THE A-OPTIMAL SUBSAMPLING FOR BIG DATA PENALIZED SPLINE SINGLE INDEX MODELS

A PREPRINT

Haixia Smithson[*] **Fang Li**[†] **and Hanxiang Peng**[‡]

April 11, 2024

ABSTRACT

Motivated by the computational burden in fitting single index models caused by high parameter dimensionality and possibly compounded by data of massive size, we construct the A-optimal subsampling estimators to approximate the full data estimators. We derive the A-optimal sampling distributions by minimizing the sum of the component variances of the subsampling estimator. For an arbitrary distribution (π_i) on the n data points with its minimum $\pi_{\min}$ satisfying $n\pi_{\min} \geq l_0 > 0$ for some constant l_0, we prove asymptotic normality of the subsampling estimator for either fixed or growing sum $p + d$ of the number p of the index parameters and the number d of basis functions as the subsample size r tends to infinity such that $p + d$ grows slowly at the rate $p + d = o(r^{1/5})$ under suitable conditions. We also construct an unweighted subsampling estimator, prove its asymptotic normality for growing dimension without the foregoing assumption on (π_i), and establish its higher efficiency than the weighted estimator. We provide the analytic formulas of the first-order bias for both estimators, and explore how the estimators and their biases are affected by the penalty λ, $p + d$, (π_i) and r. We construct a fast algorithm having running time $O(r^2(p + d))$ with r far less than n, and study the numerical behavior of the Subsampling approach using both simulated and real data. Our results indicated that the proposed approach significantly outperformed the uniform subsampling and substantially reduced the amount of computing time.

***K*eywords** A-optimality · Asymptotic normality · Big data · Infinite dimension · Inverse Probability · Penalized spline · Single index model

1 Introduction

The single index model (SIM) is a hybrid of parametric and nonparametric models. It generalizes the linear regression model by introducing a nonparametric link function, extends the generalized linear models (GLM) by allowing the link to be unknown and data-driven, and can model interactions among covariate variables through the *index*. Most importantly, the index reduces a multivariate predictor to a univariate quantity, thus avoids the "curse of dimensionality" problem in a fully nonparametric model. The numerical implementation in fitting SIM is, however, computational challenging, which is aggravated by data of massive size in the Era of Big Data. To tackle the challenge, one could use the computationally easy uniform subsampling or the popular statistical-leverage-scores-based subsampling. The two approaches, however, are not efficient in extracting important observations, see Zhang, *et al.* (2023) for a detailed discussion in a linear regression model. Motivated by this, we propose the A-optimal Subsampling approach based on the criterion of minimizing the sum of the component variances (equivalently, the trace of the variance-covariance matrix) of the subsampling estimator.

There is an extensive amount of literature on the methods of estimation for SIM, which can be grouped into the direct and indirect approaches. The former estimates the index parameter $\boldsymbol{\beta}$ without estimating the mean function $m(x)$ (see

[*] US Food and Drug Administration, 10903 New Hampshire Ave, White Oak, MD 20903, USA. haixia.smithson@fda.hhs.gov

[†] Department of Mathematical Sciences, IUPUI, 402 N. Blackford Street, Indianapolis, IN 46202, USA. fali@iupui.edu

[‡] Department of Mathematical Sciences, IUPUI, 402 N. Blackford Street, Indianapolis, IN 46202, USA. hanxpeng@iupui.edu

(2.1)), including the average derivative method (Stoker, 1986), the local linear estimation (Hristache, *et al.*, 2001), and the methods involving the conditional variance of Y (Xia, *et al.*, 2002 and Xia, 2006). The latter estimates $\boldsymbol{\beta}$ by minimizing certain objective function after $m(x)$ is estimated by some nonparametric regression method such as the kernel smoothing (Ichimura, 1993 and Hardle, *et al.*, 1993), the penalized splines (the P-splines, Yu & Ruppert, 2002), the regression splines and the B-splines (both of Antoniadis, *et al.*, 2004). In this article, we shall present the A-optimal Subsampling approach for the penalized spline SIM.

The P-splines gained popularity in 1990's as a flexible smoothing method for semiparametric regression. The idea of penalization was originally from O'Sullivan (1986) who proposed the integrated squared derivative of the fitted curve as the penalty. Then Eiler & Marx (1992) derived the difference penalty, which is purely discrete, thus much simpler as it is trivial to calculate the difference of any order. Later in 1996, they proposed a benchmark method of curve fitting by combining regressions with the B(Basic)-spline basis and their difference penalty. Subsequently, Ruppert & Carrol (1997) proposed to use the truncated power function (TPF) basis as components of penalized splines with smoothness from a ridge penalty on the coefficients of parameters. Later, Ruppert & Carroll (2000) and Yu & Ruppert (2002) used TPF in the basis with equally spaced quantiles as knots plus a partial ridge penalty on the model function, and they named their approach as the P-spline. Their work has greatly advanced the study and applications of penalized splines. Since then, penalized splines become increasingly popular and are extended to various regression models for different purposes. Recent work includes the penalized spline estimation for generalized partially linear single-index models by Yu, *et al.* (2017), the variable selections for SIM with diverging number of index parameters by Wang & Wang (2015), the multivariate single index models for longitudinal data by Wu & Tu (2016), and the large-sample estimation and inference in multivariate single-index models by Wu, *et al.* (2019). Recently, Jiang and Peng (2023) proposed a computationally efficient method for estimation in a Big Data SIM using the divide-and-conquer method, and showed that the resulting estimator possesses the optimal convergence rate with no restriction on the number of dividing blocks. Robust estimation in SIM especially for data of massive size is investigated in Jiang, *et al.* (2022). Instead of the usual quadratic loss function only, the authors adopted a weighted linear combination of several loss functions to accommodate the diverse data structures, combined with the divide-and-conquer method. They demonstrated that their approach significantly reduces the memory space and the resulting estimator attains the model efficiency.

2 Overview of the Subsampling Approach in Single Index Models

In a SIM, the response y_i and the covriate $\mathbf{x}_i$ satisfy

$$y_i = m_0(\boldsymbol{\beta}_0^t \mathbf{x}_i) + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{2.1}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is an unknown *index parameter* which satisfies $\|\boldsymbol{\beta}_0\| = 1$ with its first component $\beta_1 > 0$ for identifiability, $m_0(x)$ is an unknown nonparametric function on the reals $\mathbb{R}$, and $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed (i.i.d.) random errors with zero mean and constant variance $\sigma_0^2 = \text{Var}(\epsilon_i) > 0$. Here $\boldsymbol{a}^t$ denotes the transpose of a column vector $\boldsymbol{a}$. We assume that the covariates $\mathbf{x}_i$ are nonrandom, although the results typically hold for random covariates.

In a penalized spline SIM, the mean function $m_0(x)$ in (2.1) is approximated by

$$m(x) = \boldsymbol{\delta}^t \mathbf{B}(x), \quad x \in \mathbb{R}, \tag{2.2}$$

where $\boldsymbol{\delta} \in \mathbb{R}^d$ is an unknown parameter vector of coefficients, and $\mathbf{B}(x)$ is a vector of basis functions. See the penalized B-spline and P-spline SIM considered in Section 6. The parameter $\boldsymbol{\theta} =: (\boldsymbol{\beta}^t, \boldsymbol{\delta}^t)^t \in \mathbb{R}^{p+d}$ can be estimated by minimizing the penalized squared residuals,

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - \boldsymbol{\delta}^t \mathbf{B}(\boldsymbol{\beta}^t \mathbf{x}_i)\big)^2 + \lambda P(\boldsymbol{\theta}), \tag{2.3}$$

subject to the constraints

$$\|\boldsymbol{\beta}\| = 1 \quad \text{and} \quad \beta_1 > 0, \tag{2.4}$$

where λ is a *penalty (or tuning) parameter* and $P(\boldsymbol{\theta})$ is a *penalty function.*

Many penalties can be found in literature including the partial ridge penalty on $\boldsymbol{\delta}$ (Yu & Ruppert, 2002), the penalty on the integrated second derivative of fitted curve (Osullivan, 1986 & 1988), the difference penalty (Eilers & Marx, 1996), and the SCAD penalty for variable selection (Fan & Li, 2001). The value of the turning parameter λ is found through a grid search based on some criteria such as minimizing the cross-validation (CV) score, the generalized cross-validation (GCV) score, or Akaike's information criterion (AIC). In this paper, we shall select, suggested by Yu & Ruppert (2002),

Algorithm 1

1. Initialization. Calculate the LSE $\hat{\boldsymbol{\beta}}_0$ via linear regression of $\mathbf{y}$ on $\mathbf{X}$, compute $\hat{\boldsymbol{\phi}}_0$ by (2.8), and find $\hat{\boldsymbol{\delta}}_0$ as the minimizer of $Q_n(\boldsymbol{\theta}_{\boldsymbol{\phi}})$ in (2.10) after substituting $\boldsymbol{\phi} = \hat{\boldsymbol{\phi}}_0$.
2. Optimization. Calculate $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}} = (\hat{\boldsymbol{\phi}}^t,\, \hat{\boldsymbol{\delta}}^t)^t$ as the minimizer of $Q_n(\boldsymbol{\theta}_{\boldsymbol{\phi}})$ in (2.10) initialized at $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}_0} = (\hat{\boldsymbol{\phi}}_0^t,\, \hat{\boldsymbol{\delta}}_0^t)^t$.
3. Reparametrization. Calculate $\hat{\boldsymbol{\beta}}$ by (2.7) and obtain $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^t,\, \hat{\boldsymbol{\delta}}^t)^t$.
4. Cross validation. To each value of λ on the grid, repeat 1–3, find the value of λ which minimizes the GCV score in (2.5), and choose the corresponding estimator as the final estimator.

a grid of 30-points in which $\log_{10}(\lambda)$ are equally spaced quantiles in the interval $[-6, 7]$. We then choose the value of λ which minimizes the GCV score,

$$\text{GCV}(\lambda) = \frac{n^{-1}\sum_{i=1}^{n}\big(y_i - \boldsymbol{\delta}^t\mathbf{B}(\boldsymbol{\beta}^t\mathbf{x}_i)\big)^2}{\big(1 - n^{-1}\text{Tr}(\mathbb{A}(\lambda))\big)^2}, \tag{2.5}$$

where $\mathbb{A}(\lambda)$ is the hat matrix of the penalized spline SIM, which is defined by

$$\mathbb{A}(\lambda) = \mathbb{B}(\mathbb{B}^t\mathbb{B} + n\lambda\mathbb{D})^{-1}\mathbb{B}^t,$$

Here $\mathbb{B}$ is an $n \times d$ matrix with its ith row equal to $\mathbf{B}^t(\boldsymbol{\beta}^t\mathbf{x}_i)$, and $\mathbb{D}$ is some positive definite matrix (see two examples in (6.2) and (6.6) below). Hence the fitted value is

$$\hat{\mathbf{y}} = \mathbb{A}(\lambda)\mathbf{y}, \tag{2.6}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_n)^t$. Writing $(\mathbf{X}, \mathbf{y})$ with $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^t$ for the full data, the trace of $\mathbb{A}(\lambda)$ can be expressed as $\text{Tr}(\mathbb{A}(\lambda)) = \text{Tr}((\mathbb{B}^t\mathbb{B} + n\lambda\mathbb{D})^{-1}\mathbb{B}^t\mathbb{B})$, where we performed the cyclic permutation for the trace function, which reduces the matrix dimension and saves the memory for storing big matrices. As is customary, the constraints in (2.4) on $\boldsymbol{\beta}$ can be handled by the reparametrization,

$$\boldsymbol{\beta}(\boldsymbol{\phi}) = \frac{(1, \boldsymbol{\phi}^t)^t}{\sqrt{1 + \|\boldsymbol{\phi}\|^2}}, \quad \boldsymbol{\phi} \in \mathbb{R}^{p-1}. \tag{2.7}$$

This has the inverse given by

$$\boldsymbol{\phi} =: \boldsymbol{\phi}(\boldsymbol{\beta}) = \beta_1^{-1}(\beta_2, \ldots, \beta_p)^t. \tag{2.8}$$

The parameters to be estimated become $\boldsymbol{\theta}_{\boldsymbol{\phi}} = (\boldsymbol{\phi}^t, \boldsymbol{\delta}^t)^t \in \mathbb{R}^{p+d-1}$, losing one dimension from the original $\boldsymbol{\theta}$. The Jacobian matrix of transformation $\boldsymbol{\theta} = (\boldsymbol{\beta}(\boldsymbol{\phi})^t, \boldsymbol{\delta}^t)^t \mapsto \boldsymbol{\theta}_{\boldsymbol{\phi}}$ can readily be found as

$$\frac{\partial\boldsymbol{\theta}}{\partial(\boldsymbol{\phi}^t, \boldsymbol{\delta}^t)} = \begin{pmatrix} -\frac{\boldsymbol{\phi}^t}{(1+\|\boldsymbol{\phi}\|^2)^{3/2}} & \mathbf{0}_{1\times d} \\ \frac{\mathbf{I}_{p-1}}{\sqrt{1+\|\boldsymbol{\phi}\|^2}} - \frac{\boldsymbol{\phi}^{\otimes 2}}{(1+\|\boldsymbol{\phi}\|^2)^{3/2}} & \mathbf{0}_{(p-1)\times d} \\ \mathbf{0}_{d\times(p-1)} & \mathbf{I}_d \end{pmatrix}. \tag{2.9}$$

The objective function in (2.3) is now transformed to

$$Q_n(\boldsymbol{\theta}_{\boldsymbol{\phi}}) = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - \boldsymbol{\delta}^t\mathbf{B}(\boldsymbol{\beta}(\boldsymbol{\phi})^t\mathbf{x}_i)\big)^2 + \lambda P(\boldsymbol{\theta}_{\boldsymbol{\phi}}). \tag{2.10}$$

Let $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}} = (\hat{\boldsymbol{\phi}}^t,\, \hat{\boldsymbol{\delta}}^t)^t$ be the resulting estimator of $\boldsymbol{\theta}_{\boldsymbol{\phi}}$. Substituting it in (2.7), we obtain the plug-in estimator $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{\boldsymbol{\phi}})$ of $\boldsymbol{\beta}$. The numeric computation of the estimator can be implemented as follows:

Observe that the numerical implementation bears an enormous computational burden. Specifically, the optimization in (2.3) using Algorithm 1 in Alg. 1 through the newton or the quasi-newton method takes $O(n^2(p+d))$ running time in each of the iterations needed for numerical convergence, plus the cross validation process. As a consequence, the numerical computation can be challenging even for data of conventional sample size n and parameter dimension $p + d$, let alone for data of massive size compounded by parameters of high dimensionality.

Motivated by the computational challenge, we propose the A-optimal Subsampling approach to downsizing data and constructing both the weighted and unweighted subsampling estimators of the parameters using a subdata as a surrogate, which is feasible both economically and temporally. We investigate how the subsampling estimator is affected by the growing number $p+d$ of parameters, the subsample size r, and the sampling distribution $\boldsymbol{\pi}=(\pi_1,\pi_2,\ldots,\pi_n)$. We establish asymptotic normality of the subsampling estimator for either fixed or growing $p+d$ to infinity slowly with r such that $(p+d)/r^{1/5}\to 0$, for an arbitrary sampling distribution $\boldsymbol{\pi}=(\pi_i)$ with its minimum $\pi_{\min}=\min(\pi_i)$ satisfying $\pi_{\min}\geq l_0/n$ for some constant $l_0>0$, under additional conditions. This is detailed in Theorem 1 and Remark 1. The message here is that the number of parameters $p+d$ that can be fitted by data of subsample size r must satisfy $p+d=o(r^{1/5})$, which can't be improved for penalized SIM. See Examples 1 and 2. This number of parameters in SIM is much smaller than the optimal number $p+d=o(r^{1/2})$, which is typical in the analysis of Big Data, see, e.g. page 29 of Bühlmann, *et al.* (2016), and Portnoy (1984, 1985, 1988). In practice, the value of d is often chosen to be from the integers between 5 and 10 as suggested by Ruppert (2002). Our consideration of allowing d to grow with r sheds some light on the approximation of the unknown link function $m_0(x)$ in (2.1) by the expansion $m(x)$ in (2.2) along a sequence of basis functions. Results on growing dimension in the literature are abundant, see the above Portnory, Mammen (1993) and Chatterjee and Bose (2005) among others.

Full sample randomly weighted bootstrap estimators were well studied in the literature, see the monograph by Barbe and Bertail (1995). Unlike the construction of weighted bootstrap estimators, our subsampling estimators is constructed based on *the scheme of importance sampling* in which the weights are inversely used. The scheme is commonly used in survey sampling and recently in the causal inference for models with latent variables. Our results contribute to the scheme in high dimensional regression models including GLM, and provide the theoretical guarantee for various nonuniform sampling distributions such as the leverage-scores-based distribution and the A-optimal distributions given below.

We conducted extensive simulations and real data applications to investigate the numerical behavior of the A-optimal subsampling estimator, and our results indicated that the proposed Subsampling approach significantly outperformed the uniform subsampling by the criterion of mean squared errors, and used substantially less computing time than the full-data estimator.

We have not applied our Subsampling approach to conduct statistical inference in the SIM. It is apparent that our optimal Subsampling estimator will yield shorter confidence intervals or more powerful tests than those using other subsampling estimators including the uniform (the bootstrap) and the leverage-scores-based subsampling estimators. Here the confidence intervals and the tests are based on asymptotic normality. An application of the A-optimal subsampling approach in statistical inference in real data can be found in Tan, *et al.* (2023), where the P-values for many covariate variables in Poisson regression model using the A-optimal subsampling approach were significant while those using the uniform subsampling were not. This doesn't suggest, however, that the A-optimal Subsampling approach will always yield "optimal" results in statistical inference, because the A-optimality is *estimate-dependent*, that is, an A-optimal distribution for one estimator is not A-optimal for another different estimator. For example, the prediction made based on the model whose coefficients are estimated by the A-optimal subsampling estimator is not "optimal" in terms of EMSE in general, although one can expect that the prediction based on the A-optimal subsampling estimator has smaller EMSE than the one based on the uniform subsampling estimator.

We have employed the B-spline and the P-spline SIM to investigate the numerical performance of the A-optimal Subsampling approach. But we haven't specifically examined the two models as our goal is not to compare them although the comparison would certainly aid to further our understanding of the approach. Examining the TPF (truncated power functions) and the B-spline basis, one can see that the TPF is simpler and practically useful for understanding the spline regression, but it was not numerically stable in the optimizations considered in the case of either a large number of knots, the penalty parameter λ close to zero, or datasets of large size. In these cases, typical algorithms such as the Gauss-Newton algorithm suggested by Yu and Ruppert (2002) do not work well. The B-splines, however, are easy to calculate and numerically superior. For a close look of the difference of the two splines, see e.g. Sharif and Kamal (2018). Both splines have problems for large datasets as the dimension of the basis needs to increase accordingly for precision, which increases the dimensionality of the optimization procedure. From this viewpoint, our choice of the Subsampling approach is quite suitable, hence the meaningfulness of the study of the optimal Subsampling follows.

The article is structured as follows: In Section 3, we construct the subsampling estimator in a penalized spline SIM, and presents the asymptotic normality of the subsampling estimator for both fixed and growing dimension. In Section 4, the A-optimal distributions are derived, the Scoring Algorithm is constructed, and numerical implementation and truncation are discussed. Section 6 contains the simulations. Section 7 reports two real data applications. The proof is provided in Section 8, and supplementary tables in Section 9.

3 The Subsampling Estimator and Asymptotic Normality

In this section, we construct the subsampling estimator and prove the asymptotic normality for growing dimension.

Algorithm 2

1. Using $\boldsymbol{\pi}$, take a random subsample $(\mathbf{X}^*, \mathbf{y}^*)$ of size $r << n$ with replacement from the full sample $(\mathbf{X}, \mathbf{y})$, and formulate the sampling probability vector $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_r^*)$.
2. Calling **Algorithm 1** to the subdata $(\mathbf{X}^*, \mathbf{y}^*)$ with the vector $1/(r\boldsymbol{\pi}^*)$ as weights, calculate the subsampling estimator $\hat{\boldsymbol{\theta}}^* = (\hat{\boldsymbol{\beta}}^{*t},\, \hat{\boldsymbol{\delta}}^{*t})^t$.

3.1 The (Weighted) Subsampling Estimator

Let $\boldsymbol{\pi} = (\pi_i)$ be a sampling distribution. The uniform sampling corresponds to $\boldsymbol{\pi} = \mathbf{1}/n$, where $\mathbf{1}$ denotes the vector of all components equal to 1. Later, we shall derive the A-optimal distributions. Using it, take a random subsample of size $r(r << n)$ from the full data $(\mathbf{X}; \mathbf{y})$ with replacement, denoted it by $(\mathbf{X}^*, \mathbf{y}^*)$ with the corresponding sampling probabilities $\boldsymbol{\pi}^* = (\pi_1^*, \pi_2^*, \ldots, \pi_r^*)$. The subsampling estimator $\hat{\boldsymbol{\theta}}^*_{\boldsymbol{\phi}} = (\hat{\boldsymbol{\phi}}^{*t}, \hat{\boldsymbol{\delta}}^{*t})^t$ then minimizes the weighted objective function,

$$Q_n^*(\boldsymbol{\theta}_{\boldsymbol{\phi}}) = \frac{1}{r}\sum_{j=1}^{r} \frac{\left(y_j^* - \boldsymbol{\delta}^t \mathbf{B}(\boldsymbol{\beta}(\boldsymbol{\phi})^t \mathbf{x}_j^*)\right)^2}{n\pi_j^*} + \lambda P(\boldsymbol{\theta}_{\boldsymbol{\phi}}). \tag{3.1}$$

By (2.7), we obtain $\hat{\boldsymbol{\theta}}^* = (\boldsymbol{\beta}(\hat{\boldsymbol{\phi}}^*)^t,\, \hat{\boldsymbol{\delta}}^{*t})^t$. Notice that (3.1) is the weighted version of (2.10) based on the subsample $(\mathbf{X}^*, \mathbf{y}^*)$ with the random vector $1/(r\boldsymbol{\pi}^*)$ as weights. For the uniform sampling, the weight vector reduces to $\mathbf{1}$, we obtain the bootstrap subsampling estimator. Note that $Q_n^*(\boldsymbol{\theta}_{\boldsymbol{\phi}})$ is the Hansen-Hurwitz estimator of $Q_n(\boldsymbol{\theta}_{\boldsymbol{\phi}})$, that is, $Q_n^*(\boldsymbol{\theta}_{\boldsymbol{\phi}})$ is an unbiased estimator of $Q_n(\boldsymbol{\theta}_{\boldsymbol{\phi}})$, $\mathrm{E}^* Q_n^*(\boldsymbol{\theta}_{\boldsymbol{\phi}}) = Q_n(\boldsymbol{\theta}_{\boldsymbol{\phi}})$, where E^* denotes the conditional expectation given the data.

