ in six different distribution scenarios across 1000 runs, with standard error presented in parentheses

	uni	mmse	mV_c
Uniform	0.91 (0.33)	0.64 (0.22)	0.84 (0.38)
Normal	0.88 (0.35)	0.66 (0.25)	0.78 (0.33)
Exponential	0.99 (0.39)	0.87 (0.35)	1.00 (0.48)
Gamma	0.87 (0.31)	0.68 (0.24)	0.81 (0.34)
Poisson	0.84(0.32)	0.66 (0.23)	0.76 (0.32)
Weibull	0.88(0.33)	0.65 (0.25)	0.79 (0.35)

To investigate the impact of subsample size on the performance of the proposed method, we evaluate the average 2-norm error of optimal subsample estimators while varying the subsample size in stage two, ranging from $r = 250, 300, 350, 400, 450, 500, 550,$ across 1000 simulation runs. Additionally, we compute the corresponding error of the uniform subsampling estimator $\widehat{\boldsymbol{\beta}}^{\text{uni}}$ from a subsample of size $r_0 + r$. The results are presented in Fig. 1.

Figure 1 depicts a consistent decrease in the average 2-norm errors of all estimators as the subsample size r increases. $\breve{\boldsymbol{\beta}}^{\text{mmse}}$ consistently outperforms the other two estimators across all scenarios, exhibiting the lowest errors. In contrast, $\widehat{\boldsymbol{\beta}}^{\text{uni}}$ from uniform random subsampling demonstrates higher errors in almost all six scenarios. Notably, in the exponential distribution scenario, as the subsample size increases, $\breve{\boldsymbol{\beta}}^{\text{mV}_c}$ exhibits slightly larger errors than $\widehat{\boldsymbol{\beta}}^{\text{uni}}$. This phenomenon may be attributed to the highly skewed covariate distribution in the exponential scenario, resulting in numerous outliers. The covariate effect in the subsampling probability computation poses challenges in obtaining a robust optimal SSP for $\breve{\boldsymbol{\beta}}^{\text{mV}_c}$. In comparison, $\breve{\boldsymbol{\beta}}^{\text{mmse}}$ leverages additional information from the penalized likelihood objective function, specifically the second-order derivative $\boldsymbol{D}_x^{-1}$, resulting in a more effective optimal subsampling probability and, consequently, a more accurate coefficient estimate. However, when r is small, and the total subsample

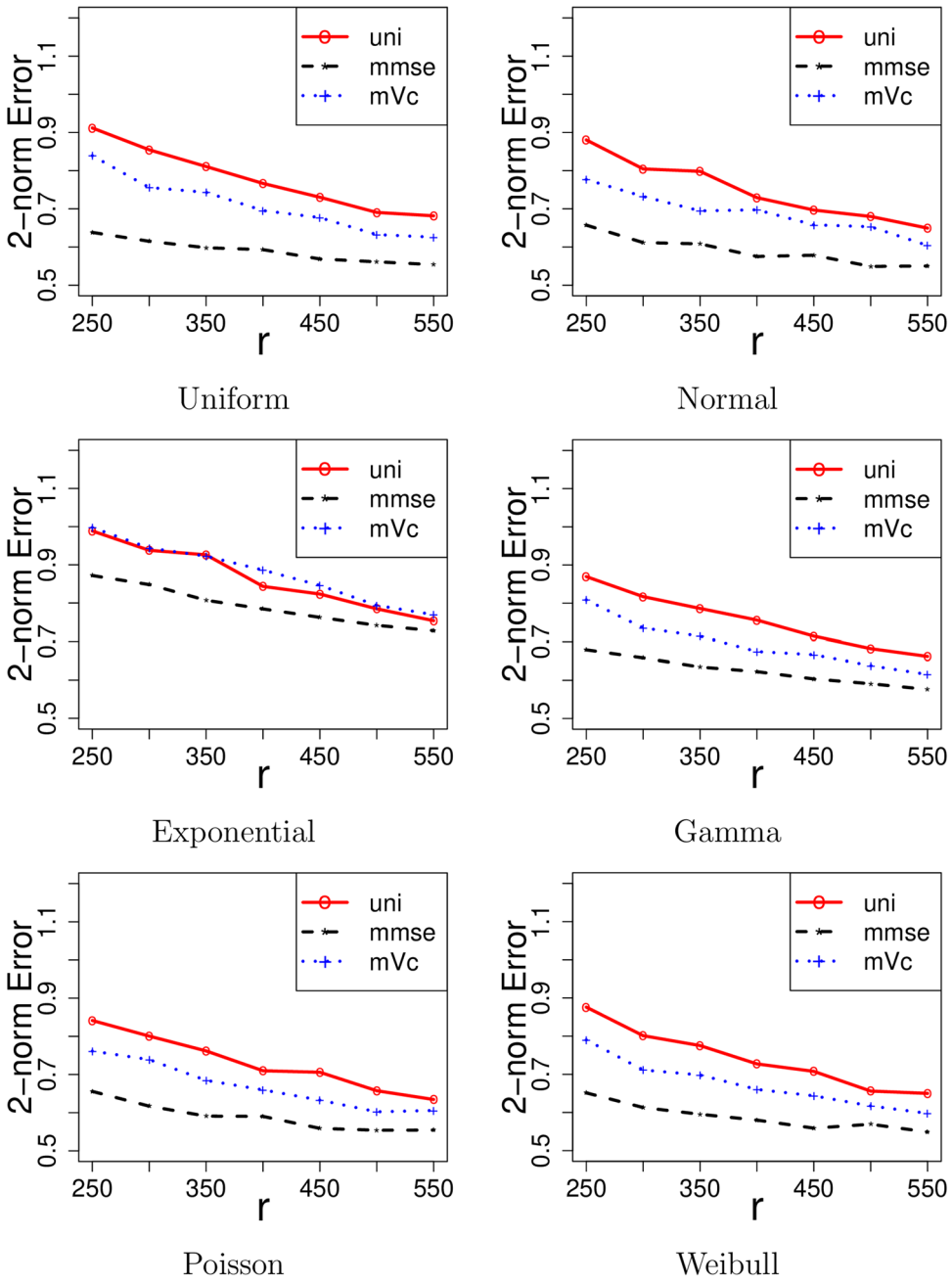


Fig. 1 Average 2-norm errors of subsampling estimators from uniform random sampling (uni), optimal subsampling under A-optimality (mmse) and optimal subsampling under L-optimality (mV_c) computed at fixed $r_0 = 250$ and different $r = \{250, 300, 350, 400, 450, 500, 550\}$ in six different distribution scenarios across 1000 runs

size is limited, both proposed estimators, $\breve{\boldsymbol{\beta}}^{\mathrm{mV_c}}$ and $\breve{\boldsymbol{\beta}}^{\mathrm{mmse}}$ exhibit smaller average 2-norm errors.

The computation time of the proposed methods is another study of interest. Figure 2 presents the average CPU computing time (in seconds) for the three subsampling methods at various r values in each of the six scenarios across 1000 runs. As r increases, the CPU computing time for all three subsampling estimators also increases. $\breve{\boldsymbol{\beta}}^{\mathrm{mmse}}$ requires the most time for computing which is expected as demonstrated by its computational complexity. The uniform random sampling method generally requires the least computing time since it only estimates parameters on the subsample without the need to re-compute the subsampling probability for the second step, as the proposed methods do. However, in the Exponential distribution scenario, the computing time for $\widehat{\boldsymbol{\beta}}^{\mathrm{uni}}$ is similar to and even slightly longer than that of the proposed estimators on average when r increases. Outliers in the exponential distribution of covariates may influence the convergence of the estimator, leading to more iterations in parameter estimation and, consequently, longer computing time for $\widehat{\boldsymbol{\beta}}^{\mathrm{uni}}$. Moreover, the average CPU computing time for parameter estimation on the full dataset in Scenarios 1 to 6 is 228.79, 229.12, 228.78, 228.60, 171.47, and 227.44 s,

