Scale-invariant Optimal Sampling for Rare-events Data with Sparse Models

Jing Wang
Department of Statistics
University of Connecticut
Storrs, CT 06269
jing.7.wang@uconn.edu

HaiYing Wang
Department of Statistics
University of Connecticut
Storrs, CT 06269
haiying.wang@uconn.edu

Hao Helen Zhang
Department of Mathematics
University of Arizona
hzhang@math.arizona.edu

Abstract

Subsampling is effective in tackling computational challenges for massive data with rare events. Overly aggressive subsampling may adversely affect estimation efficiency, and optimal subsampling is essential to mitigate the information loss. However, existing optimal subsampling probabilities depends on data scales, and some scaling transformations may result in inefficient subsamples. This problem is more significant when there are inactive features, because their influence on the subsampling probabilities can be arbitrarily magnified by inappropriate scaling transformations. We tackle this challenge and introduce a scale-invariant optimal subsampling function in the context of sparse models, where inactive features are commonly assumed. Instead of focusing on estimating model parameters, we define an optimal subsampling function to minimize the prediction error, using adaptive lasso as an example to outline the estimation procedure and study its theoretical guarantee. We first introduce the adaptive lasso estimator for rare-events data and establish its oracle properties, thereby validating the use of subsampling. Then we derive a scale-invariant optimal subsampling function that minimizes the prediction error of the inverse probability weighted (IPW) adaptive lasso. Finally, we present an estimator based on the maximum sampled conditional likelihood (MSCL) to further improve the estimation efficiency. We conduct numerical experiments using both simulated and real-world data sets to demonstrate the performance of the proposed methods.

1 Introduction

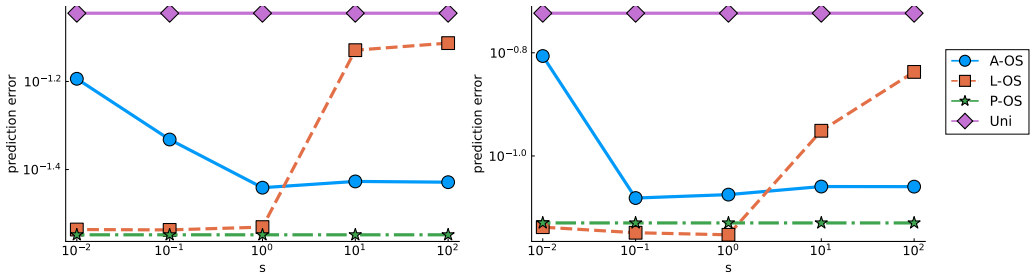
Rare-events data refer to binary-response data that are highly imbalanced, i.e., the number of zeros (a.k.a "controls" or "negative instances") are possibly hundreds or thousands of times as large as the number of ones (a.k.a. "cases" or "positive instances"). This type of data is common in various fields, such as medicine, natural science, political science, and social science, where examples of rare events can be rare diseases, natural disasters, wars, and financial crises, respectively. Modern technologies also prompt us to pay more attention to rare-events data. For example, in modern online recommendation systems, clicks are usually rare events compared with nonclicks. Statistical analyses, including parameter estimation and inferences, pose unique challenges for rare-events data because of high imbalance. In addition, rare-events data often involve sparse models. For instance,

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-3122

rare diseases might be linked to a limited number of key genes. Therefore, researchers frequently adopt sparse models in genome-wide association studies for analyzing rare diseases. A different yet related example is the use of deep neural networks to predict click-through rates in modern online recommendation systems. These networks are typically overparameterized, necessitating methods that balance rare-events data with the sparsity of the underlying models. Data balancing is a popular approach to overcome challenges caused by imbalanced data and is usually accomplished through subsampling the zeros [5, 15] or oversampling the ones [3, 12, 16, 4]. In addition, rare-events data are often massive in order to obtain an adequate number of ones, and computation is demanding. Therefore, we focus on the subsampling approach since it addresses the imbalance issue and reduce the computational burden simultaneously.

It is shown in [20] that the efficiency of parameter estimation is essentially determined by the number of ones for rare-events logistic regression, and subsampling does not reduce the estimation efficiency as long as sufficient zeros are kept. In case of excessive removal of zeros, [22] developed an optimal sampling approach to minimize information loss. However, the optimal sampling probabilities in [22] are scale-dependent, which may lead to inefficient results. Figure 1 illustrates the issue using a simulated example, with details in Section D.1 of the appendix. We generate the data from the same logistic regression model and transform one of the covariates with different scales $s = 0.01, 0.1, 1, 10,$ and 100. Then we apply two optimal subsampling methods in [22], labeled with "A-OS" and "L-OS" in Figure 1. It is observed that the prediction errors of A-OS and L-OS are significantly impacted by the data scaling. The A-OS may perform similarly to the Uni (simple random sampling or uniform sampling) in Figure 1a when $s = 0.01$; so is the L-OS in Figure 1b when $s = 100$. This scale-dependent issue is not specific to logistic regression and rare-events data in [22]; it is a wide concern in literature for various data types and models, including but not limited to [1, 29, 21, 14, 26, 25, 24]. In this paper, we propose a scale-invariant optimal subsampling method to overcome the issue. It is labeled "P-OS" in Figure 1.



(a) Non-sparse parameter

(b) Sparse parameter

Figure 1: Prediction errors with different scale transformation of the same model. (a): with non-sparse parameter $(-1, -1, -0.01, -0.01, -0.01, -0.01)^{\mathrm{T}}$. (b): with sparse parameter $(-1, 0, 0, 0, 0, 0)^{\mathrm{T}}$.

The scale-dependence issue can seriously impact variable selection results for sparse models, where true parameters are zero for inactive covariates. In this case, inactive variables may be arbitrarily transformed without changing the underlying model, but the A-OS or L-OS would be highly influenced and may lead to misleading results. To resolve this issue, we investigate scale-invariant optimal subsampling in the context of variable selection, for which one main goal is to distinguish active and inactive features.

Penalty-based feature selection methods are widely used. Specifically, the adaptive lasso is a popular choice due to its oracle properties, convexity, and practical ease of implementation [see 30, 28]. While penalization methods have been used for bias reduction in rare-events analysis [7], variable selection for rare-events data has not been investigated. Conducting effective variable selection is difficult in the context of rare-events data analysis, mainly due to the scarcity of information available for ones. An inaccurate variable selection result can subsequently impact both the effectiveness of optimal subsampling and the efficiency of parameter estimation. In this paper, we address the challenge of variable selection in the context of rare-events data. First, we propose the full data adaptive lasso and study its theoretical properties. Next, we introduce a novel subsampling estimator that seamlessly combines penalty-based variable selection and optimal sampling into one unified framework for rare-events data. The implementation of the adaptive lasso requires a pilot estimator to construct

data-dependent weights for covariates. Given that optimal sampling also relies on pilot estimates [see 23, 1], the adaptive lasso emerges as a natural choice for conducting variable selection method in the context of subsampled rare-events data. We validate the new estimators by proving their oracle properties and also develop an efficient algorithm to facilitate their practical implementation when handling massive real-world data sets. In summary, our main contributions are listed as follows:

- We propose scale-invariant optimal subsampling to enhance parameter estimation and variable selection. Existing optimal subsampling methods are scale-dependent, which may lead to unreliable or misleading results.
- We define adaptive lasso and establish its oracle properties for rare-events data, which show that the asymptotic variances are determined by the number of ones in the data and the active features in the model.
- We present a practical subsampling algorithm based on optimal probabilities that significantly reduces the computational burden and accelerates the optimization for penalty-based feature selection methods.

The rest of the paper is organized as follows. Section 2 introduces the model setup. Section 3 investigates nonuniform sampling and variable selection tailored for rare-events data. We propose new methods to construct scale-invariant optimal probabilities. Section 4 discusses theoretical properties of the MSCL estimator and presents a two-step algorithm to implement the proposed methods. Section 5 conducts numerical experiments on simulated and real data sets. Section 6 concludes the paper. Proofs and mathematical details are presented in the appendix.

2 Background and model setup

We use the subscript $_{\mathrm{t}}$ to indicate the true parameters. For a p-dimensional vector $\boldsymbol{x}$, we use $x_{(i)}$ to represent its i-th element. For an index subset $\mathcal{A} \subset \{i : 1, 2, ..., p\}$, we use $\boldsymbol{x}_{(\mathcal{A})}$ to denote the subvector of $\boldsymbol{x}$, whose elements correspond to the indexes in $\mathcal{A}$. Furthermore, we use $\boldsymbol{x}^{\otimes 2}$ to denote $\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}$, use "$\rightsquigarrow$" to denote convergence in distribution, use "$\xrightarrow{P}$" to denote convergence in probability, and use "$\xrightarrow{a.s.}$" to denote convergence almost surely. We use $\boldsymbol{I}$ to denote an identity matrix of a suitable dimension and use $\mathbf{0}$ to denote a vector of zeros of a suitable dimension.

Let $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_N, y_N)$ denote N sample points from the joint distribution of $(\boldsymbol{x}, y)$, where $\{\boldsymbol{x}_i\}_{i=1}^N$ denote the p-dimensional predictors and $\{y_i\}_{i=1}^N$ the binary responses. Assume that the probability of y being a one ($y = 1$) given $\boldsymbol{x}$ is

$$p(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}}) := \mathbb{P}(y = 1|\boldsymbol{x}) = \frac{e^{\alpha_{\mathrm{t}} + f(\boldsymbol{x}; \boldsymbol{\beta}_{\mathrm{t}})}}{1 + e^{\alpha_{\mathrm{t}} + f(\boldsymbol{x}; \boldsymbol{\beta}_{\mathrm{t}})}} = \frac{e^{g(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}})}}{1 + e^{g(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}})}},$$

where $\boldsymbol{\theta}_{\mathrm{t}} = (\alpha_{\mathrm{t}}, \boldsymbol{\beta}_{\mathrm{t}}^{\mathrm{T}})^{\mathrm{T}}$ is the vector of true parameters and $f(\boldsymbol{x}; \boldsymbol{\beta}_t)$ is a smooth function of $\boldsymbol{\beta}_t$. For rare-events data, $N_1 \ll N_0$, where $N_1 = \sum_{i=1}^N y_i$ is the number of ones (i.e. $y_i = 1$) and $N_0 = N - N_1$ is the number of zeros (i.e. $y_i = 0$). Following the model setup used in [22], we assume that $\alpha_{\mathrm{t}} \to -\infty$ as $N \to \infty$, which implies that, under appropriate moment conditions,

$$\frac{N_1}{N_0} = \frac{\mathbb{E}\{p(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}})\}}{1 - \mathbb{E}\{p(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}})\}} + o(1) = \mathbb{E}\{e^{\alpha_{\mathrm{t}} + f(\boldsymbol{x}; \boldsymbol{\beta}_{\mathrm{t}})}\} + o(1) \to 0, \text{ almost surely.} \tag{1}$$

Under this assumption, the asymptotic variance of the full data maximum likelihood estimator (MLE) is of order $1/N_1$ instead of $1/N$, indicating that the estimation efficiency is determined by the number of rare ones. Therefore, we can keep all the ones and sample the zeros to save computational costs. There could be a variance inflation due to aggressive subsampling, and [22] developed optimal subsampling functions to reduce the variance inflation. Specifically, the authors proposed non-uniform optimal sampling functions under the A- and L-optimality criteria, respectively, as follows: $\varphi_{\mathrm{A-OS}}^{\mathrm{scale}}(\boldsymbol{x}) \propto p(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}}) \|\boldsymbol{M}^{-1} \dot{g}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}})\|$ and $\varphi_{\mathrm{L-OS}}^{\mathrm{scale}}(\boldsymbol{x}) \propto p(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}}) \|\dot{g}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}})\|$, where $\boldsymbol{M} = \mathbb{E}\{e^{f(\boldsymbol{x}; \boldsymbol{\beta}_{\mathrm{t}})} \dot{g}^{\otimes 2}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}})\}$ and $\dot{g}(\boldsymbol{x}; \boldsymbol{\theta})$ denotes the derivative of $g(\boldsymbol{x}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. However, the sampling functions $\varphi_{\mathrm{A-OS}}^{\mathrm{scale}}(\boldsymbol{x})$ and $\varphi_{\mathrm{L-OS}}^{\mathrm{scale}}(\boldsymbol{x})$ proposed in [22] depend on the scale of $\boldsymbol{x}$, and may not perform well for certain measurement scale of $\boldsymbol{x}$. For example, if $g(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}}) = \alpha_{\mathrm{t}} + \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\theta}_{\mathrm{t}}$, then $\varphi_{\mathrm{L-OS}}^{\mathrm{scale}}(\boldsymbol{x})$ is proportional to $1 + \|\boldsymbol{x}\|$, which will be influenced by the scale of $\boldsymbol{x}$. Similarly, scale

https://doi.org/10.52202/079017-3122

changes in $\boldsymbol{x}$ may also change $\varphi^{\text{scale}}_{\text{A-OS}}(\boldsymbol{x})$, although the impact may not be in the same direction, as demonstrated in Figure 1. Besides parameter estimation, variable selection is another important topic, which has not been studied in the literature on rare-events data. This work aims to fill this gap.

3 Nonuniform sampling with variable selection for rare-events data

The adaptive lasso [30, 28] is a popular variable selection method because it has oracle properties and is easy to implement. We define the full data adaptive lasso for rare-events data as

$$\hat{\boldsymbol{\theta}}^{\text{adp}}_{\text{mle}} := \arg\max_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{N} [y_i g(\boldsymbol{x}_i;\boldsymbol{\theta}) - \log\{1+e^{g(\boldsymbol{x}_i;\boldsymbol{\theta})}\}] - \lambda_N \sum_{j=1}^{p} \frac{|\beta_{(j)}|}{|\hat{\beta}_{\text{pl}(j)}|^{\gamma}} \right\}, \tag{2}$$

where λ_N and γ are tuning parameters, and $\hat{\boldsymbol{\beta}}_{\text{pl}}$ is a consistent pilot estimator of $\boldsymbol{\beta}_t$. In practice, it is common to set $\gamma = 1$. In the literature, iterative algorithms such as coordinate descent are commonly used to solve the adaptive lasso [8]. However, their computational demand can become prohibitive when dealing with massive data. It is feasible to alleviate the computational burden by subsampling zeros and create a smaller subset of data for adaptive lasso. To be specific, consider Algorithm 1.

Algorithm 1 Poisson Subsampling algorithm

For $i = 1, ..., N$:
if $y_i = 1$ **then**
  include $(\boldsymbol{x}_i, y_i)$ in the subsample;
else
  Compute $\varphi(\boldsymbol{x}_i)$ and generate $u_i \sim U[0,1]$;
  if $u_i \leq \pi(\boldsymbol{x}_i, y_i)$ **then**
    include $(\boldsymbol{x}_i, y_i)$ and record $\rho\varphi(\boldsymbol{x}_i)$ in the subsample;
  end if
end if

The inclusion probability in Algorithm 1 for the ith observation is $\pi(\boldsymbol{x}_i, y_i) = y_i + (1-y_i)\rho\varphi(\boldsymbol{x}_i)$, where ρ is the baseline sampling rate for the zeros and $\varphi(\boldsymbol{x}) > 0$ satisfies $\mathbb{E}\{\varphi(\boldsymbol{x})\} = 1$. Let the subsample from Algorithm 1 be $\{\boldsymbol{x}^{\text{sub}}_i, y^{\text{sub}}_i\}_{i=1}^{N^*_{\text{sub}}}$, which is biased since $\pi(\boldsymbol{x}_i, y_i)$'s depend on the responses. We introduce an Inverse Probability Weighting (IPW) adaptive lasso estimator to correct for the bias, defined as

$$\hat{\boldsymbol{\theta}}^{\text{adp}}_{\text{w}} := \arg\max_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{N^*_{\text{sub}}} \frac{[y^{\text{sub}}_i g(\boldsymbol{x}^{\text{sub}}_i;\boldsymbol{\theta}) - \log\{1+e^{g(\boldsymbol{x}^{\text{sub}}_i;\boldsymbol{\theta})}\}]}{\pi(\boldsymbol{x}^{\text{sub}}_i, y^{\text{sub}}_i)} - \lambda_N \sum_{j=1}^{p} \frac{|\beta_{(j)}|}{|\hat{\beta}_{\text{pl}(j)}|^{\gamma}} \right\}. \tag{3}$$

To save space, we put the general assumptions used throughout this paper in Section B.1 of the appendix. We use $\mathcal{A}$ to denote the set of indexes of active variables, i.e., $\mathcal{A} = \{j : \beta_{\text{t}(j)} \neq 0\}$ and $\mathcal{A}^c$ to denote the set of indexes of inactive variables, i.e., $\mathcal{A}^c = \{j : \beta_{\text{t}(j)} = 0\}$. We first study the asymptotic properties of $\hat{\boldsymbol{\theta}}^{\text{adp}}_{\text{w}}$ in the following theorem.

Theorem 1. *Let $\hat{\boldsymbol{\beta}}_{\text{pl}}$ be a consistent pilot estimate such that $\lambda_N/(\sqrt{N_1}|\hat{\beta}_{\text{pl}(j)}|^{\gamma}) \xrightarrow{P} \infty$ for $j \in \mathcal{A}^c$. Under Assumptions 1-4, if $\lambda_N/\sqrt{N_1} \to 0$, then the IPW adaptive lasso estimator defined in* (3) *has the following properties:*

1. *Consistency in variable selection: The estimated active set $\hat{\mathcal{A}}_{\text{w}} := \{j : \hat{\beta}^{\text{adp}}_{\text{w}(j)} \neq 0\}$ satisfies that $\lim_{N\to\infty} \mathbb{P}(\hat{\mathcal{A}}_{\text{w}} = \mathcal{A}) = 1$.*
2. *Asymptotic normality: The estimator of the active parameter vector satisfies that*

$$\sqrt{N_1}\boldsymbol{V}^{-1/2}_{\text{w}(\mathcal{A})}(\hat{\boldsymbol{\theta}}^{\text{adp}}_{\text{w}(\mathcal{A})} - \boldsymbol{\theta}_{\text{t}(\mathcal{A})}) \rightsquigarrow \mathbb{N}(\boldsymbol{0}, \boldsymbol{I}),$$

where $\boldsymbol{V}_{\text{w}(\mathcal{A})} = \mathbb{E}\left\{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\text{t}})}\right\}\boldsymbol{M}^{-1}_{(\mathcal{A})}\boldsymbol{M}_{\text{w}(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})} = \mathbb{E}\left\{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\text{t}})}\right\}\left\{\boldsymbol{M}^{-1}_{(\mathcal{A})} + c\boldsymbol{V}_{\text{sub}(\mathcal{A})}\right\}$, $\boldsymbol{M}_{(\mathcal{A})} = \mathbb{E}\left\{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\text{t}})}\dot{g}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\text{t}})\right\}$, $\boldsymbol{V}_{\text{sub}(\mathcal{A})} = \boldsymbol{M}^{-1}_{(\mathcal{A})}\mathbb{E}\left\{\frac{e^{2f(\boldsymbol{x};\boldsymbol{\beta}_{\text{t}})}}{\varphi(\boldsymbol{x})}\dot{g}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\text{t}})\right\}\boldsymbol{M}^{-1}_{(\mathcal{A})}$, $c =$

$\lim_{N\to\infty} e^{\alpha_t}/\rho$, *and* $\dot{g}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_t)$ *consists of the elements of gradient vector* $\dot{g}(\boldsymbol{x};\boldsymbol{\theta}_t)$ *with indexes in the active set* $\mathcal{A}$.

Remark 1. *Theorem 1 shows that the estimation efficiency of* $\hat{\boldsymbol{\theta}}^{\mathrm{adp}}_{\mathrm{w}(\mathcal{A})}$ *is predominantly determined by the number of ones instead of the full data size. The term* $c\boldsymbol{V}_{\mathrm{sub}(\mathcal{A})}$ *is the variation inflation due to subsampling. The full data adaptive lasso in* (2) *correspond to the scenario with* $\rho = 1$ *and* $\varphi(\boldsymbol{x}) = 1$*, for which* $c = \lim_{N\to\infty} e^{\alpha_t}/\rho = 0$*. Intuitively,* c *can be interpreted as the imbalance rate in the subsample. If we include sufficient zeros (*$c = 0$*), the subsampling does not reduce the estimation efficiency of* $\hat{\boldsymbol{\theta}}^{\mathrm{adp}}_{\mathrm{w}(\mathcal{A})}$.