3.2 Asymptotic Normality

We shall suppress $\boldsymbol{\phi}$ and write $\boldsymbol{\theta} = \boldsymbol{\theta}_{\boldsymbol{\phi}} = (\boldsymbol{\phi}^t, \boldsymbol{\delta}^t)^t$ unless otherwise specified. Let $f_i(\boldsymbol{\theta}) = \boldsymbol{\delta}^t \mathbf{B}(\mathbf{x}_i^t \boldsymbol{\beta}(\boldsymbol{\phi}))$, and let

$$e_i(\boldsymbol{\theta}) = y_i - f_i(\boldsymbol{\theta}), \quad i = 1, 2, \ldots, n.$$

For a continuously differentiable penalty function $P(\boldsymbol{\theta})$ with gradient $P'(\boldsymbol{\theta})$, the minimizer $\hat{\boldsymbol{\theta}} =: \hat{\boldsymbol{\theta}}_n$ of $Q_n(\boldsymbol{\theta})$ in (2.10) satisfies $\boldsymbol{\Phi}_n(\boldsymbol{\theta}) = Q_n'(\boldsymbol{\theta}) = 0$. Specifically, $\hat{\boldsymbol{\theta}}$ solves the equation

$$\boldsymbol{\Phi}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\phi}_i(\boldsymbol{\theta}) =: \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\boldsymbol{\theta}) + \lambda P'(\boldsymbol{\theta}) = 0. \tag{3.2}$$

where $\mathbf{g}_i(\boldsymbol{\theta}) = -2e_i(\boldsymbol{\theta}) f_i'(\boldsymbol{\theta})$. Likewise, the subsampling estimator $\hat{\boldsymbol{\theta}}^* =: \hat{\boldsymbol{\theta}}_r^*(\boldsymbol{\pi})$ satisfies $\boldsymbol{\Psi}_n^*(\boldsymbol{\theta}) =: Q_n'^*(\boldsymbol{\theta}) = 0$, i.e.,

$$\boldsymbol{\Psi}_n^*(\boldsymbol{\theta}) = \frac{1}{r}\sum_{j=1}^{r} \boldsymbol{\psi}_{nj}^*(\boldsymbol{\theta}) =: \frac{1}{r}\sum_{j=1}^{r} \frac{\mathbf{g}_j^*(\boldsymbol{\theta})}{n\pi_j^*} + \lambda \dot{P}(\boldsymbol{\theta}) = 0. \tag{3.3}$$

Using the scaled multinomial rv $\mathbf{w} = (w_1, \ldots, w_n) \sim \mathrm{smultn}(r, \boldsymbol{\pi}, \boldsymbol{\pi})$ defined in (8.1), a stochastic equivalent expression for $\boldsymbol{\Psi}_n^*(\boldsymbol{\theta})$ in terms of the full data is

$$\boldsymbol{\Psi}_n^*(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} w_i \mathbf{g}_i(\boldsymbol{\theta}) + \lambda \dot{P}(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \mathbb{R}^{p+d-1}. \tag{3.4}$$

As $\boldsymbol{\Phi}_n(\hat{\boldsymbol{\theta}}) = 0$, we obtain, with $\hat{\mathbf{g}}_i = \mathbf{g}_i(\hat{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{\psi}}_{ni} = \hat{\mathbf{g}}_i/(n\pi_i)$,

$$\hat{\boldsymbol{\Psi}}_n^* =: \boldsymbol{\Psi}_n^*(\hat{\boldsymbol{\theta}}) = \frac{1}{r}\sum_{j=1}^{r} \hat{\boldsymbol{\psi}}_{nj}^* = \frac{1}{r}\sum_{j=1}^{r} \frac{\hat{\mathbf{g}}_j^*}{n\pi_j^*} = \frac{1}{n}\sum_{i=1}^{n} \bar{w}_i \hat{\mathbf{g}}_i, \tag{3.5}$$

where $\bar{w}_i = w_i - \mathrm{E}(w_i) = w_i - 1$. Such a stochastic representation is useful, which decouples the random scheme from the data and is commonly used in developing the bootstrap theory. One has

$$\mathrm{E}(\mathbf{w}) = \mathbf{1}, \quad \mathrm{Cov}(\mathbf{w}) = r^{-1}(\mathrm{Diag}(1/\boldsymbol{\pi}) - \mathbf{1}\mathbf{1}^t). \tag{3.6}$$

See Zhang, *et al.* (2023) for more details about the scaled multinomial distribution. As a result, we readily calculate

$$\mathrm{E}^*(\boldsymbol{\Psi}_n^*(\boldsymbol{\theta})) = \mathrm{E}^*(\boldsymbol{\psi}_{nj}^*(\boldsymbol{\theta})) = \boldsymbol{\Phi}_n(\boldsymbol{\theta}),\ \hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi}) = r\mathrm{Var}^*(\hat{\boldsymbol{\Psi}}_n^*) = \sum_{i=1}^{n} \pi_i \Big(\frac{\hat{\mathbf{g}}_i}{n\pi_i} - \bar{\hat{\mathbf{g}}}\Big)^{\otimes 2} = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\mathbf{g}}_i^{\otimes 2}}{n\pi_i} - \lambda^2 P'(\hat{\boldsymbol{\theta}})^{\otimes 2}. \tag{3.7}$$

noting $\bar{\hat{\mathbf{g}}} =: \mathrm{E}^*(\hat{\mathbf{g}}_i^*/(n\pi_i)) = n^{-1}\sum_{i=1}^n \hat{\mathbf{g}}_i = -\lambda P'(\hat{\boldsymbol{\theta}})$. As a consequence, $\mathrm{E}^*(\boldsymbol{\psi}_{nj}^*(\hat{\boldsymbol{\theta}})) = 0$ for all j. This manifests that $\hat{\boldsymbol{\Psi}}_n^*$ is a sum of influence functions. Let $\mathbf{H}_n(\boldsymbol{\theta}) = \dot{\boldsymbol{\Phi}}_n(\boldsymbol{\theta})$ be the Hessian matrix. Set

$$\hat{\mathbf{H}}_n =: \mathbf{H}_n(\hat{\boldsymbol{\theta}}), \quad \hat{\boldsymbol{\Sigma}}_n =: \hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi}), \quad \hat{\lambda}_n =: \hat{\lambda}_n(\boldsymbol{\pi}) = \lambda_{\max}(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi})).$$

We need the following assumptions.

A1 The condition number of $\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi})$ is bounded in probability, i.e., $\lambda_{\max}(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi}))/\lambda_{\min}(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi})) = O_P(1)$.

A2 There exists a constant b_0 such that it holds in probability that

$$\lambda_{\min}(\hat{\mathbf{H}}_n)/\hat{\lambda}_n(\boldsymbol{\pi}) \geq b_0 > 0, \quad n = 1, 2, \ldots.$$

A3 With $\mathbf{g}_i(\boldsymbol{\theta}) = -2e_i(\boldsymbol{\theta})f_i'(\boldsymbol{\theta}) = \boldsymbol{\phi}_i(\boldsymbol{\theta}) - \lambda P'(\boldsymbol{\theta})$,

$$\frac{1}{n}\sum_{i=1}^n \frac{\|\hat{\boldsymbol{\phi}}_i\|^2}{n\pi_i} = O_P((p+d)\hat{\lambda}_n(\boldsymbol{\pi})), \quad \frac{p+d}{r}\frac{1}{n}\sum_{i=1}^n \frac{\|\dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}})\|^2}{n\pi_i} = O_P(\hat{\lambda}_n^2(\boldsymbol{\pi})).$$

A4 There are a neighborhood $\mathbb{N}_0$ of $\boldsymbol{\theta}_0$ and rv η_i such that $P(\boldsymbol{\theta})$ and $\mathbf{g}_i(\boldsymbol{\theta})$ satisfy

$$|\ddot{P}(\boldsymbol{\theta}) - \ddot{P}(\boldsymbol{\theta}_0)| \leq h\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|, \quad |\dot{\mathbf{g}}_i(\boldsymbol{\theta}) - \dot{\mathbf{g}}_i(\boldsymbol{\theta}_0)|_o \leq \eta_i\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|, \quad \boldsymbol{\theta} \in \mathbb{N}_0, \quad i = 1, \ldots, n,$$

where $\eta_1, ..., \eta_n$ satisfy

$$\frac{(p+d)^2}{r}\frac{1}{n}\sum_{i=1}^n \Big(1 + \frac{1}{rn\pi_i}\Big)\eta_i^2 + \frac{\lambda^2(p+d)^2h^2}{r} = o_P(\hat{\lambda}_n^3(\boldsymbol{\pi})). \tag{3.8}$$

A5 For an arbitrary $\mathbf{u}$ with $\|\mathbf{u}\| = 1$, the double array $z_{ni}(\boldsymbol{\pi}) = s_n^{-1}(\boldsymbol{\pi})\mathbf{u}^t\hat{\mathbf{H}}_n^{-1}\hat{\boldsymbol{\psi}}_{ni}$, $i = 1, 2, \ldots, n$, $n \geq 1$ with $s_n^2(\boldsymbol{\pi}) =: s_n^2(\boldsymbol{\pi}; \mathbf{u}) = \mathbf{u}^t\hat{\mathbf{H}}_n^{-1}\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi})\hat{\mathbf{H}}_n^{-t}\mathbf{u}$ satisfies Lindeberg's condition: for any $\epsilon > 0$

$$\frac{1}{n}\sum_{i=1}^n z_{ni}^2(\boldsymbol{\pi})\mathbf{1}\left[|z_{ni}(\boldsymbol{\pi})| \geq \sqrt{r}\epsilon\right] = o_P(1), \quad r \to \infty.$$

For later use, we denote (A5) by (A5') when $\boldsymbol{\pi} = \mathbf{1}/n$ (the uniform sampling).

Theorem 1 *(Asymptotic Normality) Suppose that $\hat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}_0$. Assume (A1)–(A4). Then it holds in probability that there exists a sequence of rv $\hat{\boldsymbol{\theta}}_r^*(\boldsymbol{\pi})$ which minimizes $Q_n^*(\boldsymbol{\theta}_{\boldsymbol{\phi}})$ in* (3.1)*, and that if $p + d = o_P(r\hat{\lambda}_n(\boldsymbol{\pi}))$,*

$$(p+d)^{-1/2}\hat{\lambda}_n^{1/2}\sqrt{r}(\hat{\boldsymbol{\theta}}_r^*(\boldsymbol{\pi}) - \hat{\boldsymbol{\theta}}_n) = O_{P^*}(1), \tag{3.9}$$

$$\hat{\mathbf{H}}_n\sqrt{r}(\hat{\boldsymbol{\theta}}_r^*(\boldsymbol{\pi}) - \hat{\boldsymbol{\theta}}_n) = -\frac{1}{\sqrt{r}}\sum_{j=1}^r \hat{\boldsymbol{\psi}}_{nj}^* + o_{P^*}(\hat{\lambda}_n^{1/2}(\boldsymbol{\pi})). \tag{3.10}$$

If, furthermore, $(A5)$ is met, then for any unit vector $\mathbf{u}$, it holds in probability that

$$s_n^{-1}(\boldsymbol{\pi})\sqrt{r}\boldsymbol{u}^t(\hat{\boldsymbol{\theta}}_r^*(\boldsymbol{\pi}) - \hat{\boldsymbol{\theta}}_n) \Rightarrow N(0,1), \quad r \to \infty. \tag{3.11}$$

Example 1 *(The bootstrap)* For the uniform sampling $\boldsymbol{\pi} = \mathbf{1}/n$, (A3)-(A4) boil down to

$$\frac{1}{n}\sum_{i=1}^n \|\hat{\boldsymbol{\phi}}_i\|^2 = O_P(\hat{\lambda}_n(p+d)), \quad \frac{p+d}{r}\frac{1}{n}\sum_{i=1}^n \|\dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}})\|^2 = o_P(\hat{\lambda}_n^2), \quad \frac{(p+d)^2}{r}\frac{1}{n}\sum_{i=1}^n \eta_i^2 + \frac{\lambda^2(p+d)^2h^2}{r} = o_P(\hat{\lambda}_n^3).$$

For the first equation to be fulfilled in a typical case, assume $\hat{\lambda}_n \geq c_0 > 0$ for some constant c_0. Furthermore, assume

$$\max_i \|\hat{\boldsymbol{\phi}}_i\| = O_P(\sqrt{p+d}), \quad \max_i \|\dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}})\| = O_P(p+d), \quad \max_i \eta_i = O_P((p+d)^{3/2}) = h. \tag{3.12}$$

The preceding equations are then satisfied if $p + d = o(r^{1/5})$ and $\sqrt{\lambda}(p+d) = o(r^{1/5})$. This shows that (1) the dimension $p + d$ is allowed to grow with r at the rate $r^{1/5}$, which is much slower than the optimal rate $r^{1/2}$, and (2) the penalty λ grows to infinity at the rate $\sqrt{r}/(p+d)^{5/2}$, which depends on both r and $p + d$. And for fixed r, a higher dimension $p + d$ leads to a slower growing rate for λ. As a consequence, the results in Theorem 1 hold for the bootstrap estimator if, moreover, (A1)-(A2) and (A5') are met. It is noteworthy that (A2)-(A3) restrict the rate for $\hat{\lambda}_n$ to grow to infinity and to diminish to zero, respectively.

Remark 1 Assume that the last equation in (3.12) holds and $\hat{\lambda}_n \geq c_0 > 0$ for some constant c_0. Then (A4) is implied by $(p+d)^5 = o_P(\min(r,\, r^2(n\pi_{\min})))$ and $\lambda^2(p+d)^5 = o(r)$, where $\pi_{\min} = \min(\pi_i)$. If, furthermore, the first two equalities in (3.12) holds, then (A3) is satisfied if $n\pi_{\min} \geq l_0$ for some constant $l_0 > 0$. The latter condition together with (A5') also implies (A5). (A3)-(A4) then hold if the dimension $p+d$ and the penality λ satisfy $p+d = o(r^{1/5})$ and $\sqrt{\lambda}(p+d) = o(r^{1/5})$. The results in Theorem 1, therefore, hold for any distribution $\boldsymbol{\pi}$ if, moreover, (A1)-(A2) and (A5') hold. As a consequence, the number $p+d = o(r^{1/5})$ of parameters in SIM is much smaller than the optimal number $p+d = O(r^{1/2})$, see Portnoy (1985, 1987); and the penalty λ depends on both r and $p+d$ and is allowed to grow to infinity at a much slow rate. Note that $p+d$ is the number of parameters that can be fitted by subdata of size r when the sampling distribution $\boldsymbol{\pi} = (\pi_i)$ satisfies $n\pi_i \geq l_0$ for all i.

Consider the unpenalized case, $\lambda = 0$. In this case, the dimensionality assumption for η_i can be relaxed to $\max_i \eta_i = O_P(1)$. For example, in the case of GLM, η_i can be taken as the spectral norm of the second derivative matrix $\ddot{\phi}_i(\boldsymbol{\theta})$, while some common structure of the matrices can be used to relax the dimension assymption. In this case, (3.8) is equivalent to $p+d = o_P(\min(r^{1/2},\, r(n\pi_{\min})^{1/2}))$. Thus $p+d = o_P(r^{1/2})$ (the optimal rate) provided $n\pi_i \geq l_0$ for all i. It is clear that such relaxation is invalid for the penalization function $P(\boldsymbol{\theta})$, suggesting that the rate $r^{1/5}$ can't be improved for the penalized SIM.

Example 2 *(The leverage scores)* The scores induce a distribution $\boldsymbol{\ell} = (h_{i,i}/p) =: (\ell_i)$, where $h_{i,i}$ are the diagonal entries of the hat matrix $\mathbb{H}_n = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, with $\mathbf{X}$ denoting the $n \times p$ covariate matrix with $\mathbf{x}_i^t$ as its ith rows. $\boldsymbol{\ell}$ is widely used in the development of stochastic algorithms, see e.g. Ma, *et al.* (2015). Assume $\lambda_{\max}(n^{-1}\mathbf{X}^t\mathbf{X}) \leq c_1$ and $\|\mathbf{x}_i\|/\sqrt{p} \geq c_2$ uniformly in i for some positive constants c_1, c_2. From $h_{i,i} = \mathbf{x}_i^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_i$ it thus follows

$$\ell_i \geq \|\mathbf{x}_i\|^2/(pnc_1) \geq c_1^{-1}c_2/n, \quad i = 1, 2, \ldots, n.$$

As a consequence, by Remark 1, the results in Theorem 1 hold for $\boldsymbol{\ell}$ if, furthermore, (A1)-(A2) and (A5') hold.

Remark 2 Examples 1–2 and Remark 1 demonstrate that truncation of the A-optimal and the leverage-scores based distributions is indispensable, see Subsection 4.3 for more details.

Remark 3 The commonly used cubic spline is Lipschitz continuous and hence satisfies (A4).

Remark 4 Analogous to the proof of Theorem 1, one can prove that the same results hold for the full sample estimator $\hat{\boldsymbol{\theta}}$ under similar conditions for growing dimension.

3.3 The Unweighted Subsampling Estimator and Asymptotic Normality

Given a subsample $(\mathbf{X}^*,\, \mathbf{y}^*)$, consider the unweighted objective function (cf. the weighted objective $Q_n^*(\boldsymbol{\theta}_{\boldsymbol{\phi}})$ (3.1)),

$$\tilde{Q}_n^*(\boldsymbol{\theta}_{\boldsymbol{\phi}}) = \frac{1}{r}\sum_{j=1}^{r}\big(y_j^* - \boldsymbol{\delta}^t\boldsymbol{B}(\boldsymbol{\beta}(\boldsymbol{\phi})^t\mathbf{x}_j^*)\big)^2 + \lambda P(\boldsymbol{\theta}_{\boldsymbol{\phi}}). \tag{3.13}$$

Minimizing $\tilde{Q}_n^*(\boldsymbol{\theta}_{\boldsymbol{\phi}})$, we obtain an unweighted subsampling estimator $\tilde{\boldsymbol{\theta}}_{\boldsymbol{\phi}}^* = (\tilde{\boldsymbol{\phi}}^{*t},\, \tilde{\boldsymbol{\delta}}^{*t})^t$. Note that $\tilde{Q}_n^*(\boldsymbol{\theta}_{\boldsymbol{\phi}})$ is a biased estimator of $Q_n(\boldsymbol{\theta}_{\boldsymbol{\phi}})$ because $\mathrm{E}^*\tilde{Q}_n^*(\boldsymbol{\theta}_{\boldsymbol{\phi}}) \neq Q_n(\boldsymbol{\theta}_{\boldsymbol{\phi}})$, where

$$\tilde{Q}_n(\boldsymbol{\theta}_{\boldsymbol{\phi}}) =: \mathrm{E}^*\tilde{Q}_n^*(\boldsymbol{\theta}_{\boldsymbol{\phi}}) = \sum_{i=1}^{n}\pi_i\big(y_i - \boldsymbol{\delta}^t\boldsymbol{B}(\boldsymbol{\beta}(\boldsymbol{\phi})^t\mathbf{x}_i)\big)^2 + \lambda P(\boldsymbol{\theta}_{\boldsymbol{\phi}}). \tag{3.14}$$

Recalling $\boldsymbol{\theta} = \boldsymbol{\theta}_{\boldsymbol{\phi}} = (\boldsymbol{\phi}^t, \boldsymbol{\delta}^t)^t$, let $\tilde{\boldsymbol{\theta}} =: \tilde{\boldsymbol{\theta}}_n(\boldsymbol{\pi})$ be the minimizer of $\tilde{Q}_n(\boldsymbol{\theta})$, so that it solves the equation $\tilde{\boldsymbol{\Phi}}_n(\boldsymbol{\theta}) = \partial\tilde{Q}_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = 0$, specifically,

$$\tilde{\boldsymbol{\Phi}}_n(\boldsymbol{\theta}) = \tilde{\boldsymbol{\Phi}}_n(\boldsymbol{\theta}; \boldsymbol{\pi}) = \sum_{i=1}^{n}\tilde{\boldsymbol{\phi}}_i(\boldsymbol{\theta}) =: \sum_{i=1}^{n}\pi_i\mathbf{g}_i(\boldsymbol{\theta}) + \lambda P'(\boldsymbol{\theta}) = 0. \tag{3.15}$$

This is a *penalized generalized bootstrap estimator* for estimating equations. Chatterjee and Bose (2002) studied dimension asymptotics for generalized bootstrap estimators, that is, they established asymptotic normality of the estimator for growing parameter dimension.