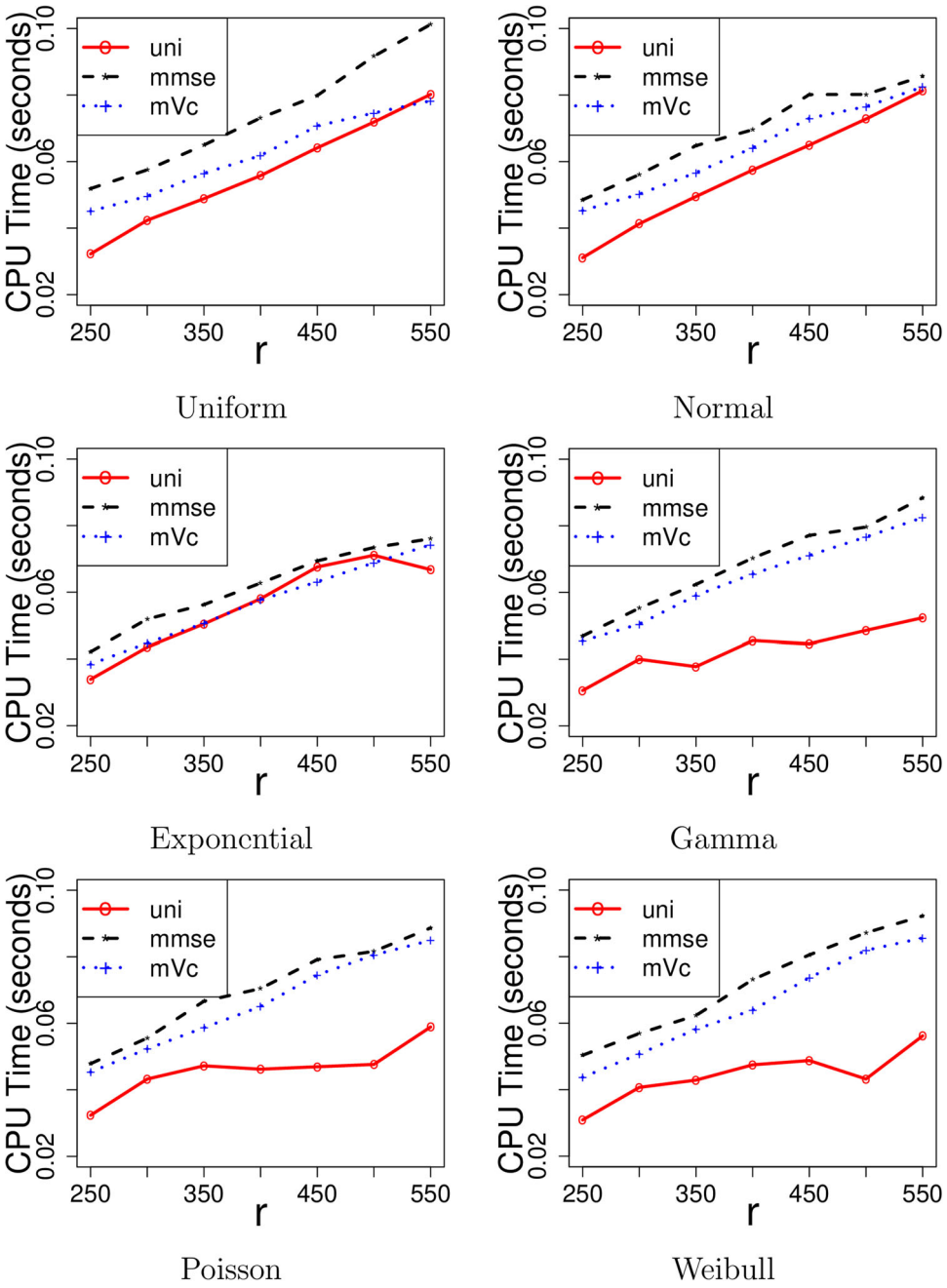


Fig. 2 Averaged CPU time (in seconds) of subsampling estimators from uniform random sampling (uni), optimal subsampling under A-optimality (mmse) and optimal subsampling under L-optimality (mV_c) computed at fixed $r_0 = 250$ and different $r = \{250, 300, 350, 400, 450, 500, 550\}$ across 1000 runs in six different distribution scenarios

respectively. Our proposed methods are much faster in computation benefitting from subsampling.

Overall, the proposed optimal subsampling estimators outperform the uniform random sampling estimator in estimating the full-data estimator with higher accuracy. Furthermore, the computing time of the proposed estimators is significantly reduced compared to estimation using the full dataset.

4 Real data application

This study is particularly motivated by the need to analyze electricity consumption data to predict household active power at specific times. Accurate predictions of electric consumption are vital for optimizing energy production, enhancing distribution efficiency, and promoting energy conservation. Given the continuous and voluminous nature of such data, combining subsampling with semi-parametric models, which allow for fixed functional forms for some variables while relaxing linear or parametric assumptions for

others (like time), emerges as a suitable strategy for this analysis. In this section, we employ uniform random sampling, as well as optimal subsampling under the A- and L-optimality criterion, with GAM to analyze a dataset containing individual household electric power consumption introduced by Hebrail and Berard (2006).

The dataset comprises the variable "global active power", representing the electric power consumption in a household located in Sceaux, France, over a period spanning December 2006 to November 2010. The dataset covers 47 months with a one-minute sampling rate, resulting in a dataset of size $n = 2,049,280$ after removing records with missing values. It includes a date variable capturing the date of the year and a time variable providing time information on the minute level. Our analysis aims to understand how the household's global minute-averaged active power (in kilowatts) varies with the month of the year and the time (in minutes) of the day.

To facilitate this analysis, we first engineered month indicator variables from the "date" variable, denoted as "month$_j$" for $j = 2, \ldots, 12$ such that month$_j = 1$ if the observed electric consumption occurs in the j-th month of the year. Next, we standardized the original time information, denoted as variable time. Denote "global active power" by Y as the response variable which is modeled by regressing it on month indicators and a smooth function of the standardized variable "time", as depicted in the following model:

$$\begin{aligned} Y =& \beta_0 + \beta_1 \text{month}_2 + \beta_2 \text{month}_3 + \beta_3 \text{month}_4 \\ &+ \beta_4 \text{month}_5 + \beta_5 \text{month}_6 + \beta_6 \text{month}_7 \\ &+ \beta_7 \text{month}_8 + \beta_8 \text{month}_9 + \beta_9 \text{month}_{10} \\ &+ \beta_{10} \text{month}_{11} + \beta_{11} \text{month}_{12} + f(\text{time}) + \epsilon. \end{aligned}$$

Variable "time" represents the standardized time in minutes of the day when the observed electric consumption is recorded. Smooth function f, under the sum-to-zero constraint, is modeled by a linear combination of basis functions. Eventually, degree three polynomial basis functions are chosen in this study for simplicity.

Similar to the simulation study, we set the first-stage sample size fixed to be $r_0 = 300$ and vary the second-stage sample size, r, across the values for $r = 300, 600, 900, 1200, 1500, 1800$. The 2-norm error of the parameters estimated using the full dataset versus subsamples sampled by uniform random sampling and our two proposed methods is computed. To ensure robust estimates, we repeat the sampling process for 1000 runs. Figure 3 presents the average 2-norm error of the parameters estimated using the three sampling methods mentioned above and their associated CPU computing time at the fixed r_0 and different r.

The 2-norm errors from all three methods show decreasing trends, dropping from about 0.7–0.4 as r increases from 300 to 1800. However, it is evident that the proposed subsampling estimators are associated with smaller errors than those of uniform random sampling. Moreover, as r increases, the CPU computing time of all three methods increases almost linearly. $\breve{\boldsymbol{\beta}}^{\text{mmse}}$ requires the most computing time, and $\breve{\boldsymbol{\beta}}^{\text{mV}_c}$ requires less computing time. $\widehat{\boldsymbol{\beta}}^{\text{uni}}$ requires the least computing time. All results are consistent with the theorems derived above.

5 Conclusions

In this paper, we introduced two optimal subsampling estimators for Generalized Additive Models, grounded in the principles of A-optimality and L-optimality. Our methodology involves a two-stage process. Initially, naive sampling is first conducted to estimate optimal subsampling probabilities, effectively guiding data selection. Subsequently, comprehensive subsampling is performed to maximize information gain and achieve a consistent estimator with minimized estimation variance. We support our approach with theorems and proofs, affirming the convergence and asymptotic variance of these estimators.

Our assessment, through both simulation and real data studies, validates the superiority of our proposed estimators compared with the uniform random subsampling estimator. While the computational cost of our method is marginally higher compared to uniform random subsampling, it remains significantly more efficient than processing the entire dataset. Thus, our proposed estimators present a promising approach for handling large-scale datasets in situations where traditional full-data analysis is computationally impractical.

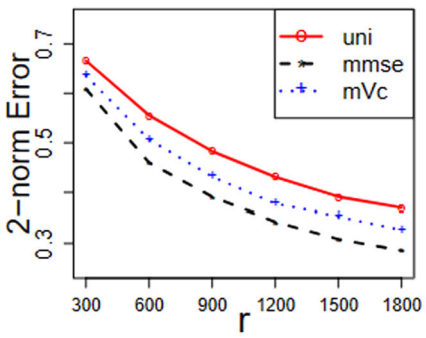
Appendix A Proofs

Proof of Lemma 1

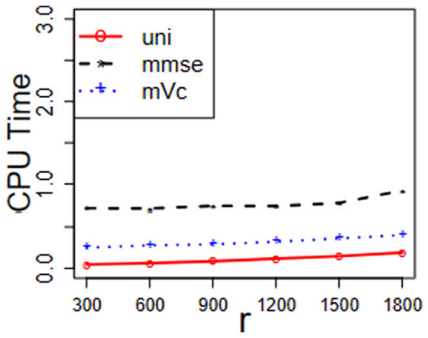
$$\begin{aligned} \mathrm{E}(\widetilde{\boldsymbol{D}}_x|\mathcal{F}_n) =& \mathrm{E}\left\{\frac{1}{n}\left(\frac{1}{r}\sum_{i=1}^{r}\frac{1}{\pi_i^*}w_i^*\boldsymbol{X}_i^*\boldsymbol{X}_i^{*\top} - \boldsymbol{S}_\lambda\right)|\mathcal{F}_n\right\} \\ =& \frac{1}{n}\left(\sum_{i=1}^{n} w_i \boldsymbol{X}_i \boldsymbol{X}_i^\top - \boldsymbol{S}_\lambda\right) \\ =& \boldsymbol{D}_x. \end{aligned}$$