From Theorem 1, we see that there maybe information loss reflected as an inflated variance if $c \neq 0$. To minimize the information loss due to sampling, we derive optimal functions as follows, where $\varphi^{\mathrm{adp}}_{\mathrm{A-OS}}(\boldsymbol{x})$ corresponds to the A-optimality criterion [17] and $\varphi^{\mathrm{adp}}_{\mathrm{L-OS}}(\boldsymbol{x})$ corresponds to the L-optimality criterion [17] in design of experiments. Here, the A-optimality minimizes the trace of the asymptotic variance of $\hat{\boldsymbol{\theta}}^{\mathrm{adp}}_{\mathrm{w}(\mathcal{A})}$; the L-optimality focuses on the asymptotic variance of a linearly transformed estimator $\boldsymbol{M}_{(\mathcal{A})}\hat{\boldsymbol{\theta}}^{\mathrm{adp}}_{\mathrm{w}(\mathcal{A})}$, which is proportional to $\boldsymbol{M}_{\mathrm{w}(\mathcal{A})}$. The A-optimality criterion has a more direct interpretation, while an advantage of the L-optimality criterion is that the resulting optimal function is often faster to calculate.

Proposition 1. *The A-optimal function that minimizes tr*$(\boldsymbol{V}_{\mathrm{w}(\mathcal{A})})$ *is*

$$\varphi^{\mathrm{adp}}_{\mathrm{A-OS}}(\boldsymbol{x}) = \frac{p(\boldsymbol{x};\boldsymbol{\theta}_t)\|\boldsymbol{M}^{-1}_{(\mathcal{A})}\dot{g}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_t)\|}{\mathbb{E}\left\{p(\boldsymbol{x};\boldsymbol{\theta}_t)\|\boldsymbol{M}^{-1}_{(\mathcal{A})}\dot{g}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_t)\right\}}. \tag{4}$$

The L-optimal function that minimizes tr$(\boldsymbol{M}_{\mathrm{w}(\mathcal{A})})$ *is*

$$\varphi^{\mathrm{adp}}_{\mathrm{L-OS}}(\boldsymbol{x}) = \frac{p(\boldsymbol{x};\boldsymbol{\theta}_t)\|\dot{g}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_t)\|}{\mathbb{E}\left\{p(\boldsymbol{x};\boldsymbol{\theta}_t)\|\dot{g}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_t)\right\}}. \tag{5}$$

Unlike the optimal sampling function in [22], $\varphi^{\mathrm{adp}}_{\mathrm{A-OS}}(\boldsymbol{x})$ (or $\varphi^{\mathrm{adp}}_{\mathrm{L-OS}}(\boldsymbol{x})$) relies only on the active variables. This implies that a first-step pilot estimator given by the adaptive lasso algorithm can benefit from sparse estimation methods when calculating optimal probabilities. For example, employing the standard lasso can effectively eliminate a large number of inactive variables to facilitate the computation of optimal $\varphi^{\mathrm{adp}}_{\mathrm{A-OS}}(\boldsymbol{x})$ and $\varphi^{\mathrm{adp}}_{\mathrm{L-OS}}(\boldsymbol{x})$. However, in practice, pilot estimators are often obtained from a small subsample size, introducing additional uncertainty. Therefore, it becomes crucial to exercise caution and be conservative by over-selecting variables during the first step to prevent the exclusion of important variables. As a consequence, although theoretically $\varphi^{\mathrm{adp}}_{\mathrm{A-OS}}(\boldsymbol{x})$ and $\varphi^{\mathrm{adp}}_{\mathrm{L-OS}}(\boldsymbol{x})$ do not depend on inactive variables, they are affected by inactive variables in practical implementations.

3.1 Scale invariant optimal function

As discussed in Section 1, scaling dependent optimal probabilities may impact the performance of variable selection in practice. To address the issue, we propose to construct a scale invariant optimal function by focusing on the prediction error of an estimator $\hat{\boldsymbol{\theta}}$, defined below.

$$\mathrm{MSPE}(\hat{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{x}}\left[\left\{p(\boldsymbol{x};\hat{\boldsymbol{\theta}}) - p(\boldsymbol{x};\boldsymbol{\theta}_t)\right\}^2\right] = \int\left\{p(\boldsymbol{x};\hat{\boldsymbol{\theta}}) - p(\boldsymbol{x};\boldsymbol{\theta}_t)\right\}^2 \mathrm{d}\mathbb{P}_{\boldsymbol{x}},$$

where $\mathbb{P}_{\boldsymbol{x}}$ is the probability measure of $\boldsymbol{x}$. The probability term $p(\boldsymbol{x};\boldsymbol{\theta}_t)$ involves both the covariates $\boldsymbol{x}$ and the parameter vector $\boldsymbol{\theta}_t$, and it often does not depend on the scale of $\boldsymbol{x}$. For example, in the logistic regression model, the value $p(\boldsymbol{x};\boldsymbol{\theta}_t)$ is only related to $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_t$. If we change the scale of $x_{(j)}$, the value of $\boldsymbol{\theta}_t$ would change accordingly under the same data-generating model and so $p(\boldsymbol{x};\boldsymbol{\theta}_t)$ remains the same. Thus, re-scaling covariates would not affect this criterion. In the following, we give an optimal function that minimizes the prediction error.

Theorem 2. *Under the assumptions of Theorem 1, for the IPW adaptive lasso estimator defined in* (3)*, its prediction error satisfies*

$$N_1 e^{-2\alpha_t}\mathrm{MSPE}(\hat{\boldsymbol{\theta}}^{\mathrm{adp}}_{\mathrm{w}(\mathcal{A})}) \rightsquigarrow \mathbb{E}^{-1}\left\{e^{f(\boldsymbol{x};\boldsymbol{\beta}_t)}\right\}\boldsymbol{Z}^{\mathrm{T}}_{(\mathcal{A})}\boldsymbol{M}^{1/2}_{\mathrm{w}(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\boldsymbol{\Omega}_{(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\boldsymbol{M}^{1/2}_{\mathrm{w}(\mathcal{A})}\boldsymbol{Z}_{(\mathcal{A})}. \tag{6}$$

https://doi.org/10.52202/079017-3122

where $\boldsymbol{Z}_{(\mathcal{A})} \sim \mathbb{N}(\boldsymbol{0}, \boldsymbol{I})$*, and* $\boldsymbol{\Omega}_{(\mathcal{A})} = \mathbb{E}\left[e^{2f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}\dot{g}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x},\boldsymbol{\theta}_{\mathrm{t}})\right]$*. The optimal function that minimizes the asymptotic mean of the prediction error in (6) is given as*

$$\varphi^{\mathrm{adp}}_{\mathrm{P-OS}}(\boldsymbol{x}) = \frac{p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\boldsymbol{\Omega}^{\frac{1}{2}}_{(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\dot{g}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|}{\mathbb{E}\left[p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\boldsymbol{\Omega}^{\frac{1}{2}}_{(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\dot{g}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\right]}. \tag{7}$$

We refer this prediction oriented criterion as P-optimality criterion. As we expect, the optimal function in (7) is unaffected by the scale of $\boldsymbol{x}$ for a class of functions g. The following proposition proves that $\varphi^{\mathrm{adp}}_{\mathrm{P-OS}}(\boldsymbol{x})$ is invariant to rescaling of $\boldsymbol{x}$.

Proposition 2. *If* $g(\boldsymbol{x};\boldsymbol{\theta})$ *satisfies that for every non-singular matrix* $\boldsymbol{A}$ *there exists a non-singular matrix* $\boldsymbol{B}$*, such that*

$$g(\boldsymbol{A}\boldsymbol{x};\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\theta}) = g(\boldsymbol{x};\boldsymbol{\theta}), \tag{8}$$

then, $\varphi^{\mathrm{adp}}_{\mathrm{P-OS}}(\boldsymbol{x})$ *is invariant to scale changes of* $\boldsymbol{x}$*.*

Remark 2. *The condition in* (8) *is not restrictive and it is quite easy to satisfy. One simple example of* $g(\boldsymbol{x};\boldsymbol{\theta})$ *that satisfies the condition is a linear function* $g(\boldsymbol{x};\boldsymbol{\theta}) = \alpha + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$*, which corresponds to the logistic regression. The condition is also satisfied by more complex models. For example, consider an* L*-layer neural network*

$$g(\boldsymbol{x};\boldsymbol{W}^1,\boldsymbol{W}^2,...,\boldsymbol{W}^L,\boldsymbol{b}^1,...,\boldsymbol{b}^L) = f^L(f^{L-1}(...f^1(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}^1+\boldsymbol{b}^1))^{\mathrm{T}}\boldsymbol{W}+\boldsymbol{b}^L),$$

where $\boldsymbol{W}^l$ *are the weights and* $\boldsymbol{b}^l$ *are the biases in each layer,* $l = 1, 2, ..., L$*. If* $\boldsymbol{x}$ *is rescaled to* $\boldsymbol{A}\boldsymbol{x}$*, we can change* $\boldsymbol{W}^1$ *to* $(\boldsymbol{A}^T)^{-1}\boldsymbol{W}^1$ *so that the value of* g *does not change. That is*

$$\begin{aligned} &g(\boldsymbol{A}\boldsymbol{x};(\boldsymbol{A}^T)^{-1}\boldsymbol{W}^1,\boldsymbol{W}^2,...,\boldsymbol{W}^L,\boldsymbol{b}^1,...,\boldsymbol{b}^L) = f^L(f^{L-1}(...f^1(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}^1+\boldsymbol{b}^1))^{\mathrm{T}}\boldsymbol{W}+\boldsymbol{b}^L)\\ &= g(\boldsymbol{x};\boldsymbol{W}^1,\boldsymbol{W}^2,...,\boldsymbol{W}^L,\boldsymbol{b}^1,...,\boldsymbol{b}^L). \end{aligned}$$

4 Penalized MSCL estimator

The IPW estimator in (3) is not the most efficient estimator, because it assigns smaller weights for more informative data points with larger sampling probabilities. To improve the estimation efficiency, we propose the penalized MSCL estimator for variable selection given as

$$\hat{\boldsymbol{\theta}}^{\mathrm{adp}}_{\mathrm{mscl}} := \arg\max_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{N^*_{\mathrm{sub}}} [y^{\mathrm{sub}}_i g(\boldsymbol{x}^{\mathrm{sub}}_i;\boldsymbol{\theta}) - \log\{1+e^{g(\boldsymbol{x}^{\mathrm{sub}}_i;\boldsymbol{\theta})+l^{\mathrm{sub}}_i}\}] - \lambda_N \sum_{j=1}^{p} \frac{|\beta_{(j)}|}{|\hat{\beta}_{\mathrm{pl}(j)}|^{\gamma}} \right\}, \tag{9}$$

where $l^{\mathrm{sub}}_i = -\log\left\{\rho\varphi(\boldsymbol{x}^{\mathrm{sub}}_i)\right\}$. The MSCL estimator introduced in [22] is defined as the minimizer of the objective function in (9), excluding the penalization term. In this paper, we extend this approach by proposing a penalized MSCL estimator to ensure model sparsity. We present the oracle properties of the penalized MSCL estimator in the following theorem.

Theorem 3. *Let* $\hat{\boldsymbol{\beta}}_{\mathrm{pl}}$ *be a consistent pilot estimate such that* $\lambda_N/(\sqrt{N_1}|\hat{\beta}_{\mathrm{pl}(j)}|^{\gamma}) \xrightarrow{P} \infty$ *for* $j \in \mathcal{A}^c$*. Under Assumptions 1-3 and 5, if* $\lambda_N/\sqrt{N_1} \to 0$*, the estimator based on MSCL function with adaptive lasso penalty defined in* (9) *have the following properties:*

1. *Consistency in variable selection: The estimated active set* $\hat{\mathcal{A}}_{\mathrm{mscl}} := \{j : \hat{\beta}^{\mathrm{adp}}_{\mathrm{mscl}(j)} \neq 0\}$ *satisfies that* $\lim_{N\to\infty}\mathbb{P}(\hat{\mathcal{A}}_{\mathrm{mscl}} = \mathcal{A}) = 1$
2. *Asymptotic normality: The estimator of the active parameter vector satisfies that*

$$\sqrt{N_1}\boldsymbol{V}^{-1/2}_{\mathrm{mscl}(\mathcal{A})}(\hat{\boldsymbol{\theta}}^{\mathrm{adp}}_{\mathrm{mscl}(\mathcal{A})} - \boldsymbol{\theta}_{\mathrm{t}(\mathcal{A})}) \rightsquigarrow \mathbb{N}(\boldsymbol{0},\boldsymbol{I}), \tag{10}$$

where $\boldsymbol{V}_{\mathrm{mscl}(\mathcal{A})} = \mathbb{E}\left\{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}\right\}\boldsymbol{\Lambda}^{-1}_{\mathrm{mscl}(\mathcal{A})}$ *and* $\boldsymbol{\Lambda}_{\mathrm{mscl}(\mathcal{A})} = \mathbb{E}\left[\frac{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}\dot{g}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}{1+c\varphi^{-1}(\boldsymbol{x})e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}}\right]$.

The penalized MSCL estimator has the same asymptotic variance as the MSCL estimator under the true model, indicating that it is more efficient than the penalized IPW estimator [22]. We prove this by comparing the asymptotic variances and present the result in the following theorem.

Theorem 4. *If the asymptotic variances $\boldsymbol{V}_{\mathrm{w}(\mathcal{A})}$ for $\hat{\boldsymbol{\theta}}^{\mathrm{adp}}_{\mathrm{w}(\mathcal{A})}$ in (2) and $\boldsymbol{V}_{\mathrm{mscl}(\mathcal{A})}$ for $\hat{\boldsymbol{\theta}}^{\mathrm{adp}}_{\mathrm{mscl}(\mathcal{A})}$ in (9), are finite, i.e., $0 < \boldsymbol{V}_{\mathrm{w}(\mathcal{A})}, \boldsymbol{V}_{\mathrm{mscl}(\mathcal{A})} < \infty$, then $\boldsymbol{V}_{\mathrm{mscl}(\mathcal{A})} \leq \boldsymbol{V}_{\mathrm{w}(\mathcal{A})}$, where the inequalities hold in the sense of Loewner ordering.*

Thus, we give a practical two-step algorithm based on the penalized MSCL estimator. Since the optimal sampling functions contain unknown values and the adaptive lasso penalty also requires a consistent pilot estimator to build weights, it is natural to combine optimal sampling and the adaptive lasso into one unified framework. We recommend to use the lasso for pilot estimation. One reason is that it does estimation and variable selection simultaneously, and excluding some inactive variables improves the estimation accuracy of optimal probabilities. This also reduces the computational burden for subsequent steps. Another reason is that the lasso estimator tends to include more variables in practice and therefore has a low risk of excluding important variables in the pilot step. We present an outline of the practical implementation in Algorithm 2. More details are given in Section C.

Algorithm 2 Two-step subsampling adaptive lasso algorithm

1: First stage screening:
- Take a pilot sample of expected sample size N_{pl} using $\{\pi(y_i) = \rho_0 + y_i(\rho_1 - \rho_0)\}_{i=1}^{N}$ and obtain a lasso penalized MSCL pilot estimator and an estimated active set $\hat{\mathcal{A}}_{\mathrm{pl}}$.
- Calculate approximate optimal sampling probabilities $\{\hat{\pi}(\boldsymbol{x}_i, y_i) = y_i + (1 - y_i)\rho\hat{\varphi}(x_i)\}_{i=1}^{N}$ based on (4), (5), or (7).

2: Second stage screening: Use Algorithm 1 with the estimated optimal sampling probabilities to obtain a subsample of expected sample size N_{sub} and compute the adaptive lasso penalized MSCL estimator based on $\hat{\mathcal{A}}_{\mathrm{pl}}$.

5 Numerical experiments

In this section, we use numerical experiments on both simulated and real data to investigate the performances of our proposed optimal subsampling and variable selection procedures.

5.1 Simulation design

We consider a logistic regression with $g(\boldsymbol{x}; \boldsymbol{\theta}) = \alpha + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$ and the following three true parameters $\boldsymbol{\beta}_{\mathrm{t}}$ of dimension 50. We set different α_{t} so that the proportion of ones is 0.005:

(1) **Case A:** $\boldsymbol{\beta}_{\mathrm{t}} = (0.75, 0.75, \mathbf{0}_7^{\mathrm{T}}, 0.75, 0, 0.75, 0.75, \mathbf{0}_{37}^{\mathrm{T}})^{\mathrm{T}}$ and $\alpha_{\mathrm{t}} = -5.8$.

(2) **Case B:** $\boldsymbol{\beta}_{\mathrm{t}} = (3, -2, \mathbf{0}_7^{\mathrm{T}}, 0.85, 0, -0.75, \mathbf{0}_{38}^{\mathrm{T}})^{\mathrm{T}}$ and $\alpha_{\mathrm{t}} = -6.2$.

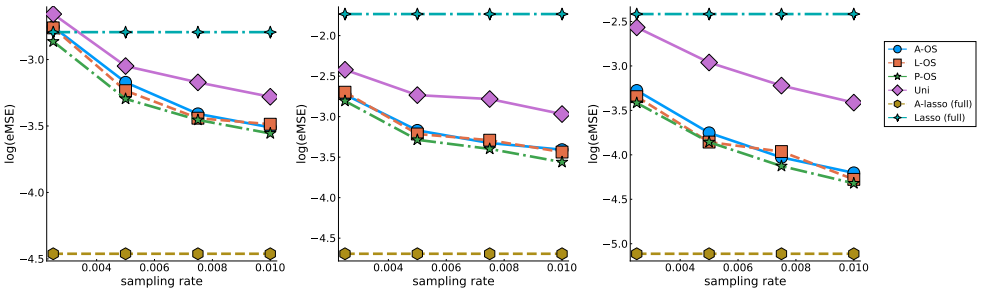
(3) **Case C:** $\boldsymbol{\beta}_{\mathrm{t}} = (3, 2, \mathbf{0}_7^{\mathrm{T}}, 0.85, \mathbf{0}_{40}^{\mathrm{T}})^{\mathrm{T}}$ and $\alpha_{\mathrm{t}} = -7.5$.

Here, $\mathbf{0}_d$ denotes the zero vector of dimension d. We use $p_{\mathcal{A}}$ and $p_{\mathcal{A}^c}$ to denote the number of active and inactive variables, respectively, and assume that $\boldsymbol{x}$ is a normal random vector. The active components $x_{(\mathcal{A},j)}$, $1 \leq j \leq p_{\mathcal{A}}$ of $\boldsymbol{x}$ have variances 0.25 and the inactive components $x_{(\mathcal{A}^c,j)}$, $1 \leq j \leq p_{\mathcal{A}^c}$of $\boldsymbol{x}$ have variances $100/p^3_{\mathcal{A}^c}, 100/(p_{\mathcal{A}^c} - 1)^3, ..., 100/3^3, 100/2^3, 100/1^3$. The correlation between the i-th and j-th elements of $\boldsymbol{x}$ is $0.5^{|i-j|}, 1 \leq i, j \leq p$. We repeat our experiments $S = 500$ times generating $N = 500000$ data points in each run and use a pilot sample of size $N_{\mathrm{pl}} = 500$ for obtaining pilot estimates based on the lasso. We consider uniform sampling, the full data lasso, and the full data adaptive lasso for comparison. We use the 5-fold cross-validation and Bayesian information criterion (BIC) to determine the tuning parameter λ for the lasso and the adaptive lasso, and choose $\gamma = 1$ for the adaptive lasso.

