Scale-invariant optimal sampling and variable selection with rare-events data [1] [2]

HaiYing Wang

University of Connecticut

mODa, July 10, 2023

[1] Collaborators: Jing Wang, Kun Chen, Hao Hellen Zhang

[2] Supported by NSF grant CCF 2105571

Outline

1. Introduction
2. Subsampling with Variable selection
3. Penalized MSCL estimator and its theoretical analysis
4. Numerical experiments

Outline

1. Introduction
2. Subsampling with Variable selection
3. Penalized MSCL estimator and its theoretical analysis
4. Numerical experiments

Introduction[3]

- Imbalanced data are ubiquitous in scientific fields and applications with binary response outputs, where the number of instances in the positive class is much smaller than that in the negative class.
- For an online recommendation system in ByteDance, there are over 10 billion impressions each day, but only about 1.25% are clicked (negative/positives $\approx$ 80:1.).
- Due to limited storage and computational resources, the goal is to ignore some non-clicks and reduce negative/positive to 4:1, i.e., remove about 95% of the negative instances.

[3] Wang, H., Zhang, A., and Wang, C. (2021). Nonuniform negative sampling and log odds correction with rare events data. In *NeurIPS 2021.*

Big binary imbalanced data

- Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ be training data that satisfies

$$\mathbb{P}(y = 1 \mid \boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}) := \frac{1}{1 + e^{-g(\boldsymbol{x};\boldsymbol{\theta})}}, \tag{1}$$

  where $y \in \{0, 1\}$ is the class label, 1 for the case and 0 for the control; $\boldsymbol{x}$ is the feature vector; and $\boldsymbol{\theta}$ is the parameter.
- Let N_1 be the number of cases, and N_0 be the number of controls. For imbalanced data, i.e., $N_1 \ll N_0$.
- When N_1 is much smaller than N_0, it is more appropriate to assume that N_1 increases in a slower rate compared with N_0, (Wang, 2020; Wang *et al.*, 2021) i.e.,

$$\frac{N_1}{N_0} \xrightarrow{P} 0 \quad \text{and} \quad N_1 \xrightarrow{P} \infty \qquad \text{as} \quad N \to \infty. \tag{2}$$

- This requires $\mathbb{P}(y = 1) \to 0$ as $N \to \infty$ on the model side.

Model that allows $\mathbb{P}(y = 1) \to 0$

- Let $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ and write the log odds as

$$g(\boldsymbol{x}; \boldsymbol{\theta}) := \log\left\{\frac{p(\boldsymbol{x}; \boldsymbol{\theta})}{1 - p(\boldsymbol{x}; \boldsymbol{\theta})}\right\} = \alpha + f(\boldsymbol{x}; \boldsymbol{\beta}) \tag{3}$$

- Here $f(\boldsymbol{x}; \boldsymbol{\beta})$ is a smooth function of $\boldsymbol{\beta}$, such as a neural net. If it is linear, the model reduces to the logistic regression.
- Denoted the true parameter as $\boldsymbol{\theta}^* = (\alpha^*, \boldsymbol{\beta}^{*\mathrm{T}})^{\mathrm{T}}$, and assume that $\alpha^* \to -\infty$ as $N \to \infty$ and $\boldsymbol{\beta}^*$ is fixed.
- A diverging α^* and a fixed $\boldsymbol{\beta}^*$ indicates that the both the marginal and conditional probabilities for a positive instance are small.
- This means a covariate change does not convert a small-probability-event to a large-probability-event.
- We can also let $\boldsymbol{\beta}^*$ change with N, but as long as $\boldsymbol{\beta}^*$ has a finite limit, the problem is essentially the same as a fixed $\boldsymbol{\beta}^*$.

How much information do we really have?