For any integers $j_1, j_2 \in [1, p]$, let $\widetilde{\boldsymbol{D}}_x^{j_1,j_2} = \frac{1}{n}\left\{\frac{1}{r}\sum_{i=1}^{r}\frac{1}{\pi_i^*(\boldsymbol{\beta}_0)}w_i^* x_{ij_1}^* x_{ij_2}^* - \boldsymbol{S}_\lambda^{j_1 j_2}\right\}$ be the component of $\widetilde{\boldsymbol{D}}_x$. We have

$$\mathrm{Var}(\widetilde{\boldsymbol{D}}_x^{j_1 j_2}|\mathcal{F}_n)$$



(a) Average 2-norm Error across 1000 runs



(b) Averaged CPU Time across 1000 runs

Fig. 3 Average 2-norm error and CPU time (in seconds) of the estimated parameters from uni, mmse and mV_c computed at fixed $r_0 = 300$ and different subsample size $r = \{300, 600, 900, 1200, 1500, 1800\}$ across 1000 runs

$$
\begin{aligned}
&= \text{Var}\left\{\frac{1}{n}\left(\frac{1}{r}\sum_{i=1}^{r} w_i^* \frac{x_{ij_1}^* x_{ij_2}^*}{\pi_i^*} - \boldsymbol{S}_\lambda^{j_1 j_2}\right)|\mathcal{F}_n\right\} \\
&= \text{Var}\left\{\frac{1}{nr}\sum_{i=1}^{r} w_i^* \frac{x_{ij_1}^* x_{ij_2}^*}{\pi_i^*}|\mathcal{F}_n\right\} \\
&= \frac{1}{r}\sum_{i=1}^{n}\pi_i\left(\sum_{i=1}^{n} w_i \frac{x_{ij_1}x_{ij_2}}{n\pi_i} - \frac{1}{rn}\sum_{i=1}^{n} w_i x_{ij_1}x_{ij_2}\right)^2 \\
&= \frac{1}{rn^2}\sum_{i=1}^{n}\frac{(w_i x_{ij_1}x_{ij_2})^2}{\pi_i} - \frac{1}{r}\left(\frac{1}{rn}\sum_{i=1}^{n} w_i x_{ij_1}x_{ij_2}\right)^2 \\
&\le \frac{1}{rn^2}\sum_{i=1}^{n}\frac{w_i^2||\boldsymbol{X}_i||^4}{\pi_i} - \frac{1}{r}\left(\frac{1}{rn}\sum_{i=1}^{n} w_i x_{ij_1}x_{ij_2}\right)^2.
\end{aligned}
$$

Since w_i^2 is bounded, we denote its upper bound as M_1, i.e., $w_i^2 \le M_1$. From Assumptions 2 and 4, we have

$$
\begin{aligned}
\text{Var}(\widetilde{\boldsymbol{D}}_x^{j_1 j_2}|\mathcal{F}_n) &\le \frac{M_1}{rn^2}\sum_{i=1}^{n}\frac{\|\boldsymbol{X}_i\|^4}{\pi_i} - \frac{1}{r}\left(\frac{1}{rn}\sum_{i=1}^{n} w_i x_{ij_1}x_{ij_2}\right)^2 \\
&= O_{P|\mathcal{F}_n}(r^{-1}).
\end{aligned}
$$

According to Markov's Inequality, we have

$$
\begin{aligned}
&P\{(\widetilde{\boldsymbol{D}}_x - \boldsymbol{D}_x) \ge a\} \le \frac{M_1}{rn^2}\sum_{i=1}^{n}\frac{\|\boldsymbol{X}_i\|^4}{\pi_i} \\
&\quad - \frac{1}{r}\left(\frac{1}{rn}\sum_{i=1}^{n} w_i x_{ij_1}x_{ij_2}\right)^2 \\
&= O_{P|\mathcal{F}_n}(r^{-1}).
\end{aligned}
$$

Thus, $(\widetilde{\boldsymbol{D}}_x - \boldsymbol{D}_x)^2 = O_{P|\mathcal{F}_n}(r^{-1})$. Moreover,

$$
\frac{1}{n}\frac{\partial l^*(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = \frac{1}{n}\left\{\frac{1}{r}\sum_{i=1}^{r}\frac{1}{\pi_i^*} w_i^*(y_i^* - \mu_i^*)\boldsymbol{X}_i^* - \boldsymbol{S}_\lambda\widehat{\boldsymbol{\beta}}\right\}, \quad \text{(A1)}
$$

Thus,

$$
\begin{aligned}
\text{E}\left\{\frac{1}{n}\frac{\partial l^*(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}|\mathcal{F}_n\right\} &= \frac{1}{n}\left\{\sum_{i=1}^{n} w_i(y_i - \mu_i)\boldsymbol{X}_i - \boldsymbol{S}_\lambda\widehat{\boldsymbol{\beta}}\right\} \\
&= \frac{1}{n}\frac{\partial l(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = 0, \\
\text{Var}\left\{\frac{1}{n}\frac{\partial l^*(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}|\mathcal{F}_n\right\} &= \text{Var}\left[\frac{1}{n}\left\{\frac{1}{r}\sum_{i=1}^{r}\frac{1}{\pi_i^*} w_i^*(y_i^* - \mu_i)\boldsymbol{X}_i^*\right.\right. \\
&\quad \left.\left. - \boldsymbol{S}_\lambda\widehat{\boldsymbol{\beta}}\right\}|\mathcal{F}_n\right] \\
&= \text{Var}\left\{\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*} w_i^* \right. \\
&\quad \left. (y_i^* - \mu_i^*)\boldsymbol{X}_i^*|\mathcal{F}_n\right\} \\
&= \frac{1}{rn^2}\sum_{i=1}^{n}\frac{w_i^2(y_i - \boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}})^2\boldsymbol{X}_i\boldsymbol{X}_i^\top}{\pi_i} \\
&\le \frac{1}{rn^2}\sum_{i=1}^{n}\frac{w_i^2(y_i - \mu_i)^2\|\boldsymbol{X}_i\|^2}{\pi_i}.
\end{aligned}
$$

Let $|y_i - \mu_i|^2 \le M_2$. Given that $w_i^2 \le M_1$, from Assumption 2, we have

$$
\begin{aligned}
&\text{Var}\left\{\frac{1}{n}\frac{\partial l^*(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}|\mathcal{F}_n\right\} \\
&\le \frac{M_1 M_2}{rn^2}\sum_{i=1}^{n}\frac{\|\boldsymbol{X}_i\|^2}{\pi_i} = O_{P|\mathcal{F}_n}(r^{-1}).
\end{aligned}
$$

From Markov's Inequality, we have

$$
\begin{aligned}
&P\left[\left\{\frac{1}{n}\frac{\partial l^*(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} - \frac{1}{n}\frac{\partial l(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\right\}^2 \ge a\right] \\
&\le \frac{\text{Var}\left\{\frac{1}{n}\frac{\partial l^*(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\right\}}{a} = O_{P|\mathcal{F}_n}(r^{-1}).
\end{aligned}
$$

Thus,

$$\frac{1}{n}\frac{\partial l^*(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = O_{P|\mathcal{F}_n}(r^{-1/2}).$$

□

Proof of Theorem 1 Denote the first order derivative of $l^*(\widetilde{\boldsymbol{\beta}})$ to be $\dot{l}^*(\widetilde{\boldsymbol{\beta}})$. For $\forall\ 0 < u, v < 1$, the Tayler expansion of $\dot{l}^*(\widetilde{\boldsymbol{\beta}})$ at $\widehat{\boldsymbol{\beta}}_j$ is

$$\begin{aligned}\dot{l}^*(\widetilde{\boldsymbol{\beta}}_j) \approx \dot{l}^*(\widehat{\boldsymbol{\beta}}_j) + \frac{\partial \dot{l}^*(\widehat{\boldsymbol{\beta}}_j)}{\partial \boldsymbol{\beta}_j^\top}(\widetilde{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\beta}}_j) + R_j = 0,\\ R_j = (\widetilde{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\beta}}_j)^\top \\ \times \int_0^1\int_0^1 \frac{\partial^2 \dot{l}^*\{\widehat{\boldsymbol{\beta}}_j + uv(\widetilde{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\beta}}_j)\}}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j^\top} v\mathrm{d}u\mathrm{d}v \\ \times (\widetilde{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\beta}}_j). \end{aligned} \tag{A2}$$

According to Chapter 4 of Ferguson (1996), $\|\frac{\partial^2 \dot{l}^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top}\| = 0$ is true for $\forall \boldsymbol{\beta}$. Thus,

$$\begin{aligned}&\left\| \int_0^1\int_0^1 \frac{\partial^2 \dot{l}^*\{\widehat{\boldsymbol{\beta}}_j + uv(\widetilde{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\beta}}_j)\}}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j^\top} v\mathrm{d}u\mathrm{d}v \right\| \\ &\le \int_0^1\int_0^1 \left\| \frac{\partial^2 \dot{l}^*\{\widehat{\boldsymbol{\beta}}_j + uv(\widetilde{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\beta}}_j)\}}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j^\top} \right\| v\mathrm{d}u\mathrm{d}v = 0.\end{aligned}$$