5.1.1 Estimation and prediction efficiency

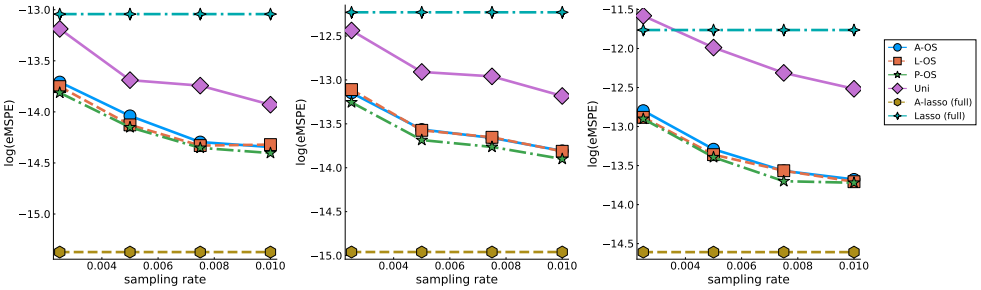
We present the empirical median squared error (eMSE) for parameter estimation in Figure 2. All optimal sampling estimators outperform the uniform sampling. As the sampling rate increases, sampling estimators outperform the full data lasso estimator eventually in all of the three cases. Among the three optimal subsampling methods, $\hat{\boldsymbol{\beta}}^{\mathrm{adp}}_{\mathrm{P-OS}}$ performs better than the other two subsampling methods.

https://doi.org/10.52202/079017-3122



(a) Case A (b) Case B (c) Case C

Figure 2: eMSE for different true parameters with different sampling rates.



(a) Case A (b) Case B (c) Case C

Figure 3: eMPSE of estimated probability with different sampling rates.

Figure 3 shows the results of the empirical median squared prediction error (eMSPE). Similarly to the results of eMSE, optimal sampling estimators perform better than the uniform sampling, meaning optimal sampling results in less information loss. It is possible that sampling estimators outperform the full data lasso estimator as the sampling rate increases, despite that the latter uses all of the data. In general, $\hat{\boldsymbol{\beta}}^{\text{adp}}_{\text{P}-\text{OS}}$ performs the best among the three optimal subsampling algorithms.

5.1.2 Variable selection and computational complexity

In this section, we discuss the results of variable selection in terms of the first stage screening and the second stage screening. Table 1 presents the mean numbers of selected variables in Case C, where the numbers in the parentheses are the corresponding standard errors. Results for Cases A and B are similar so are put in Table 4 of the appendix.

Table 1: Mean number of selected variables in Case C

ρ	first-stage	Uni	A-OS	L-OS	P-OS
0.0025	13.27(0.34)	2.84(0.02)	2.97(0.02)	2.96(0.02)	2.96(0.02)
0.005	12.46(0.32)	2.94(0.02)	3.04(0.03)	3.05(0.03)	3.06(0.03)
0.0075	12.76(0.33)	2.97(0.01)	3.04(0.02)	3.03(0.02)	3.03(0.02)
0.01	12.81(0.34)	2.96(0.01)	3.03(0.02)	3.02(0.01)	3.02(0.01)

While the first stage screening significantly reduces the dimension in Table 1, it indeed includes inactive variables as expected. In the second stage screening, the mean numbers of selected variables are close to the true numbers of active variables for all subsampling methods. However, the mean number of selected variables from uniform sampling is smaller than the true number of active variables especially when the sampling rate is low. This indicates that the second-stage screening of uniform sampling may exclude active variable. We present the rates of missing active variables in Table 2 for Case C. It shows that uniform sampling has higher rates of excluding active variables than optimal subsampling procedures, so optimal sampling may be preferable in practice. Results for Cases A and B are similar and are put in Section E.1. We also investgate the rates of selecting the true model in that section.

Table 2: Rates of excluding active variables (false negative rate) in Case C

ρ	Uni	A-OS	L-OS	P-OS
0.0025	0.168(0.017)	0.086(0.013)	0.088(0.013)	0.084(0.013)
0.005	0.100(0.013)	0.068(0.011)	0.066(0.011)	0.066(0.011)
0.0075	0.066(0.011)	0.046(0.009)	0.048(0.010)	0.046(0.009)
0.01	0.068(0.011)	0.052(0.010)	0.054(0.010)	0.054(0.010)

5.1.3 Computational time

We present the mean computational times of different algorithms in Table 3. Our codes are written in the *julia* programming language [2] and implemented on a Linux workstation. The lasso pathes are solved with *Lasso.jl* [13]. As shown in Table 3, subsampling algorithms significantly reduce the computational times compared with full data estimators. Although optimal sampling requires to calculate sampling probabilities, they use only about 0.77% of the computational time that the full data adaptive lasso requires. As we discussed in Section C.2, optimal sampling algorithms reduce both sample size and the data dimension. Therefore, the computational cost of the coordinate decent algorithm, which often requires a large number of iterations, is significantly reduced.

Table 3: Mean computational time (seconds)

Case	Uni	A-OS	L-OS	P-OS	A-lasso (full)	Lasso (full)
A	0.29	1.09	0.91	1.06	129.62	112.97
B	0.31	1.23	1.20	1.27	129.89	122.40
C	0.31	1.02	0.93	1.00	130.33	121.29

5.2 Real data

We evaluate the performances of proposed estimators on two real data sets.

(i) **Covtype data set:** It is available at `https://archive.ics.uci.edu/ml/datasets/covertype`, with $N = 581012$ observations and 54 covariates – 10 being quantitative and 44 being qualitative with dummy coding. We drop the 14th and 54th columns to avoid exact colinearity of the dummy variables. Our goal is to classify whether the forest cover type is Cottonwood/Willow (labeled as 1) or not (labeled as 0). The proportion of Cottonwood/Willow is 0.473%, which is highly imbalanced.

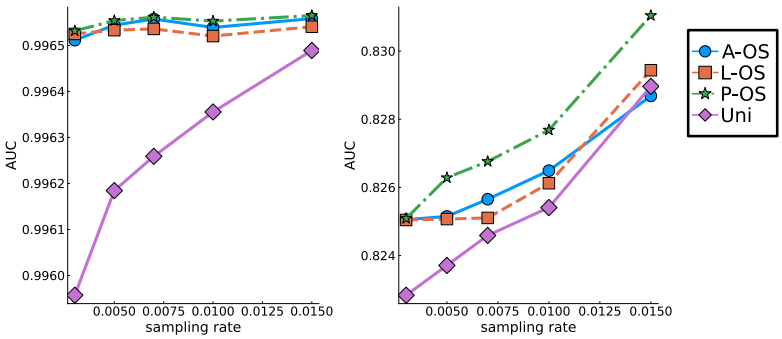
(ii) **Font data set:** It is available at `https://archive.ics.uci.edu/ml/datasets/Character+Font+Images`, with 0.50% of the $N = 832670$ responses being the GADUGI font. The first 10 covariates are about the value, size, and style of the characters and there are additional 400 pixel values of the 20×20 images. We remove the 4th, 9th, and 10th covariates because they are constants.

For both data sets, we apply Algorithm 2 on the logarithmic-transformed data. We use pilot samples of size $N_{\mathrm{pl}} = 1000$ for the covtype data and $N_{\mathrm{pl}} = 1500$ for the font data due to its higher dimension. Since we do not know the true parameter for real data, we use area under the curve (AUC) to measure the performances of subsampling algorithms. We repeat the experiment for $S = 500$ and compute the empirical median AUC using the full data. The results are summarized in Figure 4. As shown in Figure 4, nonuniform sampling outperforms uniform sampling in general. There is one case for font data set that $\hat{\boldsymbol{\beta}}^{\mathrm{adp}}_{\mathrm{A-OS}}$ is worse than the uniform sampling when the sampling rate is high. For the covtype data set, among the three estimators based on optimal sampling, $\hat{\boldsymbol{\beta}}^{\mathrm{adp}}_{\mathrm{P-OS}}$ performs the best and $\hat{\boldsymbol{\beta}}^{\mathrm{adp}}_{\mathrm{L-OS}}$ is worst. For the font data set, $\hat{\boldsymbol{\beta}}^{\mathrm{adp}}_{\mathrm{A-OS}}$ and $\hat{\boldsymbol{\beta}}^{\mathrm{adp}}_{\mathrm{L-OS}}$ are similar, and $\hat{\boldsymbol{\beta}}^{\mathrm{adp}}_{\mathrm{P-OS}}$ based on the scale invariant optimal sampling function is significantly better.

6 Conclusion and limitations

In this paper, we investigated the problem of scale-invariant optimal subsampling in the context of variable selection for rare-events data. We derived optimal probabilities based on the A- and L-

https://doi.org/10.52202/079017-3122



(a) AUCs of the covtype data set (b) AUCs of the font data set

Figure 4: Empirical median AUCs for two real data sets

optimality criteria, and discussed their limitations. Furthermore, we proposed scale-invariant optimal probabilities based on prediction errors to overcome the limitations. Both analytical and numerical results show the desirable properties of the proposed methods.

Our investigation has the following limitations.

- Our proposed criterion optimizes the probabilities by minimizing the asymptotic mean squared error in estimating rare-event probabilities. While this prioritizes the accuracy of estimation, it puts less emphasis on the quality of variable selection. Further research is needed to devise optimal probabilities that focus on variable selection performance metrics.
- Our theoretical analysis is based on asymptotic properties, with optimal probabilities defined through the asymptotic normality. Although our results may hold for sufficiently sparse models, they may not generalize to cases where the model is dense or over-parameterized, because asymptotic normality may no longer be applicable. Therefore, an important direction for future research is to study the non-asymptotic properties of our estimators, such as prediction error bounds. Non-asymptotic behaviors are particularly of interest in high-dimensional regimes.
- We employ Lasso as the pilot estimator. However, other variable selection methodologies, such as sure independence screening, can also be considered. Exploring the impact of different pilot estimators on our method's performance represents another avenue for future investigations.
- We assume that the underlying full model is correctly specified and possesses a sparse structure. Our analysis does not account for model misspecification. Further research is required to address scenarios where the model is possibly misspecified or where the number of features vastly exceeds the number of observations.

Acknowledgments and Disclosure of Funding

The authors are grateful to Professor Kun Chen for the insightful comments and suggestions on the development of the manuscript. Funding in direct support of this work: NEI grant R21EY035710, NSF grant 2105571, UConn CLAS Research Funding in Academic Themes, GPUs donated by NVIDIA.

References

[1] M. Ai, J. Yu, H. Zhang, and H. Wang. Optimal subsampling algorithms for big data regressions. *Statistica Sinica*, 31(2):749–772, 2021.

[2] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[4] G. Douzas and F. Bacao. Self-organizing map oversampling (somo) for imbalanced data set learning. *Expert systems with Applications*, 82:40–52, 2017.

[5] C. Drummond, R. C. Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, 2003.

[6] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[7] D. Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 03 1993.

[8] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302 – 332, 2007.

[9] J. H. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[10] W. Fu and K. Knight. Asymptotics for lasso-type estimators. *The Annals of statistics*, 28(5):1356–1378, 2000.

[11] C. J. Geyer. On the asymptotics of constrained m-estimation. *The Annals of statistics*, pages 1993–2010, 1994.

[12] H. Han, W.-Y. Wang, and B.-H. Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.

[13] JuliaStats. Lasso.jl. `https://github.com/JuliaStats/Lasso.jl`, 2022.

[14] N. Keret and M. Gorfine. Analyzing big ehr data—optimal cox regression subsampling procedure with rare events. *Journal of the American Statistical Association*, 118(544):2262–2275, 2023.

[15] X.-Y. Liu, J. Wu, and Z.-H. Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2008.

[16] J. Mathew, C. K. Pang, M. Luo, and W. H. Leong. Classification of imbalanced data by oversampling in kernel space of support vector machines. *IEEE transactions on neural networks and learning systems*, 29(9):4065–4076, 2017.

[17] F. Pukelsheim. *Optimal design of experiments*. SIAM, 2006.

[18] J. Shao. *Mathematical Statistics, 2nd*. Springer-Verlag, New York, 2003.

[19] G. Tripathi. A matrix extension of the cauchy-schwarz inequality. *Economics Letters*, 63(1):1–3, 1999.

[20] H. Wang. Logistic regression for massive data with rare events. In *International Conference on Machine Learning*, pages 9829–9836. PMLR, 2020.

[21] H. Wang and Y. Ma. Optimal subsampling for quantile regression in big data. *Biometrika*, 108(1):99–112, 2021.

[22] H. Wang, A. Zhang, and C. Wang. Nonuniform negative sampling and log odds correction with rare events data. *Advances in Neural Information Processing Systems*, 34, 2021.

[23] H. Wang, R. Zhu, and P. Ma. Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522):829–844, 2018.

 https://doi.org/10.52202/079017-3122

[24] J. Wang, J. Zou, and H. Wang. Sampling with replacement vs poisson sampling: a comparative study in optimal subsampling. *IEEE Transactions on Information Theory*, 68(10):6605–6630, 2022.

[25] Y. Yao and H. Wang. Optimal subsampling for softmax regression. *Statistical Papers*, pages 585–599, 12 2018.

[26] J. Yu, H. Wang, M. Ai, and H. Zhang. Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association*, 0(0):1–12, 2020. DOI:10.1080/01621459.2020.1773832.

[27] G.-X. Yuan, C.-H. Ho, and C.-J. Lin. An improved glmnet for l1-regularized logistic regression. *J. Mach. Learn. Res.*, 13:1999–2030, 2012.

[28] H. H. Zhang and W. Lu. Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 94(3):691–703, 05 2007.

[29] T. Zhang, Y. Ning, and D. Ruppert. Optimal sampling for generalized linear models under measurement constraints. *Journal of Computational and Graphical Statistics*, 30(1):106–114, 2021.

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

A Appendix / supplemental material

In this appendix, we present the details of the proof, the practical algorithm and simuation settings in the paper. Details of mathematical proofs are provided in Section B. Details of the practical algorithm are provided in Section C. We present the details of simulation settings in Section D and in Section E, we give some additional simulation results.

B Details of mathematical proofs

In this section, we provide details of mathematical proofs.

B.1 General assumptions in the main paper

We begin with some general assumptions used throughout this paper.

Assumption 1. *The first, second and third derivatives of $f(\boldsymbol{x};\boldsymbol{\theta})$ and $e^{f(\boldsymbol{x};\boldsymbol{\beta})}f(\boldsymbol{x};\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ are bouned by a square intergrable random variable $B(\boldsymbol{x})$.*

Assumption 2. *The matrix $\mathbb{E}\left\{\dot{g}^{\otimes 2}(\boldsymbol{x};\boldsymbol{\theta})\right\}$ is finite and positive definite.*

Assumption 3. *The subsampling rate ρ satisfies that $c_N = e^{\alpha_{\mathrm{t}}}/\rho \to c$, where $0 \leq c < \infty$ is a constant.*

Assumption 4. *The integral $\mathbb{E}\left[\left\{\varphi(\boldsymbol{x}) + \varphi^{-1}(\boldsymbol{x})\right\}B^2(\boldsymbol{x})\right]$ is finite, where $B(\boldsymbol{x})$ is a square-integrable function that dominates the first, second, and third derivatives of $f(\boldsymbol{x};\boldsymbol{\theta})$ and $e^{f(\boldsymbol{x};\boldsymbol{\beta})}f(\boldsymbol{x};\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$.*

Assumption 5. *The integral $\mathbb{E}\left\{e^{f(\boldsymbol{x};\boldsymbol{\beta})}\varphi^{-1}(\boldsymbol{x})B(\boldsymbol{x})\right\}$ is finite.*

These assumptions are the same assumptions used in [22]. Here, we remind some notations used in the main paper:

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{e^{\alpha + f(\boldsymbol{x};\boldsymbol{\beta})}}{1 + e^{\alpha + f(\boldsymbol{x};\boldsymbol{\beta})}},$$

$$\phi(\boldsymbol{x};\boldsymbol{\theta}) = p(\boldsymbol{x};\boldsymbol{\theta})\left\{1 - p(\boldsymbol{x};\boldsymbol{\theta})\right\},$$

$$\boldsymbol{M} = \mathbb{E}\left\{e^{f(\boldsymbol{x}_i;\boldsymbol{\beta}_{\mathrm{t}})}\dot{g}^{\otimes 2}(\boldsymbol{x}_i,\boldsymbol{\theta}_{\mathrm{t}})\right\},$$

and

$$\boldsymbol{\Lambda}_{\text{mscl}} = \mathbb{E}\left[\frac{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\text{t}})}\dot{g}^{\otimes 2}(\boldsymbol{x};\boldsymbol{\beta}_{\text{t}})}{1+c\varphi^{-1}(\boldsymbol{x})e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\text{t}})}}\right].$$

To ease the presentation in the following sections, we denote

$$\boldsymbol{M}_{\text{w}(\mathcal{A})} = \mathbb{E}\left[\left\{1+\frac{ce^{f(\boldsymbol{x};\boldsymbol{\beta}_{\text{t}})}}{\varphi(\boldsymbol{x})}\right\}e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\text{t}})}\dot{g}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\text{t}})\right],$$

and $a_N = \sqrt{Ne^{\alpha_{\text{t}}}}$ in the appendix. Note that

$$\begin{aligned} N_1 &= \sum_{i=1}^{N} y_i = N\mathbb{E}\left\{\frac{e^{\alpha_{\text{t}}+f(\boldsymbol{x};\boldsymbol{\beta}_{\text{t}})}}{1+e^{\alpha_{\text{t}}+f(\boldsymbol{x};\boldsymbol{\beta}_{\text{t}})}}\right\}\{1+o_P(1)\} \\ &= Ne^{\alpha_{\text{t}}}\mathbb{E}\left\{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\text{t}})}\right\}\{1+o_P(1)\} = a_N^2\mathbb{E}\left\{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\text{t}})}\right\}\{1+o_P(1)\}. \end{aligned}$$

B.2 Proof of Theorem 1

Proof of Theorem 1. We consider the target of IPW adaptive lasso estimator:

$$\begin{aligned} Q_{\text{w}}(\boldsymbol{\theta}) &= -\sum_{i=1}^{N}\frac{\delta_i}{\pi(\boldsymbol{x}_i,y_i)}[y_i g(\boldsymbol{x}_i;\boldsymbol{\theta}) - \log\{1+e^{g(\boldsymbol{x}_i;\boldsymbol{\theta})}\}] + \lambda_N\sum_{j=1}^{p}\hat{w}_j|\beta_{(j)}| \\ &= -\ell_{\text{w}}(\boldsymbol{\theta}) + \lambda_N\sum_{j=1}^{p}\hat{w}_j|\beta_{(j)}|, \end{aligned}$$

where $\hat{w}_j = 1/|\hat{\beta}_{\text{pl}(j)}|, 1\le j\le p$. Then, we have that $\hat{\boldsymbol{u}}_N = a_N(\hat{\boldsymbol{\theta}}_{\text{w}} - \boldsymbol{\theta}_{\text{t}})$ is the minimizer of

$$\gamma_{\text{w}}^N(\boldsymbol{u}) = Q_{\text{w}}(\boldsymbol{\theta}_{\text{t}} + a_N^{-1}\boldsymbol{u}) - Q_{\text{w}}(\boldsymbol{\theta}_{\text{t}}).$$

Asymptotic normality: We prove the asymptotic normality part in this paragraph. By Taylor's expansion,

$$\begin{aligned} \gamma_{\text{w}}^N(\boldsymbol{u}) = &-\frac{1}{a_N}\boldsymbol{u}^{\text{T}}\dot{\ell}_{\text{w}}(\boldsymbol{\theta}_{\text{t}}) + \frac{1}{2a_N^2}\sum_{i=1}^{N}\frac{\delta_i}{\pi(\boldsymbol{x}_i,y_i)}\phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{\text{t}})\{\boldsymbol{u}^{\text{T}}\dot{g}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\text{t}})\}^2 - \Delta_{\text{w}} + R_{\text{w}} \\ &+ \frac{\lambda_N}{a_N}\sum_{j=1}^{p}\hat{w}_j a_N\left(\left|\beta_{\text{t}(j)} + \frac{u_{(j)}}{a_N}\right| - |\beta_{\text{t}(j)}|\right). \end{aligned}$$

We first consider the limit behavior of the IPW target function by prove the asymptotic normality. In [22], the authors established that under Assumptions 1 to 3,

$$a_N^{-1}\dot{\ell}_{\text{w}}(\boldsymbol{\theta}_{\text{t}}) \rightsquigarrow \boldsymbol{M}_{\text{w}}^{1/2}\boldsymbol{Z},$$

$$\frac{1}{a_N^2}\sum_{i=1}^{N}\frac{\delta_i}{\pi(\boldsymbol{x}_i,y_i)}\phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{\text{t}})\dot{g}^{\otimes 2}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\text{t}}) \xrightarrow{P} \boldsymbol{M},$$

and

$$\Delta_{\text{w}} = o_P(1), \quad R_{\text{w}} = o_P(1).$$

Thus,

$$-\ell_{\text{w}}(\boldsymbol{\theta}_{\text{t}}) \rightsquigarrow -\boldsymbol{u}^{\text{T}}\boldsymbol{M}_{\text{w}}^{1/2}\boldsymbol{Z} + \frac{1}{2}\boldsymbol{u}^{\text{T}}\boldsymbol{M}\boldsymbol{u}.$$