The unweighted subsampling estimator $\tilde{\boldsymbol{\theta}}^* =: \tilde{\boldsymbol{\theta}}^*(\boldsymbol{\pi})$ satisfies $\boldsymbol{\Phi}_n^*(\boldsymbol{\theta}) = 0$, that is,

$$\boldsymbol{\Phi}_n^*(\boldsymbol{\theta}) = \frac{1}{r}\sum_{j=1}^{r}\boldsymbol{\phi}_j^*(\boldsymbol{\theta}) =: \frac{1}{r}\sum_{j=1}^{r}\mathbf{g}_j^*(\boldsymbol{\theta}) + \lambda P'(\boldsymbol{\theta}) = 0. \tag{3.16}$$

Using the scaled multinomial rv $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^t \sim \mathrm{smultn}(r, \mathbf{1}, \boldsymbol{\pi})$ defined in (8.1), a stochastic equivalent expression for $\boldsymbol{\Phi}_n^*(\boldsymbol{\theta})$ in terms of the full data is

$$\boldsymbol{\Phi}_n^*(\boldsymbol{\theta}) = \sum_{i=1}^n \omega_i \mathbf{g}_i(\boldsymbol{\theta}) + \lambda P'(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \mathbb{R}^{p+d-1}. \tag{3.17}$$

As $\tilde{\boldsymbol{\Phi}}_n(\tilde{\boldsymbol{\theta}}) = 0$, we obtain

$$\tilde{\boldsymbol{\Phi}}_n^* =: \boldsymbol{\Phi}_n^*(\tilde{\boldsymbol{\theta}}) = \frac{1}{r} \sum_{j=1}^r \boldsymbol{\phi}_j^*(\tilde{\boldsymbol{\theta}}) = \sum_{i=1}^n \bar{\omega}_i \tilde{\mathbf{g}}_i, \tag{3.18}$$

where $\bar{\omega}_i = \omega_i - \mathrm{E}(\omega_i) = \omega_i - \pi_i$, and $\tilde{\mathbf{g}}_i = \mathbf{g}_i(\tilde{\boldsymbol{\theta}})$. One verifies for $\boldsymbol{\omega} \sim \mathrm{smultn}(r, \mathbf{1}, \boldsymbol{\pi})$ that

$$\mathrm{E}(\boldsymbol{\omega}) = \boldsymbol{\pi}, \quad \mathrm{Cov}(\boldsymbol{\omega}) = r^{-1}(\mathrm{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^t). \tag{3.19}$$

As a result, by (3.19), we readily calculate, noting $\bar{\tilde{\mathbf{g}}} =: \mathrm{E}^*(\tilde{\mathbf{g}}_i^*) = \sum_{i=1}^n \pi_i \tilde{\mathbf{g}}_i = -\lambda P'(\tilde{\boldsymbol{\theta}})$, that

$$\mathrm{E}^*(\boldsymbol{\Phi}_n^*(\boldsymbol{\theta})) = \tilde{\boldsymbol{\Phi}}_n(\boldsymbol{\theta}), \quad \tilde{\boldsymbol{\Sigma}}_n = \tilde{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi}) = r\mathrm{Var}^*(\tilde{\boldsymbol{\Phi}}_n^*) = \sum_{i=1}^n \pi_i (\tilde{\mathbf{g}}_i - \bar{\tilde{\mathbf{g}}})^{\otimes 2} = \sum_{i=1}^n \pi_i \tilde{\mathbf{g}}_i^{\otimes 2} - \lambda^2 P'(\tilde{\boldsymbol{\theta}})^{\otimes 2}. \tag{3.20}$$

Thus $\mathrm{E}^*(\boldsymbol{\phi}_j^*(\tilde{\boldsymbol{\theta}})) = 0$ for all j, and $\tilde{\boldsymbol{\Phi}}_n^*$ is a sum of influence functions. Let $\tilde{\mathbf{H}}_n(\boldsymbol{\theta}) = \tilde{\mathbf{H}}_n(\boldsymbol{\theta}; \boldsymbol{\pi})$ be the Hessian. Set

$$\tilde{\boldsymbol{\phi}}_i = \tilde{\boldsymbol{\phi}}_i(\tilde{\boldsymbol{\theta}}), \quad \tilde{\mathbf{H}}_n = \tilde{\mathbf{H}}_n(\boldsymbol{\pi}) = \tilde{\mathbf{H}}_n(\tilde{\boldsymbol{\theta}}; \boldsymbol{\pi}), \quad \tilde{\lambda}_n = \tilde{\lambda}_n(\boldsymbol{\pi}) = \lambda_{\max}(\tilde{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi})).$$

Analogous to Theorem 1, under similar assumptions ($\tilde{\mathbf{A}}$1)–($\tilde{\mathbf{A}}$5) stated in the last section, we prove

Theorem 2 *(ASN) Suppose that $\tilde{\boldsymbol{\theta}}_n(\boldsymbol{\pi})$ is a consistent estimator of $\boldsymbol{\theta}_0$. Assume ($\tilde{\mathbf{A}}$1)–($\tilde{\mathbf{A}}$4). Then it holds in probability that there exists a sequence of rv $\tilde{\boldsymbol{\theta}}^*(\boldsymbol{\pi})$ which minimizes $\tilde{Q}_n^*(\boldsymbol{\theta}_{\boldsymbol{\phi}})$ in* (3.1)*, and that if $p + d = o_P(r\tilde{\lambda}_n(\boldsymbol{\pi}))$,*

$$(p+d)^{-1/2} \tilde{\lambda}_n^{1/2} \sqrt{r}(\tilde{\boldsymbol{\theta}}_r^*(\boldsymbol{\pi}) - \tilde{\boldsymbol{\theta}}_n(\boldsymbol{\pi})) = O_{P^*}(1), \tag{3.21}$$

$$\tilde{\mathbf{H}}_n(\boldsymbol{\pi}) \sqrt{r}(\tilde{\boldsymbol{\theta}}_r^*(\boldsymbol{\pi}) - \tilde{\boldsymbol{\theta}}_n(\boldsymbol{\pi})) = -\frac{1}{\sqrt{r}} \sum_{j=1}^r \boldsymbol{\phi}_j^*(\tilde{\boldsymbol{\theta}}) + o_{P^*}(\tilde{\lambda}_n^{1/2}(\boldsymbol{\pi})). \tag{3.22}$$

If, further, ($\tilde{\mathbf{A}}$5) is met with $\tilde{s}_n^2(\boldsymbol{\pi}) = \mathbf{u}^t \tilde{\mathbf{H}}_n^{-1}(\boldsymbol{\pi}) \tilde{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi}) \tilde{\mathbf{H}}_n^{-t}(\boldsymbol{\pi}) \mathbf{u}$, then for any unit vector $\mathbf{u}$, it holds in probability,

$$\tilde{s}_n^{-1}(\boldsymbol{\pi}) \sqrt{r} \boldsymbol{u}^t (\tilde{\boldsymbol{\theta}}_r^*(\boldsymbol{\pi}) - \tilde{\boldsymbol{\theta}}_n(\boldsymbol{\pi})) \Rightarrow N(0,1), \quad r \to \infty. \tag{3.23}$$

Remark 5 For the uniform $\boldsymbol{\pi} = \mathbf{1}/n$, both the weighted and the unweighted subsamplining estimators simplify to the same bootstrap estimators. Therefore, the results in Example 1 still hold for the unweighted estimators. The results in Remark 1 with $n\pi_i$ replaced by one hold as well and, in particular, $p + d = o(r^{1/5})$. Here the sampling distribution is arbitrary, manifesting that truncation is not needed for the unweighted estimators.

4 The A-optimal Distributions, Implementation and Truncation

In this Section, we derive the optimal distributions and discuss numerical computation.

4.1 The Weighted-Estimator-Based A-optimal Distributions

By (3.10), the (asymptotic) covariance matrix of $\hat{\boldsymbol{\theta}}^*(\boldsymbol{\pi})$ is

$$\hat{\mathbf{V}}_n(\boldsymbol{\pi}) = \hat{\mathbf{H}}_n^{-1} \hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi}) \hat{\mathbf{H}}_n^{-t}. \tag{4.1}$$

The criterion of A-optimality aims to minimize the sum of the component variances of the subsampling estimator. The criterion has been used to study data of massive size in linear regression (Zhu, *et al.*, 2015; Zhang, *et al.*, 2023), logistic regression (Wang, *et al.*, 2017), and count data regression (Tan, *et al.*, 2023) among others. In the case of SIM, the estimation of the index parameters $\boldsymbol{\beta}$ and the coefficients $\boldsymbol{\delta}$ of the basis functions is interconnected. Luckily, a quality estimator of $\boldsymbol{\beta}$ often results in a good plug-in estimator of the link function. Moreover, the knots of a spline basis depend on the quantiles of the indices $\boldsymbol{\beta}^t \mathbf{x}_i$. Consequently, the estimation of $\boldsymbol{\beta}$ plays a central role, and we shall derive the sampling distributions for estimating $\boldsymbol{\beta}$, but through those for estimating $\boldsymbol{\phi}$ due to technical considerations of the constraints in the optimization. Specifically, we seek $\boldsymbol{\pi}$ which minimizes the trace norm $\mathrm{Tr}(\hat{\mathbf{V}}_n(\boldsymbol{\pi}))$ of the asymptotic covariance matrix $\hat{\mathbf{V}}_n(\boldsymbol{\pi})$ in (4.1) subject to the constraint $\sum_{i=1}^n \pi_i = 1$. To this end, let

$$L(\boldsymbol{\pi}, \tau) = \mathrm{Tr}(\hat{\mathbf{V}}_n(\boldsymbol{\pi})) + \tau\Big(\sum_{i=1}^n \pi_i - 1\Big).$$

Using the identity $\mathrm{Tr}(\mathbf{a}\mathbf{a}^t) = \|\mathbf{a}\|^2$, we calculate

$$\mathrm{Tr}(\hat{\mathbf{V}}_n(\boldsymbol{\pi})) = c_1 \mathrm{Tr}\Big(\hat{\mathbf{H}}_n^{-1} \sum_{i=1}^n \Big(\frac{\hat{\mathbf{g}}\hat{\mathbf{g}}_i^t}{\pi_i} + c_2\Big)\hat{\mathbf{H}}_n^{-t}\Big) = c_1 \sum_{i=1}^n \frac{\|\hat{\mathbf{H}}_n^{-1}\hat{\mathbf{g}}_i\|^2}{\pi_i} + c_3, \tag{4.2}$$

where c_1, c_2, c_3 are constants independent of $\boldsymbol{\pi}$. By the Lagrange multiplier method,

$$\frac{\partial L(\boldsymbol{\pi}; \tau)}{\partial \pi_i} = \frac{-c_1\|\hat{\mathbf{H}}_n^{-1}\hat{\mathbf{g}}_i\|^2}{\pi_i^2} + \tau = 0, \quad i = 1, 2, \ldots, n.$$

If $\hat{\mathbf{g}}_j = 0$, then we take $\pi_j = 0$; otherwise we solve the equations for the rest of π_i using the constraint $\sum_{i=1}^n \pi_i = 1$. Write $\boldsymbol{\pi} \propto (a_i)$ if $\pi_i = a_i / \sum_{i=1}^n a_i$ for all i. Let $\hat{e}_i = y_i - f_i(\hat{\boldsymbol{\theta}})$.

Theorem 3 *(The $\hat{A}$-optimal Distribution) Assume that $\hat{\mathbf{H}}_n^{-1}$ exists. Then there exists a probability distribution $\hat{\boldsymbol{\pi}}$ that minimizes the trace of the asymptotic variance-covariance matrix $\hat{\mathbf{V}}_n(\boldsymbol{\pi})$ of the subsampling estimator $\hat{\boldsymbol{\phi}}^*$, given by*

$$\hat{\boldsymbol{\pi}} \propto (\|\hat{\mathbf{H}}_n^{-1} f_i'(\hat{\boldsymbol{\theta}})\| |\hat{e}_i|). \tag{4.3}$$

Consider the conditional covariance given $\mathbf{X}$, $\bar{\boldsymbol{\Sigma}}_0(\boldsymbol{\pi}) = r\mathrm{Var}(\boldsymbol{\Psi}_n^* | \mathbf{X})$, where $\boldsymbol{\Psi}_n^* = \boldsymbol{\Psi}_n^*(\boldsymbol{\theta}_0)$. Write $f_i = f_i(\boldsymbol{\theta}_0)$, $e_i = e_i(\boldsymbol{\theta}_0)$, and $\mu_i = \mathrm{E}(e_i|\mathbf{X})$, and $\sigma_i^2 = \mathrm{E}(e_i^2|\mathbf{x}_i) = \sigma_0^2 + (f_i - m_0(\mathbf{x}_i^t \boldsymbol{\beta}_0))^2$. Then $\boldsymbol{\Psi}_n^* - \mathrm{E}(\boldsymbol{\Psi}_n^*|\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n -2(\bar{w}_i e_i + (e_i - \mu_i))\dot{f}_i$, so that

$$\bar{\boldsymbol{\Sigma}}_0(\boldsymbol{\pi}) = \frac{4}{n^2} \sum_{i=1}^n \frac{\sigma_i^2}{\pi_i} f_i'^{\otimes 2} + C_n,$$

where C_n is a constant independent of $\boldsymbol{\pi}$. Let $\bar{\mathbf{V}}_0(\boldsymbol{\pi}) = \mathbf{H}_n^{-1}\bar{\boldsymbol{\Sigma}}_0\mathbf{H}_n^{-t}$. Analogously, minimizing $\mathrm{Tr}(\bar{\mathbf{V}}_0(\boldsymbol{\pi}))$, we get

Theorem 4 *(The $\bar{A}$-optimal Distribution) Assume that $\mathbf{H}_n^{-1}$ exists. Then there exists a probability distribution $\bar{\boldsymbol{\pi}}$ that minimizes the trace of the asymptotic conditional covariance matrix $\bar{\mathbf{V}}_0(\boldsymbol{\pi})$ of the subsampling estimator $\hat{\boldsymbol{\phi}}^*$, given by*

$$\bar{\boldsymbol{\pi}} \propto (\sigma_i \|\mathbf{H}_n^{-1} f_i'\|). \tag{4.4}$$

Remark 6 While in a simulation study we choose $\boldsymbol{\beta}_0$, σ_0^2, $m_0(x)$ and $m(x)$ and calculate σ_i^2, we would take $m_0(x) = m(x) = \boldsymbol{\delta}^t \mathbf{B}(x)$ in an analysis of real world data, which leads to $\sigma_i^2 = \sigma_0^2$ for all i. Thus, we estimate $\bar{\boldsymbol{\pi}}$ by

$$\underset{\sim}{\boldsymbol{\pi}} \propto (\|\hat{\mathbf{H}}_n^{-1} f_i'(\hat{\boldsymbol{\theta}})\|). \tag{4.5}$$

Remark 7 The A-optimal distributions for $\hat{\boldsymbol{\phi}}^*$ to approximate $\hat{\boldsymbol{\phi}}$ are given by

$$\hat{\boldsymbol{\pi}} \propto (\|\boldsymbol{\Lambda}_1\hat{\mathbf{H}}_n^{-1} f_i'(\hat{\boldsymbol{\theta}})\| |\hat{e}_i|), \quad \bar{\boldsymbol{\pi}} \propto (\sigma_i \|\boldsymbol{\Lambda}_1\mathbf{H}_n^{-1} f_i'\|), \quad \underset{\sim}{\boldsymbol{\pi}} \propto (\|\boldsymbol{\Lambda}_1\hat{\mathbf{H}}_n^{-1} f_i'(\hat{\boldsymbol{\theta}})\|), \tag{4.6}$$

where $\boldsymbol{\Lambda}_1$ be the matrix with all the entries equal to zero except the first $p - 1$ diagonal entries are equal to 1. These are the sampling distributions used in our simulations and real data applications.

Remark 8 It is worth to mention that, on one hand, the residuals $\hat{e}_i$ in $\hat{\boldsymbol{\pi}}$ bring more information than that in $\bar{\boldsymbol{\pi}}$ as the residuals contain the information about the response y_i. See Zhang, *et al.* (2023) for more discussion, where their extensive simulations in a linear model exhibited the gain of efficiency in the subsampling estimator in terms of empirical mean squared errors. On the other hand, the zero values of the residuals $\hat{e}_i$ necessitate truncation for $\hat{\boldsymbol{\pi}}$.

Moreover, as $\underset{\sim}{\boldsymbol{\pi}}$ doesn't contain the residuals $\hat{\varepsilon}_i$, it satisfies ($\tilde{\mathbf{A}}$6). As a result, it can be used in the unweighted subsampling estimator $\tilde{\boldsymbol{\theta}}^*$, which is more efficient than the weighted subsampling estimator $\hat{\boldsymbol{\theta}}^*$, see Theorem 5.1.

Remark 9 If $\hat{\mathrm{g}}_i = \mathrm{g}_i(\hat{\boldsymbol{\theta}}) = -2\hat{e}_i f_i'(\hat{\boldsymbol{\theta}}) = y_i$ in (4.3), then $\pi_i = 0$, and the ith observation is dropped as it is "unimportant". This implies that an observation closer to $\hat{\mathrm{g}}_i = 0$ (a hypersurface in the $(\mathbf{x}, y)$-coordinate system) has smaller probability to be chosen in the subsampling.

Remark 10 We showcased the A-optimal Subsampling approach for the SIM above. Similarly, one can obtain the A-optimal distributions for other SIM. For example, robust estimation in SIM via using the loss function of a linear combination of several loss functions, see e.g. Jiang, *et al.* (2022). We believe that conclusions similar to those in this paper can be drawn based on our results about the Subsamping approach for various models in this and other papers.

Remark 11 It must be noted that the A-optimal sampling distribution for $\hat{\boldsymbol{\phi}}^*$ to approximate $\hat{\boldsymbol{\phi}}$ is not A-optimal for $\boldsymbol{\beta}(\hat{\boldsymbol{\phi}}^*)$ to approximate $\boldsymbol{\beta}(\hat{\boldsymbol{\phi}})$, see Remark 5 in Zhang, *et al.* (2023).

Remark 12 It is not clear how much loss of efficiency for the subsampling estimator resulted from using the A-optimal distribution for estimating $\boldsymbol{\phi}$. The simulated EMSE ratios of the subsampling estimator of using the A-optimal subsampling to using the uniform were about 10% in Table 1, and about 28% for r equal to 1% of sample size n in Table 3. Similar results can be seen in those tables in the **Supplementary Material**. For the two real datasets, when r was 5% of n, the ratios were less than 40% in Tables 5 and 7. For r as low as 0.03% of n, the ratios were about 80% in Tables 9 and 11. These results indicated that the gain of efficiency was significant when using the A-optimal distribution for estimating $\boldsymbol{\phi}$, although further gain is likely at the price of additional mathematical and algorithmic operations.

4.2 The Unweighted-Estimator-Based A-optimal Sampling Distributions

By (3.22), the (asymptotic) covariance matrix of the unweighted estimator $\tilde{\boldsymbol{\theta}}^*$ is

$$\tilde{\mathbf{V}}_n(\boldsymbol{\pi}) = \tilde{\mathbf{H}}_n^{-1}(\boldsymbol{\pi})\tilde{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi})\tilde{\mathbf{H}}_n(\boldsymbol{\pi})^{-t}. \tag{4.7}$$

As the trace norm $\tilde{\tau}(\boldsymbol{\pi}) = \mathrm{Tr}(\tilde{\mathbf{V}}_n(\boldsymbol{\pi}))$ is a continuous function on the probability simplex $\boldsymbol{\pi} \in [0,\ 1]^n$ with $\pi_1 + \cdots + \pi_n = 1$, there exists a sampling distribution $\tilde{\boldsymbol{\pi}} \in [0,\ 1]^n$ which minimizes $\tilde{\tau}(\boldsymbol{\pi})$. Apparently, there is no explicit formula for $\tilde{\boldsymbol{\pi}}$, and an algorithm must be employed to find the numerical solution. For one dimension, $\tilde{\boldsymbol{\pi}} =: (\tilde{\boldsymbol{\pi}}_{n-1}, \tilde{\pi}_n)$ can be explicitly found as

$$\tilde{\boldsymbol{\pi}}_{n-1} = (\mathbf{A}^{-1}\mathbf{b})_+, \quad \tilde{\pi}_n = 1 - \mathbf{1}^t\tilde{\boldsymbol{\pi}}_{n-1},$$

where $\mathbf{a}_+$ denotes the component-wise positive part of a vector $\mathbf{a}$, $\mathbf{A} = (a_{k,i})$ and $\mathbf{b} = (b_1, \ldots, b_{n-1})^t$ with

$$a_{k,i} = [g_k^2(g_n' - g_i') - 2\dot{g}_k(g_n^2 - g_i^2)]_{\tilde{\boldsymbol{\theta}}}, \quad b_k = [g_k^2 g_n' - 2g_k' g_n^2 + \lambda(g_k^2\ddot{P} - 2\dot{g}_k P'^2)]_{\tilde{\boldsymbol{\theta}}}, \quad k, i = 1, \ldots, n-1.$$

For $n = 2$, one has $\tilde{\pi}_1 \propto [g_2'/(g_1' - g_2')|_{\tilde{\theta}} - 2g_2^2/(g_1^2 - g_2^2)|_{\tilde{\theta}}]_+$ and $\tilde{\pi}_2 = 1 - \tilde{\pi}_1$. Consider the 'expected version' of the sampling distribution, $\tilde{\pi}_{01} \propto [\mathrm{E}(g_2')/\mathrm{E}(g_1' - g_2') - 2\mathrm{E}(g_2^2)/\mathrm{E}(g_1^2 - g_2^2)]_+$ and $\tilde{\pi}_{02} = 1 - \pi_{01}$. One calculates

$$\tilde{\pi}_{01} \propto \frac{a_2^2 x_2^2}{a_2^2 x_2^2 - a_1^2 x_1^2}\mathbf{1}[|a_2 x_2| > |a_1 x_1|], \quad \tilde{\pi}_{02} \propto \frac{a_1^2 x_1^2}{a_1^2 x_1^2 - a_2^2 x_2^2}\mathbf{1}[|a_1 x_1| > |a_2 x_2|],$$

where $a_i = \boldsymbol{\delta}^t\dot{\mathbf{B}}(\beta(\phi)x_i)\beta'(\phi)$. Oberve that (1) $\tilde{\boldsymbol{\pi}}_0 = (\tilde{\pi}_{01}, \tilde{\pi}_{02})$ is on the boundary of the probability simplex, whereas the weighted-estimator-based A-optimal distributions $\hat{\boldsymbol{\pi}}$ and $\bar{\boldsymbol{\pi}}$ are in the interior of the simplex. This implies that some observations will be sampled with zero probabilities (dropped) by the $\tilde{\boldsymbol{\pi}}_0$- subsampling, while all observations will be sampled with positive probabilities by the $\hat{\boldsymbol{\pi}}$- or $\bar{\boldsymbol{\pi}}$- subsampling at least for large n. Specifically, by the $\tilde{\boldsymbol{\pi}}_0$-subsampling, x_2 will be dropped if $|a_2 x_2| > |a_1 x_1|$; (2) each $\tilde{\pi}_i$ is an increasing function of $a_i^2 x_i^2$, just like $\hat{\pi}_i$ or $\bar{\pi}_i$ which are increasing functions of the ith observations given in (4.3) and (4.4) although such functions are quite different. The sampling mechanism for $n \geq 3$ and high parameter dimension appears much more complicated.