By Lemma 1, we have

$$\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}} = -\widetilde{\boldsymbol{D}}_x^{-1}\frac{\dot{l}^*(\widehat{\boldsymbol{\beta}})}{n} = O_{P|\mathcal{F}_n}\left(r^{-1/2}\right). \tag{A3}$$

□

Proof of Theorem 2 Note that

$$\begin{aligned}\frac{\dot{l}^*(\widehat{\boldsymbol{\beta}})}{n} &= \frac{1}{r}\sum_{i=1}^r \frac{w_i^*(y_i^* - \mu_i)\boldsymbol{X}_i^*}{n\pi_i^*} - \frac{1}{n}\boldsymbol{S}_\lambda\widehat{\boldsymbol{\beta}} \\ &= \frac{1}{r}\sum_{i=1}^r \left\{\frac{w_i^*(y_i^* - \mu_i)\boldsymbol{X}_i^*}{n\pi_i^*} - \frac{1}{n}\boldsymbol{S}_\lambda\widehat{\boldsymbol{\beta}}\right\} \\ &:= \frac{1}{r}\sum_{i=1}^r \boldsymbol{\eta}_i.\end{aligned}$$

Furthermore, by Lemma 1,

$$\mathrm{E}\left\{\frac{1}{n}\dot{l}^*(\widehat{\boldsymbol{\beta}})\right\} = 0, \tag{A4}$$

$$\mathrm{Var}\left\{\frac{1}{n}\dot{l}^*(\widehat{\boldsymbol{\beta}})\right\} = \boldsymbol{V}_c = O_{P|\mathcal{F}_n}(r^{-1}). \tag{A5}$$

Conditional on σ-field $\mathcal{F}_n$, with $\boldsymbol{\eta}_i$ i.i.d., we can know from Equation (A4) that $\mathrm{E}(\boldsymbol{\eta}_i|\mathcal{F}_n) = 0$, $i = 1, 2, ..., r$. Thus, one can further derive that $\mathrm{Var}(\boldsymbol{\eta}_i|\mathcal{F}_n) = r\boldsymbol{V}_c = O_{P|\mathcal{F}_n}(1)$ for $i = 1, 2, ..., r$. Moreover, denote identity function as $I(\cdot)$. For $\forall\varepsilon > 0$, there exists $\tau > 0$ such that

$$\begin{aligned}&\sum_{i=1}^r \mathrm{E}\left\{\|r^{-1/2}\boldsymbol{\eta}_i\|^2 I(\|r^{-1/2}\boldsymbol{\eta}_i\| > \varepsilon)|\mathcal{F}_n\right\} \\ &= \sum_{i=1}^r \mathrm{E}\left\{\|r^{-1/2}\boldsymbol{\eta}_i\|^2 I(\|\boldsymbol{\eta}_i\| > r^{1/2}\varepsilon)|\mathcal{F}_n\right\} \\ &\le \sum_{i=1}^r \mathrm{E}\left\{\|r^{-1/2}\boldsymbol{\eta}_i\|^2 \left(\frac{\|\boldsymbol{\eta}_i\|}{r^{1/2}\varepsilon}\right)^\tau I(\|\boldsymbol{\eta}_i\| > r^{1/2}\varepsilon)|\mathcal{F}_n\right\} \\ &\le \frac{1}{r^{1+\frac{\tau}{2}}\varepsilon^\tau}\mathrm{E}\left(\sum_{i=1}^r \|\boldsymbol{\eta}_i\|^{2+\tau}\right) \\ &= \frac{1}{r^{1+\frac{\tau}{2}}\varepsilon^\tau}\mathrm{E}\left\{\sum_{i=1}^r \left\|\frac{w_i^*(y_i^* - \mu_i)\boldsymbol{X}_i^*}{n\pi_i^*} - \frac{1}{n}\boldsymbol{S}_\lambda\widehat{\boldsymbol{\beta}}\right\|^{2+\tau}\right\} \\ &\le \frac{1}{r^{1+\frac{\tau}{2}}\varepsilon^\tau}\sum_{i=1}^r \left\{\frac{\|w_i^*(y_i^* - \mu_i)\boldsymbol{X}_i^*\|^{2+\tau}}{n^{2+\tau}(\pi_i^*)^{2+\tau}} + \frac{\|\boldsymbol{S}_\lambda\widehat{\boldsymbol{\beta}}\|^{2+\tau}}{n^{2+\tau}}\right\} \\ &= \frac{1}{r^{\frac{\tau}{2}}\varepsilon^\tau}\sum_{i=1}^n \frac{\|w_i(y_i - \mu_i)\boldsymbol{X}_i\|^{2+\tau}}{n^{2+\tau}(\pi_i)^{1+\tau}} \\ &\quad + \frac{1}{r^{1+\frac{\tau}{2}}\varepsilon^\tau}\sum_{i=1}^r \frac{\|\boldsymbol{S}_\lambda\widehat{\boldsymbol{\beta}}\|^{2+\tau}}{n^{2+\tau}} \\ &\le \frac{1}{r^{\frac{\tau}{2}}\varepsilon^\tau}\sum_{i=1}^n \frac{|w_i|^{2+\tau}\|\boldsymbol{X}_i\|^{2+\tau}\|y_i - \mu_i\|^{2+\tau}}{n^{2+\tau}\pi_i^{1+\tau}} \\ &\quad + \frac{1}{r^{\frac{\tau}{2}}\varepsilon^\tau}\frac{\|\boldsymbol{S}_\lambda\|^{2+\tau}\|\widehat{\boldsymbol{\beta}}\|^{2+\tau}}{n^{2+\tau}}.\end{aligned}$$

There exists $M_3, M_4, M_5, M_6 > 0$, such that $|w_i|^{2+\tau} \le M_3$, $|y_i - \mu_i|^{2+\tau} \le M_4$, $\|\boldsymbol{S}_\lambda\|^{2+\tau} \le M_5$, $\|\widehat{\boldsymbol{\beta}}\|^{2+\tau} \le M_6$. Thus,

$$\begin{aligned}&\sum_{i=1}^r \mathrm{E}\left\{\|r^{-1/2}\boldsymbol{\eta}_i\|^2 I(\|r^{-1/2}\boldsymbol{\eta}_i\| > \varepsilon)|\mathcal{F}_n\right\} \\ &\le \frac{M_3M_4}{r^{\frac{\tau}{2}}\varepsilon^\tau}\sum_{i=1}^n \frac{\|\boldsymbol{X}_i\|^{2+\tau}}{n^{2+\tau}\pi_i^{1+\tau}} + \frac{M_5M_6}{r^{\frac{\tau}{2}}\varepsilon^\tau n^{2+\tau}}.\end{aligned}$$

From Assumption 3, when $n \to \infty$ and $r \to \infty$,

$$\sum_{i=1}^r \mathrm{E}\left\{\|r^{-1/2}\boldsymbol{\eta}_i\|^2 I(\|r^{-1/2}\boldsymbol{\eta}_i\| > \varepsilon)|\mathcal{F}_n\right\} = o_{P|\mathcal{F}_n}(1).$$

Thus, by the Lindeberg-Feller central limit theorem,

$$\frac{1}{n}\boldsymbol{V}_c^{-1/2}\dot{l}^*(\widehat{\boldsymbol{\beta}})$$

$$= r^{-1/2}\mathrm{Var}(\boldsymbol{\eta}_i|\mathcal{F}_n)^{-1/2}\sum_{i=1}^{r}\boldsymbol{\eta}_i \xrightarrow{d} N(\mathbf{0}, \boldsymbol{I}). \tag{A6}$$

From Lemma 1 and Theorem 1,

$$\widetilde{\boldsymbol{D}}_x^{-1} - \boldsymbol{D}_x^{-1} = -\boldsymbol{D}_x^{-1}(\widetilde{\boldsymbol{D}}_x - \boldsymbol{D}_x)\widetilde{\boldsymbol{D}}_x^{-1} = O_{P|\mathcal{F}_n}(r^{-1/2}).$$

From Equation (A5), $r\boldsymbol{V}_c = O_p(1)$. We can further derive that

$$\boldsymbol{V} = \boldsymbol{D}_x^{-1}\boldsymbol{V}_c\boldsymbol{D}_x^{-1} = O_{P|\mathcal{F}_n}(r^{-1}).$$

Thus, by Theorem 1, we have

$$\boldsymbol{V}^{-1/2}(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}) = O_{P|\mathcal{F}_n}(1).$$