Next, we consider the limit behavior of the adaptive lasso penalty. Since we assume $\hat{\boldsymbol{\beta}}_{\text{pl}}$ to be a consistent estimator, we know that when $j\in\mathcal{A}$, i.e., $\boldsymbol{\beta}_{\text{t}(j)}\neq 0$,

$$\hat{w}_j = |\hat{\beta}_{\text{pl}(j)}|^{-\gamma} \xrightarrow{P} |\beta_{\text{t}(j)}|^{-\gamma} > 0,$$

and

$$a_N\left(\left|\beta_{\text{t}(j)} + \frac{u_{(j)}}{a_N}\right| - |\beta_{\text{t}(j)}|\right) \to \text{sgn}(\beta_{\text{t}(j)})u_{(j)}.$$

https://doi.org/10.52202/079017-3122

Therefore, for $j \in \mathcal{A}$, we have that

$$\frac{\lambda_N}{a_N}\hat{w}_j a_N \left(\left| \beta_{\mathrm{t}(j)} + \frac{u_{(j)}}{a_N} \right| - |\beta_{\mathrm{t}(j)}| \right) = o_P(1)$$

since $\lambda_N / a_N = \lambda_N / \sqrt{N e^{\alpha_{\mathrm{t}}}} \to 0$. On the other hand, when $j \in \mathcal{A}^c$, i.e., $\beta_{\mathrm{t}(j)} = 0$, we have that for $u_{(j)} \neq 0$,

$$\frac{\lambda_N}{a_N}\hat{w}_j a_N \left(\left| \beta_{\mathrm{t}(j)} + \frac{u_{(j)}}{a_N} \right| - |\beta_{\mathrm{t}(j)}| \right) = \frac{\lambda_N}{a_N}\hat{w}_j |u_{(j)}| = \frac{\lambda_N}{a_N |\hat{\beta}_{\mathrm{pl}(j)}|^{\gamma}} |u_{(j)}| \xrightarrow{P} \infty,$$

since $\lambda_N / (\sqrt{N e^{\alpha_{\mathrm{t}}}} |\hat{\beta}_{\mathrm{pl}(j)}|^{\gamma}) \xrightarrow{P} \infty$. Then, we have that $\gamma_{\mathrm{w}}^N(\boldsymbol{u}) \rightsquigarrow \gamma_{\mathrm{w}}(\boldsymbol{u})$, where

$$\gamma_{\mathrm{w}}(\boldsymbol{u}) = \begin{cases} \frac{1}{2}\boldsymbol{u}_{(\mathcal{A})}^{\mathrm{T}} \boldsymbol{M}_{(\mathcal{A})} \boldsymbol{u}_{(\mathcal{A})} - \boldsymbol{u}_{(\mathcal{A})}^{\mathrm{T}} \boldsymbol{M}_{\mathrm{w}}^{1/2} \boldsymbol{Z}_{(\mathcal{A})} & \text{if } u_{(j)} = 0, \forall j \in \mathcal{A}^c \\ \infty & \text{otherwise.} \end{cases}$$

Note that the unique minimizer of $\gamma_{\mathrm{w}}^N(\boldsymbol{u})$ is $(\boldsymbol{M}_{(\mathcal{A})}^{-1} \boldsymbol{M}_{\mathrm{w}}^{1/2} \boldsymbol{Z}_{(\mathcal{A})}^{\mathrm{T}}, \boldsymbol{0})^{\mathrm{T}}$ if we put all the indexes of active variables in front. Thus, following the results of [11] and [10], we have the minimizer of $\gamma_{\mathrm{w}}^N(\boldsymbol{u})$, i.e., $\hat{\boldsymbol{u}}_N$, satisfies that

$$\hat{\boldsymbol{u}}_{N(\mathcal{A})} \rightsquigarrow \boldsymbol{M}_{(\mathcal{A})}^{-1} \boldsymbol{M}_{\mathrm{w}}^{1/2} \boldsymbol{Z}_{(\mathcal{A})} \text{ and } \hat{\boldsymbol{u}}_{N(\mathcal{A}^c)} \rightsquigarrow \boldsymbol{0}.$$

Thus,

$$\hat{\boldsymbol{u}}_{N(\mathcal{A})} = a_N (\hat{\boldsymbol{\theta}}_{\mathrm{w}(\mathcal{A})} - \boldsymbol{\theta}_{\mathrm{t}(\mathcal{A})}) \rightsquigarrow \mathbb{N}(\boldsymbol{0}, \boldsymbol{M}_{(\mathcal{A})}^{-1} \boldsymbol{M}_{\mathrm{w}(\mathcal{A})} \boldsymbol{M}_{(\mathcal{A})}^{-1}).$$

Since

$$\sqrt{N_1} = a_N \mathbb{E}^{1/2} \left\{ e^{f(\boldsymbol{x}; \boldsymbol{\beta}_{\mathrm{t}})} \right\} \{1 + o_P(1)\},$$

applying Slusky's theorem, we have

$$\sqrt{N_1} \boldsymbol{V}_{\mathrm{w}(\mathcal{A})}^{-1/2} (\hat{\boldsymbol{\theta}}_{\mathrm{w}(\mathcal{A})} - \boldsymbol{\theta}_{\mathrm{t}(\mathcal{A})}) \rightsquigarrow \mathbb{N}(\boldsymbol{0}, \boldsymbol{I}).$$

Consistency in variable selection We prove the consistency in variable selection in this paragraph. From the result of asymptotic normality, we know that $\hat{\beta}_{\mathrm{w}(j)} \xrightarrow{P} \beta_{\mathrm{t}(j)}$ for every $j \in \mathcal{A}$ and therefore $\mathbb{P}(j \in \hat{\mathcal{A}}_{\mathrm{w}}) \to 1$. Thus, we only consider $j' \in \mathcal{A}^c$. When $j' \in \hat{\mathcal{A}}_{\mathrm{w}}$, we know that by K-K-T optimality conditions, we have

$$\lambda_N \hat{w}_{j'} \mathrm{sgn}(\hat{\beta}_{(j')}) = \dot{\ell}_{\mathrm{w}}(\hat{\boldsymbol{\theta}}_{\mathrm{w}}),$$

which means

$$\begin{aligned} \frac{\lambda_N \hat{w}_{j'} \mathrm{sgn}(\hat{\beta}_{(j')})}{a_N} &= \frac{\dot{\ell}_{\mathrm{w}}(\hat{\boldsymbol{\theta}}_{\mathrm{w}})}{a_N} \\ &= \frac{\dot{\ell}_{\mathrm{w}}(\boldsymbol{\theta}_{\mathrm{t}})}{a_N} + \frac{a_N \left\{ \dot{\ell}_{\mathrm{w}}(\hat{\boldsymbol{\theta}}_{\mathrm{w}}) - \dot{\ell}_{\mathrm{w}}(\boldsymbol{\theta}_{\mathrm{t}}) \right\}}{a_N^2} =: I_1 + I_2. \end{aligned}$$

We have known that $I_1 = \dot{\ell}_{\mathrm{w}}(\boldsymbol{\theta}_{\mathrm{t}}) / a_N \rightsquigarrow \boldsymbol{Z}_{\mathrm{w}}$. We now prove that proof that $I_2 = O_P(1)$. We apply Taylor expansion to the k-th element of $\dot{\ell}_{\mathrm{w}}(\hat{\boldsymbol{\theta}}_{\mathrm{w}})$ and have that

$$\frac{a_N \left\{ \dot{\ell}_{(k)}(\hat{\boldsymbol{\theta}}_{\mathrm{w}}) - \dot{\ell}_{(k)}(\boldsymbol{\theta}_{\mathrm{t}}) \right\}}{a_N^2} = -\frac{1}{a_N^2} \sum_{i=1}^{N} \frac{\delta_i}{\pi(\boldsymbol{x}, y_i)} \phi(\boldsymbol{x}_i; \boldsymbol{\theta}_{\mathrm{t}}) \dot{g}_{(k)}(\boldsymbol{x}_i; \boldsymbol{\theta}_{\mathrm{t}}) \dot{g}^{\mathrm{T}}(\boldsymbol{x}_i; \boldsymbol{\theta}_{\mathrm{t}}) \hat{\boldsymbol{u}}_N + \tilde{\Delta}_{\mathrm{w}(k)} + \tilde{R}_{\mathrm{w}(k)},$$

where,

$$\hat{\boldsymbol{u}}_N = a_N (\hat{\boldsymbol{\theta}}_{\mathrm{w}} - \boldsymbol{\theta}_{\mathrm{t}}) = O_P(1),$$

$$\tilde{\Delta}_{\mathrm{w}(k)} = \frac{1}{a_N^2} \sum_{i=1}^{N} \frac{\delta_i}{\pi(\boldsymbol{x}_i, y_i)} \{ y_i - p(\boldsymbol{x}_i; \boldsymbol{\theta}_{\mathrm{t}}) \} \sum_{j=1}^{d} \ddot{g}_{(kj)}(\boldsymbol{x}_i; \boldsymbol{\theta}_{\mathrm{t}}) \hat{u}_{N(j)},$$

and

$$\begin{aligned}\tilde{R}_{\mathrm{w}(k)} = &-\frac{1}{2a_N^3}\sum_{i=1}^{N}\frac{\delta_i}{\pi(\boldsymbol{x}_i,y_i)}\phi(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\left\{1-2p(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right\}\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\hat{\boldsymbol{u}}_N^{\mathrm{T}}\dot{g}^{\otimes 2}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\hat{\boldsymbol{u}}_N\\ &-\frac{2}{2a_N^3}\sum_{i=1}^{N}\frac{\delta_i}{\pi(\boldsymbol{x}_i,y_i)}\phi(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\left\{\hat{\boldsymbol{u}}_N^{\mathrm{T}}\frac{\partial\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)}{\partial\boldsymbol{\theta}}\right\}\left\{\hat{\boldsymbol{u}}_N^{\mathrm{T}}\dot{g}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right\}\\ &-\frac{1}{2a_N^3}\sum_{i=1}^{N}\frac{\delta_i}{\pi(\boldsymbol{x}_i,y_i)}\phi(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\left\{\hat{\boldsymbol{u}}_N^{\mathrm{T}}\ddot{g}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\hat{\boldsymbol{u}}_N\right\}\\ &+\frac{1}{2a_N^3}\sum_{i=1}^{N}\frac{\delta_i}{\pi(\boldsymbol{x}_i,y_i)}\left\{y_i-p(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right\}\hat{\boldsymbol{u}}_N^{\mathrm{T}}\frac{\partial^2\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)}{\partial\boldsymbol{\theta}^2}\hat{\boldsymbol{u}}_N.\end{aligned}$$

where $\acute{\boldsymbol{\theta}}_k$ is between $\hat{\boldsymbol{\theta}}_{\mathrm{mle}}$ and $\boldsymbol{\theta}_{\mathrm{t}}$. First, we prove that $\tilde{R}_{(k)}$ is $o_P(1)$. We have that

$$\begin{aligned}|\tilde{R}_{\mathrm{w}(k)}| \le &\frac{\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^3}\sum_{i=1}^{N}\frac{\delta_i}{\pi(\boldsymbol{x}_i,y_i)}\phi(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\left|1-2p(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right|\left|\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right|\left\|\dot{g}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right\|^2\\ &+\frac{2\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^3}\sum_{i=1}^{N}\frac{\delta_i}{\pi(\boldsymbol{x}_i,y_i)}\phi(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\left\|\frac{\partial\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)}{\partial\boldsymbol{\theta}}\right\|\left\|\dot{g}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right\|\\ &+\frac{\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^2}\sum_{i=1}^{N}\frac{\delta_i}{\pi(\boldsymbol{x}_i,y_i)}\phi(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\left|\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right|\left\|\ddot{g}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right\|\\ &+\frac{\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^3}\sum_{i=1}^{N}\frac{\delta_i}{\pi(\boldsymbol{x}_i,y_i)}p(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\left\|\frac{\partial^2\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)}{\partial\boldsymbol{\theta}^2}\right\|\\ &+\frac{\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^3}\sum_{i=1}^{N}\frac{\delta_i}{\pi(\boldsymbol{x}_i,y_i)}y_i\left\|\frac{\partial^2\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)}{\partial\boldsymbol{\theta}^2}\right\|\\ \le &\frac{\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^3}\sum_{i=1}^{N}\frac{\delta_i}{\pi(\boldsymbol{x}_i,y_i)}p(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)C(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)+\frac{\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^3}\sum_{i=1}^{N}\frac{\delta_i}{\pi(\boldsymbol{x}_i,y_i)}y_iB(\boldsymbol{x}_i)\\ \le &\frac{\|\hat{\boldsymbol{u}}_N\|^2e^{\acute{\alpha}_k-\alpha_{\mathrm{t}}}e^{\alpha_{\mathrm{t}}}}{2a_N^3}\sum_{i=1}^{N}\frac{\delta_i}{\pi(\boldsymbol{x}_i,y_i)}e^{f(\boldsymbol{x}_i;\acute{\boldsymbol{\beta}}_k)}C(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)+\frac{\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^3}\sum_{i=1}^{N}\frac{\delta_i}{\pi(\boldsymbol{x}_i,y_i)}y_iB(\boldsymbol{x}_i)\\ \le &\frac{\|\hat{\boldsymbol{u}}_N\|^2e^{\acute{\alpha}_k-\alpha_{\mathrm{t}}}}{2Na_N}\sum_{i=1}^{N}\frac{\delta_i}{\pi(\boldsymbol{x}_i,y_i)}e^{f(\boldsymbol{x}_i;\acute{\boldsymbol{\beta}}_k)}C(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)+\frac{\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^3}\sum_{i=1}^{N}\frac{\delta_i}{\pi(\boldsymbol{x}_i,y_i)}y_iB(\boldsymbol{x}_i)\\ = &\,o_P(1),\end{aligned}$$

where

$$\begin{aligned}C(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}) = &\left|\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}})\right|\left\{\left\|\dot{g}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right\|^2+\left\|\ddot{g}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}})\right\|\right\}\\ &+\left\|\frac{\partial\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)}{\partial\boldsymbol{\theta}}\right\|\left\|\dot{g}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right\|+\left\|\frac{\partial^2\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)}{\partial\boldsymbol{\theta}^2}\right\|.\end{aligned}$$

Therefore, we proved that $\tilde{R}_{\mathrm{w}(k)} = o_P(1)$. Next, we prove that $\tilde{\Delta}_{\mathrm{w}(k)} = o_P(1)$. We know that $\mathbb{E}\left[a_N^{-2}\sum_{i=1}^{N}\delta_i/\pi(\boldsymbol{x}_i,y_i)\left\{y_i-p(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}})\right\}\ddot{g}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}})\right] = \mathbf{0}$. We also have that for the every element of $a_N^{-2}\sum_{i=1}^{N}\delta_i/\pi(\boldsymbol{x}_i,y_i)\left\{y_i-p(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}})\right\}\ddot{g}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}})$, we have

$$\mathbb{V}\left[a_N^{-2}\sum_{i=1}^{N}\frac{\delta_i}{\pi(\boldsymbol{x}_i,y_i)}\left\{y_i-p(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}})\right\}\ddot{g}_{(jl)}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}})\right]$$

https://doi.org/10.52202/079017-3122

$$\leq \frac{1}{a_N^4}\sum_{i=1}^{N}\mathbb{E}\left\{p(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}})\ddot{g}^2_{(jl)}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}})\right\} \leq \frac{1}{a_N^2}\mathbb{E}[e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}\|\ddot{g}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|^2] \to 0.$$

Thus, due to Chebyshev's inequality, we know that $\tilde{\Delta}_{\mathrm{w}} = o_P(1)$. Since we know that $\frac{1}{a_N^2}\sum_{i=1}^{N}\delta_i/\pi(\boldsymbol{x}_i,y_i)\phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}})\dot{g}^{\otimes 2}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}}) = H_{\mathrm{w}} = O_P(1)$. Hence, we have that

$$\frac{\dot{\ell}_{\mathrm{w}}(\hat{\boldsymbol{\theta}}_{\mathrm{w}})}{a_N} = O_P(1).$$

Note that we also have

$$\frac{\lambda_N\hat{w}_{j'}}{a_N} = \frac{\lambda_N}{a_N}\frac{1}{|\hat{\beta}_{\mathrm{pl}(j')}|^{\gamma}} \xrightarrow{P} \infty.$$

Therefore,

$$\begin{aligned}\mathbb{P}(j' \in \hat{\mathcal{A}}_{\mathrm{w}}) &\leq \mathbb{P}\left\{\lambda_N\hat{w}_{j'}\mathrm{sgn}(\hat{\beta}_{(j')}) = \dot{\ell}_{\mathrm{w}}(\hat{\boldsymbol{\theta}}_{\mathrm{w}})\right\}\\ &= \mathbb{P}\left\{\frac{\lambda_N\hat{w}_{j'}\mathrm{sgn}(\hat{\beta}_{(j')})}{a_N} = \frac{\dot{\ell}_{\mathrm{w}}(\hat{\boldsymbol{\theta}}_{\mathrm{w}})}{a_N}\right\} \to 0.\end{aligned}$$

Thus, we prove the part of consistency of variable selection. □

B.3 Proof of Proposition 1

We first give a lemma for general optimal functions.

Lemma 1. *Assume that $h(\boldsymbol{x})^2$ and $\varphi(\boldsymbol{x})$ are integrable function with $\mathbb{E}\{\varphi(\boldsymbol{x})\} = 1$. The optimal function $\varphi^{**}(\boldsymbol{x})$ that minimize the value $\mathbb{E}\left\{\frac{h^2(\boldsymbol{x})}{\varphi(\boldsymbol{x})}\right\}$ is given as $\varphi^{**}(\boldsymbol{x}) = \frac{h(\boldsymbol{x})}{\mathbb{E}\{h(\boldsymbol{x})\}}$.*

Proof. Appying Cauchy-Schwartz inequality, we have that

$$\mathbb{E}\{h(\boldsymbol{x})\}^2 = \mathbb{E}\left\{\frac{h(\boldsymbol{x})}{\sqrt{\varphi(\boldsymbol{x})}}\sqrt{\varphi(\boldsymbol{x})}\right\}^2 \leq \mathbb{E}\left\{\frac{h^2(\boldsymbol{x})}{\varphi(\boldsymbol{x})}\right\}\mathbb{E}\{\varphi(\boldsymbol{x})\} = \mathbb{E}\left\{\frac{h^2(\boldsymbol{x})}{\varphi(\boldsymbol{x})}\right\}.$$

Therefore, we have that $\mathbb{E}\left\{\frac{h^2(\boldsymbol{x})}{\varphi(\boldsymbol{x})}\right\} \geq \mathbb{E}\{h(\boldsymbol{x})\}^2$ and the equality holds if and only if $\sqrt{\varphi(\boldsymbol{x})} = Kh(\boldsymbol{x})/\sqrt{\varphi(\boldsymbol{x})}$, where K is a constant. Therefore, $\varphi^{**}(\boldsymbol{x}) = Kh(\boldsymbol{x})$, and since $\mathbb{E}\{\varphi^{**}(\boldsymbol{x})\} = 1$, we know that $\varphi^{**}(\boldsymbol{x}) = h(\boldsymbol{x})/\mathbb{E}\{h(\boldsymbol{x})\}$. □

Now, we prove Proposition 1.