4.3 Implementation, Presampling and Truncation

Since one of the bottlenecks for computing the sampling distribution $\boldsymbol{\pi}$ is to compute the Hessian matrix $\hat{\mathbf{H}} =: \mathbf{H}_n(\hat{\boldsymbol{\theta}})$, we shall approximate it by a diagonal block matrix, suggested in Le Cun (1987), with the blocks equal to $\hat{\mathbf{H}}_\phi = \partial^2 Q_n(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t$ and $\hat{\mathbf{H}}_\delta = \partial^2 Q_n(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^t$. That is, the diagonal blocks are kept, while the information from other entries is skipped. Denote the first $(p-1)$ elements of f_i' by $f_{\phi,i}'$. When only $\boldsymbol{\delta}$ is penalized, $P(\boldsymbol{\theta}) = \|\boldsymbol{\delta}\|^2$ (such as the ridge penalty), one obtains a computational easy version $\underline{\boldsymbol{\pi}}$ of $\hat{\boldsymbol{\pi}}$ given in (4.3),

$$\underline{\boldsymbol{\pi}} \propto (\|\hat{\mathbf{H}}_\phi^{-1} f_{\phi,i}'(\hat{\boldsymbol{\theta}})\|\,|\hat{e}_i|). \tag{4.8}$$

Although the computational complexity of $\underline{\boldsymbol{\pi}}$ is still the same as the full-sample estimator $\hat{\boldsymbol{\theta}}$, the computation of the Hessian matrix and its inverse in $\underline{\boldsymbol{\pi}}$ are time-saving.

The Scoring Algorithm 3

1. Take a uniform presample of size $r_0 << n$ from the full data, call **Algorithm 1** in Alg.1 to obtain an initial estimator $\underline{\boldsymbol{\theta}}_0$ using the presample, and compute an approximation $\underline{\boldsymbol{\pi}}_0$ to $\underline{\boldsymbol{\pi}}$ by replacing $\hat{\boldsymbol{\theta}}$ with $\underline{\boldsymbol{\theta}}_0$ in (4.8).
2. Take a subsample of size $r << n$ from the remaining sample using $\underline{\boldsymbol{\pi}}_0$, and call **Algorithm 2** in Alg.2 to obtain the subsampling estimator $\hat{\boldsymbol{\theta}}^*$ using the subsample.

Remark 13 Observe that $\underline{\boldsymbol{\pi}}$ given in (4.8) depends on the partial derivative w.r.t. $\boldsymbol{\phi}$ only (not on that w.r.t. $\boldsymbol{\delta}$). In fact, if we fix $\boldsymbol{\delta}$ so that the only parameter is $\boldsymbol{\phi}$, then we can show that $\underline{\boldsymbol{\pi}}$ is the A-optimal distribution.

As $\underline{\boldsymbol{\pi}}$ relies on $\hat{\boldsymbol{\theta}}$, which is not available in reality, we shall follow the Scoring Algorithm proposed in Zhang, *et al.* (2023) to approximate it by *presampling*.

Truncation Note that each probability $\underline{\pi}_i$ in (4.8) is proportional to the absolute value of the residual $e_i(\hat{\boldsymbol{\theta}})$. More generally, π_i in (4.3)is proportional to the norm of the linear transformation of $\Phi_i(\hat{\boldsymbol{\theta}})$. As in a typical case of importance sampling, the sampling probabilities are inversely used as weights in calculating the subsampling estimator. The probabilities that are close to zero violate the assumptions which grantee appropriate properties of the subsampling estimators, see Section 3.2. Moreover, the Hessian matrix can be poorly conditioned, which may lead to small values in the probabilities. As pointed out in Remark 9, such observations will be selected with small probabilities in the subsampling. Following Zhang, *et al.* (2023), we truncate $\boldsymbol{\pi} = (\pi_i)$ from below as follows:

$$\boldsymbol{\pi}_{\rm trunc} \propto (\pi_i \mathbf{1}[\pi_i > L/n] + (l/n)\mathbf{1}[\pi_i \leq L/n]),$$

where L is a threshold value. One drops unimportant observations by taking $l = 0$, otherwise $l = \#\{\pi_i > L/n\}$ (the number of truncated observations). In our simulations in Section 6 and real data applications in Section 7, we truncated up to 25%, as zero values often happened before the 25% quantile of the sampling distributions considered. In the cases in which the uniform sampling outperformed the optimal sampling, the truncation resulted in a remarkable improvement. In the cases in which the optimal sampling probability outperformed the uniform, the truncation didn't yield significant improvement.

5 The Biases and Efficiency Comparison

In this Section, we compare efficiency of the two subsampling estimators and give the analytic formulas of the first order biases. It is standard that the formulas can be used to constructed bias-corrected estimators, and the details can be found in textbooks.

5.1 The Biases

Peng, *et al.* (2024) gave the analytic formulas in Section 2 for the first-order bias of the zero estimators $\hat{\boldsymbol{\beta}}$ of parameters $\boldsymbol{\beta}$ in general estimating equations $\boldsymbol{\Psi}_n(\boldsymbol{\theta}) = 0$, and rigorously proved the rate for the reminder. The formula is given, see (5.1), using $\mathbf{J}_n = \mathrm{E}(\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0)^{\otimes 2})$, $\mathbf{H}_n = \mathrm{E}(\dot{\boldsymbol{\Psi}}_n(\boldsymbol{\theta}_0))$, and $\mathbf{G}_{n,k} = \mathrm{E}(\ddot{\Psi}_{n,k}(\boldsymbol{\theta}_0))$ of the components $\Psi_{n,k}(\boldsymbol{\theta}_0)$ of $\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0)$. In our case, $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}$ and the expectation is calculated given the data, i.e., $\mathrm{E} = \mathrm{E}^*$. Let $P_k'(\boldsymbol{\theta})$ be the kth component of $P'(\boldsymbol{\theta})$ and $\dddot{P}_k(\boldsymbol{\theta}) =: \partial^2 P_k'(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^t$.

The Biases for the Weighted Estimators Noting $\mathbf{J}_n =: \mathbf{J}_n(\boldsymbol{\pi}) = \mathrm{E}^*(\hat{\boldsymbol{\Psi}}_n^{*\otimes 2})$, we have

$$\mathbf{J}_n = r^{-1}\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi}), \quad \mathbf{H}_n = \mathrm{E}^*(\dot{\boldsymbol{\Psi}}_n^*(\hat{\boldsymbol{\theta}})) = \frac{1}{n}\sum_{i=1}^n \dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}}) + \lambda\ddot{P}(\hat{\boldsymbol{\theta}}), \quad \mathbf{G}_{n,k} = \mathrm{E}^*(\ddot{\boldsymbol{\Psi}}_{n,k}^*(\hat{\boldsymbol{\theta}})) = \frac{1}{n}\sum_{i=1}^n \ddot{g}_{i,k}(\hat{\boldsymbol{\theta}}) + \lambda\dddot{P}_k(\hat{\boldsymbol{\theta}}),$$

where $\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi})$ is given in (3.7). The first-order bias for $\hat{\boldsymbol{\theta}}^* = (\hat{\boldsymbol{\phi}}^*, \hat{\boldsymbol{\delta}}^*)$ is then given by

$$\mathrm{Bias}(\hat{\boldsymbol{\phi}}^*, \hat{\boldsymbol{\delta}}^*; \boldsymbol{\pi}) = r^{-1}\mathbf{H}_n^{-1}(\bar{\mathbf{b}}_{n1}(\boldsymbol{\pi}) - 2^{-1}\bar{\mathbf{q}}_n(\boldsymbol{\pi})), \tag{5.1}$$

where $\bar{\mathbf{q}}_n(\boldsymbol{\pi}) = (\bar{q}_{n,k}(\boldsymbol{\pi}))$ with $\bar{q}_{n,k}(\boldsymbol{\pi}) = r\mathrm{Tr}(\mathbf{H}_n^{-\top}\mathbf{G}_{n,k}\mathbf{H}_n^{-1}\mathbf{J}_n) = \mathrm{Tr}(\mathbf{H}_n^{-\top}\mathbf{G}_{n,k}\mathbf{H}_n^{-1}\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi}))$, and by (3.5),

$$\bar{\mathbf{b}}_{n1} =: \bar{\mathbf{b}}_{n1}(\boldsymbol{\pi}) = r\mathrm{E}^*(\dot{\boldsymbol{\Psi}}_n^*(\hat{\boldsymbol{\theta}})\mathbf{H}_n^{-1}\hat{\boldsymbol{\Psi}}_n^*) = \frac{1}{n}\sum_{i=1}^n \frac{1}{n\pi_i}\dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}})\mathbf{H}_n^{-1}\hat{\mathbf{g}}_i + \lambda\Big(\frac{1}{n}\sum_{i=1}^n \dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}})\Big)\mathbf{H}_n^{-1}P'(\hat{\boldsymbol{\theta}}).$$

As a consequence, the first-order bias for $\hat{\boldsymbol{\beta}}^* = \boldsymbol{\beta}^*(\hat{\boldsymbol{\phi}}^*)$ is determined by $\mathrm{Bias}(\hat{\boldsymbol{\beta}}^*) = \mathbf{Jac}(\hat{\boldsymbol{\phi}})\mathrm{Bias}(\hat{\boldsymbol{\phi}}^*)$, where $\mathbf{Jac}^\top(\boldsymbol{\phi}) = (-\boldsymbol{\phi},\, (1+\|\boldsymbol{\phi}\|^2)\mathbf{I} - \boldsymbol{\phi}^{\otimes 2})(1+\|\boldsymbol{\phi}\|^2)^{-3/2}$ is the Jacobian matrix of the transformation $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\phi})$.

Observe that $\mathrm{Bias}(\hat{\boldsymbol{\phi}}^*, \hat{\boldsymbol{\delta}}^*)$ is influenced by the sampling distribution $\boldsymbol{\pi}$ through the form of the reciprocals $1/(n\pi_i)$. As a consequence, truncation is necessary for the $\hat{A}$-optimal distribution $\hat{\boldsymbol{\pi}}$ in (4.3), whereas truncation may not be needed for the $\bar{A}$-optimal distribution $\bar{\boldsymbol{\pi}}$ in (4.4).

Let $\mathbf{J}^0, \mathbf{H}^0, \mathbf{G}^0$ denote the values of $\mathbf{J}_n, \mathbf{H}_n, \mathbf{G}_n$ when $\lambda = 0$ (unpenalized), respectively. Then

$$\mathbf{J}_n(\boldsymbol{\pi}) = \mathbf{J}^0(\boldsymbol{\pi}) + \lambda^2\mathbf{J}^1, \quad \mathbf{H}_n = \mathbf{H}^0 + \lambda\mathbf{H}^1, \quad \mathbf{G}_{n,k} = \mathbf{G}^0_k + \lambda\mathbf{G}^1_k,$$

where $\mathbf{J}^1 = -\dot{P}^{\otimes 2}(\hat{\boldsymbol{\theta}})$, $\mathbf{H}^1 = \ddot{P}(\hat{\boldsymbol{\theta}})$ and $\mathbf{G}^1_k = \dddot{P}_k(\hat{\boldsymbol{\theta}})$. Under suitable conditions, $\mathbf{H}_n^{-1} = (\mathbf{H}^0)^{-1} + \lambda\mathbf{H}^1_- + o(\lambda)$ as λ tends to zero for some matrix $\mathbf{H}^1_-$. It then follows from the trace expression of $q_{n,k}$ that

$$\bar{q}_{n,k}(\boldsymbol{\pi}) = q^0_k(\boldsymbol{\pi}) + q^1_k(\lambda;\boldsymbol{\pi}) + \alpha(\lambda, p+d, \boldsymbol{\pi}),$$

where $q^0_k(\boldsymbol{\pi})$ is the value of $q_{n,k}$ when $\lambda = 0$, and $q^1_k(\lambda;\boldsymbol{\pi}) = \mathrm{Tr}(\lambda\mathbf{C}_1(\boldsymbol{\pi}) + \lambda^2\mathbf{C}_2(\boldsymbol{\pi}) + \lambda^3\mathbf{C}_3(\boldsymbol{\pi}))$ where $\mathbf{C}_k =: C_k(\boldsymbol{\pi})$ are independent of λ, and $\alpha(\lambda, p+d, \boldsymbol{\pi})$ is the remainder. Since each $\mathbf{C}_k$ is a product of four square matrices of dimension $p+d-1$, it follows $\mathrm{Tr}(\mathbf{C}_k) = O((p+d)^4)$, so that $q^1_k(\lambda;\boldsymbol{\pi}) = (\lambda+\lambda^2+\lambda^3)O((p+d)^4)$ and $\alpha(\lambda, p+d, \boldsymbol{\pi}) = o(\lambda)\Omega((p+d)^4)$ (the asymptotic lower bound). Hence

$$\bar{\mathbf{q}}_n(\boldsymbol{\pi}) = (\lambda+\lambda^2+\lambda^3)O((p+d)^{9/2}) + o(\lambda)\Omega((p+d)^{9/2}).$$

Similarly, by (5.1),

$$\bar{\mathbf{b}}_{n1}(\boldsymbol{\pi}) = \mathbf{b}^0_1(\boldsymbol{\pi}) + \lambda\mathbf{b}^1_1(\boldsymbol{\pi}) + o(\lambda)\Omega((p+d)^{5/2}),$$

where $\mathbf{b}^1_1(\boldsymbol{\pi}) = O((p+d)^{5/2})$. Consequently, the first-order bias satisfies

$$\mathrm{Bias}(\hat{\boldsymbol{\phi}}^*, \hat{\boldsymbol{\delta}}^*;\boldsymbol{\pi}) = r^{-1}\mathbf{H}_0^{-1}(\mathbf{b}^0_1(\boldsymbol{\pi}) - 2^{-1}\mathbf{q}^0(\boldsymbol{\pi})) + r^{-1}(\lambda+\lambda^4)O((p+d)^{11/2}) + r^{-1}o(\lambda)\Omega((p+d)^{9/2}). \tag{5.2}$$

Note that the first term on the left-hand side is the main term of the first-order bias for the unpenalized estimator $\hat{\boldsymbol{\theta}}^*$, which is of order of magnitude $O(r^{-1}(p+d)^{11/2})$. It is celebrated in literature that the optimal rate for the penalty is $\lambda = O(\sqrt{r^{-1}\log(p+d)})$, see, e.g., Bickel, *et al.* (2009). Consequently, for the main term of the first-order bias to be negligible at the optimal rate, the dimension $p+d$ must grow with r at such a slow rate that $(p+d)\sqrt[11]{\log(p+d)} = o(r^{3/11})$.

In the ridge regression, $P(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|^2$, hence $\dddot{P}_k(\boldsymbol{\theta}_0) = 0$ for all k and $\mathbf{q}_n \equiv 0$. Accordingly, $\bar{\mathbf{b}}_{n1}(\boldsymbol{\pi}) = \mathbf{b}^0_1(\boldsymbol{\pi}) + \lambda\mathbf{b}^1_1(\boldsymbol{\pi})$. As $\mathbf{b}^1_1(\boldsymbol{\pi}) = O((p+d)^{5/2})$, we have

$$\mathrm{Bias}(\hat{\boldsymbol{\phi}}^*, \hat{\boldsymbol{\delta}}^*;\boldsymbol{\pi}) = r^{-1}\mathbf{H}_0^{-1}\mathbf{b}^0_1(\boldsymbol{\pi}) + r^{-1}(\lambda+\lambda^2)O((p+d)^{5/2}) + r^{-1}o(\lambda(p+d)^{5/2}), \tag{5.3}$$

In this case, for the main term of the first-order bias to be negligible at the optimal rate, the dimension $p+d$ grows at a faster rate $(p+d)\sqrt[5]{\log(p+d)} = o(r^{3/5})$. The message is that a less smooth penalty function leads to more biased estimates, high dimension causes tremendous biases, and a bigger penalty λ results in higher biases.

The Biases for the Unweighted Estimators. Similarly, we have

$$\tilde{\mathbf{H}}_n(\boldsymbol{\pi}) = \mathrm{E}^*(\dot{\tilde{\boldsymbol{\Phi}}}{}^*_n(\tilde{\boldsymbol{\theta}})) = \sum_{i=1}^n \pi_i\dot{\mathbf{g}}_i(\tilde{\boldsymbol{\theta}}) + \lambda\ddot{P}(\tilde{\boldsymbol{\theta}}), \quad \tilde{\mathbf{G}}_{n,k}(\boldsymbol{\pi}) = \mathrm{E}^*(\ddot{\tilde{\boldsymbol{\Psi}}}{}^*_{n,k}(\tilde{\boldsymbol{\theta}})) = \sum_{i=1}^n \pi_i\ddot{g}_{i,k}(\tilde{\boldsymbol{\theta}}) + \lambda\dddot{P}_k(\tilde{\boldsymbol{\theta}}).$$

Also, $\tilde{\mathbf{J}}_n(\boldsymbol{\pi}) = \mathrm{E}^*(\tilde{\boldsymbol{\Psi}}^{*\otimes 2}_n) = r^{-1}\tilde{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi})$ with $\tilde{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi})$ given in (3.20). The first-order bias for $\tilde{\boldsymbol{\theta}}^* = (\tilde{\boldsymbol{\phi}}^*, \tilde{\boldsymbol{\delta}}^*)$ is

$$\mathrm{Bias}(\tilde{\boldsymbol{\phi}}^*, \tilde{\boldsymbol{\delta}}^*;\boldsymbol{\pi}) = r^{-1}\tilde{\mathbf{H}}_n^{-1}(\boldsymbol{\pi})(\tilde{\mathbf{b}}_{n1}(\boldsymbol{\pi}) - 2^{-1}\tilde{\mathbf{q}}_n(\boldsymbol{\pi})), \tag{5.4}$$

where $\tilde{\mathbf{q}}_n = (\tilde{q}_{n,k}(\boldsymbol{\pi}))$ with

$$\tilde{q}_{n,k}(\boldsymbol{\pi}) = r\mathrm{Tr}(\tilde{\mathbf{H}}_n^{-\top}\tilde{\mathbf{G}}_{n,k}\tilde{\mathbf{H}}_n^{-1}) = \mathrm{Tr}(\tilde{\mathbf{H}}_n^{-\top}(\boldsymbol{\pi})\tilde{\mathbf{G}}_{n,k}(\boldsymbol{\pi})\tilde{\mathbf{H}}_n^{-1}(\boldsymbol{\pi})\tilde{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi})),$$

$$\tilde{\mathbf{b}}_{n1}(\boldsymbol{\pi}) = r\mathrm{E}^*(\dot{\tilde{\boldsymbol{\Phi}}}{}^*_n(\tilde{\boldsymbol{\theta}})\tilde{\mathbf{H}}_n^{-1}\tilde{\boldsymbol{\Phi}}^*_n) = \sum_{i=1}^n \pi_i\dot{\mathbf{g}}_i(\tilde{\boldsymbol{\theta}})\tilde{\mathbf{H}}_n^{-1}(\boldsymbol{\pi})\tilde{\mathbf{g}}_i + \lambda\Big(\sum_{i=1}^n \pi_i\dot{\mathbf{g}}_i(\tilde{\boldsymbol{\theta}})\Big)\tilde{\mathbf{H}}_n^{-1}(\boldsymbol{\pi})P'(\tilde{\boldsymbol{\theta}}).$$

As a result, the first-order bias for $\tilde{\boldsymbol{\beta}}^*(\boldsymbol{\pi}) = \tilde{\boldsymbol{\beta}}^*(\tilde{\boldsymbol{\phi}}^*(\boldsymbol{\pi}))$ is $\mathrm{Bias}(\tilde{\boldsymbol{\beta}}^*)(\boldsymbol{\pi}) = \mathbf{Jac}(\tilde{\boldsymbol{\phi}}(\boldsymbol{\pi}))\mathrm{Bias}(\tilde{\boldsymbol{\phi}}^*(\boldsymbol{\pi}))$.

Unlike the weighted subsampling estimators in which the $\mathrm{Bias}(\hat{\boldsymbol{\phi}}^*, \hat{\boldsymbol{\delta}}^*)$ in (5.1) are *inversely* influenced by the sampling distribution $\boldsymbol{\pi}$, the $\mathrm{Bias}(\tilde{\boldsymbol{\phi}}^*, \tilde{\boldsymbol{\delta}}^*)$ of the unweighted estimators *directly* depend on $\boldsymbol{\pi}$. As a result, truncation for a sampling distribution $\boldsymbol{\pi}$ may not be needed, whereas the influence of the penalty λ on the biases is similar to the aforementioned result for the weighted estimators.

5.2 The Efficiency Comparison

By the A-optimality, we have

Remark 14 The $\hat{A}(\bar{A})$-optimal sampling distribution $\hat{\boldsymbol{\pi}}$ ($\bar{\boldsymbol{\pi}}$) given in (4.3) ((4.4)) minimizes the trace norm of the variance-covariance matirx $\hat{\mathbf{V}}_n(\boldsymbol{\pi})(\bar{\mathbf{V}}_n(\boldsymbol{\pi}))$ of the weighted subsampling estimator $\hat{\boldsymbol{\theta}}^*$, that is, $\mathrm{Tr}(\hat{\mathbf{V}}_n(\hat{\boldsymbol{\pi}})) \leq \mathrm{Tr}(\hat{\mathbf{V}}_n(\boldsymbol{\varpi}))$ $(\mathrm{Tr}(\bar{\mathbf{V}}_n(\bar{\boldsymbol{\pi}})) \leq \mathrm{Tr}(\bar{\mathbf{V}}_n(\boldsymbol{\varpi})))$ for any sampling distribution $\boldsymbol{\varpi}$.