From Equation (A3),

$$\begin{aligned}
&\boldsymbol{V}^{-1/2}(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})\\
&= -\boldsymbol{V}^{-1/2}\widetilde{\boldsymbol{D}}_x^{-1}\frac{\dot{l}^*(\widehat{\boldsymbol{\beta}})}{n}\\
&= -\boldsymbol{V}^{-1/2}\left\{\boldsymbol{D}_x^{-1} + (\widetilde{\boldsymbol{D}}_x^{-1} - \boldsymbol{D}_x^{-1})\right\}\frac{\dot{l}^*(\widehat{\boldsymbol{\beta}})}{n}\\
&= -\boldsymbol{V}^{-1/2}\boldsymbol{D}_x^{-1}\frac{\dot{l}^*(\widehat{\boldsymbol{\beta}})}{n} +\\
&\quad\left\{-\boldsymbol{V}^{-1/2}(\widetilde{\boldsymbol{D}}_x^{-1} - \boldsymbol{D}_x^{-1})\right\}\frac{\dot{l}^*(\widehat{\boldsymbol{\beta}})}{n}\\
&= -\boldsymbol{V}^{-1/2}\boldsymbol{D}_x^{-1}\frac{\dot{l}^*(\widehat{\boldsymbol{\beta}})}{n} + O_{P|\mathcal{F}_n}(r^{-1/2})\\
&= -\boldsymbol{V}^{-1/2}\boldsymbol{D}_x^{-1}\boldsymbol{V}_c^{1/2}\boldsymbol{V}_c^{-1/2}\frac{\dot{l}^*(\widehat{\boldsymbol{\beta}})}{n} + O_{P|\mathcal{F}_n}(r^{-1/2}),
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\mathrm{Var}(\boldsymbol{V}^{-1/2}\boldsymbol{D}_x^{-1}\boldsymbol{V}_c^{1/2})\\
&= \boldsymbol{V}^{-1/2}\boldsymbol{D}_x^{-1}\boldsymbol{V}_c^{1/2}(\boldsymbol{V}^{-1/2}\boldsymbol{D}_x^{-1}\boldsymbol{V}_c^{1/2})^\top\\
&= \boldsymbol{V}^{-1/2}\boldsymbol{D}_x^{-1}\boldsymbol{V}_c^{1/2}\boldsymbol{V}_c^{1/2}\boldsymbol{D}_x^{-1}\boldsymbol{V}^{-1/2} = \boldsymbol{I}.
\end{aligned}$$

Thus,

$$\mathrm{Var}(\boldsymbol{V}^{-1/2}(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})) = \boldsymbol{I}.$$

By Slutsky Theorem and Eq. (A6),

$$\boldsymbol{V}^{-1/2}(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{I}).$$

□

Proof of Theorem 3 From Theorem 2,

$$\begin{aligned}
\mathrm{tr}(\boldsymbol{V}) &= \mathrm{tr}(\boldsymbol{D}_x^{-1}\boldsymbol{V}_c\boldsymbol{D}_x^{-1})\\
&= \mathrm{tr}\left\{\boldsymbol{D}_x^{-1}\frac{1}{rn^2}\sum_{i=1}^{n}\frac{w_i^2(y_i-\mu_i)^2\boldsymbol{X}_i\boldsymbol{X}_i^\top}{\pi_i}\boldsymbol{D}_x^{-1}\right\}\\
&= \frac{1}{rn^2}\sum_{i=1}^{n}\mathrm{tr}\left\{\frac{w_i^2(y_i-\mu_i)^2}{\pi_i}\boldsymbol{D}_x^{-1}\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{D}_x^{-1}\right\}\\
&= \frac{1}{rn^2}\sum_{i=1}^{n}\left\{\frac{w_i^2(y_i-\mu_i)^2}{\pi_i}\|\boldsymbol{D}_x^{-1}\boldsymbol{X}_i\|^2\right\}\\
&= \frac{1}{rn^2}\sum_{i=1}^{n}(\sqrt{\pi_i})^2\sum_{i=1}^{n}\left\{\frac{|w_i||y_i-\mu_i|}{\sqrt{\pi_i}}\|\boldsymbol{D}_x^{-1}\boldsymbol{X}_i\|\right\}^2\\
&\geq \frac{1}{rn^2}\sum_{i=1}^{n}\left(\sqrt{\pi_i}\frac{|w_i||y_i-\mu_i|}{\sqrt{\pi_i}}\|\boldsymbol{D}_x^{-1}\boldsymbol{X}_i\|\right)^2.
\end{aligned}$$

By Cauchy–Schwarz inequality, the equality holds if and only if $\sqrt{\pi_i} \propto \frac{|w_i||y_i-\mu_i|}{\sqrt{\pi_i}}\|\boldsymbol{D}_x^{-1}\boldsymbol{X}_i\|$. Thus, when SSP $\pi_i^{\mathrm{mmse}} = \frac{\pi_i}{\sum_{j=1}^n\pi_j} = \frac{|w_i||y_i-\mu_i|\|\boldsymbol{D}_x^{-1}\boldsymbol{X}_i\|}{\sum_{j=1}^n|w_j||y_j-\mu_j|\|\boldsymbol{D}_x^{-1}\boldsymbol{X}_j\|}$, $\mathrm{tr}(\boldsymbol{V})$ achieves minimum. □

Proof of Theorem 4 From Theorem 2,

$$\begin{aligned}
&\mathrm{tr}(\boldsymbol{V}_c)\\
&= \mathrm{tr}\left\{\frac{1}{rn^2}\sum_{i=1}^{n}\frac{w_i^2(y_i-\mu_i)^2\boldsymbol{X}_i\boldsymbol{X}_i^\top}{\pi_i}\right\}\\
&= \frac{1}{rn^2}\sum_{i=1}^{n}\mathrm{tr}\left\{\frac{w_i^2(y_i-\mu_i)^2}{\pi_i}\boldsymbol{X}_i\boldsymbol{X}_i^\top\right\}\\
&= \frac{1}{rn^2}\sum_{i=1}^{n}\left\{\frac{w_i^2(y_i-\mu_i)^2}{\pi_i}\|\boldsymbol{X}_i\|^2\right\}\\
&= \frac{1}{rn^2}\sum_{i=1}^{n}(\sqrt{\pi_i})^2\sum_{i=1}^{n}\left\{\frac{|w_i||y_i-\mu_i|}{\sqrt{\pi_i}}\|\boldsymbol{X}_i\|\right\}^2\\
&\geq \frac{1}{rn^2}\sum_{i=1}^{n}\left\{\sqrt{\pi_i}\frac{|w_i||y_i-\mu_i|}{\sqrt{\pi_i}}\|\boldsymbol{X}_i\|\right\}^2.
\end{aligned}$$

According to Cauchy–Schwarz inequality, the equality holds if and only if $\sqrt{\pi_i} \propto \frac{|w_i||y_i-\mu_i|\|\boldsymbol{X}_i\|}{\sqrt{\pi_i}}$. Thus, when subsampling probability $\pi_i^{\mathrm{mVc}} = \frac{\pi_i}{\sum_{j=1}^n\pi_j} = \frac{|w_i||y_i-\mu_i|\|\boldsymbol{X}_i\|}{\sum_{j=1}^n|w_j||y_j-\mu_j|\|\boldsymbol{X}_j\|}$, $\mathrm{tr}(\boldsymbol{V}_c)$ achieves minimum. □

Proof of Lemma 2 For any integers $j_1, j_2 \in [1, p]$, let

$$\widetilde{\boldsymbol{D}}_x^{j_1,j_2} = \frac{1}{n}\left\{\frac{1}{r}\sum_{i=1}^{r}\frac{1}{\pi_i^*(\widetilde{\boldsymbol{\beta}}_0)}w_i^*x_{ij_1}^*x_{ij_2}^* + \boldsymbol{S}_\lambda^{j_1j_2}\right\}$$

be the component of $\widetilde{\boldsymbol{D}}_x$.

$$
\begin{aligned}
&\mathrm{Var}(\widetilde{\boldsymbol{D}}_x^{j_1 j_2}|\mathcal{F}_n, \widetilde{\boldsymbol{\beta}}_0) \\
&= \mathrm{Var}\left[\frac{1}{n}\left\{\frac{1}{r}\sum_{i=1}^{r}\frac{1}{\pi_i^*(\widetilde{\boldsymbol{\beta}}_0)} w_i^* x_{ij_1}^* x_{ij_2}^* - \boldsymbol{S}_\lambda^{j_1 j_2}\right\}|\mathcal{F}_n\right] \\
&= \mathrm{Var}\left\{\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*(\widetilde{\boldsymbol{\beta}}_0)} w_i^* x_{ij_1}^* x_{ij_2}^*\right\} \\
&\le \frac{1}{r}\sum_{i=1}^{n}\pi_i(\widetilde{\boldsymbol{\beta}}_0)\left\{\frac{1}{n\pi_i(\widetilde{\boldsymbol{\beta}}_0)} w_i x_{ij_1} x_{ij_2}\right. \\
&\left. - \frac{1}{nr}\sum_{i=1}^{n} w_i x_{ij_1} x_{ij_2}\right\}^2 \\
&= \frac{1}{rn^2}\sum_{i=1}^{n}\frac{w_i^2 (x_{ij_1} x_{ij_2})^2}{\pi_i(\widetilde{\boldsymbol{\beta}}_0)} - \frac{1}{r}\left(\frac{1}{nr}\sum_{i=1}^{n} w_i x_{ij_1} x_{ij_2}\right)^2 \\
&\le \frac{1}{rn^2}\sum_{i=1}^{n}\frac{w_i^2 \|\boldsymbol{X}_i\|^4}{\pi_i(\widetilde{\boldsymbol{\beta}}_0)} - \frac{1}{r}\left(\frac{1}{nr}\sum_{i=1}^{n} w_i x_{ij_1} x_{ij_2}\right)^2.
\end{aligned}
$$