Proof. We first calculate the optimal function that minimizes $\mathrm{tr}(\boldsymbol{V}_{\mathrm{w}(\mathcal{A})})$. We have that

$$\begin{aligned}\mathrm{tr}(\boldsymbol{V}_{\mathrm{w}(\mathcal{A})}) &= \mathrm{tr}\left\{\boldsymbol{M}^{-1}_{(\mathcal{A})}\boldsymbol{M}_{\mathrm{w}(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\right\}\\ &= \mathrm{tr}\left\{\boldsymbol{M}^{-1}_{(\mathcal{A})}\mathbb{E}\left[\left\{1+\frac{ce^{f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}}{\varphi(\boldsymbol{x})}\right\}e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}\dot{g}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right]\boldsymbol{M}^{-1}_{(\mathcal{A})}\right\}.\end{aligned}$$

We focus on the values that related to $\varphi(\boldsymbol{x})$. We know that

$$\mathbb{E}\left\{\varphi^{-1}(\boldsymbol{x})e^{2f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}\dot{g}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right\} = e^{-2\alpha_{\mathrm{t}}}\{1+o_P(1)\}\mathbb{E}\left\{\varphi^{-1}(\boldsymbol{x})p^2(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\dot{g}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right\}.$$

Therefore, we need to minimize

$$\begin{aligned}&\mathrm{tr}\left[\mathbb{E}\left\{\frac{p^2(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\boldsymbol{M}^{-1}_{(\mathcal{A})}\dot{g}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\boldsymbol{M}^{-1}_{(\mathcal{A})}}{\varphi(\boldsymbol{x})}\right\}\right]\\ &= \mathbb{E}\left[\mathrm{tr}\left\{\frac{p^2(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\boldsymbol{M}^{-1}_{(\mathcal{A})}\dot{g}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\boldsymbol{M}^{-1}_{(\mathcal{A})}}{\varphi(\boldsymbol{x})}\right\}\right] = \mathbb{E}\left[\frac{p^2(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\boldsymbol{M}^{-1}_{(\mathcal{A})}\dot{g}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|^2}{\varphi(\boldsymbol{x})}\right].\end{aligned}$$

Appying Lemma 1. We know that the minimizer is given as

$$\varphi_{\mathrm{A-OS}}(\boldsymbol{x}) = \frac{p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\boldsymbol{M}_{(\mathcal{A})}^{-1}\dot{g}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|}{\mathbb{E}\left\{p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\boldsymbol{M}_{(\mathcal{A})}^{-1}\dot{g}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\right\}}.$$

Next, we calculate the optimal function that minimize $\mathrm{tr}(\boldsymbol{M}_{\mathrm{w}(\mathcal{A})})$. We have that

$$\mathrm{tr}(\boldsymbol{M}_{\mathrm{w}(\mathcal{A})}) = \mathrm{tr}\left\{\mathbb{E}\left[\left\{1+\frac{ce^{f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}}{\varphi(\boldsymbol{x})}\right\}e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}\dot{g}_{(\mathcal{A})}^{\otimes 2}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right]\right\}.$$

Therefore, we need to minimize

$$\mathrm{tr}\left[\mathbb{E}\left\{\frac{p^2(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\dot{g}_{(\mathcal{A})}^{\otimes 2}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})}{\varphi(\boldsymbol{x})}\right\}\right] = \mathbb{E}\left[\frac{p^2(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\dot{g}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|^2}{\varphi(\boldsymbol{x})}\right].$$

Appying Lemma 1. We know that the minimizer is given as

$$\varphi_{\mathrm{L-OS}}(\boldsymbol{x}) = \frac{p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\dot{g}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|}{\mathbb{E}\left\{p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\dot{g}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\right\}}.$$

□

B.4 Proof of Theorem 2

Proof. In the proof of Thereom 1, we know that

$$a_N(\hat{\boldsymbol{\theta}}_{\mathrm{w}(\mathcal{A})}^{\mathrm{adp}} - \boldsymbol{\theta}_{\mathrm{t}(\mathcal{A})}) \rightsquigarrow \boldsymbol{M}_{(\mathcal{A})}^{-1}\boldsymbol{M}_{\mathrm{w}(\mathcal{A})}^{1/2}\boldsymbol{Z}_{(\mathcal{A})}.$$

To simplify the representation, We define a function

$$h(\boldsymbol{\theta}) = e^{-2\alpha_{\mathrm{t}}}\mathrm{MSPE}(\boldsymbol{\theta}) = e^{-2\alpha_{\mathrm{t}}}\mathbb{E}\left[\{p(\boldsymbol{x};\boldsymbol{\theta}) - p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\}^2\right].$$

We have that

$$\frac{\partial\left\{p(\boldsymbol{x};\boldsymbol{\theta}) - p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right\}^2}{\partial\boldsymbol{\theta}} = 2\left\{p(\boldsymbol{x};\boldsymbol{\theta}) - p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right\}\phi(\boldsymbol{x};\boldsymbol{\theta})\dot{g}(\boldsymbol{x};\boldsymbol{\theta}),$$

and

$$\begin{aligned}\frac{\partial^2\left\{p(\boldsymbol{x};\boldsymbol{\theta}) - p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right\}^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}} &= 2\phi^2(\boldsymbol{x};\boldsymbol{\theta})\dot{g}^{\otimes 2}(\boldsymbol{x};\boldsymbol{\theta}) + 2\left\{p(\boldsymbol{x};\boldsymbol{\theta}) - p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right\}\frac{\partial\phi(\boldsymbol{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\dot{g}(\boldsymbol{x};\boldsymbol{\theta}) \\ &\quad + 2\left\{p(\boldsymbol{x};\boldsymbol{\theta}) - p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right\}\phi(\boldsymbol{x};\boldsymbol{\theta})\ddot{g}(\boldsymbol{x};\boldsymbol{\theta}).\end{aligned}$$

Note that $|p(\boldsymbol{x};\boldsymbol{\theta}) - p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})| \leq 2$ and thus,

$$\left|\frac{\partial\left\{p(\boldsymbol{x};\boldsymbol{\theta}) - p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right\}^2}{\partial\boldsymbol{\theta}}\right| \leq 2\left|p(\boldsymbol{x};\boldsymbol{\theta}) - p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right|\left|\phi(\boldsymbol{x};\boldsymbol{\theta})\dot{g}(\boldsymbol{x};\boldsymbol{\theta})\right| \leq 4B(\boldsymbol{x}),$$

and

$$\begin{aligned}&\left|\frac{\partial^2\left\{p(\boldsymbol{x};\boldsymbol{\theta}) - p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right\}^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\right| \\ &\leq 2\left|\phi^2(\boldsymbol{x};\boldsymbol{\theta})\dot{g}^{\otimes 2}(\boldsymbol{x};\boldsymbol{\theta})\right| + 2\left|\{p(\boldsymbol{x};\boldsymbol{\theta}) - p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\}\right|\left|\frac{\partial\phi(\boldsymbol{x};\boldsymbol{\theta})}{\partial\boldsymbol{t}}\dot{g}(\boldsymbol{x};\boldsymbol{\theta})\right| \\ &\quad + 2\left|\{p(\boldsymbol{x};\boldsymbol{\theta}) - p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\}\right|\left|\phi(\boldsymbol{x};\boldsymbol{\theta})\ddot{g}(\boldsymbol{x};\boldsymbol{\theta})\right| \leq 10B(\boldsymbol{x}).\end{aligned}$$

Hence, due to donimating convergence theorem, we know that the expectation and derivitive are exchangable. Thus, we have that

$$\frac{\partial h(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = e^{-2\alpha_{\mathrm{t}}}\mathbb{E}\left[2\left\{p(\boldsymbol{x};\boldsymbol{\theta}) - p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right\}\phi(\boldsymbol{x};\boldsymbol{\theta})\dot{g}(\boldsymbol{x};\boldsymbol{\theta})\right],$$

https://doi.org/10.52202/079017-3122

and

$$\frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\mathrm{T}}} = e^{-2\alpha_{\mathrm{t}}}\mathbb{E}\Big[2\phi^2(\boldsymbol{x};\boldsymbol{\theta})\dot{g}^{\otimes 2}(\boldsymbol{x};\boldsymbol{\theta}) + 2\left\{p(\boldsymbol{x};\boldsymbol{\theta}) - p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right\}\frac{\partial \phi(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\dot{g}(\boldsymbol{x};\boldsymbol{\theta}) + 2\left\{p(\boldsymbol{x};\boldsymbol{\theta}) - p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right\}\phi(\boldsymbol{x};\boldsymbol{\theta})\ddot{g}(\boldsymbol{x};\boldsymbol{\theta})\Big].$$

Due to donimating convergence theorem, we also know that the first and second derivitive of $h(\boldsymbol{\theta})$ are continous. We have that

$$\frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{t}}} = \mathbf{0}, \frac{\partial^2 h(\boldsymbol{t})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\mathrm{T}}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{t}}} = 2e^{-2\alpha_{\mathrm{t}}}\mathbb{E}\left[\phi^2(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\dot{g}^{\otimes 2}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right]$$

Note that $e^{-2\alpha_{\mathrm{t}}}\mathbb{E}\left[\phi^2(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\dot{g}^{\otimes 2}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right] \to \boldsymbol{\Omega}$ due to donimating covergence theorem. Now applying Theorem 1.12(ii) in [18], we have that

$$\begin{aligned}
&a_N^2\left\{h(\hat{\boldsymbol{\theta}}^{\mathrm{adp}}_{\mathrm{w}(\mathcal{A})}) - h(\boldsymbol{\theta}_{\mathrm{t}(\mathcal{A})})\right\} = a_N^2 e^{-2\alpha_{\mathrm{t}}}\mathbb{E}\left[\left\{p(\boldsymbol{x};\hat{\boldsymbol{\theta}}_{\mathrm{w}(\mathcal{A})}) - p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}(\mathcal{A})})\right\}^2\right] \\
&\rightsquigarrow \frac{1}{2!}2\boldsymbol{Z}^{\mathrm{T}}_{(\mathcal{A})}\boldsymbol{M}^{1/2}_{\mathrm{w}(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\boldsymbol{\Omega}_{(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\boldsymbol{M}^{1/2}_{\mathrm{w}(\mathcal{A})}\boldsymbol{Z}_{(\mathcal{A})} \\
&= \boldsymbol{Z}^{\mathrm{T}}_{(\mathcal{A})}\boldsymbol{M}^{1/2}_{\mathrm{w}(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\boldsymbol{\Omega}_{(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\boldsymbol{M}^{1/2}_{\mathrm{w}(\mathcal{A})}\boldsymbol{Z}_{(\mathcal{A})}.
\end{aligned}$$

Considering $N_1 = a_N^2\mathbb{E}\left\{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}\right\}\{1 + o_P(1)\}$, applying Slutsky's theorem, we have that

$$\begin{aligned}
&N_1 e^{-2\alpha_{\mathrm{t}}}\mathbb{E}\left[\left\{p(\boldsymbol{x};\hat{\boldsymbol{\theta}}_{\mathrm{w}(\mathcal{A})}) - p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}(\mathcal{A})})\right\}^2\right] \\
&\rightsquigarrow \mathbb{E}^{-1}\left\{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}\right\}\boldsymbol{Z}^{\mathrm{T}}_{(\mathcal{A})}\boldsymbol{M}^{1/2}_{\mathrm{w}(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\boldsymbol{\Omega}_{(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\boldsymbol{M}^{1/2}_{\mathrm{w}(\mathcal{A})}\boldsymbol{Z}_{(\mathcal{A})}.
\end{aligned}$$

Since $\boldsymbol{Z}_{\mathrm{w}(\mathcal{A})} = \mathbb{N}(\mathbf{0}, \boldsymbol{I})$, we have

$$\begin{aligned}
&\mathbb{E}\left\{\boldsymbol{Z}^{\mathrm{T}}_{(\mathcal{A})}\boldsymbol{M}^{1/2}_{\mathrm{w}(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\boldsymbol{\Omega}_{(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\boldsymbol{M}^{1/2}_{\mathrm{w}(\mathcal{A})}\boldsymbol{Z}_{(\mathcal{A})}\right\} = \mathrm{tr}\left\{\boldsymbol{M}^{-1}_{(\mathcal{A})}\boldsymbol{\Omega}_{(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\boldsymbol{M}_{\mathrm{w}(\mathcal{A})}\right\} \\
&= \mathrm{tr}\left\{\boldsymbol{M}^{-1}_{(\mathcal{A})}\boldsymbol{\Omega}_{(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\mathbb{E}\left[\left\{1 + \frac{ce^{f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}}{\varphi(\boldsymbol{x})}\right\}e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}\dot{g}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right].\right\}
\end{aligned}$$

We focus on the values that related to $\varphi(\boldsymbol{x})$. We know that

$$\mathbb{E}\left\{\varphi^{-1}(\boldsymbol{x})e^{2f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}\dot{g}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right\} = e^{-2\alpha_{\mathrm{t}}}\{1 + o_P(1)\}\mathbb{E}\left\{\varphi^{-1}(\boldsymbol{x})p^2(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\dot{g}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right\}.$$

Therefore, we need to minimize

$$\begin{aligned}
&\mathrm{tr}\left[\mathbb{E}\left\{\boldsymbol{M}^{-1}_{(\mathcal{A})}\boldsymbol{\Omega}_{(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\frac{p^2(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\dot{g}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})}{\varphi(\boldsymbol{x})}\right\}\right] \\
&= \mathbb{E}\left[\mathrm{tr}\left\{\boldsymbol{M}^{-1}_{(\mathcal{A})}\boldsymbol{\Omega}_{(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\frac{p^2(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\dot{g}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})}{\varphi(\boldsymbol{x})}\right\}\right] \\
&= \mathbb{E}\left[\mathrm{tr}\left\{\frac{p^2(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\boldsymbol{\Omega}^{1/2}_{(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\dot{g}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\boldsymbol{M}^{-1}_{(\mathcal{A})}\boldsymbol{\Omega}^{1/2}_{(\mathcal{A})}}{\varphi(\boldsymbol{x})}\right\}\right] \\
&= \mathbb{E}\left[\frac{p^2(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\boldsymbol{\Omega}^{1/2}_{(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\dot{g}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|^2}{\varphi(\boldsymbol{x})}\right].
\end{aligned}$$

Appying Lemma 1. We know that the minimizer is given as

$$\varphi_{\mathrm{P-OS}}(\boldsymbol{x}) = \frac{p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\boldsymbol{\Omega}^{1/2}_{(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\dot{g}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|}{\mathbb{E}\left\{p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\boldsymbol{\Omega}^{1/2}_{(\mathcal{A})}\boldsymbol{M}^{-1}_{(\mathcal{A})}\dot{g}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\right\}}.$$

□

B.5 Proof of Proposition 2

Proof. First, we know that $g(\boldsymbol{x};\boldsymbol{\theta}) = g(\boldsymbol{A}\boldsymbol{x};\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\theta})$. Since the equation holds for all $\boldsymbol{x}$ and $\boldsymbol{\theta}$, if we take derivitive with respect to $\boldsymbol{\theta}$ on both sides, the equation still holds. Thus, we have that

$$\dot{g}(\boldsymbol{x};\boldsymbol{\theta}) = \boldsymbol{B}\dot{g}(\boldsymbol{A}\boldsymbol{x};\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\theta}).$$

If we scale the whole covariate variable $\boldsymbol{x}$ to $\tilde{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x}$, we need to reparameterize $\boldsymbol{\theta}_{\mathrm{t}}$ to $\tilde{\boldsymbol{\theta}}_{\mathrm{t}} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{\theta}_{\mathrm{t}}$ to remain the problem invariant. We have that for $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{\theta}}_{\mathrm{t}}$

$$\dot{g}(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{\theta}}_{\mathrm{t}}) = \dot{g}(\boldsymbol{A}\boldsymbol{x};\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\theta}_{\mathrm{t}}) = \boldsymbol{B}^{-1}\dot{g}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}}).$$

Now, we know that

$$\tilde{\boldsymbol{M}} = \mathbb{E}\{e^{f(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{\beta}}_{\mathrm{t}})}\dot{g}^{\otimes 2}(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{\theta}}_{\mathrm{t}})\} = \boldsymbol{B}^{-1}\mathbb{E}\{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}\dot{g}^{\otimes 2}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\}(\boldsymbol{B}^{\mathrm{T}})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{M}(\boldsymbol{B}^{\mathrm{T}})^{-1},$$

and

$$\tilde{\boldsymbol{\Omega}} = \mathbb{E}\{e^{2f(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{\beta}}_{\mathrm{t}})}\dot{g}^{\otimes 2}(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{\theta}}_{\mathrm{t}})\} = \boldsymbol{B}^{-1}\mathbb{E}\{e^{2f(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})}\dot{g}^{\otimes 2}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\}(\boldsymbol{B}^{\mathrm{T}})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{\Omega}(\boldsymbol{B}^{\mathrm{T}})^{-1}.$$

Thus, we have that

$$\begin{aligned}
&\|\tilde{\boldsymbol{\Omega}}^{1/2}\tilde{\boldsymbol{M}}^{-1}\dot{g}(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{\theta}}_{\mathrm{t}})\|^2 = \dot{g}^{\mathrm{T}}(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{\theta}}_{\mathrm{t}})\tilde{\boldsymbol{M}}^{-1}\tilde{\boldsymbol{\Omega}}\tilde{\boldsymbol{M}}^{-1}\dot{g}(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{\theta}}_{\mathrm{t}}) \\
&= \dot{g}^{\mathrm{T}}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})(\boldsymbol{B}^{-1})^{\mathrm{T}}(\boldsymbol{B}^{\mathrm{T}})\boldsymbol{M}^{-1}\boldsymbol{B}\boldsymbol{B}^{-1}\boldsymbol{\Omega}(\boldsymbol{B}^{\mathrm{T}})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{B}\boldsymbol{B}^{-1}\dot{g}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}}) \\
&= \dot{g}^{\mathrm{T}}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\boldsymbol{M}^{-1}\boldsymbol{\Omega}\boldsymbol{M}^{-1}\dot{g}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}}) = \|\boldsymbol{\Omega}^{1/2}\boldsymbol{M}^{-1}\dot{g}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|^2.
\end{aligned}$$

Therefore, the leveraging term is invariant. For the probability term, we know that is only related to value $g(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}}) = g(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{\theta}}_{\mathrm{t}})$, we know that it does not change after scaling inactive variables. This complete the proof. □

B.6 Proof of Theorem 3

Proof of Theorem 3. Consider maximum sampled conditional likeihood function with adaptive lasso penalty:

$$\begin{aligned}
Q^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}}(\boldsymbol{\theta}) &= -\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}[y_i g(\boldsymbol{x}_i;\boldsymbol{\theta}) - \log\{1+e^{g(\boldsymbol{x}_i;\boldsymbol{\theta})+l_i}\}] + \lambda_N\sum_{j=1}^{p}\hat{w}_j|\beta_{(j)}| \\
&= -\ell^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}}(\boldsymbol{\theta}) + \lambda_N\sum_{j=1}^{p}\hat{w}_j|\beta_{(j)}|,
\end{aligned}$$

where $\hat{w}_j = 1/|\hat{\beta}_{\mathrm{pl}(j)}|^{\gamma}, 1 \le j \le p$. Then, we have that $\hat{\boldsymbol{u}}_N = a_N(\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}} - \boldsymbol{\theta}_{\mathrm{t}})$ is the minimizer of

$$\gamma^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}}(\boldsymbol{u}) = Q^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}}(\boldsymbol{\theta}_{\mathrm{t}} + a_N^{-1}\boldsymbol{u}) - Q^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}}(\boldsymbol{\theta}_{\mathrm{t}}).$$

Asymptotic normality: We prove the asymptotic normality part in this paragraph. By Taylor's expansion,

$$\begin{aligned}
\gamma^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}}(\boldsymbol{u}) = &-\frac{1}{a_N}\boldsymbol{u}^{\mathrm{T}}\dot{\ell}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}}(\boldsymbol{\theta}_{\mathrm{t}}) + \frac{1}{2a_N^2}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}\phi_{\pi}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}})\{\boldsymbol{u}^{\mathrm{T}}\dot{g}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}})\}^2 - \Delta^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}} + R^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}} \\
&+ \frac{\lambda_N}{a_N}\sum_{j=1}^{p}\hat{w}_j a_N\left(\left|\beta_{\mathrm{t}(j)} + \frac{u_{(j)}}{a_N}\right| - |\beta_{\mathrm{t}(j)}|\right).
\end{aligned}$$

First, we consider the limit behavior of the MSCL function. In [22], the authors proved that under Assumptions 1 and 3,

$$a_N^{-1}\dot{\ell}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}}(\boldsymbol{\theta}_{\mathrm{t}}) \rightsquigarrow (\boldsymbol{\Lambda}^{\mathrm{pl}}_{\mathrm{mscl}})^{1/2}\boldsymbol{Z}.$$

$$\frac{1}{a_N^2}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}\phi_{\pi}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}})\dot{g}^{\otimes 2}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}}) \xrightarrow{P} \boldsymbol{\Lambda}^{\mathrm{pl}}_{\mathrm{mscl}},$$

https://doi.org/10.52202/079017-3122

and

$$\Delta_{\text{mscl}}^{\hat{\boldsymbol{\theta}}_{\text{pl}}} = o_P(1), \quad R_{\text{mscl}}^{\hat{\boldsymbol{\theta}}_{\text{pl}}} = o_P(1).$$