To clarify our efficiency comparison of the sampling distributions based on different estimators, we introduce

Definition 1 *Let $\hat{\boldsymbol{\theta}}_{kn}, k = 1, 2$ be two consistent estimators of $\boldsymbol{\theta}_0$. Given a subsample of size r and for each k, let $\hat{\boldsymbol{\theta}}^*_{kn}$ be a subsampling estimator approximating $\hat{\boldsymbol{\theta}}_{kn}$ such that $\sqrt{r}(\hat{\boldsymbol{\theta}}^*_{kn} - \hat{\boldsymbol{\theta}}_{kn})$ is asymptotically normal in probability with zero mean and covariance matrix $\mathbf{V}_{kn}(\hat{\boldsymbol{\theta}}_n) = (\mathbf{H}_{kn}^{-1}\boldsymbol{\Sigma}_{kn}\mathbf{H}_{kn}^{-t})(\hat{\boldsymbol{\theta}}_n)$ for some invertible matrix $\mathbf{H}_{kn}(\hat{\boldsymbol{\theta}}_n)$ and positive definite matrix $\boldsymbol{\Sigma}_{kn}(\hat{\boldsymbol{\theta}}_{kn})$. We say that $\hat{\boldsymbol{\theta}}^*_{1n}$ is (asymptotically) more efficient than $\hat{\boldsymbol{\theta}}^*_{2n}$ if $\bar{\mathbf{V}}_{kn} = \mathrm{E}(\mathbf{H}_{kn}(\boldsymbol{\theta}_0)|\mathbf{X})^{-1}\mathrm{E}(\boldsymbol{\Sigma}_{kn}(\boldsymbol{\theta}_0)|\mathbf{X})\mathrm{E}(\mathbf{H}_{kn}(\boldsymbol{\theta}_0)|\mathbf{X})^{-t}$ are well defined and invertible such that for any compatible $\mathbf{u}$,*

$$\mathbf{u}^t(\bar{\mathbf{V}}_{1n}^{-1} - \bar{\mathbf{V}}_{2n}^{-1})\mathbf{u} \geq 0, \quad n \to \infty. \tag{5.5}$$

Typically, there is a sequence of positive numbers c_n (the same c_n may mean different values) such that

$$\boldsymbol{\Sigma}_{kn}(\hat{\boldsymbol{\theta}}_{kn}) = \mathrm{E}(\boldsymbol{\Sigma}_{kn}(\boldsymbol{\theta}_0)|\mathbf{X}) + o_P(c_n), \quad \mathbf{H}_{kn}(\hat{\boldsymbol{\theta}}_{kn}) = \mathrm{E}(\mathbf{H}_{kn}(\boldsymbol{\theta}_0)|\mathbf{X}) + o_P(c_n), \quad k = 1, 2. \tag{5.6}$$

Assume that there are constants $0 < l_0 \leq u_0$ and a neighborhood $\mathbf{N}_0$ of $\boldsymbol{\theta}_0$ such that $\mathbf{H}_{kn}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_{kn}(\boldsymbol{\theta}), k = 1, 2$ are continuous in $\boldsymbol{\theta} \in \mathbf{N}_0$, and that

$$l_0 c_n \leq \lambda_{\min}\mathbf{H}_{kn}(\boldsymbol{\theta}) \leq \lambda_{\max}\mathbf{H}_{kn}(\boldsymbol{\theta}) \leq u_0 c_n, \quad l_0 c_n \leq \lambda_{\min}\boldsymbol{\Sigma}_{kn}(\boldsymbol{\theta}) \leq \lambda_{\max}\boldsymbol{\Sigma}_{kn}(\boldsymbol{\theta}) \leq u_0 c_n, \ \boldsymbol{\theta} \in \mathbf{N}_0. \tag{5.7}$$

These boundedness conditions can be relaxed to bounded in probability. Let $\mathbf{V}_{kn} = \mathbf{H}_{kn}^{-1}(\hat{\boldsymbol{\theta}}_{kn})\boldsymbol{\Sigma}_{kn}(\hat{\boldsymbol{\theta}}_{kn})\mathbf{H}_{kn}^{-t}(\hat{\boldsymbol{\theta}}_{kn})$. Then for any unit vector $\mathbf{u}$,

$$\mathbf{u}^t(\mathbf{V}_{1n}^{-1} - \mathbf{V}_{2n}^{-1})\mathbf{u} \geq o_P(c_n). \tag{5.8}$$

Recalling $\hat{\boldsymbol{\Sigma}}_n$ in (3.7) and $\hat{\mathbf{H}}_n$ therein, we write $\hat{\boldsymbol{\Sigma}}_n = \hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi}; \hat{\boldsymbol{\theta}}, \lambda)$, $\hat{\mathbf{H}}_n = \hat{\mathbf{H}}_n(\hat{\boldsymbol{\theta}}, \lambda)$, etc. to stress the dependence on $\boldsymbol{\pi}, \lambda$, etc.. We shall study the case that the penalty $\lambda = \lambda_n \to 0$ as $n \to \infty$. To simplify the presentation, consider the unpenalized case of $\lambda = 0$. Set $\hat{\boldsymbol{\Sigma}}_0 = \hat{\boldsymbol{\Sigma}}_0(\boldsymbol{\pi}) = \mathrm{E}(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi}; \boldsymbol{\theta}_0, 0)|\mathbf{X})$, $\tilde{\boldsymbol{\Sigma}}_0 = \tilde{\boldsymbol{\Sigma}}_0(\boldsymbol{\pi}) = \mathrm{E}(\tilde{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi}; \boldsymbol{\theta}_0, 0)|\mathbf{X})$, $\hat{\mathbf{H}}_0 = \mathrm{E}(\hat{\mathbf{H}}_n(\boldsymbol{\theta}_0, 0)|\mathbf{X})$, and $\tilde{\mathbf{H}}_0 = \tilde{\mathbf{H}}_0(\boldsymbol{\pi}) = \mathrm{E}(\tilde{\mathbf{H}}_n(\boldsymbol{\pi}; \boldsymbol{\theta}_0, 0)|\mathbf{X})$. We then have

$$\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi}; \boldsymbol{\theta}_0, 0) = \frac{1}{n}\sum_{i=1}^n \frac{\mathbf{g}_i^{\otimes 2}}{n\pi_i}, \quad \tilde{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi}; \boldsymbol{\theta}_0, 0) = \sum_{i=1}^n \pi_i \mathbf{g}_i^{\otimes 2}, \quad \hat{\mathbf{H}}_n(\boldsymbol{\theta}_0, 0) = \frac{1}{n}\sum_{i=1}^n \dot{\mathbf{g}}_i, \quad \tilde{\mathbf{H}}_n(\boldsymbol{\pi}; \boldsymbol{\theta}_0, 0) = \sum_{i=1}^n \pi_i \dot{\mathbf{g}}_i.$$

To compare $\hat{\mathbf{V}}_0 =: \hat{\mathbf{V}}_0(\boldsymbol{\pi}) = \hat{\mathbf{H}}_0^{-1}\hat{\boldsymbol{\Sigma}}_0(\boldsymbol{\pi})\hat{\mathbf{H}}_0^{-t}$ and $\tilde{\mathbf{V}}_0 =: \tilde{\mathbf{V}}_0(\boldsymbol{\pi}) = \tilde{\mathbf{H}}_0^{-1}(\boldsymbol{\pi})\tilde{\boldsymbol{\Sigma}}_0(\boldsymbol{\pi})\tilde{\mathbf{H}}_0^{-t}(\boldsymbol{\pi})$, we impose

$\tilde{\mathbf{A}}$6 $\boldsymbol{\pi} = \boldsymbol{\pi}(\mathbf{X})$ depends on the covariates $\mathbf{X} = (\mathbf{x}_i)$ and is independent of the random errors $\boldsymbol{\epsilon} = (\epsilon_i)$.

Remark 15 It is obvious that the uniform and the leverage-scores-based distributions satisfy ($\tilde{\mathbf{A}}$6). Let $\underset{\sim}{\boldsymbol{\pi}}_0$ be the distribution $\underset{\sim}{\boldsymbol{\pi}}$ given in (4.5) but with $\mathbf{H}_n$ replaced by the conditional expected value given $\mathbf{X}$ (i.e. $\hat{\mathbf{H}}_0$), so that $\underset{\sim}{\boldsymbol{\pi}}_0 \propto (\hat{\mathbf{H}}_0^{-1} f_i')$. Then $\underset{\sim}{\boldsymbol{\pi}}_0$ satisfies ($\tilde{\mathbf{A}}$6).

As $\mathbf{g}_i = -2e_i f_i'$ and $\dot{\mathbf{g}}_i = 2f_i'^{\otimes 2} - 2\mathbf{e}_i\ddot{f}_i$, we get $\mathrm{E}(\mathbf{g}_i^{\otimes 2}|\mathbf{X}) = 4\sigma_0^2 f_i'^{\otimes 2}$ and $\mathrm{E}(\dot{\mathbf{g}}_i|\mathbf{X}) = 2f_i'^{\otimes 2}$. Let $\check{\boldsymbol{\Sigma}}_0 = n^{-1}\sum_{i=1}^n \mathbf{g}_i^{\otimes 2}$. Then

$$\mathrm{E}(\check{\boldsymbol{\Sigma}}_0|\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n 4\sigma_0^2 f_i'^{\otimes 2} = 2\sigma_0^2\hat{\mathbf{H}}_0, \quad \tilde{\boldsymbol{\Sigma}}_0(\boldsymbol{\pi}) = \sum_{i=1}^n \pi_i 4\sigma_0^2 f_i'^{\otimes 2} = 2\sigma_0^2\tilde{\mathbf{H}}_0(\boldsymbol{\pi}). \tag{5.9}$$

The first formula is the *generalized conditional information matrix equality* for the objective function $Q_n(\boldsymbol{\phi})$ in (2.10) and the second for the objective $\tilde{Q}_n(\boldsymbol{\phi})$ in (3.14).

Let $\mathbf{d}_i = \sqrt{\pi_i} f_i'$ and $\mathbf{D}^t = (\mathbf{d}_1, \ldots, \mathbf{d}_n)$. Under ($\tilde{\mathbf{A}}$6), $\tilde{\boldsymbol{\Sigma}}_0 = 4\sigma_0^2\sum_{i=1}^n \mathbf{d}_i^{\otimes 2} = 4\sigma_0^2\mathbf{D}^t\mathbf{D}$, so that $\sigma_0^2\tilde{\mathbf{V}}_0^{-1} = \mathbf{D}^t\mathbf{D}$ by the second equality in (5.9). Let $\mathbf{b}_i = \dot{f}_i/\sqrt{\pi_i}$ and $\mathbf{B}^t = (\mathbf{b}_1, \ldots, \mathbf{b}_n)$. By ($\tilde{\mathbf{A}}$6) again, we get $\sum_{i=1}^n \mathbf{b}_i^{\otimes 2} = \mathbf{B}^t\mathbf{B} =$

$4^{-1}\sigma_0^{-2}n^2\hat{\mathbf{\Sigma}}_0$. By the first equality in (5.9), we get $\sum_{i=1}^n \mathbf{d}_i\mathbf{b}_i^t = \mathbf{D}^t\mathbf{B} = 2^{-1}n\hat{\mathbf{H}}_0$. Let $\mathbf{M}$ be the block matrix consisting column-wise of blocks $\mathbf{D}$ and $\mathbf{B}$. Then $\mathbf{M}^t\mathbf{M}$ is semi-positive definite, so that

$$\sigma_0^2\mathbf{u}^t(\tilde{\mathbf{V}}_0^{-1} - \hat{\mathbf{V}}_0^{-1})\mathbf{u} = \mathbf{u}^t(\mathbf{D}^t\mathbf{D} - (\mathbf{D}^t\mathbf{B})(\mathbf{B}^t\mathbf{B})^{-1}(\mathbf{D}^t\mathbf{B})^t)\mathbf{u} \geq 0, \quad \forall \mathbf{u}.$$

Summarizing the above derivations, we prove

Theorem 5.1 *Consider the unpenalized case of $\lambda = 0$. Let $\boldsymbol{\pi}$ be a distribution on data points such that ($\tilde{\mathbf{A}}6$) is met. Suppose that the assumptions in Theorems 1 and 2 hold. Assume that $\hat{\boldsymbol{\theta}}_n$ and $\tilde{\boldsymbol{\theta}}_n(\boldsymbol{\pi})$ are consistent estimators of $\boldsymbol{\theta}_0$. Then the unweighted subsampling estimator $\tilde{\boldsymbol{\theta}}_n^*(\boldsymbol{\pi})$ which approximates $\tilde{\boldsymbol{\theta}}_n(\boldsymbol{\pi})$ is asymptotically more efficient than the weighted estimator $\hat{\boldsymbol{\theta}}_n^*(\boldsymbol{\pi})$ which approximates $\hat{\boldsymbol{\theta}}_n$.*

Consider the penalized case of $\lambda = \lambda_n = o(1)$. Under suitable conditions, $\tilde{\boldsymbol{\theta}}_n(\boldsymbol{\pi}) = \boldsymbol{\theta}_0 + o_P(1)$ and $\hat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_0 + o_P(1)$. It then follows from (3.7) and (3.20) that (5.6) is met with $c_n = 1$, that is,

$$\tilde{\mathbf{\Sigma}}_n(\boldsymbol{\pi}) = \tilde{\mathbf{\Sigma}}_n(\boldsymbol{\pi}; \tilde{\boldsymbol{\theta}}_n(\boldsymbol{\pi}), \lambda_n) = \tilde{\mathbf{\Sigma}}_n(\boldsymbol{\pi}; \boldsymbol{\theta}_0, 0) + o_P(1), \quad \tilde{\mathbf{H}}_n(\boldsymbol{\pi}) = \tilde{\mathbf{H}}_n(\boldsymbol{\pi}, \tilde{\boldsymbol{\theta}}_n(\boldsymbol{\pi}), \lambda_n) = \tilde{\mathbf{H}}_n(\boldsymbol{\pi}, \boldsymbol{\theta}_0, 0) + o_P(1),$$
$$\hat{\mathbf{\Sigma}}_n(\boldsymbol{\pi}) = \hat{\mathbf{\Sigma}}_n(\boldsymbol{\pi}; \hat{\boldsymbol{\theta}}_n, \lambda_n) = \hat{\mathbf{\Sigma}}_n(\boldsymbol{\pi}; \boldsymbol{\theta}_0, 0) + o_P(1), \quad \hat{\mathbf{H}}_n = \hat{\mathbf{H}}_n(\hat{\boldsymbol{\theta}}_n, \lambda_n) = \hat{\mathbf{H}}_n(\boldsymbol{\theta}_0, 0) + o_P(1).$$

Assume that the above quantities are bounded as spelt in (5.7). Assume also that the penalty function $P(\boldsymbol{\theta})$ and its first and second partial derivatives $\dot{P}(\boldsymbol{\theta})$ and $\ddot{P}(\boldsymbol{\theta})$ are bounded in $\boldsymbol{\theta} \in \mathbf{N}_0$. By (5.8), for any $\boldsymbol{\pi}$ that satisfies ($\tilde{\mathbf{A}}6$) and any unit vector $\mathbf{u}$,

$$\mathbf{u}^t\big(\tilde{\mathbf{V}}_n^{-1}(\boldsymbol{\pi}, \tilde{\boldsymbol{\theta}}_n(\boldsymbol{\pi}), \lambda_n) - \hat{\mathbf{V}}_n^{-1}(\boldsymbol{\pi}, \hat{\boldsymbol{\theta}}_n, \lambda_n)\big)\mathbf{u} \geq o_P(1). \tag{5.10}$$

This exhibits that the unweighted subsampling penalized estimator $\tilde{\boldsymbol{\theta}}_n^*$ is asymptotically more efficient than the weighted penalized estimator $\hat{\boldsymbol{\theta}}_n^*$, which holds on an event whose probability goes to one as n tends to infinity.

6 A Large Simulation Study

In this section, we use two penalized spline SIM (both with $p = 12$ and $p + d = 26$) and three simulated datasets to numerically investigate the proposed A-optimal Subsampling approach.

The sampling distribution $\underline{\boldsymbol{\pi}}$ given in (4.8) was calculated using the Scoring Algorithm 3. For comparison, we also reported the results of the uniform sampling $\boldsymbol{\pi}_i = \mathbf{1}/n$ based on $B = 500$ repetitions for the sake of computational ease. From a practical viewpoint, a large value of B would be needed, see a systematic study of sample size determination in Zhang, *et al.* (2023).

Dataset 1. Generate i.i.d. random errors ϵ_i from the standard normal $N(0, 1)$ and covariates $\mathbf{x}_i$ from the p-variate normal $N(0, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = (\Sigma_{j,k}) = (0.5^{|j-k|})$ (treated as nonrandom), choose $\boldsymbol{\beta}_0$ equal to the vector consisting of $(1, 0.001, 0.001)$ repeated $p/3$ times normalized to satisfy the constraint (2.4), and generate y_i from

$$y_i = (\mathbf{x}_i^t\boldsymbol{\beta}_0)^2\exp(\mathbf{x}_i^t\boldsymbol{\beta}_0) + \sigma_0\epsilon_i, \;\; i = 1, 2, \ldots, n = 10^5, \; \sigma_0 = 1, \; p = 12.$$

Dataset 2. Same as Dataset 1 except for $\mathbf{x_i}$'s generated from the normal mixture $0.8N(0, \mathbf{\Sigma}) + 0.2N(0, 10\mathbf{\Sigma})$.

Dataset 3. Same as Dataset 1 except for $\mathbf{x_i}$'s generated from the multivariate t- distribution with 8 degrees of freedom.

Next, we apply the Subsampling approach to the following two penalized spline SIM.

6.1 The Penalized B-spline SIM

Let $\mathbf{t} = \{t_i\}_{i=0}^{\kappa+1}$ be κ interior knots with $t_0 \leq t_1 \leq \cdots \leq t_{\kappa+1}$. Define the augmented knots set $\{t_i\}_{i=1-m}^{\kappa+m}$ by

$$t_{-(m-1)} = \cdots = t_{-1} = t_0 \leq t_1 \leq \cdots \leq t_\kappa \leq t_{\kappa+1} = \cdots = t_{\kappa+m}.$$

Rearrange to get $\{t_i\}_{i=0}^{\kappa+2m-1}$. For the B-spline of order m (of degree $m - 1$), define $\{B_{i,j}\}_{j=0,1,\ldots,m-1}$ by recurrence:

$$B_{i0}(t) := \begin{cases} 1, & \text{if } t_i \leq t < t_{i+1}, \\ 0, & \text{otherwise,} \end{cases} \qquad B_{i,j+1} = \omega_{i,j+1}B_{i,j} + (1 - \omega_{i+1,j+1})B_{i+1,j}, \tag{6.1}$$

where $\omega_{ij}(t) = (t - t_i)/(t_{i+j} - t_i)$ if $t_i \neq t_{i+j}$ and 0 otherwise. Note that these functions are right-continuous. By definition, a B-spline of order m (degree $N = m - 1$) with knots $\mathbf{t}$ of length $\kappa + 2$ (i.e. κ interior knots) is a linear combination of the B-splines B_{iN}, $\boldsymbol{\delta}^t\mathbf{B}(t) = \sum_{i=0}^{\kappa+N} \delta_i B_{iN}(t)$, as described in Section 1.

In the literature, knots are selected as equally spaced quantiles of indices. The larger the number of knots, the more flexible the curve fitting is to a dataset. To avoid overfitting, O'Sullivan (1986, 1988) proposed the roughness penalty,

$$P(\boldsymbol{\theta}) = \int_L^U \Big(\sum_{i=0}^{\kappa+N} \delta_i B''_{iN}(s) \Big)^2 \, ds = \boldsymbol{\delta}^t \mathbb{D} \boldsymbol{\delta}, \tag{6.2}$$

where $\mathbb{D} = \mathbb{D}_{(\kappa+m)\times(\kappa+m)}$ with $\mathbb{D}_{ij} = \int_L^U B''_{iN}(s) B''_{jN}(s)\, ds$ and $U = \max(\boldsymbol{\beta}^t \mathbf{x}_i)$ and $L = \min(\boldsymbol{\beta}^t \mathbf{x}_i)$. It is well known that this is equivalent to the second order difference penalty in Eiler and Marx (1996). We shall use the B-spline SIM with this penalty. Clearly, the Lipschitz condition in A4 is met.

The minimizer $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^t,\, \hat{\boldsymbol{\delta}}^t)^t$ of the objective function in (2.10) satisfies $\dot{Q}_n(\hat{\boldsymbol{\theta}}) = 0$. This is a ridge regression, so that the estimator $\boldsymbol{\delta}(\boldsymbol{\phi})$ of the model parameter vector $\boldsymbol{\delta}$ can be expressed as $\hat{\boldsymbol{\delta}}(\boldsymbol{\phi}) = (\mathbb{B}^t \mathbb{B} + n\lambda \mathbb{D})^{-1} \mathbb{B}\mathbf{y}$, where $\mathbb{B} =: \mathbb{B}(\boldsymbol{\phi})$ is an $n \times d$ matrix with its ith row equal to $\mathbf{B}^t(\mathbf{x}_i^t \boldsymbol{\beta}(\boldsymbol{\phi}))$. The optimization is a $p-1$ dimensional problem.

The Hessian matrix is usually unavailable as it is computationally expensive to calculate, and the quasi-newton method is a popular optimization method for approximating the Hessian matrix requiring only gradient information. We use the quasi-newton method instead of the traditional newton's method. This is implemented in the BFGS algorithm in "optim" package in R. See Dennis and Schnabel (1983) for the properties of BFGS.

We employ the cubic spline, i.e., $N = 3$ (so $m = 4$), and choose the number of interior knots to be $\kappa = 10$, so that $d = 14$ and $p + d = 26$. Cubic splines are commonly used for its simplicity and smoothness properties (Lipschitz continuity of the second order derivative).

Presample Size Determination Before we proceeded, we investigated how different values of the pre-subsample size r_0 used in the Scoring Algorithm affected the performance of the subsampling estimator. Practically, r_0 should be as small as possible in comparison to the subsample size $r << n$, while maintaining a reasonable efficiency of the pilot estimator used in the approximation to the subsampling distributions. To this goal, we choose r_0 to be

$$100(0.1\%n),\ 300(0.3\%n),\ 500(0.5\%n),\ 1000(1\%n),\ 5000(5\%n).$$

The simulations (not reported here) suggested that $r_0 = 500(0.5\%n)$ was reasonable. Recently, Zhang, *et al.* (2023) provided the formulas for sample size determination in the case of parameter vectors.