Since $w_i^2 \le M_1$,

$$
\begin{aligned}
&\mathrm{Var}(\widetilde{\boldsymbol{D}}_x^{j_1 j_2}|\mathcal{F}_n, \widetilde{\boldsymbol{\beta}}_0) \\
&\le \frac{M_1}{rn^2}\sum_{i=1}^{n}\frac{\|\boldsymbol{X}_i\|^4}{\pi_i(\widetilde{\boldsymbol{\beta}}_0)} - \frac{1}{r}\left(\frac{1}{nr}\sum_{i=1}^{n} w_i x_{ij_1} x_{ij_2}\right)^2 \\
&= O_{P|\mathcal{F}_n}\left(r^{-1/2}\right).
\end{aligned}
$$

Under Assumptions 1, 2 and 3, by Markov's Inequality,

$$
\begin{aligned}
&P\left\{(\widetilde{\boldsymbol{D}}_x^{\widetilde{\boldsymbol{\beta}}_0} - \boldsymbol{D}_x) \ge a\right\} \\
&\le \frac{\mathrm{E}\left(\widetilde{\boldsymbol{D}}_x^{\widetilde{\boldsymbol{\beta}}_0} - \boldsymbol{D}_x\right)^2}{a} = \frac{\mathrm{Var}\left(\widetilde{\boldsymbol{D}}_x^{\widetilde{\boldsymbol{\beta}}_0}|\mathcal{F}_n\right)}{a} \\
&= O_{P|\mathcal{F}_n}\left(r^{-1/2}\right),
\end{aligned}
$$

Because

$$
\begin{aligned}
&\frac{1}{n}\frac{\partial l_{\widetilde{\boldsymbol{\beta}}_0}^*(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \\
&= \frac{1}{n}\left\{\frac{1}{r}\sum_{i=1}^{r}\frac{1}{\pi_i^*(\widetilde{\boldsymbol{\beta}}_0)} w_i^*(y_i^* - \mu_i)\boldsymbol{X}_i^* - \boldsymbol{S}_\lambda\hat{\boldsymbol{\beta}}\right\}, \\
&\mathrm{E}\left\{\frac{1}{n}\frac{\partial l_{\widetilde{\boldsymbol{\beta}}_0}^*(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}|\mathcal{F}_n\right\} = E_{\widetilde{\boldsymbol{\beta}}_0}\left[\mathrm{E}\left\{\frac{1}{n}\frac{\partial l^*(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}|\mathcal{F}_n, \widetilde{\boldsymbol{\beta}}_0\right\}\right] \\
&= E_{\widetilde{\boldsymbol{\beta}}_0}\left\{\frac{1}{n}\frac{\partial l(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}|\mathcal{F}_n\right\} \\
&= \frac{1}{n}\frac{\partial l(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \\
&= 0.
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
&\mathrm{Var}\left\{\frac{1}{n}\frac{\partial l_{\widetilde{\boldsymbol{\beta}}_0}^*(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}|\mathcal{F}_n\right\} \\
&= \mathrm{Var}\left[\frac{1}{n}\left\{\frac{1}{r}\sum_{i=1}^{r}\frac{1}{\pi_i^*(\widetilde{\boldsymbol{\beta}}_0)} w_i^*(y_i^* - \mu_i)\boldsymbol{X}_i^*\right.\right. \\
&\left.\left. - \boldsymbol{S}_\lambda\hat{\boldsymbol{\beta}}\right\}|\mathcal{F}_n\right] \\
&= \mathrm{Var}\left\{\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*(\widetilde{\boldsymbol{\beta}}_0)} w_i^*(y_i^* - \mu_i)\boldsymbol{X}_i^*|\mathcal{F}_n\right\} \\
&= \frac{1}{rn^2}\sum_{i=1}^{n}\frac{1}{\pi_i(\widetilde{\boldsymbol{\beta}}_0)} w_i^2(y_i - \mu_i)^2\boldsymbol{X}_i\boldsymbol{X}_i^\top \\
&\le \frac{1}{rn^2}\sum_{i=1}^{n}\frac{w_i^2|y_i - \mu_i|^2\|\boldsymbol{X}_i\|^2}{\pi_i(\widetilde{\boldsymbol{\beta}}_0)}.
\end{aligned}
$$

Note that $w_i^2 \le M_1$ and $|y_i - \mu_i|^2 \le M_2$. Thus,

$$
\mathrm{Var}\left\{\frac{1}{n}\frac{\partial l_{\widetilde{\boldsymbol{\beta}}_0}^*(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}|\mathcal{F}_n\right\} \le \frac{M_1 M_2}{rn^2}\sum_{i=1}^{n}\frac{\|\boldsymbol{X}_i\|^2}{\pi_i(\widetilde{\boldsymbol{\beta}}_0)} = O_{P|\mathcal{F}_n}(r^{-1}).
$$

Given Assumption 2 holds, from Markov's Inequality,

$$
\begin{aligned}
&P\left[\left\{\frac{1}{n}\frac{\partial l_{\widetilde{\boldsymbol{\beta}}_0}^*(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} - \frac{1}{n}\frac{\partial l(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\right\}^2 \ge a\right] \\
&\le \frac{\mathrm{Var}\left\{\frac{1}{n}\frac{\partial l_{\widetilde{\boldsymbol{\beta}}_0}^*(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\right\}}{a} = O_{P|\mathcal{F}_n}(r^{-1}).
\end{aligned}
$$

Equation (14) is proved. Thus, Lemma 2 is proved. □

Proof of Theorem 5 For $\forall\, 0 < u, v < 1$, the Tayler expansion of $\dot{l}_{\widetilde{\boldsymbol{\beta}}_0}^*(\breve{\boldsymbol{\beta}}_j)$ at $\hat{\boldsymbol{\beta}}$ is:

$$
\dot{l}_{\widetilde{\boldsymbol{\beta}}_0}^*(\breve{\boldsymbol{\beta}}_j) = \dot{l}_{\widetilde{\boldsymbol{\beta}}_0}^*(\hat{\boldsymbol{\beta}}_j) + \frac{\partial \dot{l}_{\widetilde{\boldsymbol{\beta}}_0}^*(\hat{\boldsymbol{\beta}}_j)}{\partial \boldsymbol{\beta}_j}(\breve{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_j) + R_{\widetilde{\boldsymbol{\beta}}_0} = 0,
$$

where

$$
\begin{aligned}
R_{\widetilde{\boldsymbol{\beta}}_0} =& (\breve{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_j)^\top \\
&\times \int_0^1\int_0^1 \frac{\partial^2 \dot{l}_{\widetilde{\boldsymbol{\beta}}_0, j}^*\{\hat{\boldsymbol{\beta}}_j + uv(\breve{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_j)\}}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j^\top} v\mathrm{d}u\mathrm{d}v \\
&\times (\breve{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_j).
\end{aligned}
$$

According to Chapter 4 of Ferguson (1996), $\left\| \frac{\partial^2 \dot{l}^*_{\widetilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \right\| = 0$ is true for $\forall \boldsymbol{\beta}$. Thus,

$$\begin{aligned}
&\left\| \int_0^1 \int_0^1 \frac{\partial^2 \dot{l}^*_{\widetilde{\boldsymbol{\beta}}_0}\{\hat{\boldsymbol{\beta}}_j + uv(\check{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_j)\}}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j^\top} v \mathrm{d}u \mathrm{d}v \right\| \\
&\leq \int_0^1 \int_0^1 \left\| \frac{\partial^2 \dot{l}^*_{\widetilde{\boldsymbol{\beta}}_{0,j}}\{\hat{\boldsymbol{\beta}}_j + uv(\check{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_j)\}}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j^\top} \right\| v \mathrm{d}u \mathrm{d}v \\
&= 0.
\end{aligned}$$

Using the result from Lemma 2, we can get

$$\check{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} = -\left(\widetilde{\boldsymbol{D}}_x^{\widetilde{\boldsymbol{\beta}}_0}\right)^{-1} \frac{\dot{l}^*_{\widetilde{\boldsymbol{\beta}}_{0,j}}(\hat{\boldsymbol{\beta}})}{n} = O_{P|\mathcal{F}_n}(r^{-1/2}).$$

□

Proof of Theorem 6 Note that

$$\begin{aligned}
&\frac{\dot{l}^*_{\widetilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}})}{n} \\
&= \frac{1}{r} \sum_{i=1}^r \frac{1}{n\pi_i^*(\widetilde{\boldsymbol{\beta}}_0)} w_i^*(y_i^* - \mu_i) \boldsymbol{X}_i^* - \frac{1}{n} \boldsymbol{S}_\lambda \hat{\boldsymbol{\beta}} \\
&= \frac{1}{r} \sum_{i=1}^r \left\{ \frac{1}{n\pi_i^*(\widetilde{\boldsymbol{\beta}}_0)} w_i^*(y_i^* - \mu_i) \boldsymbol{X}_i^* - \frac{1}{n} \boldsymbol{S}_\lambda \hat{\boldsymbol{\beta}} \right\} \\
&:= \frac{1}{r} \sum_{i=1}^r \boldsymbol{\eta}_i^{\widetilde{\boldsymbol{\beta}}_0},
\end{aligned}$$