Thus,

$$-\ell_{\text{w}}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}(\boldsymbol{\theta}_{\text{t}}) \rightsquigarrow -\boldsymbol{u}^{\text{T}}(\boldsymbol{\Lambda}_{\text{mscl}}^{\text{pl}})^{1/2}\boldsymbol{Z} + \frac{1}{2}\boldsymbol{u}^{\text{T}}\boldsymbol{\Lambda}_{\text{mscl}}^{\text{pl}}\boldsymbol{u} + o_P(1).$$

Next, we consider the limit behavior of the adaptive lasso penalty. Since we assume $\hat{\boldsymbol{\beta}}_{\text{pl}}$ tp be a consistent estimator, we know that when $j \in \mathcal{A}$, i.e., $\beta_{\text{t}(j)} \neq 0$,

$$\hat{w}_j = |\hat{\beta}_{\text{pl}(j)}|^{-\gamma} \xrightarrow{P} |\beta_{\text{t}(j)}|^{-\gamma} > 0,$$

and

$$a_N\left(\left|\beta_{\text{t}(j)} + \frac{u_{(j)}}{a_N}\right| - |\beta_{\text{t}(j)}|\right) \to \text{sgn}(\beta_{\text{t}(j)})u_{(j)}.$$

Therefore, for $j \in \mathcal{A}$, we have that

$$\frac{\lambda_N}{a_N}\hat{w}_j a_N\left(\left|\beta_{\text{t}(j)} + \frac{u_{(j)}}{a_N}\right| - |\beta_{\text{t}(j)}|\right) \xrightarrow{P} 0,$$

since $\lambda_N/a_N = \lambda_N/\sqrt{Ne^{\alpha_{\text{t}}}} \to 0$. On the other hand, when $j \in \mathcal{A}^c$, i.e., $\beta_{\text{t}(j)} = 0$, we have that for $u_{(j)} \neq 0$,

$$\frac{\lambda_N}{a_N}\hat{w}_j a_N\left(\left|\beta_{\text{t}(j)} + \frac{u_{(j)}}{a_N}\right| - |\beta_{\text{t}(j)}|\right) = \frac{\lambda_N}{a_N}\hat{w}_j|u_{(j)}| = \frac{\lambda_N}{a_N|\hat{\beta}_{\text{pl}(j)}|^{\gamma}}|u_{(j)}| \xrightarrow{P} \infty,$$

since $\lambda_N/(\sqrt{Ne^{\alpha_{\text{t}}}}|\hat{\beta}_{\text{pl}(j)}|^{\gamma}) \xrightarrow{P} \infty$. Then, we have that $\gamma_{\text{mscl}}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}(\boldsymbol{u}) \rightsquigarrow \gamma_{\text{mscl}}(\boldsymbol{u})$, where

$$\gamma_{\text{mscl}}(\boldsymbol{u}) = \begin{cases} \frac{1}{2}\boldsymbol{u}_{(\mathcal{A})}^{\text{T}}\boldsymbol{\Lambda}_{\text{mscl}}^{\text{pl}}\boldsymbol{u}_{(\mathcal{A})} - \boldsymbol{u}_{(\mathcal{A})}^{\text{T}}(\boldsymbol{\Lambda}_{\text{mscl}}^{\text{pl}})^{1/2}\boldsymbol{Z}_{(\mathcal{A})} & \text{if } u_{(j)} = 0, \forall j \notin \mathcal{A} \\ \infty & \text{otherwise.} \end{cases}$$

Note that the unique minimizer of $\gamma_{\text{mscl}}(\boldsymbol{u})$ is $((\boldsymbol{\Lambda}_{\text{mscl}}^{\text{pl}})^{-1}\boldsymbol{Z}_{(\mathcal{A})}^{\text{T}}, \boldsymbol{0})^{\text{T}}$ if we put all the indexes of active variables in front. Thus, following the results of [11] and [10], we have the minimizer of $\gamma_{\text{mscl}}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}(\boldsymbol{u})$, $\hat{\boldsymbol{u}}_N$, satisfies that

$$\hat{\boldsymbol{u}}_{N(\mathcal{A})} \rightsquigarrow (\boldsymbol{\Lambda}_{\text{mscl}}^{\text{pl}})^{1/2}\boldsymbol{Z}_{(\mathcal{A})} \text{ and } \hat{\boldsymbol{u}}_{N(\mathcal{A}^c)} \rightsquigarrow \boldsymbol{0}.$$

Thus,

$$\hat{\boldsymbol{u}}_{N(\mathcal{A})} = a_N(\hat{\boldsymbol{\theta}}_{\text{mscl}(\mathcal{A})} - \boldsymbol{\theta}_{\text{t}(\mathcal{A})}) \rightsquigarrow \mathbb{N}(\boldsymbol{0}, (\boldsymbol{\Lambda}_{\text{mscl}}^{\text{pl}})^{-1}).$$

We know that

$$\sqrt{N_1}\boldsymbol{V}_{\text{mscl}}^{-1/2}(\boldsymbol{\Lambda}_{\text{mscl}}^{\text{pl}})^{-1} = a_N\mathbb{E}^{1/2}\left\{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\text{t}})}\right\}\mathbb{E}^{-1/2}\left\{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\text{t}})}\right\}\{1 + o_P(1)\} = a_N\{1 + o_P(1)\}.$$

Hence, applying Slusky's theorem, we have

$$\sqrt{N_1}\boldsymbol{V}_{\text{mscl}(\mathcal{A})}^{-1/2}(\hat{\boldsymbol{\theta}}_{\text{mscl}(\mathcal{A})} - \boldsymbol{\theta}_{\text{t}(\mathcal{A})}) \rightsquigarrow \mathbb{N}(\boldsymbol{0}, \boldsymbol{I}).$$

Therefore, we prove the part of aymptotic normality.

Consistency in variable selection We prove the consistency in variable selection in this paragraph. From the result of asymptotic normality, we know that $\hat{\beta}_{\text{mscl}(j)} \xrightarrow{P} \beta_{\text{t}(j)}$ for every $j \in \mathcal{A}$ and therefore $\mathbb{P}(j \in \hat{\mathcal{A}}_{\text{mscl}}) \to 1$. Thus, we only consider $j' \in \mathcal{A}^c$. When $j' \in \hat{\mathcal{A}}_{\text{mscl}}$, we know that by K-K-T optimality conditions, we have

$$\lambda_N\hat{w}_{j'}\text{sgn}(\hat{\beta}_{\text{mscl}(j')}) = \dot{\ell}_{\text{mscl}}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}(\hat{\boldsymbol{\theta}}_{\text{mscl}}),$$

which means

$$\frac{\lambda_N\hat{w}_{j'}\text{sgn}(\hat{\beta}_{\text{mscl}(j')})}{a_N} = \frac{\dot{\ell}_{\text{mscl}}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}(\hat{\boldsymbol{\theta}}_{\text{mscl}})}{a_N}$$

$$= \frac{\dot{\ell}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}}(\boldsymbol{\theta}_{\mathrm{t}})}{a_N} + \frac{a_N\left\{\dot{\ell}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}}(\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}}) - \dot{\ell}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}}(\boldsymbol{\theta}_{\mathrm{t}})\right\}}{a_N^2} = I_1 + I_2.$$

We have known that $I_1 = \dot{\ell}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}}(\boldsymbol{\theta}_{\mathrm{t}})/a_N \rightsquigarrow \boldsymbol{Z}_{\mathrm{mscl}}$. We now prove that proof that $I_2 = O_P(1)$. We apply Taylor expansion to the k-th element of $\dot{\ell}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}}(\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}})$ and have that

$$\frac{a_N\left\{\dot{\ell}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{(k)}(\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}}) - \dot{\ell}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{(k)}(\boldsymbol{\theta}_{\mathrm{t}})\right\}}{a_N^2} = -\frac{1}{a_N^2}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}\phi_\pi^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}})\dot{g}_{(k)}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}})\dot{g}^{\mathrm{T}}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}})\hat{\boldsymbol{u}}_N + \tilde{\Delta}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{(k)} + \tilde{R}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{(k)},$$

where,

$$\hat{\boldsymbol{u}}_N = a_N(\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}} - \boldsymbol{\theta}_{\mathrm{t}}) = O_P(1),$$

$$\tilde{\Delta}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}(k)} = \frac{1}{a_N^2}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}\left\{y_i - p_\pi^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}})\right\}\sum_{j=1}^{d}\ddot{g}_{(kj)}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}})\tilde{u}_{(j)},$$

and

$$\begin{aligned}
\tilde{R}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}(k)} = &-\frac{1}{2a_N^3}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}\phi_\pi^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\left\{1 - 2p_\pi^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right\}\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\hat{\boldsymbol{u}}_N^{\mathrm{T}}\dot{g}^{\otimes 2}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\hat{\boldsymbol{u}}_N \\
&- \frac{2}{2a_N^3}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}\phi_\pi^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\left\{\hat{\boldsymbol{u}}_N^{\mathrm{T}}\frac{\partial\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)}{\partial\boldsymbol{\theta}}\right\}\left\{\hat{\boldsymbol{u}}_N^{\mathrm{T}}\dot{g}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right\} \\
&- \frac{1}{2a_N^2}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}\phi_\pi^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\left\{\hat{\boldsymbol{u}}_N^{\mathrm{T}}\ddot{g}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\hat{\boldsymbol{u}}_N\right\} \\
&+ \frac{1}{2a_N^3}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}\left\{y_i - p_\pi^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right\}\hat{\boldsymbol{u}}_N^{\mathrm{T}}\frac{\partial^2\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)}{\partial\boldsymbol{\theta}^2}\hat{\boldsymbol{u}}_N.
\end{aligned}$$

where $\acute{\boldsymbol{\theta}}_k$ is between $\hat{\boldsymbol{\theta}}_{\mathrm{mscl}}$ and $\boldsymbol{\theta}_{\mathrm{t}}$. First, we prove that $\tilde{R}_{\mathrm{mscl}(k)}$ is $o_P(1)$. We have that

$$\begin{aligned}
|\tilde{R}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}_{\mathrm{mscl}(k)}| \le &\frac{\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^3}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}\phi_\pi^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\left|1 - 2p_\pi^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right|\left|\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right|\left\|\dot{g}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right\|^2 \\
&+ \frac{2\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^3}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}\phi_\pi^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\left\|\frac{\partial\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)}{\partial\boldsymbol{\theta}}\right\|\left\|\dot{g}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right\| \\
&+ \frac{\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^3}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}\phi_\pi^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\left|\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right|\left\|\ddot{g}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\right\| \\
&+ \frac{\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^3}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}p_\pi^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)\left\|\frac{\partial^2\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)}{\partial\boldsymbol{\theta}^2}\right\| + \frac{\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^3}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}y_i\left\|\frac{\partial^2\dot{g}_{(k)}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)}{\partial\boldsymbol{\theta}^2}\right\| \\
\le &\frac{\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^3}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}p_\pi^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k)C(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k) + \frac{\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^3}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}y_iB(\boldsymbol{x}_i) \\
\le &\frac{\|\hat{\boldsymbol{u}}_N\|^2 e^{\acute{\alpha}_k - \alpha_{\mathrm{t}}}e^{\alpha_{\mathrm{t}}}}{2a_N^3}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}e^{f(\boldsymbol{x}_i;\acute{\boldsymbol{\beta}}_k) - \log\{\rho\varphi(\boldsymbol{x}_i)\}}C(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k) + \frac{\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^3}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}y_iB(\boldsymbol{x}_i) \\
\le &\frac{\|\hat{\boldsymbol{u}}_N\|^2 e^{\acute{\alpha}_k - \alpha_{\mathrm{t}}}}{2Na_N}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}e^{f(\boldsymbol{x}_i;\acute{\boldsymbol{\beta}}_k) - \log\{\rho\varphi(\boldsymbol{x}_i)\}}C(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k) + \frac{\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^3}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}y_iB(\boldsymbol{x}_i) \\
= &\frac{\|\hat{\boldsymbol{u}}_N\|^2 e^{\acute{\alpha}_k - \alpha_{\mathrm{t}}}}{2Na_N\rho}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}\varphi^{-1}(\boldsymbol{x}_i)e^{f(\boldsymbol{x}_i;\acute{\boldsymbol{\beta}}_k)}C(\boldsymbol{x}_i;\acute{\boldsymbol{\theta}}_k) + \frac{\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^3}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}y_iB(\boldsymbol{x}_i)
\end{aligned}$$

https://doi.org/10.52202/079017-3122

$$\leq \frac{\|\hat{\boldsymbol{u}}_N\|^2 e^{\acute{\alpha}_k - \alpha_{\mathrm{t}}}}{2Na_N\rho} \sum_{i=1}^{N} \delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}} \varphi^{-1}(\boldsymbol{x}_i) B(\boldsymbol{x}_i) + \frac{\|\hat{\boldsymbol{u}}_N\|^2}{2a_N^3} \sum_{i=1}^{N} y_i B(\boldsymbol{x}_i)$$
$$= o_P(1),$$

where

$$C(\boldsymbol{x}_i; \acute{\boldsymbol{\theta}}) = \left| \dot{g}_{(k)}(\boldsymbol{x}_i; \acute{\boldsymbol{\theta}}) \right| \left\{ \left\| \dot{g}(\boldsymbol{x}_i; \acute{\boldsymbol{\theta}}_k) \right\|^2 + \left\| \ddot{g}(\boldsymbol{x}_i; \acute{\boldsymbol{\theta}}) \right\| \right\}$$
$$+ \left\| \frac{\partial \dot{g}_{(k)}(\boldsymbol{x}_i; \acute{\boldsymbol{\theta}}_k)}{\partial \boldsymbol{\theta}} \right\| \left\| \dot{g}(\boldsymbol{x}_i; \acute{\boldsymbol{\theta}}_k) \right\| + \left\| \frac{\partial^2 \dot{g}_{(k)}(\boldsymbol{x}_i; \acute{\boldsymbol{\theta}}_k)}{\partial \boldsymbol{\theta}^2} \right\|.$$

Therefore, we proved that $\tilde{R}_{\mathrm{mscl}(k)} = o_P(1)$. Next, we prove that $\tilde{\Delta}_{\mathrm{mscl}(k)} = o_P(1)$. We know that $\mathbb{E}\left[a_N^{-2} \sum_{i=1}^{N} \delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}} \left\{ y_i - p_\pi^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i; \boldsymbol{\theta}_{\mathrm{t}}) \right\} \ddot{g}(\boldsymbol{x}_i; \boldsymbol{\theta}_{\mathrm{t}}) \;\middle|\; \hat{\boldsymbol{\theta}}_{\mathrm{pl}} \right] = \mathbf{0}$. We also have that for the every element of $a_N^{-2} \sum_{i=1}^{N} \{ y_i - p(\boldsymbol{x}_i; \boldsymbol{\theta}_{\mathrm{t}}) \} \ddot{g}(\boldsymbol{x}_i; \boldsymbol{\theta}_{\mathrm{t}})$, we have

$$\mathbb{V}\left[a_N^{-2} \sum_{i=1}^{N} \delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}} \left\{ y_i - p_\pi^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i; \boldsymbol{\theta}_{\mathrm{t}}) \right\} \ddot{g}_{(jl)}(\boldsymbol{x}_i; \boldsymbol{\theta}_{\mathrm{t}}) \;\middle|\; \hat{\boldsymbol{\theta}}_{\mathrm{pl}} \right]$$
$$\leq \frac{1}{a_N^4} \sum_{i=1}^{N} \mathbb{E}\left\{ \delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}} p_\pi^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i; \boldsymbol{\theta}_{\mathrm{t}}) \ddot{g}_{(jl)}^2(\boldsymbol{x}_i; \boldsymbol{\theta}_{\mathrm{t}}) \;\middle|\; \hat{\boldsymbol{\theta}}_{\mathrm{pl}} \right\} \leq \frac{1}{a_N^2} \mathbb{E}[e^{f(\boldsymbol{x}; \boldsymbol{\beta}_{\mathrm{t}})} \|\ddot{g}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}})\|^2] \to 0.$$

Thus, due to Chebyshev's inequality, we know that $\tilde{\Delta}_{\mathrm{mscl}}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}} = o_P(1)$. Since we know that $\frac{1}{a_N^2} \sum_{i=1}^{N} \delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}} \phi_\pi^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i; \boldsymbol{\theta}_{\mathrm{t}}) \dot{g}^{\otimes 2}(\boldsymbol{x}_i; \boldsymbol{\theta}_{\mathrm{t}}) = O_P(1)$. Hence, we have that

$$\frac{\dot{\ell}_{\mathrm{mscl}}(\hat{\boldsymbol{\theta}}_{\mathrm{mscl}})}{a_N} = O_P(1).$$

Note that we also have

$$\frac{\lambda_N \hat{w}_{j'}}{a_N} = \frac{\lambda_N}{a_N} \frac{1}{|\hat{\beta}_{\mathrm{pl}(j')}|^\gamma} \xrightarrow{P} \infty.$$

Therefore,

$$\mathbb{P}(j' \in \mathcal{A}_N) \leq \mathbb{P}\left\{ \lambda_N \hat{w}_{j'} \mathrm{sgn}(\hat{\beta}_{\mathrm{mscl}(j')}) = \dot{\ell}_{\mathrm{mscl}}^{\hat{\boldsymbol{\theta}}_p}(\hat{\boldsymbol{\theta}}_{\mathrm{mscl}}) \right\}$$
$$= \mathbb{P}\left\{ \frac{\lambda_N \hat{w}_{j'} \mathrm{sgn}(\hat{\beta}_{\mathrm{mscl}(j')})}{a_N} = \frac{\dot{\ell}_{\mathrm{mscl}}^{\hat{\boldsymbol{\theta}}_p}(\hat{\boldsymbol{\theta}}_{\mathrm{mscl}})}{a_N} \right\} \to 0.$$

Thus, we prove the part of consistency of variable selection. □

B.7 Proof of Theorem 4

Proof. Letting $h = 1 + c\{\varphi(\boldsymbol{x})\}^{-1} e^{f(\boldsymbol{x}; \boldsymbol{\beta}_{\mathrm{t}})}$, $\boldsymbol{v} = \sqrt{e^{f(\boldsymbol{x}; \boldsymbol{\beta}_{\mathrm{t}})}} \dot{g}_{(\mathcal{A})}(\boldsymbol{x}; \tilde{\boldsymbol{\theta}})$, $\boldsymbol{f} = h^{\frac{1}{2}} \boldsymbol{v}$, and $\boldsymbol{g} = h^{-\frac{1}{2}} \boldsymbol{v}$, we have that

$$\mathbb{E}(\boldsymbol{g}\boldsymbol{f}^{\mathrm{T}}) = \mathbb{E}(\boldsymbol{f}\boldsymbol{g}^{\mathrm{T}}) = \mathbb{E}(\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}) = \mathbb{E}\left\{ e^{f(\boldsymbol{x}; \boldsymbol{\beta}_{\mathrm{t}})} \dot{g}_{(\mathcal{A})}^{\otimes 2}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}}) \right\} = \boldsymbol{M}_{(\mathcal{A})},$$
$$\mathbb{E}(\boldsymbol{f}\boldsymbol{f}^{\mathrm{T}}) = \mathbb{E}(h\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}) = \mathbb{E}\left[\left\{ 1 + \frac{c e^{f(\boldsymbol{z}; \boldsymbol{\beta}_{\mathrm{t}})}}{\varphi(\boldsymbol{x})} \right\} e^{f(\boldsymbol{x}; \boldsymbol{\beta}_{\mathrm{t}})} \dot{g}_{(\mathcal{A})}^{\otimes 2}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}}) \right] = \boldsymbol{M}_{\mathrm{w}(\mathcal{A})},$$

and

$$\mathbb{E}(\boldsymbol{g}\boldsymbol{g}^{\mathrm{T}}) = \mathbb{E}(h^{-1}\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}) = \mathbb{E}\left[\frac{e^{f(\boldsymbol{x}; \boldsymbol{\beta}_{\mathrm{t}})} \dot{g}_{(\mathcal{A})}^{\otimes 2}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}})}{1 + c\varphi^{-1}(\boldsymbol{x}) e^{f(\boldsymbol{x}; \boldsymbol{\beta}_{\mathrm{t}})}} \right] = \boldsymbol{\Lambda}_{\mathrm{mscl}(\mathcal{A})}.$$