Empirical Mean Squared Error and Bias For each of a few subsample sizes r, we repeat the Subsampling approach $B = 500$ times, and calculate the empirical mean squared error (EMSE) of the subsampling estimator $\hat{\boldsymbol{\beta}}^*(\boldsymbol{\pi})$ under the A-optimal (opt) and the uniform (unif) subsamplings and their ratio $\mathrm{EMSE}_{\mathrm{ratio}} = \mathrm{EMSE}_{\mathrm{opt}}/\mathrm{EMSE}_{\mathrm{unif}}$, using the following formula for the EMSE,

$$\mathrm{EMSE}(\hat{\boldsymbol{\beta}}^*(\boldsymbol{\pi})) = \frac{1}{B} \sum_{b=1}^{B} \|\hat{\boldsymbol{\beta}}^*_{(b)}(\boldsymbol{\pi}) - \hat{\boldsymbol{\beta}}\|^2, \tag{6.3}$$

where $\hat{\boldsymbol{\beta}}^*_{(b)}(\boldsymbol{\pi})$ is the subsampling estimator in the bth repetition and $\hat{\boldsymbol{\beta}}$ is the full-sample estimator. Here $\hat{\boldsymbol{\beta}}$ is used instead of the true value $\boldsymbol{\beta}_0$ as $\hat{\boldsymbol{\beta}}^*(\boldsymbol{\pi})$ is an approximation to $\hat{\boldsymbol{\beta}}$. The empirical squared bias of $\hat{\boldsymbol{\beta}}^*$ for the optimal and the uniform subsamplings and their ratio $\mathrm{Bias}_{\mathrm{ratio}} = \mathrm{Bias}_{\mathrm{opt}}/\mathrm{Bias}_{\mathrm{unif}}$ are also calculated, using the following formula,

$$\mathrm{Bias}(\hat{\boldsymbol{\beta}}^*(\boldsymbol{\pi})) = \Big\| \frac{1}{B} \sum_{i=1}^{B} \hat{\boldsymbol{\beta}}^*_{(b)}(\boldsymbol{\pi}) - \hat{\boldsymbol{\beta}} \Big\|^2. \tag{6.4}$$

In the implementation of the Scoring Algorithm, we chose $r_0 = 500(.5\%n)$ and calculated an approximation to the A-optimal distribution in Step 1 and the subsampling estimator in Step 2 for various subsample sizes from $r = 100(.1n\%)$ to $r = 5,000(5n\%)$. For the uniform subsampling, go directly to step 2 with all $\pi_i = 1/n$.

Reported in Table 1 are the simulated EMSE (and their ratios) of the A-optimal and the uniform subsampling estimators using Dataset 1. One observes that EMSE was decreasing as r increased until r reached $3,000(3\%n)$, it became stable. The EMSE ratios were consistently and significantly less than 1. Also, the bias ratios were less than 1 for all r sizes considered.

Reported in Table 2 are the amount of time used (including the time for computing the sampling distribution) for computing the subsampling estimators using the Scoring Algorithm. We summarized the results as follows. First, the time taken for the smallest subsample size $r = 100$ in Table 2 was quite lengthy. This is possibly due to the convergence problem of the numerical solution, or of the number of repetitions. Similar behaviors also appeared in other cases. Second, the Subsampling approach saved significant amount of time compared to the full sample estimators for all the subsample sizes considered, while maintaining the desirable efficiency in terms of EMSE. The A-optimal subsampling estimator took a bit less time than the uniform subsampling estimator in Step 2 of the Scoring Algorithm, but needed extra time to calculate the sampling distribution in Step 1, which was 25.67 seconds. The total time needed for implementing the A-optimal subsampling estimators were still significantly less than $1,174.33$ seconds spent by the full-sample estimator.

The results for Datasets 2 and 3 are given in Tables 13–14 and 15–16, respectively, and reported in Section 9 as **Supplmentary Material**. The results are similar to those for Dataset 1. For example, the EMSE in Table 13 decreased as r increased until r reached $3\%n$ where the EMSE became stable. The EMSE ratios were consistently less than 1.

Table 1: The simulated EMSE and biases of the A-optimal and the uniform subsampling estimators $\hat{\boldsymbol{\beta}}^*(\boldsymbol{\pi})$ and their ratios for the penalized B-spline SIM with $d = 14$, $p = 12$ and $n = 100,000$, using Dataset 1.

r	EMSE_{opt}	$\text{EMSE}_{\text{unif}}$	$\text{EMSE}_{\text{ratio}}$	Bias_{opt}	$\text{Bias}_{\text{unif}}$	$\text{Bias}_{\text{ratio}}$
100(0.1%n)	0.007611	0.246294	0.030902	0.000361	0.020007	0.018049
300(0.3%n)	0.000301	0.010877	0.027708	0.000277	0.000419	0.660328
500(0.5%n)	0.000298	0.002183	0.136721	0.000279	0.000363	0.769482
1000(1%n)	0.000290	0.002175	0.133564	0.000275	0.000407	0.676333
3000(3%n)	0.000267	0.002454	0.108674	0.000257	0.000310	0.828419
5000(5%n)	0.000262	0.002598	0.100932	0.000255	0.000277	0.919580

Table 2: The average time (in seconds) taken to calculate the subsampling estimator $\hat{\boldsymbol{\beta}}^*(\underline{\boldsymbol{\pi}})$ for the penalized B-spline SIM. The full-sample estimator $\hat{\boldsymbol{\beta}}$ and $\underline{\boldsymbol{\pi}}$ (Step 1) took $1,174.33$s and 25.67s respectively, using Dataset 1.

r	100(0.1%n)	300(0.3%n)	500(0.5%n)	1000(1%n)	3000(3%n)	5000(5%n)
Time_{opt}	31.4029	5.9084	6.1495	8.3305	10.5931	9.0264
$\text{Time}_{\text{unif}}$	50.2206	7.5738	7.4472	10.0831	15.9996	15.1435

6.2 The Penalized P-spline SIM

Inroduced by Yu and Ruppert (2002), the mean function $m(\cdot)$ in this model is estimated by a P-spline,

$$m(u) = \boldsymbol{\delta}^t \mathbf{B}(u),$$

where $\boldsymbol{\delta} = (\delta_0, \delta_1, \ldots, \delta_{q+K})^t$ is the spline coefficient vector, and the spline basis is the truncated powers given by

$$\mathbf{B}(u) = (1,\, u,\, \ldots,\, u^q,\, (u - \kappa_1)_+^q,\, \ldots,\, (u - \kappa_K)_+^q)^t. \tag{6.5}$$

Here q is the order of spline basis and K is the number of knots. The knots $\kappa_1, \kappa_2, \ldots, \kappa_K$ are selected to be the equally spaced sample quantiles of the index $\boldsymbol{\beta}^t\mathbf{x}$. Note that $q > 2$ is needed to ensure the second order differentiability of the spline basis functions. The spline for $q = 3$ is the cubic spline, which has the Lipschitz-continuous second order derivatives. For the choice of number of knots K, Ruppert (2002) suggested that 5 to 10 knots are quite adequate for smooth a monotonic or unimodal regression function.

Yu and Ruppert proposed the residual sum of squares plus the partial ridge penalty as the objective function,

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \boldsymbol{\delta}^t \mathbf{B}(\beta(\boldsymbol{\phi})^t \mathbf{x}_i)\right)^2 + \lambda \boldsymbol{\delta}^t \mathbb{D} \boldsymbol{\delta}, \tag{6.6}$$

where $\mathbb{D}$ is a diagonal matrix with the last K diagonal entries equal to 1 and the rest equal to 0. This together with the spline basis in (6.5) implies that the penalty parameter λ works to avoid overfitting by penalizing the last K elements of model parameter $\boldsymbol{\delta}$, which forces the fitted curve to bend toward the data points closely through the knots. The penalty function clearly satisfies the smoothness assumptions.

We chose the cubic spline (i.e., q=3) and the number of knots $K = 10$, so that $d = 14$. For Dataset 1, the results are reported in Tables 3 – 4. One observes that the simulated EMSE values decreased with the increasing r for both the A-optimal and the uniform subsamplings. However, the EMSE of the A-optimal subsampling estimator decreased much faster for $r \geq 1000(1\%n)$, and the EMSE ratios were around 0.27, which is substantially smaller than 1, indicating that the A-optimal subsampling estimators significantly outperformed the uniform subsampling estimators.

The results for Datasets 2 and 3 are given in Tables 17–18 and 19–20, respectively, and reported in Section 9 as **Supplementary Material**. The results are similar to those for Dataset 1.

For all the datasets, the proposed Subsampling approach saved significant amount of time for all the subsmaple sizes considered. For example, in Table 20, the full-sample estimator took $4,4891.23$ seconds, while the proposed subsampling estimator took time in between 40 and 58 seconds.

Table 3: The simulated EMSE and biases of the A-optimal and the uniform subsampling estimators $\hat{\boldsymbol{\beta}}^*(\boldsymbol{\pi})$ and their ratios for the P-spline SIM with $d = 14$, $p = 12$, and $n = 100,000$, using Dataset 1.

r	EMSE_{opt}	$\text{EMSE}_{\text{unif}}$	$\text{EMSE}_{\text{ratio}}$	Bias_{opt}	$\text{Bias}_{\text{unif}}$	$\text{Bias}_{\text{ratio}}$
100(0.1%n)	0.5601926	0.2587949	2.1646196	0.1046895	0.0220413	4.7496934
300(0.3%n)	0.1042635	0.0120707	8.6377527	0.0084708	0.0011116	7.6204553
500(0.5%n)	0.0137123	0.0038783	3.5356218	0.0012988	0.0010875	1.1943317
1000(1%n)	0.0010950	0.0039514	0.2771201	0.0010885	0.0011457	0.9501409
3000(3%n)	0.0011112	0.0041497	0.2677879	0.0011090	0.0008217	1.3495989
5000(5%n)	0.0010887	0.0038210	0.2849246	0.0010871	0.0006551	1.6594412

Table 4: The average time (in seconds) taken to calculate the subsampling estimator $\hat{\boldsymbol{\beta}}^*(\underline{\boldsymbol{\pi}})$ for the penalized P-spline SIM. The full-sample estimator $\hat{\boldsymbol{\beta}}$ and $\underline{\boldsymbol{\pi}}$ (Step 1) took $3,357.25$s and 46.51s respectively, using Dataset 1.

r	100(0.1%n)	300(0.3%n)	500(0.5%n)	1000(1%n)	3000(3%n)	5000(5%n)
Time_{opt}	4.16990	4.34612	4.65214	4.86108	5.66108	8.76270
$\text{Time}_{\text{unif}}$	3.24740	3.37662	3.40784	5.02662	8.02662	14.72636

7 REAL DATA APPLICATIONS

7.1 The Video Transcoding

Video content is being produced, transported and consumed in more ways and devices than ever. Meanwhile, a seamless interaction is required between video content producing, transporting and consuming devices. The difference in device resources, network bandwidth and video representation types result in the necessary requirements for a mechanism for video content adoption. One such mechanism is called *video transcoding*. It is a process of converting one compressed video representation to another. The basic idea of video transcoding is to convert unsupported video formats into supported ones. Unsupported videos include videos that are not playable by a given device due to lack of format support or those that require relatively higher system resources than the device can offer. Currently, transcoding is being utilized for such purposes as bit-rate reduction in order to meet network bandwidth availability, resolution reduction for display size adoption, temporal transcoding for frame rate reduction, and error resilience transcoding for insuring high quality of service.

Runtime scheduling of transcoding jobs in multicore and cloud environments is hard as their resource requirements may not be known before hand, thus the prediction of the transcoding time based on the input and output video features is in demand. Consider the Youtube video transcoding time dataset from the UCI machine learning repository (*https://archive.ics.uci.edu/ml/datasets.php*). It has $n = 67,875$ observations and the features include bitrate, framerate, resolution, codec, number of i frames, and so on, which are treated as predictors, $X_1, X_2, \ldots, X_{19}$, respectively. Thus the dimension of the covariate vector is $p = 19$. The response variable is *the total transcoding time*.

We fit the data with the penalized B-spline and P-spline SIM. We compute the EMSE and bias of the subsampling estimator given in (6.3) – (6.4) and compare the Scoring A-optimal subsampling to the uniform via the ratios, $\text{EMSE}_{\text{ratio}}$ and $\text{Bias}_{\text{ratio}}$. Again, the Scoring Algorithm 3 is applied. We also compare the amount of time needed for the subsampling estimator $\hat{\boldsymbol{\beta}}^*$ with that for the full sample estimator $\hat{\boldsymbol{\beta}}$. The results are reported in Tables 5–8. Observe that the EMSE ratios are consistently smaller than 1, indicating that the A-optimal subsampling outperformed the uniform under both model settings and for all subsample sizes. For example, when $r \geq 2000(3\%n)$, the EMSE ratios are around 0.2 for the B-spline SIM, which was a substantial improvement. In addition, the Subsampling approach saved significant amount of time for all the subsmaple sizes considered.

Table 5: The EMSE and biases of the A-optimal and the uniform subsampling estimators $\hat{\boldsymbol{\beta}}^*(\boldsymbol{\pi})$ and their ratios for the penalized B-spline SIM with $d = 14$, $p = 19$ and $n = 67,875$, using the video transcoding dataset.

r	EMSE_{opt}	$\text{EMSE}_{\text{unif}}$	$\text{EMSE}_{\text{ratio}}$	Bias_{opt}	$\text{Bias}_{\text{unif}}$	$\text{Bias}_{\text{ratio}}$
200(0.3%n)	1.19111	1.79193	0.66471	0.43577	0.82717	0.52682
350(0.5%n)	1.07367	1.78433	0.60172	0.39044	0.81281	0.48036
680(1%n)	0.69675	1.66869	0.41755	0.31616	0.70462	0.44870
2000(3%n)	0.37747	1.72069	0.21937	0.32746	0.74311	0.44066
3400(5%n)	0.35149	1.64874	0.21319	0.33651	0.68150	0.49377
6800(10%n)	0.34296	1.55243	0.22092	0.33650	0.60342	0.55765
20000(30%n)	0.34001	1.22424	0.27773	0.33906	0.37494	0.90432

Table 6: The average time (in seconds) taken to calculate the subsampling estimator $\hat{\boldsymbol{\beta}}^*(\underline{\boldsymbol{\pi}})$ for the penalized B-spline SIM. The full-sample estimator $\hat{\boldsymbol{\beta}}$ and $\underline{\boldsymbol{\pi}}$ (Step 1) took $1,029.09$s and 46.51s respectively, using the video transcoding dataset.

r	200(.3%n)	350(.5%n)	680(1%n)	2000(3%n)	3400(5%n)	6800(10%n)	20000(30n%)
Time_{opt}	14.62300	4.09928	2.23788	1.28136	1.39596	2.26600	13.22626
$\text{Time}_{\text{unif}}$	2.69424	1.18566	1.07390	1.74482	2.54884	5.17372	16.17538

Table 7: The EMSE and biases of the A-optimal and the uniform subsampling estimators $\hat{\boldsymbol{\beta}}^*(\boldsymbol{\pi})$ and their ratios for the P-spline SIM with $d = 14$, $p = 19$ and $n = 67,875$, using the video transcoding dataset.

r	EMSE_{opt}	$\text{EMSE}_{\text{unif}}$	$\text{EMSE}_{\text{ratio}}$	Bias_{opt}	$\text{Bias}_{\text{unif}}$	$\text{Bias}_{\text{ratio}}$
200(0.3%n)	1.52647	1.64479	0.92806	0.63165	0.70591	0.89480
350(0.5%n)	1.39318	1.95854	0.71133	0.52222	0.97817	0.53387
680(1%n)	1.35133	1.75405	0.77041	0.47780	0.77783	0.61427
2000(3%n)	0.93755	1.69648	0.55265	0.22975	0.72307	0.31774
3400(5%n)	0.71443	1.69594	0.42126	0.13802	0.72093	0.19145
6800(10%n)	0.44914	1.46438	0.30671	0.06163	0.53695	0.11478
20000(30%n)	0.08919	1.48001	0.06026	0.01671	0.54788	0.03051

Table 8: The average time (in seconds) taken to calculate the subsampling estimator $\hat{\boldsymbol{\beta}}^*(\underline{\boldsymbol{\pi}})$ for the penalized P-spline SIM. The full-sample estimator $\hat{\boldsymbol{\beta}}$ and $\underline{\boldsymbol{\pi}}$ (Step 1) took $1,029.75$s and 11.18s respectively, using the video transcoding dataset.

r	200(0.3n%)	350(0.5n%)	680(1n%)	2000(3n%)	3400(5n%)	6800(10n%)	20000(30n%)
Time_{opt}	0.19664	0.22602	0.80252	1.13454	1.44596	2.43032	10.39942
$\text{Time}_{\text{unif}}$	0.19864	0.23694	1.01206	1.57142	2.21236	4.16576	18.06122

7.2 The Gas Sensor

Here we apply the Subsampling approach to the gas sensor array dataset from chemistry (*https://archive.ics.uci.edu/ml/datasets.php*). The dataset was collected by exposing $p = 16$ chemical sensors to a gas mixture of Ethylene and CO in air at varying concentration levels. For each gas mixture, the signals were recorded from the sensors. We exclude all the negative readings from each sensors and drop the first 20, 000 data points which correspond to the system run-in time. After the cleaning, there are totally $n = 1,605,003$ observations. The objective is to predict the concentration of enthylene with the 16 sensors readings as covariates. Note that the sensor readings are rescaled with factor 0.001. Due to the memory limitation of desktop computers, we used the super computer (Big Red II at Indiana University) to handle the full data estimation of the SIM fittings, then compare the optimal subsampling with the uniform. The A-optimal Scoring Algorithm 3 is used. In the first step, the subsample size is taken to be $r_0 = 800(.05\%n)$, while in the second step, the subsample size r ranges from $160(.01\%n)$ to $800(.05\%n)$. Due to the memory storage issue of the big data, we only take small subsample sizes. The results are reported in Tables 9 - 10 for the penalized B-spline SIM, and Tables 11 – 12 for the penalized P-spline SIM. For both models, the values of $\text{EMSE}_{\text{ratio}}$ were consistently less than 1, showing the better performance of the optimal subsampling over the uniform even when the subsample size was only at most 0.05 percent of the full sample. The Subsampling approach also saved significant amount of time for all cases considered.

Table 9: The EMSE and biases of the A-optimal and the uniform subsampling estimators $\hat{\boldsymbol{\beta}}^*(\boldsymbol{\pi})$ and their ratios for the penalized B-spline SIM with $d = 14$, $p = 16$ and $n = 1,605,003$, using the gas sensor dataset.

r	EMSE_{opt}	$\text{EMSE}_{\text{unif}}$	$\text{EMSE}_{\text{ratio}}$	Bias_{opt}	$\text{Bias}_{\text{unif}}$	$\text{Bias}_{\text{ratio}}$
160(0.01%n)	1.3914	1.4606	0.9526	0.4874	0.5690	0.8566
320(0.02%n)	0.7157	0.8887	0.8054	0.1425	0.2269	0.6281
480(0.03%n)	0.4875	0.6391	0.7628	0.0935	0.1213	0.7707
640(0.04%n)	0.2548	0.2971	0.8576	0.0840	0.0518	1.6213
800(0.05%n)	0.1640	0.2058	0.7968	0.1159	0.0932	1.2439

Table 10: The average time (in seconds) taken to calculate the subsampling estimator $\hat{\boldsymbol{\beta}}^*(\underline{\boldsymbol{\pi}})$ for the penalized B-spline SIM. The full-sample $\hat{\boldsymbol{\beta}}$ and $\underline{\boldsymbol{\pi}}$ (first step) took $128,798$s and 645.614s respectively, using the gas sensor dataset.

r	0.01%n	0.02%n	0.03%n	0.04%n	0.05%n
Time_{opt}	98.8942	148.4920	212.0116	351.6188	927.4393
$\text{Time}_{\text{unif}}$	77.4167	134.1387	166.5673	294.6663	697.4677

Table 11: The EMSE and biases of the A-optimal and the uniform subsampling estimators $\hat{\boldsymbol{\beta}}^*(\boldsymbol{\pi})$ and their ratios for the penalized P-spline SIM with $d = 14$, $p = 16$ and $n = 1,605,003$, using the gas sensor dataset.

r	EMSE_{opt}	$\text{EMSE}_{\text{unif}}$	$\text{EMSE}_{\text{ratio}}$	Bias_{opt}	$\text{Bias}_{\text{unif}}$	$\text{Bias}_{\text{ratio}}$
0.01%n	2.0226	2.2017	0.9186	1.0252	1.3364	0.7671
0.02%n	2.0918	2.5296	0.8269	1.1012	1.6506	0.6672
0.03%n	2.0877	2.6886	0.7765	1.1011	1.8544	0.5938
0.04%n	2.0646	2.7964	0.7383	1.0734	2.0118	0.5335
0.05%n	2.1808	2.6073	0.8364	1.2003	1.7441	0.6882

Table 12: The average time (in seconds) taken to calculate the subsampling estimator $\hat{\boldsymbol{\beta}}^*(\underline{\boldsymbol{\pi}})$ for the penalized P-spline SIM. The full-sample estimator $\hat{\boldsymbol{\beta}}$ and $\underline{\boldsymbol{\pi}}$ (Step 1) took $107,523.9$s and 319.458s respectively, using the gas sensor dataset.

r	0.01%n	0.02%n	0.03%n	0.04%n	0.05%n
Time_{opt}	64.1687	170.3558	263.0385	538.8632	1673.112
$\text{Time}_{\text{unif}}$	35.9975	133.9499	201.5794	364.5358	1024.964

8 (Ã1)–(Ã5) and Proof of Theorem 1

Ã1 $\tilde{\boldsymbol{\Sigma}}_n$ satisfies (A1).

Ã2 $\tilde{\mathbf{H}}_n$ and $\tilde{\lambda}_n =: \tilde{\lambda}(\boldsymbol{\pi})$ satisfy (A2).