Furthermore, by Lemma 2,

$$\begin{aligned}
&\mathrm{E}\left\{ \frac{1}{n} \dot{l}^*_{\widetilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}) \right\} = 0, \\
&\mathrm{Var}\left\{ \frac{1}{n} \dot{l}^*_{\widetilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}) \right\} = \boldsymbol{V}_c^{\widetilde{\boldsymbol{\beta}}_0} = O_{P|\mathcal{F}_n}(r^{-1}).
\end{aligned}$$

Conditional on $\mathcal{F}_n$ and $\widetilde{\boldsymbol{\beta}}_0$, $\boldsymbol{\eta}_i^{\widetilde{\boldsymbol{\beta}}_0}$ are independent and identically distributed and $\mathrm{E}\left(\boldsymbol{\eta}_i^{\widetilde{\boldsymbol{\beta}}_0} | \mathcal{F}_n, \widetilde{\boldsymbol{\beta}}_0\right) = 0$, $i = 1, 2, ..., r$. Thus, one can further derive that $\mathrm{Var}\left(\boldsymbol{\eta}_i^{\widetilde{\boldsymbol{\beta}}_0} | \mathcal{F}_n, \widetilde{\boldsymbol{\beta}}_0\right) = r\boldsymbol{V}_c^{\widetilde{\boldsymbol{\beta}}_0} = O_{P|\mathcal{F}_n}(1)$, $i = 1, 2, ..., r$. Moreover, for $\forall \varepsilon > 0$, there exists $\tau > 0$ such that

$$\begin{aligned}
&\sum_{i=1}^r \mathrm{E}\left\{ \|r^{-1/2} \boldsymbol{\eta}_i^{\widetilde{\boldsymbol{\beta}}_0}\|^2 I(\|r^{-1/2} \boldsymbol{\eta}_i^{\widetilde{\boldsymbol{\beta}}_0}\| > \varepsilon) | \mathcal{F}_n, \widetilde{\boldsymbol{\beta}}_0 \right\} \\
&= \sum_{i=1}^r \mathrm{E}\left\{ \|r^{-1/2} \boldsymbol{\eta}_i^{\widetilde{\boldsymbol{\beta}}_0}\|^2 I(\|\boldsymbol{\eta}_i^{\widetilde{\boldsymbol{\beta}}_0}\| > r^{1/2}\varepsilon) | \mathcal{F}_n, \widetilde{\boldsymbol{\beta}}_0 \right\} \\
&\leq \sum_{i=1}^r \mathrm{E}\left\{ \|r^{-1/2} \boldsymbol{\eta}_i^{\widetilde{\boldsymbol{\beta}}_0}\|^2 \left(\frac{\|\boldsymbol{\eta}_i^{\widetilde{\boldsymbol{\beta}}_0}\|}{r^{1/2}\varepsilon} \right)^\tau I(\|\boldsymbol{\eta}_i^{\widetilde{\boldsymbol{\beta}}_0}\| \right. \\
&\qquad \left. > r^{1/2}\varepsilon) | \mathcal{F}_n, \widetilde{\boldsymbol{\beta}}_0 \right\} \\
&= \frac{1}{r^{1+\frac{\tau}{2}}\varepsilon^\tau} \sum_{i=1}^r \mathrm{E}\left\{ \|\boldsymbol{\eta}_i^{\widetilde{\boldsymbol{\beta}}_0}\|^{2+\tau} I(\|\boldsymbol{\eta}_i^{\widetilde{\boldsymbol{\beta}}_0}\| \right. \\
&\qquad \left. > r^{1/2}\varepsilon) | \mathcal{F}_n, \widetilde{\boldsymbol{\beta}}_0 \right\} \\
&\leq \frac{1}{r^{1+\frac{\tau}{2}}\varepsilon^\tau} \mathrm{E}\left(\sum_{i=1}^r \|\boldsymbol{\eta}_i^{\widetilde{\boldsymbol{\beta}}_0}\|^{2+\tau} \right) \\
&= \frac{1}{r^{1+\frac{\tau}{2}}\varepsilon^\tau} \mathrm{E}\left\{ \sum_{i=1}^r \left\| \frac{1}{n\pi_i^*(\widetilde{\boldsymbol{\beta}}_0)} w_i^*(y_i^* - \mu_i) \boldsymbol{X}_i^* \right. \right. \\
&\qquad \left. \left. - \frac{1}{n} \boldsymbol{S}_\lambda \hat{\boldsymbol{\beta}} \right\|^{2+\tau} \right\} \\
&\leq \frac{1}{r^{1+\frac{\tau}{2}}\varepsilon^\tau} \sum_{i=1}^r \left\{ \frac{\|w_i^*(y_i^* - \mu_i) \boldsymbol{X}_i^*\|^{2+\tau}}{n^{2+\tau}\{\pi_i^*(\widetilde{\boldsymbol{\beta}}_0)\}^{2+\tau}} \right. \\
&\qquad \left. + \frac{\|\boldsymbol{S}_\lambda \hat{\boldsymbol{\beta}}\|^{2+\tau}}{n^{2+\tau}} \right\} \\
&= \frac{1}{r^{\frac{\tau}{2}}\varepsilon^\tau} \sum_{i=1}^n \frac{\|w_i(y_i - \mu_i) \boldsymbol{X}_i^*\|^{2+\tau}}{n^{2+\tau}\{\pi_i(\widetilde{\boldsymbol{\beta}}_0)\}^{1+\tau}} \\
&\qquad + \frac{1}{r^{\frac{\tau}{2}}\varepsilon^\tau} \frac{\|\boldsymbol{S}_\lambda \hat{\boldsymbol{\beta}}\|^{2+\tau}}{n^{2+\tau}} \\
&\leq \frac{1}{r^{\frac{\tau}{2}}\varepsilon^\tau} \sum_{i=1}^n \frac{|w_i|^{2+\tau} |y_i - \mu_i|^{2+\tau} \|\boldsymbol{X}_i\|^{2+\tau}}{n^{2+\tau}\{\pi_i(\widetilde{\boldsymbol{\beta}}_0)\}^{1+\tau}} \\
&\qquad + \frac{1}{r^{\frac{\tau}{2}}\varepsilon^\tau} \frac{\|\boldsymbol{S}_\lambda\|^{2+\tau} \|\hat{\boldsymbol{\beta}}\|^{2+\tau}}{n^{2+\tau}}.
\end{aligned}$$

Given Assumption 3, by Lindeberg-Feller central limit theorem,

$$\begin{aligned}
&\frac{1}{n} (\boldsymbol{V}_c^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2} \dot{l}_{\widetilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}) \\
&= r^{-\frac{1}{2}} \left\{ \mathrm{Var}\left(\boldsymbol{\eta}_i^{\widetilde{\boldsymbol{\beta}}_0} | \mathcal{F}_n, \widetilde{\boldsymbol{\beta}}_0 \right) \right\}^{-\frac{1}{2}} \sum_{i=1}^r \boldsymbol{\eta}_i^{\widetilde{\boldsymbol{\beta}}_0} \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{I}). \quad (7)
\end{aligned}$$

By Theorem 5, we have

$$(\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2} (\check{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) = O_{P|\mathcal{F}_n}(1).$$

By Lemma 2 and Theorem 5,

$$\begin{aligned}
&(\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2} (\check{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \\
&= (\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2} (\widetilde{\boldsymbol{D}}_x^{\widetilde{\boldsymbol{\beta}}_0})^{-1} \frac{\dot{l}^*_{\widetilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}})}{n}
\end{aligned}$$

$$
\begin{aligned}
&= -(\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2}\boldsymbol{D}_x^{-1}\frac{\dot{l}^*_{\widetilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}})}{n} \\
&\quad + \left[-(\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2}\left\{(\widetilde{\boldsymbol{D}}_x^{\widetilde{\boldsymbol{\beta}}_0})^{-1} - \boldsymbol{D}_x^{-1}\right\}\right]\frac{\dot{l}^*_{\widetilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}})}{n} \\
&= -(\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2}\boldsymbol{D}_x^{-1}\frac{\dot{l}^*_{\widetilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}})}{n} + O_{P|\mathcal{F}_n}(r^{-1/2}) \\
&= -(\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2}\boldsymbol{D}_x^{-1}(\boldsymbol{V}_c^{\widetilde{\boldsymbol{\beta}}_0})^{1/2}(\boldsymbol{V}_c^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2}\frac{\dot{l}^*_{\widetilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}})}{n} \\
&\quad + O_{P|\mathcal{F}_n}(r^{-1/2}).
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
&\mathrm{Var}\left\{(\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2}\boldsymbol{D}_x^{-1}(\boldsymbol{V}_c^{\widetilde{\boldsymbol{\beta}}_0})^{1/2}\right\} \\
&= (\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2}\boldsymbol{D}_x^{-1}(\boldsymbol{V}_c^{\widetilde{\boldsymbol{\beta}}_0})^{1/2}\left\{(\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2}\boldsymbol{D}_x^{-1}\right. \\
&\quad \left.(\boldsymbol{V}_c^{\widetilde{\boldsymbol{\beta}}_0})^{1/2}\right\}^{\top} \\
&= (\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2}\boldsymbol{D}_x^{-1}\boldsymbol{V}_c\boldsymbol{D}_x^{-1}(\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2} \\
&\quad + (\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2}\boldsymbol{D}_x^{-1}(\boldsymbol{V}_c^{\widetilde{\boldsymbol{\beta}}_0} - \boldsymbol{V}_c)\boldsymbol{D}_x^{-1}(\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2} \\
&= (\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2}(\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})(\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2} \\
&\quad + (\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2}\boldsymbol{D}_x^{-1}(\boldsymbol{V}_c^{\widetilde{\boldsymbol{\beta}}_0} - \boldsymbol{V}_c)\boldsymbol{D}_x^{-1}(\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2} \\
&= \boldsymbol{I} + (\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2}\boldsymbol{D}_x^{-1}(\boldsymbol{V}_c^{\widetilde{\boldsymbol{\beta}}_0} - \boldsymbol{V}_c)\boldsymbol{D}_x^{-1}(\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2}.
\end{aligned}
$$