Now, applying the matrix form of Cauchy-Schwartz's inequality (see [19]), we have that

$$\boldsymbol{\Lambda}_{\mathrm{mscl}(\mathcal{A})} = \mathbb{E}(\boldsymbol{g}\boldsymbol{g}^{\mathrm{T}}) \geq \mathbb{E}(\boldsymbol{g}\boldsymbol{f}^{\mathrm{T}}) \{\mathbb{E}(\boldsymbol{f}\boldsymbol{f}^{\mathrm{T}})\}^{-1} \mathbb{E}(\boldsymbol{f}\boldsymbol{g}^{\mathrm{T}})$$
$$= \boldsymbol{M}_{(\mathcal{A})} \{\boldsymbol{M}_{\mathrm{w}(\mathcal{A})}\}^{-1} \boldsymbol{M}_{(\mathcal{A})} = \mathbb{E}\left\{ e^{f(\boldsymbol{x}; \boldsymbol{\beta}_{\mathrm{t}})} \right\} \{\boldsymbol{V}_{\mathrm{w}(\mathcal{A})}\}^{-1}.$$

Therefore, simple algebra shows that

$$\boldsymbol{V}_{\mathrm{mscl}(\mathcal{A})} = \mathbb{E}\left\{ e^{f(\boldsymbol{x}; \boldsymbol{\beta}_{\mathrm{t}})} \right\} \{\boldsymbol{\Lambda}_{\mathrm{mscl}(\mathcal{A})}\}^{-1} \leq \boldsymbol{V}_{\mathrm{w}(\mathcal{A})},$$

which complete the proof □

C Details about the practical algorithm and its complexity

C.1 Two-step algorithm

We take a pilot sample by uniform sampling with the sampling rate $\rho_1 = N_{\rm pl}/2N_1$ for the ones and $\rho_0 = N_{\rm pl}/2N_0$ for the zeros. Denote a pilot sample of actual sample size $N_{\rm pl}^*$ as $\{(\boldsymbol{x}_i^{\rm pl}, y_i^{\rm pl})\}_{i=1}^{N_{\rm pl}^*}$, the pilot estimate of $\boldsymbol{\theta}$ as $\hat{\boldsymbol{\theta}}_{\rm pl}$, and the pilot estimate of the active set as $\hat{\mathcal{A}}_{\rm pl} = \{j : \hat{\beta}_{{\rm pl}(j)} \neq 0\}$. We propose the following moment estimators of $\boldsymbol{M}_{(\mathcal{A})}$ and $\boldsymbol{\Omega}_{(\mathcal{A})}$:

$$\hat{\boldsymbol{M}}_{(\hat{\mathcal{A}}_{\rm pl})}^{\rm pl} = \frac{1}{N_{\rm pl}} \sum_{i=1}^{N_{\rm pl}^*} \frac{e^{f(\boldsymbol{x}_i^{{\rm pl}T}\hat{\boldsymbol{\beta}}_{\rm pl})} \dot{g}_{(\hat{\mathcal{A}}_{\rm pl})}^{\otimes 2}(\boldsymbol{x}_i^{\rm pl}; \hat{\boldsymbol{\theta}}_{\rm pl})}{\rho_0 + y_i^{\rm pl}(\rho_1 - \rho_0)}, \tag{11}$$

$$\hat{\boldsymbol{\Omega}}_{(\hat{\mathcal{A}}_{\rm pl})}^{\rm pl} = \frac{1}{N_{\rm pl}} \sum_{i=1}^{N_{\rm pl}^*} \frac{e^{2f(\boldsymbol{x}_i^{{\rm pl}T}\hat{\boldsymbol{\beta}}_{\rm pl})} \dot{g}_{(\hat{\mathcal{A}}_{\rm pl})}^{\otimes 2}(\boldsymbol{x}_i^{\rm pl}; \hat{\boldsymbol{\theta}}_{\rm pl})}{\rho_0 + y_i^{\rm pl}(\rho_1 - \rho_0)}, \tag{12}$$

respectively. We also use the following moment estimator to estimate the denominator of (4):

$$\frac{1}{N_{\rm pl}} \sum_{i=1}^{N_{\rm pl}^*} \frac{\omega_i^{\rm A-OS}}{\rho_0 + y_i^{\rm pl}(\rho_1 - \rho_0)}, \tag{13}$$

where $\omega_i^{\rm A-OS} = p(\boldsymbol{x}_i^{\rm pl}; \hat{\boldsymbol{\theta}}_{\rm pl}) \| (\hat{\boldsymbol{M}}_{(\hat{\mathcal{A}}_{\rm pl})}^{\rm pl})^{-1} \dot{g}_{(\hat{\mathcal{A}}_{\rm pl})}(\boldsymbol{x}_i^{\rm pl}; \hat{\boldsymbol{\theta}}_{\rm pl}) \|$. If using (5) or (7), we use

$$\omega_i^{\rm L-OS} = p(\boldsymbol{x}_i^{\rm pl}; \hat{\boldsymbol{\theta}}_{\rm pl}) \| \dot{g}_{(\hat{\mathcal{A}}_{\rm pl})}(\boldsymbol{x}_i^{\rm pl}; \hat{\boldsymbol{\theta}}_{\rm pl}) \|, \text{ or}$$
$$\omega_i^{\rm P-OS} = p(\boldsymbol{x}_i^{\rm pl}; \hat{\boldsymbol{\theta}}_{\rm pl}) \| (\hat{\boldsymbol{\Omega}}_{(\hat{\mathcal{A}}_{\rm pl})}^{\rm pl})^{1/2} (\hat{\boldsymbol{M}}_{(\hat{\mathcal{A}}_{\rm pl})}^{\rm pl})^{-1} \dot{g}_{(\hat{\mathcal{A}}_{\rm pl})}(\boldsymbol{x}_i^{\rm pl}; \hat{\boldsymbol{\theta}}_{\rm pl}) \|,$$

respectively, instead of $\omega_i^{\rm A-OS}$ in (13). Now, we present the proposed two-step procedure in Algorithm 3 with more details than Algorithm 2 in Section 4.

Algorithm 3 Subsampling adaptive lasso algorithm

1:
- Take a pilot sample $\{(\boldsymbol{x}_i^{\rm pl}, y_i^{\rm pl})\}_{i=1}^{N_{\rm pl}^*}$ of expected sample size $N_{\rm pl}$ using $\{\pi(y_i) = \rho_0 + y_i(\rho_1 - \rho_0)\}_{i=1}^N$ and obtain a pilot estimator

$$\hat{\boldsymbol{\theta}}_{\rm pl} := \arg\max_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{N_{\rm pl}^*} [y_i^{\rm pl} g(\boldsymbol{x}_i^{\rm pl}; \boldsymbol{\theta}) - \log\{1 + e^{g(\boldsymbol{x}_i^{\rm pl}; \boldsymbol{\theta}) + l}\}] - \lambda_{\rm pl} \sum_{j=1}^{p} |\beta_{(j)}| \right\}, \tag{14}$$

  where $N_{\rm pl}^*$ is the actual pilot sample size and $l = \log(N_0/N_1)$. We call this first stage screening.
- Calculate approximate optimal sampling probabilities $\{\hat{\pi}(\boldsymbol{x}_i, y_i) = y_i + (1 - y_i)\rho\hat{\varphi}(x_i)\}_{i=1}^N$ by replacing $\hat{\varphi}(x_i)$ with $\varphi_{\rm A-OS}^{\rm adp}(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_{\rm pl})$, $\varphi_{\rm L-OS}^{\rm adp}(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_{\rm pl})$, or $\varphi_{\rm P-OS}^{\rm adp}(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_{\rm pl})$, based on (4), (5), or (7), respectively. The denominator of (4) is estimated using (13), and we replace $\omega_i^{\rm A-OS}$ with $\omega_i^{\rm L-OS}$ or $\omega_i^{\rm P-OS}$ for the denominator of (5) or (7), respectively. If using $\pi_{\rm A-OS}^{\rm adp}(\boldsymbol{x})$ or $\pi_{\rm P-OS}^{\rm adp}(\boldsymbol{x})$, estimate $\boldsymbol{M}_{(\mathcal{A})}$ and $\boldsymbol{\Omega}_{(\mathcal{A})}$ using the moment estimators in (11) and (12), respectively.

2: Use Algorithm 1 with the estimated optimal sampling probabilities to obtain a subsample $\{(\boldsymbol{x}_i^{\rm sub}, y_i^{\rm sub})\}_{i=1}^{N_{\rm sub}^*}$ and compute the adaptive lasso estimator:

$$\hat{\boldsymbol{\theta}}_{\rm mscl}^{\rm adp} := \arg\max_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{N_{\rm sub}^*} [y_i^{\rm sub} g(\boldsymbol{x}_i^{\rm sub}; \boldsymbol{\theta}) - \log\{1 + e^{g(\boldsymbol{x}_i^{\rm sub}; \boldsymbol{\theta}) + l_i}\}] - \lambda_N \sum_{j \in \hat{\mathcal{A}}_{\rm pl}} \frac{|\beta_{(j)}|}{|\hat{\beta}_{{\rm pl}(j)}|^\gamma} \right\},$$

where $N_{\rm sub}^*$ is the actual subsample size, based on the smaller model obtain from the first stage screening. We call this step the second stage screening.

https://doi.org/10.52202/079017-3122

Remark 3. *Our algorithm naturally integrates the MSCL function with the adaptive lasso penalty. It can also be implemented when $p > N$ as long as the dimension of selected variables is smaller than N in the first-stage screening. If the model is sparse and the data are massive, this is usually possible in practice. Screening algorithms such as sure independence screening [6] can also be used for the first stage screening to guarantee that the dimension of second-stage screening is smaller than the subsample size. Furthermore, the first stage screening can help to speed up the computation, as shown by the analysis of computational complexity in the next section.*

We consider a coordinate desent method to calculate the estimators defined in Algorithm 3. (see [8], [9] and [27]). In each cycle, we need to find an optimal direction $\boldsymbol{d}$ at a starting point $\tilde{\boldsymbol{\theta}}$. We consider the quardratic approximation of $Q^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}}(\tilde{\boldsymbol{\theta}}+\boldsymbol{d}) - Q^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}}(\tilde{\boldsymbol{\theta}})$, which is

$$
\begin{aligned}
&Q^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}}(\tilde{\boldsymbol{\theta}}+\boldsymbol{d}) - Q^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}}(\tilde{\boldsymbol{\theta}}) \\
&= \sum_{i=1}^{N} \delta_i [-y_i g(\boldsymbol{x}_i; \tilde{\boldsymbol{\theta}}+\boldsymbol{d}) + \log\{1 + e^{g(\boldsymbol{x}_i;\tilde{\boldsymbol{\theta}}+\boldsymbol{d})+l_i}\}] + \lambda_N \sum_{j=1}^{p} \hat{w}_j |\beta_{(j)} + d_{(j)}| \\
&\quad - \sum_{i=1}^{N} \delta_i [-y_i g(\boldsymbol{x}_i; \tilde{\boldsymbol{\theta}}) - \log\{1 + e^{g(\boldsymbol{x}_i;\tilde{\boldsymbol{\theta}})+l_i}\}] + \lambda_N \sum_{j=1}^{p} \hat{w}_j |\beta_{(j)}| \\
&\approx \dot{\ell}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}}(\tilde{\boldsymbol{\theta}})^T \boldsymbol{d} + \frac{1}{2} \boldsymbol{d}^T \ddot{\ell}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}}(\tilde{\boldsymbol{\theta}}) \boldsymbol{d} + \lambda_N \sum_{j=1}^{p} \left\{ \hat{w}_j |\tilde{\beta}_{(j)} + d_{(j)}| - \hat{w}_j |\tilde{\beta}_{(j)}| \right\}
\end{aligned}
$$

where $\dot{\ell}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}}(\tilde{\boldsymbol{\theta}}) = -\sum_{i=1}^{N} \delta_i^{\hat{\boldsymbol{\theta}}_{\text{pl}}} \left\{ y_i - p_\pi^{\hat{\boldsymbol{\theta}}_{\text{pl}}}(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}) \right\} \dot{g}(\boldsymbol{x}_i; \tilde{\boldsymbol{\theta}})$ and $\ddot{\ell}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}}(\tilde{\boldsymbol{\theta}}) = \sum_{i=1}^{N} \delta_i^{\hat{\boldsymbol{\theta}}_{\text{pl}}} \phi_\pi^{\hat{\boldsymbol{\theta}}_{\text{pl}}}(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}) \dot{g}^{\otimes 2}(\boldsymbol{x}_i; \tilde{\boldsymbol{\theta}})$. Thus, using coordinate desent to obtain the optimal direction, the quadratic approximation for the j-th element is given as

$$
\begin{aligned}
Q^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}}(\boldsymbol{d} + z\boldsymbol{e}_j) - Q^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}}(\boldsymbol{d}) &= \dot{\ell}_{\text{mscl}(j)}(\tilde{\boldsymbol{\theta}}) z + \left\{ \ddot{\ell}_{\text{mscl}}(\tilde{\boldsymbol{\theta}}) \boldsymbol{d} \right\}_{(j)} z + \frac{1}{2} \ddot{\ell}_{\text{mscl}(jj)}(\tilde{\boldsymbol{\theta}}) z^2 \\
&\quad + \lambda_N \hat{w}_j |\tilde{\beta}_{(j)} + d_{(j)} + z| - \lambda_N \hat{w}_j |\tilde{\beta}_{(j)} + d_{(j)}|.
\end{aligned}
$$

Then, we have the value of z that minimize $Q^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}}(\boldsymbol{d} + z\boldsymbol{e}_j) - Q^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}}(\boldsymbol{d})$ is

$$
z^{**} = \begin{cases}
\dfrac{\dot{\ell}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}(j)}(\tilde{\boldsymbol{\theta}}) + \left\{ \ddot{\ell}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}}(\tilde{\boldsymbol{\theta}}) \boldsymbol{d} \right\}_{(j)} + \lambda \hat{w}_j}{-\ddot{\ell}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}(jj)}(\tilde{\boldsymbol{\theta}})} & \text{if } \tilde{\beta}_{(j)} + d_{(j)} + z \geq 0 \\
\dfrac{\dot{\ell}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}(j)}(\tilde{\boldsymbol{\theta}}) + \left\{ \ddot{\ell}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}}(\tilde{\boldsymbol{\theta}}) \boldsymbol{d} \right\}_{(j)} - \lambda \hat{w}_j}{-\ddot{\ell}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}(jj)}(\tilde{\boldsymbol{\theta}})} & \text{if } \tilde{\beta}_{(j)} + d_{(j)} + z \leq 0 \\
-\tilde{\beta}_{(j)} - d_{(j)} & \text{otherwise,}
\end{cases}
$$

which is the same as

$$
z^{**} = \max\left\{ z_1, -\tilde{\beta}_{(j)} - d_{(j)} \right\} - \max\left\{ -z_2, \tilde{\beta}_{(j)} + d_{(j)} \right\} + \tilde{\beta}_{(j)} + d_{(j)},
$$

where

$$
z_1 = \frac{\dot{\ell}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}(j)}(\tilde{\boldsymbol{\theta}}) + \left\{ \ddot{\ell}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}}(\tilde{\boldsymbol{\theta}}) \boldsymbol{d} \right\}_{(j)} + \lambda \hat{w}_j}{-\ddot{\ell}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}(jj)}(\tilde{\boldsymbol{\theta}})},
$$

and

$$
z_2 = \frac{\dot{\ell}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}(j)}(\tilde{\boldsymbol{\theta}}) + \left\{ \ddot{\ell}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}}(\tilde{\boldsymbol{\theta}}) \boldsymbol{d} \right\}_{(j)} - \lambda \hat{w}_j}{-\ddot{\ell}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}(jj)}(\tilde{\boldsymbol{\theta}})}.
$$

For the special form of $g(\boldsymbol{x}; \boldsymbol{\theta}) = \alpha + f(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta})$, we know that

$$
\ddot{\ell}^{\hat{\boldsymbol{\theta}}_{\text{pl}}}_{\text{mscl}}(\tilde{\boldsymbol{\theta}}) = \sum_{i=1}^{N} \delta_i^{\hat{\boldsymbol{\theta}}_{\text{pl}}} \phi_\pi^{\hat{\boldsymbol{\theta}}_{\text{pl}}}(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}) \dot{g}^{\otimes 2}(\boldsymbol{x}_i; \tilde{\boldsymbol{\theta}}) = \boldsymbol{G}^{\mathrm{T}} \boldsymbol{\Phi} \boldsymbol{G},
$$

where

$$\boldsymbol{G} = \begin{pmatrix} 1 & \dot{f}(\boldsymbol{x}_1^{\mathrm{T}}\tilde{\boldsymbol{\beta}})\boldsymbol{x}_1^{\mathrm{T}} \\ 1 & \dot{f}(\boldsymbol{x}_2^{\mathrm{T}}\tilde{\boldsymbol{\beta}})\boldsymbol{x}_2^{\mathrm{T}} \\ \vdots & \vdots \\ 1 & \dot{f}(\boldsymbol{x}_N^{\mathrm{T}}\tilde{\boldsymbol{\beta}})\boldsymbol{x}_N^{\mathrm{T}} \end{pmatrix}$$

and $\boldsymbol{\Phi} = diag\{\delta_i^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}\phi_\pi^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}})\}$. Thus, we have

$$\left\{\ddot{\ell}_{\mathrm{mscl}}^{\hat{\boldsymbol{\theta}}_{\mathrm{pl}}}(\tilde{\boldsymbol{\theta}})\boldsymbol{d}\right\}_{(j)} = \left(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{G}\boldsymbol{d}\right)^{\mathrm{T}}\boldsymbol{e}_j = (\boldsymbol{G}\boldsymbol{d})^{\mathrm{T}}\boldsymbol{\Phi}(\boldsymbol{G}\boldsymbol{e}_j) = (\boldsymbol{G}\boldsymbol{d})^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{G}_{(j)}.$$

Therefore, we can store $\boldsymbol{G}\boldsymbol{d}$ and keep updating $\boldsymbol{G}\boldsymbol{d}$ with

$$\boldsymbol{G}(\boldsymbol{d} + z\boldsymbol{e}_j) = \boldsymbol{G}\boldsymbol{d} + z\boldsymbol{G}\boldsymbol{e}_j = \boldsymbol{G}\boldsymbol{d} + \boldsymbol{G}_{(j)}z.$$

Thus, we do not need to obtain the full matrix $\ddot{\ell}_{\mathrm{mscl}}(\tilde{\boldsymbol{\theta}}) = \boldsymbol{G}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{G}$. We only need to calculate the diagnoal elements: $\ddot{\ell}_{\mathrm{mscl}(jj)}(\tilde{\boldsymbol{\theta}}) = \boldsymbol{G}_{(j)}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{G}_{(j)}, j = 1, ..., p+1$ and $\left\{\dot{\ell}_{\mathrm{mscl}}(\tilde{\boldsymbol{\theta}})\boldsymbol{d}\right\}_{(j)} = (\boldsymbol{G}\boldsymbol{d})^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{G}_{(j)}, j = 1, ..., p+1$. From the analysis above, we can notice that the computaional complexity of one cycle calculating optimal direction $\boldsymbol{d}$ is $O(\zeta_{\mathrm{in}}Np)$, where ζ_{in} denotes the number of inner iteration.