Ã3 With $\mathrm{g}_i(\boldsymbol{\theta}) = -2e_i(\boldsymbol{\theta})f_i'; (\boldsymbol{\theta}) = \boldsymbol{\phi}_i(\boldsymbol{\theta}) - \lambda P'(\boldsymbol{\theta})$,

$$\sum_{i=1}^{n} \pi_i \|\tilde{\boldsymbol{\phi}}_i\|^2 = O_P((p+d)\tilde{\lambda}_n), \quad \sum_{i=1}^{n} \pi_i \|\dot{\mathbf{g}}_i(\tilde{\boldsymbol{\theta}})\|^2 = O_P((p+d)^{-1} r \tilde{\lambda}_n^2).$$

Ã4 $\eta_1, ..., \eta_n$ and λ, h introduced in (A4) satisfy

$$\sum_{i=1}^{n} \pi_i \eta_i^2 + \lambda^2 h^2 = o_P((p+d)^{-2} r \tilde{\lambda}_n^3).$$

Ã5 The array $z_{ni} = \tilde{s}_n^{-1}(\boldsymbol{\pi})\mathbf{u}\tilde{\mathbf{H}}_n^{-1}(\boldsymbol{\pi})\tilde{\boldsymbol{\phi}}_i : 1 \leq i \leq n, n \geq 1$ with $\tilde{s}_n^2 = \tilde{s}_n^2(\boldsymbol{\pi}) = \mathbf{u}^t \tilde{\mathbf{H}}_n^{-1}(\boldsymbol{\pi}) \tilde{\boldsymbol{\Sigma}}_n(\boldsymbol{\pi}) \tilde{\mathbf{H}}_n^{-t}(\boldsymbol{\pi})\mathbf{u}$ satisfies (A5).

For sequences $\{\mathbf{X}_n\}, \{\mathbf{Y}_n\}$ of rv, recall $\mathbf{Y}_n = o_P(\mathbf{X}_n)$ if $\|\mathbf{Y}_n\|/\|\mathbf{X}_n\|$ converges to zero in probability, and $\mathbf{Y}_n = O_P(\mathbf{X}_n)$ if $\|\mathbf{Y}_n\|/\|\mathbf{X}_n\|$ is bounded in probability. We shall write P^* for the conditional probability given $(\mathbf{X}, \mathbf{y})$, and define o_{P^*} and O_{P^*} similarly. Let $\lambda_{\min}(\boldsymbol{A})$ ($\lambda_{\max}(\boldsymbol{A})$) be the minimum (maximum) eigenvalue of matrix $\boldsymbol{A}$. The euclidean norm $\|\boldsymbol{A}\|$ and the operator (or spectral) norm $|\boldsymbol{A}|_o$ of matrix $\boldsymbol{A}$ are defined by

$$\|\boldsymbol{A}\|^2 = \mathrm{Tr}(\boldsymbol{A}^t\boldsymbol{A}) = \sum_{i,j} A_{ij}^2, \quad |\boldsymbol{A}|_o = \sup_{\|\boldsymbol{u}\|=1} |\boldsymbol{A}\boldsymbol{u}| = \sup_{\|\boldsymbol{u}\|=1} (\boldsymbol{u}^t\boldsymbol{A}^t\boldsymbol{A}\boldsymbol{u})^{1/2}.$$

For $f: \mathbb{R}^p \mapsto \mathbb{R}$, define $\dot{f}(\mathbf{x}) = \partial f(\mathbf{x})/\partial \mathbf{x}^t$ ($f'(\mathbf{x}) = \partial f(\mathbf{x})/\partial \mathbf{x}$) to be a row (column) vector. More generally, for $\mathbf{f}: \mathbb{R}^p \mapsto \mathbb{R}^q$ define $\dot{\mathbf{f}}(\mathbf{x}) = \partial \mathbf{f}(\mathbf{x})/\partial \mathbf{x}^t$ to be a $p \times q$ matrix.

We need to quote Theorem 6.3.4 of Ortega and Rheinboldt (1970) below for the proof. For a given set C, its closure and boundary are, respectively, denoted by $\bar{C}$ and ∂C.

Lemma 5 *Let C be an open, bounded set in $\mathbb{R}^n$ and assume that $\mathbf{F}: \bar{C} \subset \mathbb{R}^n \to \mathbf{R}^n$ is continuous and satisfies $(\mathbf{x} - \mathbf{x}_0)^t\mathbf{F}(\mathbf{x}) \geq 0$ for some $\mathbf{x}_0 \in C$ and all $\mathbf{x} \in \partial C$. Then $F(\mathbf{x}) = 0$ has a solution in $\bar{C}$.*

Let $\mathbf{p} = (p_1, \ldots, p_n)^t$ be a vector with $p_i \geq 0$. A random vector $\mathbf{w} = (w_1, \ldots, w_n)^t$ has a *scaled mutilnomial distribution* with parameters $r, \mathbf{p}, \boldsymbol{\pi}$, written $\mathbf{w} \sim \mathrm{smultn}(r, \mathbf{p}, \boldsymbol{\pi})$, if it has the probability mass function,

$$\mathrm{P}\Big(w_1 = \frac{k_1}{rp_1}, \ldots, w_n = \frac{k_n}{rp_n}\Big) = \frac{r!}{\prod_{i=1}^n k_i!} \prod_{i=1}^n \pi_i^{k_i}, \qquad k_i \geq 0, \quad \sum_{i=1}^n k_i = r. \tag{8.1}$$

Two cases of interest are $\mathbf{p} = \boldsymbol{\pi}$ and $\mathbf{p} = \mathbf{1}$, written $\mathbf{w} \sim \mathrm{smultn}(r, \boldsymbol{\pi}, \boldsymbol{\pi})$ and $\boldsymbol{\omega} \sim \mathrm{smultn}(r, \mathbf{1}, \boldsymbol{\pi})$, respectively. The former is utilized in the study of the (weighted) subsampling estimator and the latter in the unweighted estimator.

Using $\mathbf{w} \sim \mathrm{smultn}(r, \boldsymbol{\pi}, \boldsymbol{\pi})$, we can express $Q_n^*(\boldsymbol{\theta}_{\boldsymbol{\phi}})$ in terms of the full observations $(\mathbf{X}, \mathbf{y})$, and obtain a useful *stochastically equivalent* representation,

$$Q_n^*(\boldsymbol{\theta}_{\boldsymbol{\phi}}) = \frac{1}{n}\sum_{i=1}^n w_i\big(y_i - \boldsymbol{\delta}^t\mathbf{B}(\boldsymbol{\beta}(\boldsymbol{\phi})^t\mathbf{x}_i)\big)^2 + \lambda P(\boldsymbol{\theta}_{\boldsymbol{\phi}}). \tag{8.2}$$

Recall $\hat{\lambda}_n = \lambda_{\max}(\hat{\boldsymbol{\Sigma}}_n)$ in (A3), $\boldsymbol{\theta} = \boldsymbol{\theta}_{\boldsymbol{\phi}}$ and $\boldsymbol{\Psi}_r^*(\boldsymbol{\theta})$ in (3.16).

Proof of Theorem 1. Let $\mathbf{t}_n = q^{1/2}r^{-1/2}\hat{\lambda}_n^{-1/2}\mathbf{t}$ with $q = p + d$ for $\mathbf{t} \in \mathbb{R}^{p+d-1}$, and let

$$\mathbf{T}^*(\mathbf{t}) = q^{-1/2}r^{1/2}\hat{\lambda}_n^{-1/2}\big(\boldsymbol{\Psi}_r^*(\hat{\boldsymbol{\theta}} + \mathbf{t}_n) - \boldsymbol{\Psi}_r^*(\hat{\boldsymbol{\theta}})\big) - \hat{\lambda}_n^{-1}\mathbf{H}_n(\hat{\boldsymbol{\theta}})\mathbf{t}. \tag{8.3}$$

For an arbitrary constant $c > 0$, fix $\|\boldsymbol{t}\| \leq c$. By assumption, $\mathbf{t}_n = o_P(1)$. Recalling $\mathbf{H}_n(\hat{\boldsymbol{\theta}}) = n^{-1}\sum_{i=1}^n \dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}}) + \lambda\ddot{P}(\hat{\boldsymbol{\theta}})$, Taylor's theorem implies that there exists $\mathbf{t}_n^* = q^{1/2}r^{-1/2}\hat{\lambda}_n^{-1/2}\mathbf{t}^*$ with $\|\boldsymbol{t}^*\| \leq c$ such that

$$\begin{aligned}
\mathbf{T}^*(\mathbf{t}) &= q^{-1/2}r^{1/2}\hat{\lambda}_n^{-1/2}n^{-1}\sum_{i=1}^n \big(w_i\dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}} + \mathbf{t}_n^*) + \lambda\ddot{P}(\hat{\boldsymbol{\theta}} + \mathbf{t}_n^*)\big)\mathbf{t}_n - \hat{\lambda}_n^{-1}\mathbf{H}_n(\hat{\boldsymbol{\theta}})\mathbf{t} \\
&= (n\hat{\lambda}_n)^{-1}\Big(\sum_{i=1}^n \big(w_i\dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}} + \mathbf{t}_n^*) + n\lambda\ddot{P}(\hat{\boldsymbol{\theta}} + \mathbf{t}_n^*)\big) - \sum_{i=1}^n \dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}})\mathbf{t} - n\lambda\ddot{P}(\hat{\boldsymbol{\theta}})\Big)\mathbf{t} \\
&= (n\hat{\lambda}_n)^{-1}\Big(\sum_{i=1}^n \bar{w}_i\dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}}) + \sum_{i=1}^n w_i\big(\dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}} + \mathbf{t}_n^*) - \dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}})\big) + n\lambda\big(\ddot{P}(\hat{\boldsymbol{\theta}} + \mathbf{t}_n^*) - \ddot{P}(\hat{\boldsymbol{\theta}})\big)\Big)\mathbf{t},
\end{aligned}$$

where $\bar{w}_i = w_i - \mathrm{E}(w_i) = w_i - 1$. Therefore, by (A4), we have for large r and with large probability,

$$\|\mathbf{T}^*(\mathbf{t})\|^2 \leq 3c^2(n\hat{\lambda}_n)^{-2}\Big(\|\bar{\mathbf{H}}^*(\hat{\boldsymbol{\theta}})\|^2 + qr^{-1}\hat{\lambda}_n^{-1}\big(\big(\sum_{i=1}^n w_i\eta_i\big)^2 + n^2\lambda^2h^2\big)\Big), \tag{8.4}$$

where $\bar{\mathbf{H}}^*(\boldsymbol{\theta}) = \sum_{i=1}^n \bar{w}_i \dot{\mathbf{g}}_i(\boldsymbol{\theta})$. Using (3.6), it is not difficult to calculate

$$\begin{aligned}
r\mathrm{E}^*(\|\bar{\mathbf{H}}^*(\hat{\boldsymbol{\theta}})\|^2) &= r\mathrm{Tr}\Big(\mathrm{E}^*(\sum_{i=1}^n\sum_{j=1}^n \bar{w}_i\bar{w}_j \dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}})\dot{\mathbf{g}}_j(\hat{\boldsymbol{\theta}})^t\Big)\\
&= r\mathrm{Tr}\Big(\sum_{i=1}^n \mathrm{E}^*(\bar{w}_i^2)\dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}})\dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}})^t + \sum_{i\neq j}\mathrm{E}^*(\bar{w}_i\bar{w}_j)\dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}})\dot{\mathbf{g}}_j(\hat{\boldsymbol{\theta}})^t\Big)\\
&= r\mathrm{Tr}\Big(\sum_{i=1}^n \frac{(1-\pi_i)}{r\pi_i}\dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}})\dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}})^t - \sum_{i\neq j}\frac{1}{r}\dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}})\dot{\mathbf{g}}_j(\hat{\boldsymbol{\theta}})^t\Big)\\
&\le \sum_{i=1}^n \frac{\|\dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}})\|^2}{\pi_i} =: A_n,
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}^*\Big(\sum_{i=1}^n w_i\eta_i\Big)^2 = \mathrm{E}^*\Big(\sum_{i=1}^n \bar{w}_i\eta_i + \sum_i \eta_i\Big)^2 &\le 2\sum_{i=1}^n \frac{1}{r\pi_i}\eta_i^2 + 2\Big(\sum_{i=1}^n \eta_i\Big)^2\\
&\le 2\sum_{i=1}^n (n + (r\pi_i)^{-1})\eta_i^2 =: B_n.
\end{aligned}$$

Consequently, by the second equality in (A3) and (A4), we have

$$\mathrm{E}^*\Big(\sup_{\|\mathbf{t}\|\le c}\|\mathbf{T}^*(\mathbf{t})\|^2\Big) \le 3c^2r^{-1}(n\hat{\lambda}_n)^{-2}\big(A_n + q\hat{\lambda}_n^{-1}(B_n + n^2\lambda^2h^2)\big) = o_P(q^{-1}). \tag{8.5}$$

It thus follows from (8.8) that

$$\begin{aligned}
\ell^*(c) =: &\inf_{\|\mathbf{t}\|=c}\Big\{q^{-1/2}r^{1/2}\hat{\lambda}_n^{-1/2}\mathbf{t}^t\boldsymbol{\Psi}_r^*(\hat{\boldsymbol{\theta}}+\mathbf{t}_n)\Big\}\\
&\ge c^2\hat{\lambda}_n^{-1}\lambda_{\min}(\hat{\mathbf{H}}_n) - c\sup_{\|\mathbf{t}\|=c}\|\mathbf{T}^*(\mathbf{t})\| - cq^{-1/2}r^{1/2}\hat{\lambda}_n^{-1/2}\|\boldsymbol{\Psi}_r^*(\hat{\boldsymbol{\theta}})\|.
\end{aligned}$$

By (A2), $\hat{\lambda}_n^{-1}\lambda_{\min}(\hat{\mathbf{H}}_n) \ge b_0 > 0$ with large probability for large n. Fix an arbitrary $K > 0$. Using Markov's inequality and (3.7), we obtain

$$\begin{aligned}
\mathrm{P}^*(q^{-1/2}r^{1/2}\hat{\lambda}_n^{-1/2}\|\hat{\boldsymbol{\Psi}}_r^*\| > K) &\le q^{-1}r\hat{\lambda}_n^{-1}K^{-2}\mathrm{Tr}\big(\mathrm{Var}^*(\hat{\boldsymbol{\Psi}}_r^*)\big)\\
&= q^{-1}r\hat{\lambda}_n^{-1}n^{-2}K^{-2}\sum_{i=1}^n\frac{1}{r\pi_i}\|\hat{\boldsymbol{\phi}}_i\|^2 = O_P(K^{-2}),
\end{aligned}$$

where the last equality follows from the first equality in (A3). This and (8.10)-(8.11) imply that for large $K = c$,

$$\begin{aligned}
\mathrm{P}^*(\ell^*(c) > 0) &\ge 1 - \mathrm{P}^*\big(\sup_{\|\mathbf{t}\|=c}\|\mathbf{T}^*(\mathbf{t})\| > b_0c/2\big) - \mathrm{P}^*\big(q^{-1/2}r^{1/2}\hat{\lambda}_n^{-1/2}\|\hat{\boldsymbol{\Psi}}_r^*\| > b_0c/2\big)\\
&= 1 - o_P(1).
\end{aligned}$$

Therefore, by the continuity of $\boldsymbol{\Psi}_r^*(\boldsymbol{\theta})$ on Θ and Lemma 5, there exists $\mathbf{t}_*$ with $\|\mathbf{t}_*\| \le c$ such that

$$\boldsymbol{\Psi}_r^*(\hat{\boldsymbol{\theta}} + q^{1/2}r^{-1/2}\hat{\lambda}_n^{-1/2}\mathbf{t}_*) = 0.$$

Let $\hat{\boldsymbol{\theta}}^* = \hat{\boldsymbol{\theta}} + q^{1/2}r^{-1/2}\hat{\lambda}_n^{-1/2}\mathbf{t}_*$. Then $\hat{\boldsymbol{\theta}}^*$ minimizes (3.1) and satisfies

$$\mathrm{P}^*(\|q^{-1/2}r^{1/2}\hat{\lambda}_n^{1/2}(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})\| \le c) \ge 1 - o_P(1).$$

This shows (3.21). By (8.10), we also have

$$\mathbf{T}^*(\mathbf{t}_*) = o_{P^*}(q^{-1/2}). \tag{8.6}$$

This and (8.8) yield the desired (3.10). Noting $\boldsymbol{\Psi}_r^*(\hat{\boldsymbol{\theta}}^*) = 0$, we have

$$\mathbf{T}^*(\mathbf{t}_*) = -q^{-1/2}r^{1/2}\hat{\lambda}_n^{-1/2}\boldsymbol{\Psi}_r^*(\hat{\boldsymbol{\theta}}) - \hat{\lambda}_n^{-1/2}\hat{\mathbf{H}}_nq^{-1/2}r^{1/2}(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}).$$

Consequently, for any unit vector $\mathbf{u}$, we derive

$$\begin{aligned} s_n^{-1} r^{1/2} \mathbf{u}^t (\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}) &= -s_n^{-1} \mathbf{u}^t \hat{\mathbf{H}}_n^{-1} r^{1/2} \boldsymbol{\Psi}_r^*(\hat{\boldsymbol{\theta}}) - s_n^{-1} q^{1/2} \hat{\lambda}_n^{1/2} \mathbf{u}^t \hat{\mathbf{H}}_n^{-1} \mathbf{T}^*(\mathbf{t}_*) \\ &= -s_n^{-1} \mathbf{u}^t \hat{\mathbf{H}}_n^{-1} r^{1/2} \boldsymbol{\Psi}_r^*(\hat{\boldsymbol{\theta}}) + o_{P^*}(1), \end{aligned} \tag{8.7}$$

where we used $q^{1/2}\mathbf{T}^*(\mathbf{t}_*) = o_{P^*}(1)$ with the help of the inequality,

$$\frac{\hat{\lambda}_n^{1/2} \| \mathbf{u}^t \hat{\mathbf{H}}_n^{-1} \|}{s_n(\mathbf{u})} \leq \frac{\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_n) \| \mathbf{u}^t \hat{\mathbf{H}}_n^{-1} \|}{\lambda_{\min}^{1/2}(\hat{\boldsymbol{\Sigma}}_n) \| \mathbf{u}^t \hat{\mathbf{H}}_n^{-1} \|} = \frac{\lambda_{\max}^{1/2}(\hat{\boldsymbol{\Sigma}}_n)}{\lambda_{\min}^{1/2}(\hat{\boldsymbol{\Sigma}}_n)} \leq B,$$

where B is a constant from (A1). The asymptotic normality in (3.11) now follows from (8.12), (A5), and Lindeberg-Feller's theorem (e.g. Theorem 7.2.1 of Chung, 2001). Specifically, (A5) implies that it holds in probability that the first term on the last line in (8.12) has an asymptotic standard normal distribution given the data. □

Proof of Theorem 2. Let $\mathbf{t}_n = q^{1/2} r^{-1/2} (\hat{\lambda}_n^u)^{-1/2} \mathbf{t}$ with $q = p + d$ for $\mathbf{t} \in \mathbb{R}^{p+d-1}$, and let

$$\mathbf{T}^*(\mathbf{t}) = q^{-1/2} r^{1/2} (\hat{\lambda}_n^u)^{-1/2} \big(\boldsymbol{\Psi}_r^{u*}(\hat{\boldsymbol{\theta}}^u + \mathbf{t}_n) - \boldsymbol{\Psi}_r^{u*}(\hat{\boldsymbol{\theta}}^u) \big) - \hat{\lambda}_n^{-u} \mathbf{H}_n^u(\hat{\boldsymbol{\theta}})^u \mathbf{t}. \tag{8.8}$$

For notional brevity, write $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^u$, $\hat{\boldsymbol{\theta}}^* = \hat{\boldsymbol{\theta}}^{u*}$, $w_i = \omega_i$, $\boldsymbol{\Psi}_r^*(\boldsymbol{\theta}) = \boldsymbol{\Psi}_r^{u*}(\boldsymbol{\theta})$, $\mathbf{H}_n(\boldsymbol{\theta}) = \mathbf{H}_n^u(\boldsymbol{\theta})$, etc. Given an arbitrary constant $c > 0$, fix $\|\boldsymbol{t}\| \leq c$. By assumption, $\mathbf{t}_n = o_P(1)$. Recalling $\mathbf{H}_n(\hat{\boldsymbol{\theta}}) = n^{-1} \sum_{i=1}^n \pi_i \dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}}) + \lambda \ddot{P}(\hat{\boldsymbol{\theta}})$, Taylor's theorem implies that there exists $\mathbf{t}_n^* = q^{1/2} r^{-1/2} \hat{\lambda}_n^{-1/2} \mathbf{t}^*$ with $\|\boldsymbol{t}^*\| \leq c$ such that

$$\begin{aligned} \mathbf{T}^*(\mathbf{t}) &= q^{-1/2} r^{1/2} \hat{\lambda}_n^{-1/2} \sum_{i=1}^n \big(w_i \dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}} + \mathbf{t}_n^*) + \lambda \ddot{P}(\hat{\boldsymbol{\theta}} + \mathbf{t}_n^*) \big) \mathbf{t}_n - \hat{\lambda}_n^{-1} \mathbf{H}_n(\hat{\boldsymbol{\theta}}) \mathbf{t} \\ &= \hat{\lambda}_n^{-1} \Big(\sum_{i=1}^n \big(w_i \dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}} + \mathbf{t}_n^*) + \lambda \ddot{P}(\hat{\boldsymbol{\theta}} + \mathbf{t}_n^*) \big) \mathbf{t} - \sum_{i=1}^n \pi_i \dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}}) \mathbf{t} - \lambda \ddot{P}(\hat{\boldsymbol{\theta}}) \mathbf{t} \Big) \\ &= \hat{\lambda}_n^{-1} \Big(\sum_{i=1}^n \bar{w}_i \dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}}) + \sum_{i=1}^n w_i \big(\dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}} + \mathbf{t}_n^*) - \dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}}) \big) + \lambda \big(\ddot{P}(\hat{\boldsymbol{\theta}} + \mathbf{t}_n^*) - \ddot{P}(\hat{\boldsymbol{\theta}}) \big) \Big) \mathbf{t}, \end{aligned}$$

where $\bar{w}_i = w_i - \mathrm{E}(w_i) = w_i - \pi_i$. Therefore, by ($\tilde{\mathbf{A}}$4), we have for large r and with large probability,

$$\|\mathbf{T}^*(\mathbf{t})\|^2 \leq 3c^2 \hat{\lambda}_n^{-2} \Big(\|\bar{\mathbf{H}}^*(\hat{\boldsymbol{\theta}})\|^2 + q r^{-1} \hat{\lambda}_n^{-1} \big(\big(\sum_{i=1}^n w_i \eta_i \big)^2 + \lambda^2 h^2 \big) \Big), \tag{8.9}$$

where $\bar{\mathbf{H}}^*(\boldsymbol{\theta}) = \sum_{i=1}^n \bar{w}_i \dot{\mathbf{g}}_i(\boldsymbol{\theta})$. Using (3.19), it is not difficult to calculate

$$\begin{aligned} r \mathrm{E}^*(\|\bar{\mathbf{H}}^*(\hat{\boldsymbol{\theta}})\|^2) &= r \mathrm{Tr} \Big(\mathrm{E}^* \big(\sum_{i=1}^n \sum_{j=1}^n \bar{w}_i \bar{w}_j \dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}}) \dot{\mathbf{g}}_j(\hat{\boldsymbol{\theta}})^t \Big) \\ &= r \mathrm{Tr} \Big(\sum_{i=1}^n \mathrm{E}^*(\bar{w}_i^2) \dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}}) \dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}})^t + \sum_{i \neq j} \mathrm{E}^*(\bar{w}_i \bar{w}_j) \dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}}) \dot{\mathbf{g}}_j(\hat{\boldsymbol{\theta}})^t \Big) \\ &= \sum_{i=1}^n \pi_i (1 - \pi_i) \|\dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}})\|^2 - \mathrm{Tr} \Big(\sum_{i \neq j} \pi_i \pi_j \dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}}) \dot{\mathbf{g}}_j(\hat{\boldsymbol{\theta}})^t \Big) \\ &\leq \sum_{i=1}^n \pi_i \|\dot{\mathbf{g}}_i(\hat{\boldsymbol{\theta}})\|^2 =: A_n, \end{aligned}$$

$$\begin{aligned} \mathrm{E}^* \Big(\sum_{i=1}^n w_i \eta_i \Big)^2 &= \mathrm{E}^* \Big(\sum_{i=1}^n \bar{w}_i \eta_i + \sum_i \pi_i \eta_i \Big)^2 \leq 2 \sum_{i=1}^n \pi_i \eta_i^2 / r + 2 \Big(\sum_{i=1}^n \pi_i \eta_i \Big)^2 \\ &\leq 4 \sum_{i=1}^n \pi_i \eta_i^2 =: B_n. \end{aligned}$$