Note that

$$
\begin{aligned}
\|\boldsymbol{V}_c - \boldsymbol{V}_c^{\widetilde{\boldsymbol{\beta}}_0}\| &= \left\| \frac{1}{rn^2}\sum_{i=1}^{n}\frac{w_i^2(y_i-\mu_i)^2\boldsymbol{X}_i\boldsymbol{X}_i^{\top}}{\pi_i^{\mathrm{mV_c}}}\right. \\
&\quad \left. - \frac{1}{rn^2}\sum_{i=1}^{n}\frac{w_i^2(y_i-\mu_i)^2\boldsymbol{X}_i\boldsymbol{X}_i^{\top}}{\pi_i(\widetilde{\boldsymbol{\beta}}_0)}\right\| \\
&= \left\| \frac{1}{rn^2}\sum_{i=1}^{n} w_i^2(y_i-\mu_i)^2\boldsymbol{X}_i\boldsymbol{X}_i^{\top}\right. \\
&\quad \left.\left(\frac{1}{\pi_i^{\mathrm{mV_c}}} - \frac{1}{\pi_i(\widetilde{\boldsymbol{\beta}}_0)}\right)\right\| \\
&\le \frac{1}{rn^2}\sum_{i=1}^{n} w_i^2|y_i-\mu_i|^2\|\boldsymbol{X}_i\|^2\left\|\frac{1}{\pi_i^{\mathrm{mV_c}}} - \frac{1}{\pi_i(\widetilde{\boldsymbol{\beta}}_0)}\right\|.
\end{aligned}
$$

Since $w_i^2 \le M_1$ and $|y_i-\mu_i|^2 \le M_2$,

$$
\|\boldsymbol{V}_c - \boldsymbol{V}_c^{\widetilde{\boldsymbol{\beta}}_0}\| \le \frac{M_1M_2}{rn^2}\sum_{i=1}^{n}\frac{\|\boldsymbol{X}_i\|^2}{\pi_i^{\mathrm{mV_c}}}\left\|1 - \frac{\pi_i^{\mathrm{mV_c}}}{\pi_i(\widetilde{\boldsymbol{\beta}}_0)}\right\|,
$$

where

$$
\begin{aligned}
\left\|1 - \frac{\pi_i^{\mathrm{mV_c}}}{\pi_i(\widetilde{\boldsymbol{\beta}}_0)}\right\|^2 &= \frac{1}{|\pi_i(\widetilde{\boldsymbol{\beta}}_0)|^2}\|\pi_i(\widetilde{\boldsymbol{\beta}}_0) - \pi_i^{\mathrm{mV_c}}\|^2 \\
&\le \frac{\|\pi_i(\widetilde{\boldsymbol{\beta}}_0) - \pi_i^{\mathrm{mV_c}}\|^2}{k^2} = o_P(1).
\end{aligned}
$$

Thus, according to Assumption 3,

$$
\|\boldsymbol{V}_c - \boldsymbol{V}_c^{\widetilde{\boldsymbol{\beta}}_0}\| \le o_{P|\mathcal{F}_n}(r^{-1}).
$$

Thus,

$$
\mathrm{Var}\left\{(\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2}\boldsymbol{D}_x^{-1}(\boldsymbol{V}_c^{\widetilde{\boldsymbol{\beta}}_0})^{1/2}\right\} = \boldsymbol{I} + o_{P|\mathcal{F}_n}(1).
$$

Thus, by Eq. (7),

$$
\mathrm{Var}\left\{(\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2}(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})\right\} = \boldsymbol{I}.
$$

By Slutsky Theorem and Eq. (7),

$$
(\boldsymbol{V}^{\widetilde{\boldsymbol{\beta}}_0})^{-1/2}(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{I}).
$$

□

Acknowledgements The authors sincerely thank the Editor, Associate Editor, and two referees for their valuable and insightful comments, which have significantly contributed to the improvement of this work. The research of L. Li is supported by the National Social Science Fund of China (NSSFC) under Grant No. 19BTJ028. H. Shi's research is funded by the Discovery Grant (RGPIN-2021-02963) from the Natural Sciences and Engineering Research Council of Canada (NSERC). J. Cao's research is supported by the NSERC Discovery Grant (RGPIN-2023-04057) and the Canada Research Chair program.

References

Ai, M., Yu, J., Zhang, H., et al.: Optimal subsampling algorithms for big data regressions. Stat. Sin. **31**(2), 749–772 (2021)

Atkinson, A., Donev, A., Tobias, R.: Optimum Experimental Designs, with SAS, vol. 34. OUP Oxford (2007)

Buldygin, V.V., Kozachenko, Y.V.: Sub-Gaussian random variables. Ukr. Math. J. **32**, 483–489 (1980)

Drineas, P., Mahoney, M.W., Muthukrishnan, S.: Sampling algorithms for l 2 regression and applications. In: Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm, pp. 1127–1136 (2006)

Drineas, P., Mahoney, M.W., Muthukrishnan, S., et al.: Faster least squares approximation. Numerische mathematik **117**(2), 219–249 (2011)

Ferguson, T.S.: A Course in Large Sample Theory. Routledge (1996)

Hastie, T.J.: Generalized additive models. In: Statistical Models in S, pp. 249–307. Routledge (2017)

Hebrail, G. & Berard, A. (2006). Individual Household Electric Power Consumption [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C58K54.

Kiefer, J.: Optimum experimental designs. J. Roy. Stat. Soc.: Ser. B (Methodol.) **21**(2), 272–304 (1959)
Lee, J., Schifano, E.D., Wang, H.: Sampling-based Gaussian mixture regression for big data. J. Data Sci. **21**(1), 158–172 (2022)
Ma, P., Mahoney, M., Yu, B.: A statistical perspective on algorithmic leveraging. In: International Conference on Machine Learning, pp. 91–99 (2014)
Mahoney, M.W., Drineas, P.: CUR matrix decompositions for improved data analysis. Proc. Natl. Acad. Sci. **106**(3), 697–702 (2009)
Sui, Q., Ghosh, S.K.: Entropy-based subsampling methods for big data. J. Stat. Theory Prac. **18**(2), 24 (2024)
Wang, H., Zhu, R., Ma, P.: Optimal subsampling for large sample logistic regression. J. Am. Stat. Assoc. **113**(522), 829–844 (2018)
Wang, H., Yang, M., Stufken, J.: Information-based optimal subdata selection for big data linear regression. J. Am. Stat. Assoc. **114**(525), 393–405 (2019)
Wang, L., Elmstedt, J., Wong, W.K., et al.: Orthogonal subsampling for big data linear regression. Ann. Appl. Stat. **15**(3), 1273–1290 (2021)
Yao, Y., Wang, H.: A review on optimal subsampling methods for massive datasets. J. Data Sci. **19**(1), 151–172 (2021)
Yao, Y., Zou, J., Wang, H.: Model constraints independent optimal subsampling probabilities for softmax regression. J. Stat. Plan. Inference **225**, 188–201 (2023)
Yao, Y., Zou, J., Wang, H.: Optimal poisson subsampling for softmax regression. J. Syst. Sci. Complexity **36**(4), 1609–1625 (2023)
Yu, J., Wang, H., Ai, M., et al.: Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. J. Am. Stat. Assoc. **117**(537), 265–276 (2022)
Yu, J., Liu, J., Wang, H.: Information-based optimal subdata selection for non-linear models. Stat. Pap. **64**, 1069–1093 (2023)
Yu, J., Ai, M., Ye, Z.: A review on design inspired subsampling for big data. Stat. Pap. **65**(2), 467–510 (2024)
Zhang, H., Wang, H.: Distributed subdata selection for big data via sampling-based approach. Comput. Stat. Data Anal. **153**, 107072 (2021)
Zhu, R., Ma, P., Mahoney, M.W., et al.: Optimal subsampling approaches for large sample linear regression. arXiv Preprint at arXiv:1509.05111 (2015)
Zuo, L., Zhang, H., Wang, H., et al.: Optimal subsample selection for massive logistic regression with distributed data. Comput. Stat. **36**, 2535–2562 (2021)