C.2 Computational complexity

We analyze the computational complexity of the two-step algorithm. To facilitate the presentation, we consider a special case for our model when $g(\boldsymbol{x}; \boldsymbol{\theta}) = \alpha + f(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})$, and assume that the number of variables selected at the first-stage screening is q. Coordinate descent is a widely used optimization algorithm for solving lasso and adaptive lasso [see 9]. We consider the improved coordinate descent algorithm proposed in [27], which requires inner iterations to determine an optimal direction and outer iterations to update the estimator. Considering the form of $g(\boldsymbol{x}; \boldsymbol{\theta}) = \alpha + f(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})$, the computational complexity for coordinate descent with data of size N and dimension p is $O(\zeta_{\mathrm{in}}Np)$ per inner-cycle where ζ_{in} represents the number of inner iterations (detailed derivations of this complexity is presented Section C). Thus, the computational complexity of full data lasso is $O(\zeta_{\mathrm{out}\times\mathrm{in}}Np)$, where $\zeta_{\mathrm{out}\times\mathrm{in}} = \zeta_{\mathrm{out}}\zeta_{\mathrm{in}}$ and ζ_{out} is the number of outer iterations. The computational complexity of the full data adaptive lasso is $O(\zeta_{\mathrm{pl}}^{\mathrm{mle}}Np^2 + \zeta_{\mathrm{out}\times\mathrm{in}}Np)$ with the MLE as the pilot estimator, and $O(\zeta_{\mathrm{pl,out}\times\mathrm{in}}^{\mathrm{las}}Np + \zeta_{\mathrm{out}\times\mathrm{in}}Nq)$ with lasso as the pilot estimator, where $\zeta_{\mathrm{pl}}^{\mathrm{MLE}}$ and $\zeta_{\mathrm{pl,out}\times\mathrm{in}}^{\mathrm{las}}$ are the iteration numbers in the two pilot estimators, respectively. The coordinate descent algorithm often requires a large $\zeta_{\mathrm{out}\times\mathrm{in}}$ or $\zeta_{\mathrm{pl,out}\times\mathrm{in}}^{\mathrm{las}}$ while Newton's algorithm requires a small $\zeta_{\mathrm{pl}}^{\mathrm{mle}}$, so it is often the case that $\zeta_{\mathrm{pl}}^{\mathrm{mle}}p < \zeta_{\mathrm{out}\times\mathrm{in}}$. Therefore, the time complexity of the adaptive lasso is $O(\zeta_{\mathrm{out}\times\mathrm{in}}Np)$, which is the same as the full data lasso estimator.

Now, we analyze the time complexity of Algorithm 3. We start with the computational complexity of the optimal probabilities, for which the main computational cost is to approximate $\|\boldsymbol{M}_{(\mathcal{A})}^{-1}\dot{g}_{(\mathcal{A})}(\boldsymbol{x}_i; \boldsymbol{\theta})\|$ or $\|\boldsymbol{\Omega}_{(\mathcal{A})}^{1/2}\boldsymbol{M}_{(\mathcal{A})}^{-1}\dot{g}_{(\mathcal{A})}(\boldsymbol{x}_i; \boldsymbol{\theta})\|$, respectively, for $i = 1, ..., N$. Since $\dot{g}_{(\hat{\mathcal{A}}_{\mathrm{pl}})}(\boldsymbol{x}_i; \boldsymbol{\theta}) = (1, \dot{f}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{x}_{i(\hat{\mathcal{A}}_{\mathrm{pl}})}^{\mathrm{T}})^{\mathrm{T}}$, the computational complexity of calculating $\dot{g}_{(\hat{\mathcal{A}}_{\mathrm{pl}})}(\boldsymbol{x}_i; \boldsymbol{\theta})$'s is $O(Nq)$, and the computational complexity of $\hat{\boldsymbol{M}}_{(\hat{\mathcal{A}}_{\mathrm{pl}})}^{\mathrm{pl}}$ or $\hat{\boldsymbol{\Omega}}_{(\hat{\mathcal{A}}_{\mathrm{pl}})}^{\mathrm{pl}}$ is $O\left\{N_{\mathrm{pl}}(q+1)^2\right\} = O(N_{\mathrm{pl}}q^2)$. Taking the inverse $(\hat{\boldsymbol{M}}_{(\hat{\mathcal{A}}_{\mathrm{pl}})}^{\mathrm{pl}})^{-1}$ and finding the square root $(\hat{\boldsymbol{\Omega}}_{(\hat{\mathcal{A}}_{\mathrm{pl}})}^{\mathrm{pl}})^{1/2}$ both take $O(q^3)$ time. Thus, the computational complexity of calculating $\|(\hat{\boldsymbol{M}}_{(\hat{\mathcal{A}}_{\mathrm{pl}})}^{\mathrm{pl}})^{-1}\dot{g}_{(\hat{\mathcal{A}}_{\mathrm{pl}})}(\boldsymbol{x}_i; \boldsymbol{\theta})\|$'s or $\|(\hat{\boldsymbol{\Omega}}_{(\hat{\mathcal{A}}_{\mathrm{pl}})}^{\mathrm{pl}})^{1/2}(\hat{\boldsymbol{M}}_{(\hat{\mathcal{A}}_{\mathrm{pl}})}^{\mathrm{pl}})^{-1}\dot{g}_{(\hat{\mathcal{A}}_{\mathrm{pl}})}(\boldsymbol{x}_i; \boldsymbol{\theta})\|$'s is $O(Nq + N_{\mathrm{pl}}q^2 + q^3 + Nq^2) = O(Nq^2)$. Therefore, the complexity of approximating the optimal probabilities in (4) or (7) is $O(Nq^2)$. The computational complexity of approximating the optimal probabilities in (5) is only $O(Nq)$, because there is no need to compute $\hat{\boldsymbol{M}}_{(\hat{\mathcal{A}}_{\mathrm{pl}})}^{\mathrm{pl}}$ or $\hat{\boldsymbol{\Omega}}_{(\hat{\mathcal{A}}_{\mathrm{pl}})}^{\mathrm{pl}}$. Next, we analyze the complexity of parameter estimation. The average subsample size with the sampling rate ρ is on average

$$\mathbb{E}(N_1) + \rho\{N - \mathbb{E}(N_1)\} = N\mathbb{E}\{f(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}})\}\{(1-\rho)e^{\alpha_{\mathrm{t}}} + \rho + o(1)\} = O\{N(e^{\alpha_{\mathrm{t}}} + \rho)\}.$$

https://doi.org/10.52202/079017-3122

Using the coordinate descent algorithm, the computational complexity of the two-step algorithm is $O\{\zeta^{\text{las}}_{\text{pl,out}\times\text{in}}N_{\text{pl}}p + Nq^2 + \zeta_{\text{out}\times\text{in}}N(e^{\alpha_{\text{t}}}+\rho)q\}$ using optimal probabilities in (4) or (7), and it is $O\{\zeta^{\text{las}}_{\text{pl,out}\times\text{in}}N_{\text{pl}}p + Nq + \zeta_{\text{out}\times\text{in}}N(e^{\alpha_{\text{t}}}+\rho)q\}$ using the optimal probabilities in (5). For optimal probabilities in (4) or (7), when $\zeta_{\text{out}\times\text{in}} > q/(e^{\alpha}+\rho)$, the dominating term of the complexity is $\zeta_{\text{out}\times\text{in}}N(e^{\alpha_{\text{t}}}+\rho)q$. Remember that $\zeta_{\text{out}\times\text{in}} = \zeta_{\text{out}}\zeta_{\text{in}}$ and ζ_{out} is usually large for the coordinate descent algorithm. Therefore, $\zeta_{\text{out}\times\text{in}} > q/(e^{\alpha}+\rho)$ is often satisfied in practice. The dominating term for the time complexity of optimal probabilities in (5) is also $\zeta_{\text{out}\times\text{in}}N(e^{\alpha_{\text{t}}}+\rho)q$. Compared with full data estimators, both the sample size and the dimension are reduced. If we set the subsample size to be the same order of N_1, which is often the case in practice for balancing the ones and zeros, the time complexity of Algorithm 2 is of order $O(\zeta_{\text{out}\times\text{in}}Ne^{\alpha_{\text{t}}}q)$, which is significantly faster than that of the full data estimator.

D Detalis of simulation settings

In this section, we present more details of the simulation settings in the main paper. In Section D.1, we provide detailed simulation settings of the example in Section 1. In Section D.2, we present detailed settings in Section 5.

D.1 Simulation in Section 1

We first present the detailed settings of the simulations in Section 1, where we illustrate the scale-dependent issues of optimal subsampling probabilities. Our simulation based on logistic regression models with the true parameter $\boldsymbol{\beta}_{\text{t}}$ to be 6-dimentional vectors and covariates $\boldsymbol{x} \sim lognormal(\mathbf{0}, \boldsymbol{\Sigma})$ with the (i,j)-th element of $\boldsymbol{\Sigma}$ is given as $\boldsymbol{\Sigma}_{ij} = 0.5^{|i-j|}, 1 \le i,j \le 6$. We consider two cases of parameters:

(a) **Non-sparse parameter**: $\boldsymbol{\beta}_{\text{t}} = (-1,-1,-0.01,-0.01,-0.01,-0.01)$ and $\alpha_{\text{t}} = -4$.

(b) **Sparse parameter**: $\boldsymbol{\beta}_{\text{t}} = (-1,0,0,0,0,0)$ and $\alpha_{\text{t}} = -5$.

We generate full data of size $N = 500000$ according to the above logistic models. To investigate the effects of scale transformation, we multiply the $\boldsymbol{x}_{(6)}$ with s ($s = 0.01, 0.1, 1, 10, 100$) and divide $\boldsymbol{\beta}_{\text{t}(6)}$ with the same s to remain $\boldsymbol{x}^{\text{T}}\boldsymbol{\beta}_{\text{t}}$ to be the same and thus the logistics regression model does not change. We obtain subsamples with optimal subsampling probabilities described in [22] with transformed $\boldsymbol{x}$ under each s and calculate the resultant subsampling estimators. We set the nominal pilot sample size to $N_{\text{pl}} = 800$ and nominal subsample size $N_{\text{sub}} = 1000$ (see details in [22]). We repeat the experiment for 500 times under each scale and compute the mean prediction error.

D.2 Simulations in Section 5

For the estimation procedures in the second step of our two-step algorithm, we choose $\gamma = 1$, which means that the weights in the adaptive lasso penalty are $\hat{w}_j = 1/|\hat{\beta}_{\text{pl}(j)}|, 1 \le j \le q$, with q being the number of selected variables in the first stage screening. Furthermore, we consider uniform sampling, the full data lasso and the full data adaptive lasso as baselines for comparison. For the uniform sampling method, we use a similar two-step algorithm as presented in Algorithm 2 but set the sampling function in the second step as $\varphi(\boldsymbol{x}) = 1$, which means the sampling probabilities are a constant ρ. We use lasso to implement the first stage screening and adaptive lasso with $\gamma = 1$ to implement the second stage screening for a fair comparison. For full data lasso, we directly apply the lasso algorithm to the full data. For the full data adaptive lasso, we use the full data MLE estimator as the pilot estimator to construct the weights and then apply the adaptive lasso algorithm to the full data set.

E Additional simuations

In this Section, we give some addtional simulation results. More simulation results of variable selection is provided in Section E.1. We also provide addtional simulation results to compare our approach with standardization to resolve scale dependent issues in Section E.2.

E.1 Addtional variable selection results

Table 4: Mean number of selected variables in Case A and Case B

Case A (five active variables)					
ρ	first-stage	Uni	A-OS	L-OS	P-OS
0.0025	14.87(0.30)	4.99(0.02)	5.02(0.02)	5.03(0.02)	5.01(0.02)
0.005	14.68(0.28)	5.05(0.02)	5.07(0.02)	5.06(0.02)	5.08(0.02)
0.0075	14.20(0.27)	5.09(0.03)	5.08(0.02)	5.08(0.02)	5.09(0.03)
0.01	14.46(0.30)	5.13(0.02)	5.13(0.02)	5.13(0.02)	5.12(0.03)
Case B (four active variables)					
ρ	first-stage	Uni	A-OS	L-OS	P-OS
0.0025	16.81(0.33)	3.91(0.02)	4.01(0.03)	4.00(0.02)	4.01(0.03)
0.005	17.88(0.34)	4.04(0.03)	4.10(0.03)	4.08(0.04)	4.07(0.03)
0.0075	17.19(0.33)	4.06(0.02)	4.13(0.03)	4.10(0.03)	4.12(0.03)
0.01	17.51(0.34)	4.07(0.03)	4.08(0.03)	4.08(0.03)	4.08(0.03)

Table 5: Rates of excluding active variables (false negative rate) in Case A and Case B

Case A				
ρ	Uni	A-OS	L-OS	P-OS
0.0025	0.094(0.013)	0.076(0.012)	0.068(0.011)	0.070(0.012)
0.005	0.052(0.010)	0.048(0.010)	0.050(0.010)	0.044(0.010)
0.0075	0.054(0.010)	0.052(0.010)	0.052(0.010)	0.052(0.010)
0.01	0.040(0.009)	0.036(0.008)	0.036(0.008)	0.036(0.008)
Case B				
ρ	Uni	A-OS	L-OS	P-OS
0.0025	0.108(0.014)	0.074(0.012)	0.074(0.012)	0.070(0.012)
0.005	0.046(0.009)	0.034(0.008)	0.034(0.008)	0.036(0.008)
0.0075	0.062(0.011)	0.046(0.009)	0.046(0.009)	0.044(0.009)
0.01	0.056(0.010)	0.050(0.010)	0.052(0.010)	0.050(0.010)

Table 6: Rates of selecting the true model

Case A				
ρ	Uni	A-OS	L-OS	P-OS
0.0025	0.856(0.016)	0.870(0.015)	0.868(0.015)	0.878(0.015)
0.005	0.884(0.014)	0.872(0.015)	0.880(0.015)	0.872(0.015)
0.0075	0.858(0.016)	0.868(0.015)	0.868(0.015)	0.868(0.015)
0.01	0.854(0.016)	0.866(0.015)	0.862(0.015)	0.868(0.015)
Case B				
ρ	Uni	A-OS	L-OS	P-OS
0.0025	0.862(0.015)	0.872(0.015)	0.864(0.015)	0.876(0.015)
0.005	0.898(0.014)	0.888(0.014)	0.898(0.014)	0.902(0.013)
0.0075	0.848(0.016)	0.850(0.016)	0.854(0.016)	0.848(0.016)
0.01	0.874(0.015)	0.874(0.015)	0.876(0.015)	0.872(0.015)
Case C				
ρ	Uni	A-OS	L-OS	P-OS
0.0025	0.824(0.017)	0.874(0.015)	0.880(0.015)	0.884(0.014)
0.005	0.880(0.015)	0.884(0.014)	0.884(0.014)	0.882(0.014)
0.0075	0.908(0.013)	0.916(0.012)	0.910(0.013)	0.914(0.013)
0.01	0.908(0.013)	0.910(0.013)	0.908(0.013)	0.910(0.013)

It is seen in Table 6 that, no subsampling method dominates others. Table 2 and Table 5 shows that uniform sampling has higher rates of excluding active variables than optimal subsampling procedures.

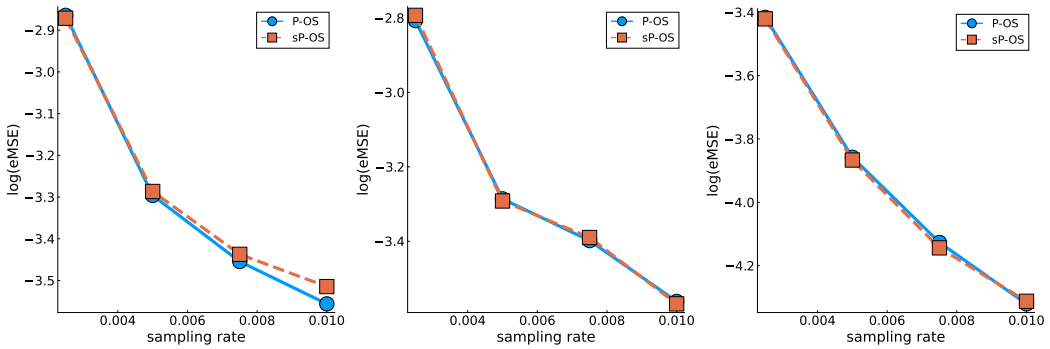
https://doi.org/10.52202/079017-3122

Although uniform sampling may have a higher rate of selecting the true model in some cases, given that it is more likely to exclude important variables, optimal sampling may be preferable in practice.

E.2 Comparison with standardization

Another approach to avoid scale-dependency is to standardize the data. We compare the proposed scale-independent optimal probabilities with the approach of data standardization here. For the data standardization approach, we standardize the data, calcualte the optimal probabilities, and then implement subsampled adpative lasso algorithm. We used the same pilot estimation methods for fair comparisons.

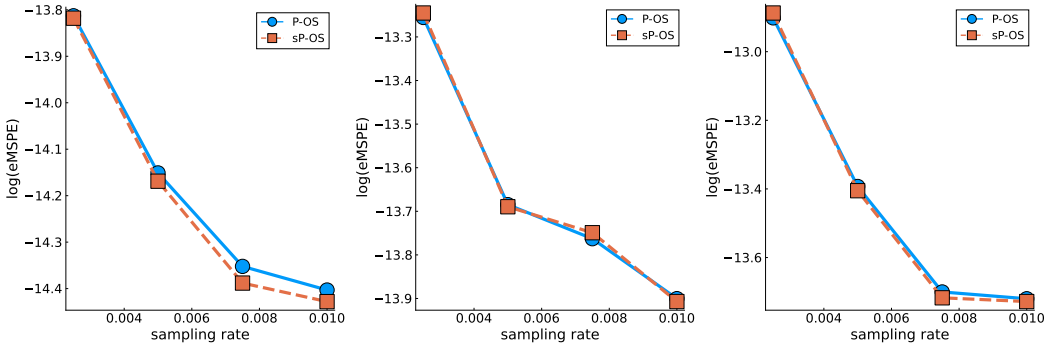
We first compare the eMSE and eMSPE in Figure 5 and Figure 6, respectively. We use sP-OS to denote the approach with data standardization and use P-OS to denote the approach without data standardization.



(a) Case A (b) Case B (c) Case C

Figure 5: Empirical median squred error of estimated probability for different parameters with different sampling rates. The same pilot sample size is $N_{\mathrm{pl}} = 500$.

In Figure 5, we notice that the performances of P-OS and sP-OS are similar, this is also true for eMSPE. However, standardization may decrease the rate of selecting the true model. We present results of variable selection in Table 7 and Table 8.



(a) Case A (b) Case B (c) Case C

Figure 6: Empirical median squred error of estimated probability for different parameters with different sampling rates. The same pilot sample size is $N_{\mathrm{pl}} = 500$.

We notice in Table 7 and Table 8 that the rates of selecting true models by $\hat{\boldsymbol{\beta}}^{\mathrm{adp}}_{\mathrm{P-OS}}$ is higher than $\hat{\boldsymbol{\beta}}^{\mathrm{adp}}_{\mathrm{sP-OS}}$ without much increase on the rates of excluding active variables. Therefore, although standardization is an approach to solve the scale-dependency issues, it may decrease the rates of selecting true models in practice.

Table 7: Rates of selecting true models

	A			B			C		
ρ	sUni	P-OS	sP-OS	sUni	P-OS	sP-OS	sUni	P-OS	sP-OS
0.0025	0.848	0.878	0.862	0.850	0.876	0.862	0.828	0.884	0.880
0.005	0.878	0.872	0.850	0.890	0.902	0.890	0.882	0.882	0.880
0.0075	0.850	0.868	0.846	0.850	0.848	0.842	0.906	0.914	0.910
0.01	0.840	0.868	0.844	0.868	0.872	0.862	0.900	0.910	0.904

Table 8: Rates of excluding active variables (false negtive rate)

	A			B			C		
ρ	sUni	P-OS	sP-OS	sUni	P-OS	sP-OS	sUni	P-OS	sP-OS
0.0025	0.088	0.070	0.068	0.106	0.070	0.070	0.164	0.084	0.084
0.005	0.056	0.044	0.044	0.046	0.036	0.036	0.100	0.066	0.066
0.0075	0.052	0.052	0.052	0.062	0.044	0.044	0.062	0.046	0.046
0.01	0.040	0.036	0.036	0.056	0.050	0.050	0.068	0.054	0.054

https://doi.org/10.52202/079017-3122

NeurIPS Paper Checklist

[Yes] [No] [NA]

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Detailed proofs and required assumptions are provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All details for numerical experiments are provided in Section 5 and in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-3122

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Codes are submitted as supplement for anonymity. They will be released in a public github repository after the review period.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: In Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide standard errors in Tables 1 and 2. We perform a large number of repetitions of the simulation experiments to calculate the empirical median squared error, so error bars are not relevant in Figures 1, 2, 3, and 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Section 5.1.3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper is theoretical research.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

https://doi.org/10.52202/079017-3122

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: URLs for real data provided in Section 5.2.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-3122