Consequently, by the second equality in ($\tilde{\mathbf{A}}$3) and ($\tilde{\mathbf{A}}$4), we have

$$\mathrm{E}^* \Big(\sup_{\|\boldsymbol{t}\| \leq c} \|\mathbf{T}^*(\mathbf{t})\|^2 \Big) \leq 3c^2 r^{-1} \hat{\lambda}_n^{-2} \big(A_n + q \hat{\lambda}_n^{-1} (B_n + \lambda^2 h^2) \big) = o_P(q^{-1}). \tag{8.10}$$

It thus follows from (8.8) that

$$\begin{aligned}\ell^*(c) =: & \inf_{\|\mathbf{t}\|=c} \left\{ q^{-1/2} r^{1/2} \hat{\lambda}_n^{-1/2} \mathbf{t}^t \mathbf{\Psi}_r^*(\hat{\boldsymbol{\theta}} + \mathbf{t}_n) \right\} \\ & \geq c^2 \hat{\lambda}_n^{-1} \lambda_{\min}(\hat{\mathbf{H}}_n) - c \sup_{\|\mathbf{t}\|=c} \|\mathbf{T}^*(\mathbf{t})\| - c q^{-1/2} r^{1/2} \hat{\lambda}_n^{-1/2} \|\mathbf{\Psi}_r^*(\hat{\boldsymbol{\theta}})\|.\end{aligned}$$

By ($\tilde{\mathbf{A}}2$), $\hat{\lambda}_n^{-1} \lambda_{\min}(\hat{\mathbf{H}}_n) \geq b_0 > 0$ with large probability for large n. Fix an arbitrary $K > 0$. Using Markov's inequality and (3.7), we obtain

$$\begin{aligned}\mathrm{P}^*(q^{-1/2} r^{1/2} \hat{\lambda}_n^{-1/2} \|\hat{\mathbf{\Psi}}_r^*\| > K) & \leq q^{-1} r \hat{\lambda}_n^{-1} K^{-2} \mathrm{Tr}\big(\mathrm{Var}^*(\hat{\mathbf{\Psi}}_r^*)\big) \\ & = q^{-1} r \hat{\lambda}_n^{-1} K^{-2} \sum_{i=1}^n \frac{\pi_i}{r} \|\hat{\boldsymbol{\phi}}_i\|^2 = O_P(K^{-2}),\end{aligned}$$

where the last equality follows from the first equality in ($\tilde{\mathbf{A}}3$). This and (8.10)-(8.11) imply that for large $K = c$,

$$\begin{aligned}\mathrm{P}^*(\ell^*(c) > 0) & \geq 1 - \mathrm{P}^*\big(\sup_{\|\mathbf{t}\|=c} \|\mathbf{T}^*(\mathbf{t})\| > b_0 c/2 \big) - \mathrm{P}^*\big(q^{-1/2} r^{1/2} \hat{\lambda}_n^{-1/2} \|\hat{\mathbf{\Psi}}_r^*\| > b_0 c/2 \big) \\ & = 1 - o_P(1).\end{aligned}$$

Therefore, by the continuity of $\mathbf{\Psi}_r^*(\boldsymbol{\theta})$ on Θ and Lemma 5, there exists $\mathbf{t}_*$ with $\|\mathbf{t}_*\| \leq c$ such that

$$\mathbf{\Psi}_r^*(\hat{\boldsymbol{\theta}} + q^{1/2} r^{-1/2} \hat{\lambda}_n^{-1/2} \mathbf{t}_*) = 0.$$

Let $\hat{\boldsymbol{\theta}}^* = \hat{\boldsymbol{\theta}} + q^{1/2} r^{-1/2} \hat{\lambda}_n^{-1/2} \mathbf{t}_*$. Then $\hat{\boldsymbol{\theta}}^*$ minimizes (3.1) and satisfies

$$\mathrm{P}^*(\|q^{-1/2} r^{1/2} \hat{\lambda}_n^{1/2} (\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})\| \leq c) \geq 1 - o_P(1).$$

This shows (3.21). By (8.10), we also have

$$\mathbf{T}^*(\mathbf{t}_*) = o_{P^*}(q^{-1/2}). \tag{8.11}$$

This and (8.8) yield the desired (3.10). Noting $\mathbf{\Psi}_r^*(\hat{\boldsymbol{\theta}}^*) = 0$, we have

$$\mathbf{T}^*(\mathbf{t}_*) = -q^{-1/2} r^{1/2} \hat{\lambda}_n^{-1/2} \mathbf{\Psi}_r^*(\hat{\boldsymbol{\theta}}) - \hat{\lambda}_n^{-1/2} \hat{\mathbf{H}}_n q^{-1/2} r^{1/2} (\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}).$$

Consequently, for any unit vector $\mathbf{u}$, we derive

$$\begin{aligned}s_n^{-1} r^{1/2} \mathbf{u}^t (\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}) & = -s_n^{-1} \mathbf{u}^t \hat{\mathbf{H}}_n^{-1} r^{1/2} \mathbf{\Psi}_r^*(\hat{\boldsymbol{\theta}}) - s_n^{-1} q^{1/2} \hat{\lambda}_n^{1/2} \mathbf{u}^t \hat{\mathbf{H}}_n^{-1} \mathbf{T}^*(\mathbf{t}_*) \\ & = -s_n^{-1} \mathbf{u}^t \hat{\mathbf{H}}_n^{-1} r^{1/2} \mathbf{\Psi}_r^*(\hat{\boldsymbol{\theta}}) + o_{P^*}(1),\end{aligned} \tag{8.12}$$

where we used $q^{1/2} \mathbf{T}^*(\mathbf{t}_*) = o_{P^*}(1)$ with the help of the inequality,

$$\frac{\hat{\lambda}_n^{1/2} \|\mathbf{u}^t \hat{\mathbf{H}}_n^{-1}\|}{s_n(\mathbf{u})} \leq \frac{\lambda_{\max}^{1/2}(\mathbf{\Sigma}_n) \|\mathbf{u}^t \hat{\mathbf{H}}_n^{-1}\|}{\lambda_{\min}^{1/2}(\mathbf{\Sigma}_n) \|\mathbf{u}^t \hat{\mathbf{H}}_n^{-1}\|} = \frac{\lambda_{\max}^{1/2}(\mathbf{\Sigma}_n)}{\lambda_{\min}^{1/2}(\mathbf{\Sigma}_n)} \leq B,$$

where B is a constant from ($\tilde{\mathbf{A}}1$). The asymptotic normality in (3.11) now follows from (8.12), ($\tilde{\mathbf{A}}5$), and Lindeberg-Feller's theorem (e.g. Theorem 7.2.1 of Chung, 2001). Specifically, ($\tilde{\mathbf{A}}5$) implies that it holds in probability that the first term on the last line in (8.12) has an asymptotic standard normal distribution given the data. □

9 Supplementary Material

The section contains the tables of the simulation results using Datasets 2 and 3.

Table 13: Same as Table 1 but using Dataset 2: The simulated EMSE and biases of the A-optimal and the uniform subsampling estimators $\hat{\boldsymbol{\beta}}^*(\boldsymbol{\pi})$ and their ratios for the penalized B-spline SIM with $d = 14$, $p = 12$, and $n = 100,000$.

r	EMSE_{opt}	$\text{EMSE}_{\text{unif}}$	$\text{EMSE}_{\text{ratio}}$	Bias_{opt}	$\text{Bias}_{\text{unif}}$	$\text{Bias}_{\text{ratio}}$
100(0.1%n)	0.7437	0.8841	0.8412	0.2171	0.2273	0.9552
300(0.3%n)	0.3003	0.4929	0.6092	0.0389	0.0847	0.4598
500(0.5%n)	0.1363	0.3776	0.3611	0.0099	0.0615	0.1616
1000(1%n)	0.0555	0.2511	0.2212	0.0061	0.0460	0.1315
3000(3%n)	0.0231	0.1333	0.1730	0.0061	0.0386	0.1571
5000(5%n)	0.0315	0.1221	0.2577	0.0070	0.0364	0.1938

Table 14: Same as Table 2 but using Dataset 2: The average time (in seconds) taken to calculate the subsampling estimator $\hat{\boldsymbol{\beta}}^*(\boldsymbol{\pi})$ for the penalized B-spline SIM. The full-sample $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\pi}$ (Step 1) took $1,577.28$s and 45.43s.

r	100(0.1%n)	300(0.3%n)	500(0.5%n)	1000(1%n)	3000(3%n)	5000(5%n)
Time_{opt}	7.2712	8.4250	9.0337	10.6185	17.1332	27.9052
$\text{Time}_{\text{unif}}$	7.9470	8.3083	8.6701	11.3170	21.2636	31.9889

Table 15: Same as Table 1 but using Dataset 3: The simulated EMSE and biases of the A-optimal and the uniform subsampling estimators $\hat{\boldsymbol{\beta}}^*(\boldsymbol{\pi})$ and their ratios for the B-spline SIM with $d = 14$, $p = 12$, and $n = 100,000$.

r	EMSE_{opt}	$\text{EMSE}_{\text{unif}}$	$\text{EMSE}_{\text{ratio}}$	Bias_{opt}	$\text{Bias}_{\text{unif}}$	$\text{Bias}_{\text{ratio}}$
100(0.1%n)	1.1369	1.1717	0.9703	0.4036	0.7062	0.5715
300(0.3%n)	0.7980	1.0016	0.7967	0.2128	0.7016	0.3033
500(0.5%n)	0.5631	1.0015	0.5622	0.1541	0.7050	0.2186
1000(1%n)	0.4542	0.9873	0.4600	0.1413	0.7172	0.1970
3000(3%n)	0.3491	0.9230	0.3782	0.1367	0.7496	0.1824
5000(5%n)	0.3373	0.8728	0.3865	0.1217	0.6829	0.1781

Table 16: Same as Table 2 but using Dataset 3: The average time (in seconds) taken to calculate the subsampling estimator $\hat{\boldsymbol{\beta}}^*(\boldsymbol{\pi})$ for the penalized B-spline SIM. The full-sample estimator $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\pi}$ (Step 1) took $1,782.5$s and 46.03s.

r	100(0.1%n)	300(0.3%n)	500(0.5%n)	1000(1%n)	3000(3%n)	5000(5%n)
Time_{opt}	11.1171	7.5731	6.1142	9.2911	26.8115	20.9743
$\text{Time}_{\text{unif}}$	6.9309	7.8669	9.2347	14.0700	45.8354	34.5437

Table 17: Same as Table 3 but using Dataset 2: The simulated EMSE and biases of the A-optimal and the uniform subsampling estimators $\hat{\boldsymbol{\beta}}^*(\boldsymbol{\pi})$ and their ratios for the penalized P-spline SIM with $d = 14$, $p = 12$ and $n = 100,000$.

r	EMSE_{opt}	$\text{EMSE}_{\text{unif}}$	$\text{EMSE}_{\text{ratio}}$	Bias_{opt}	$\text{Bias}_{\text{unif}}$	$\text{Bias}_{\text{ratio}}$
100(0.1%n)	1.28833	1.28740	1.00072	0.44064	0.43423	1.01477
300(0.3%n)	0.77724	0.96087	0.80889	0.20310	0.27307	0.74374
500(0.5%n)	0.66962	0.79219	0.84528	0.22648	0.20932	1.08198
1000(1%n)	0.49564	0.62578	0.79203	0.20184	0.17232	1.17131
3000(3%n)	0.22898	0.43788	0.52293	0.07403	0.11969	0.61848
5000(5%n)	0.15672	0.38733	0.40462	0.05543	0.10607	0.52257

Table 18: Same as Table 4 but using Dataset 2: The average time (in seconds) taken to calculate the subsampling estimator $\hat{\boldsymbol{\beta}}^*(\boldsymbol{\pi})$ for the penalized P-spline SIM. The full-sample estimator $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\pi}$ (Step 1) took $3,477.2$s and 17.71s.

r	100(0.1%n)	300(0.3%n)	500(0.5%n)	1000(1%n)	3000(3%n)	5000(5%n)
Time_{opt}	2.35708	3.66070	5.17740	10.57100	57.00426	73.81938
$\text{Time}_{\text{unif}}$	1.99488	3.38116	4.71838	8.18670	35.08972	43.59866

Table 19: Same as Table 3 but using Dataset 3: The simulated EMSE and biases of the A-optimal and the uniform subsampling estimators $\hat{\boldsymbol{\beta}}^*(\boldsymbol{\pi})$ and their ratios for the penalized P-spline SIM with $d = 14$, $p = 12$ and $n = 100,000$.

r	EMSE_{opt}	$\text{EMSE}_{\text{unif}}$	$\text{EMSE}_{\text{ratio}}$	Bias_{opt}	$\text{Bias}_{\text{unif}}$	$\text{Bias}_{\text{ratio}}$
100(0.1%n)	0.93231	1.80099	0.51767	0.42049	0.87001	0.48332
300(0.3%n)	0.64008	1.82378	0.35096	0.41384	0.88681	0.46666
500(0.5%n)	0.62072	1.82251	0.34058	0.42452	0.88969	0.47715
1000(1%n)	0.56786	1.87309	0.30317	0.43791	0.94684	0.46250
3000(3%n)	0.54606	1.85944	0.29367	0.41962	0.93361	0.44946
5000(5%n)	0.53593	1.94236	0.27591	0.41756	1.01757	0.41035

Table 20: Same as Table 4 but using Dataset 3: The average time (in seconds) taken to calculate the subsampling estimator $\hat{\boldsymbol{\beta}}^*(\underline{\boldsymbol{\pi}})$ for the penalized P-spline SIM. The full-sample estimator $\hat{\boldsymbol{\beta}}$ and $\underline{\boldsymbol{\pi}}$ (Step 1) took $4,4891.23$s and 39.83s.

r	100(0.1%n)	300(0.3%n)	500(0.5%n)	1000(1%n)	3000(3%n)	5000(5%n)
Time_{opt}	1.21686	1.84940	2.54674	4.06008	10.37228	17.91362
$\text{Time}_{\text{unif}}$	1.34630	2.62212	4.22754	7.58898	19.86076	34.76936

References

[1] Antoniadis, A., Gregoire G. and McKeague, I. W. (2004). Bayesian estimation in single-index models. *Statist. Sinica* **14**(4): 1147-1164.

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

[3] Bühlmann, P., Drineas, P., Kane, M. and van der Laan, M. (2016). *Handbook of Big Data*. CRC Press, Taylor & Francis Group, Boca Raton, FL 33487-2742.

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

[5] Chung, K. L. (2001). *A Course in Probability Theory*. Academic Press, New York.

[6] Dennis, J.E. Jr. and Schnabel, R.B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, New Jersey.

[7] Eilers, P.H.C. and Marx, B.D. (1992). Generalized linear models with P-splines. In: Proceedings of GLIM 92 and 7th International Workshop on Statistical Modelling, Munich, Germany. Lecture Notes in Statistics, Vol. 78, Advances in GLIM and Statistical Modelling, Eds. L. Fahrmeir, B. Francis, R. Gilchrist, G. Tutz. SpringerVerlag, New York, 72-77.

[8] Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing using B-splines and penalized likelihood (with Comments and Rejoinder). *Statist. Sci.* **11**(2): 89-121.

[9] Eilers, P.H.C, Marx, B.D. and Durban, M. (2015). Twenty years of P-splines. *Statistics and Operations Research Transactions* **39**(2): 1-38.

[10] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**(456): 1348-1360.

[11] Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21**(1): 157-178.

[12] Hristache, M., Juditsky, A., Polzehl, J. and Spokoiny V. (2001). Structure adaptive approach for dimension reduction. *Ann. Statist.* **29**(6): 1537-1566.

[13] Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econom.* **58**(1-2): 71-120.

[14] Jiang, R., Meng-Fan Guo, M-F and Liu, X. (2022). Composite quasi-likelihood for single-index models with massive datasets. *Commun. Stat. Simul. Comput.* **51**(9): 5024-5040. DOI: 10.1080/03610918.2020.1753074.

[15] Jiang, R. and Peng, Y. (2023). A short note on fitting a singleindex model with massive data. *Stat. Theory Relat. Fields* **7**(1): 49-60. DOI: 10.1080/24754269.2022.2135807

[16] Y. Le Cun (1987). Modeles Connexionnistes de l'Apprentissage. PhD thesis, Universite Pierre et Marie Curie, Paris, France.

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

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

[19] Mammen, E. (1993). Bootstrap and wild bootstrap for high-dimensional linear models. *Ann. Statist.* **21**: 255–285.

[20] O'Sullivan, F. (1986). A Statistical Perspective on Ill-Posed Inverse Problems (with Discussion). *Statist. Sci.* **1**: 505-527.

[21] O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal of Scientic and Statistical Computation* **9**: 363-379.

[22] Ortega, J. M. and Rheinboldt, W. G. (1970). *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York.

[23] PENG, H. (2024). Biases Of Z-Estimators Of Parameters In General Estimating Equations For Both Fixed And Growing Dimension. Available at `http://math.iupui.edu/~feitan/`

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

[25] PORTNOY, S. (1985). Asymptotic Behavior of M Estimators of p Regression Parameters when p^2/n is Large; II. Normal Approximation *Ann. Statist.* **13** (4): 1403-1417.

[26] PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16**: 356-366.

[27] Ruppert, D. (2002). Selecting the number of knots for penalized splines. *J. Amer. Statist. Assoc.* **11**(4): 735-757.

[28] Ruppert, D. and Carroll, R. (1997). Penalized Regression Splines. working paper, Cornell University, School of Operations Research and Industrial Engineering. (available at www.orie.cornell.edu/ davidr/papers).

[29] Ruppert, D. and Carroll, R. (2000). Spatially adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics* **42**: 205-223.

[30] Ruppert D., Wand M. and Carroll R. (2003). *Semiparametric Regression*. Cambridge University Press, New York.

[31] Sharif S. and Kamal S. (2018). Comparison of significant approaches of penalized spline regression (P-splines). *Pakistan Journal of Statistics and Operation Research* **14**(2): 310-315.

[32] Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica* **54**: 1461-1481.

[33] Tan, F., Zhao, X. and Peng, H. (2023). A-Optimal Subsampling Approach To The Analysis of Count Data of Massive Size. Available at `https://math.iupui.edu/~feitan/TZP_subs_glm23.pdf`.

[34] Wang, G. and Wang, L. (2015). Spline estimation and variable selection for single index prediction models with diverging number of index parameters. *J. Statist. Plann. Infer.* **162**: 1-19.

[35] Wang, H., Zhu, R. and Ma, P. (2018). Optimal subsampling for large sample logistic regression. *J. Amer. Statist. Assoc.* **113**(522): 829-844.

[36] Wu, J. and Tu, W. (2016). A multivariate single-index model for longitudinal data. *Statistical Modelling* **16**(5): 392-408.

[37] Wu, J., Peng, H. and Tu, W. (2019). Large-sample estimation and inference in multivariate single-index models. *J. Multivar. Anal.* **171**: 382-396. PMID 31588153 and DOI: 10.1016/J.Jmva.2019.01.003

[38] Xia, Y. C. and Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models. *J. Multivar. Anal.* **97**: 1162-1184.

[39] Xia, Y. C., Tong, H., LI, W. K. and Zhu, L. (2002). An adaptive estimation of dimension reduction space (with discussion). *J. R. Stat. Soc. Series B Stat. Methodol.* **64**: 363-410.

[40] Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *J. Amer. Statist. Assoc.* **97**(460): 1042-1054.

[41] Yu, Y., Wu, C. and Zhang, Y. (2017). Penalised spline estimation for generalised partially linear single-index models. *Statistics and Computing* **27**: 571-582.

[42] Zhang, S., Tan, F. and Peng, H. (2023). Sample Size Determination For Multidimensional Parameters And Optimal Subsampling In A Big Data Linear Regression Model. Available at `http://math.iupui.edu/~feitan/ZTP_SSD-23(002).pdf`

[43] Zhu, R., Ma, P., Mahoney, M.W. and Yu, B. (2015). Optimal sub-sampling approaches for large sample linear regression. arXiv: 1509.0511. v1[stat.ME].