Under some moment assumptions, as $N \to \infty$,

$$\sqrt{N_1}(\widehat{\boldsymbol{\theta}}_{\mathrm{mle}} - \boldsymbol{\theta}^*) \longrightarrow \mathbb{N}(\mathbf{0},\ \mathbf{V}_{\mathrm{mle}}), \quad \text{in distribution.} \tag{4}$$

Table 1: Numerical illustration

	Correct model			Mis-sprcified model		
$(N,\ N_1^a)$	$\mathrm{tr}(\hat{\mathbf{V}}_{\mathrm{e}})$	$N_1^a\mathrm{tr}(\hat{\mathbf{V}}_{\mathrm{e}})$	$N\mathrm{tr}(\hat{\mathbf{V}}_{\mathrm{e}})$	$\mathrm{tr}(\hat{\mathbf{V}}_{\mathrm{e}})$	$N_1^a\mathrm{tr}(\hat{\mathbf{V}}_{\mathrm{e}})$	$N\mathrm{tr}(\hat{\mathbf{V}}_{\mathrm{e}})$
$(10^3,\ 32)$	0.169	5.41	169.17	0.969	30.99	968.70
$(10^4,\ 64)$	0.097	6.20	969.29	0.322	20.59	3217.12
$(10^5,\ 128)$	0.045	5.76	4497.24	0.135	17.32	13527.60
$(10^6,\ 256)$	0.018	4.62	18048.40	0.046	11.74	45847.40

Here, $\hat{\mathbf{V}}_{\mathrm{e}}$ is the empirical variance of $\widehat{\boldsymbol{\theta}}_{\mathrm{mle}}$ and $N_1^a = \mathbb{E}(N_1)$.

General negative sampling algorithm

- ρ: sampling rate on the negative class.
- $\varphi(\boldsymbol{x}) > 0$: a function with $\mathbb{E}\{\varphi(\boldsymbol{x})\} = 1$.
- $\pi(\boldsymbol{x}) = \rho\varphi(\boldsymbol{x})$: sampling probability for the negative class.
- $\pi(\boldsymbol{x}_i, y_i) = y_i + (1 - y_i)\pi(\boldsymbol{x}_i)$: inclusion probability of $(\boldsymbol{x}_i, y_i)$.
- $\delta_i = 1$ if the i-th data point is selected and $\delta_i = 0$ otherwise.

Algorithm 1 Negative sampling

```
For $i = 1, ..., N$:
  1 if $y_i = 1$,  record $\{\boldsymbol{x}_i, y_i, \pi(\boldsymbol{x}_i, y_i = 1) = 1\}$ in the sample;
  2 if $y_i = 0$,  with probability $\pi(\boldsymbol{x}_i, y_i = 0)$,
                       include $\{\boldsymbol{x}_i, y_i, \pi(\boldsymbol{x}_i, y_i = 0) = \rho\varphi(\boldsymbol{x}_i)\}$ in the sample.
```

Subsample estimator and A-optimal sampling function

The subsample inverse probability weighted (IPW) estimator of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}}_{\mathrm{w}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \delta_i \frac{y_i g(\boldsymbol{x}_i; \boldsymbol{\theta}) - \log\{1 + e^{g(\boldsymbol{x}_i;\boldsymbol{\theta})}\}}{\pi(\boldsymbol{x}_i, y_i)}. \tag{5}$$

Under some moment assumptions, as $N \to \infty$,

$$\sqrt{N_1}(\widehat{\boldsymbol{\theta}}_{\mathrm{w}} - \boldsymbol{\theta}^*) \longrightarrow \mathbb{N}(\mathbf{0},\ \mathbf{V}_{\mathrm{w}}), \quad \text{in distribution}, \tag{6}$$

where $\mathbf{V}_{\mathrm{w}} = \mathbf{V}_{\mathrm{mle}} + \mathbf{V}_{\mathrm{sub}}$ and $\mathbf{V}_{\mathrm{sub}}$ depends on $\varphi(\cdot)$.

Optimal sampling of MLE of rare-events data

The A-optimal function that minimize $\mathbf{V}_{\mathrm{sub}}$ is

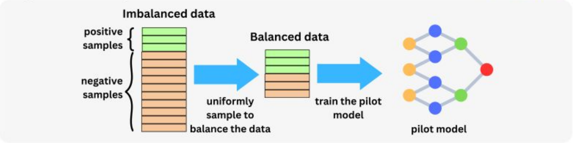
$$\varphi_{\mathrm{A-OS}}^{\mathrm{mle}}(\boldsymbol{x}) = \frac{p(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}})\|\mathbf{M}^{-1}\dot{\boldsymbol{g}}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}})\|}{\mathbb{E}\left\{p(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}})\|\mathbf{M}^{-1}\dot{\boldsymbol{g}}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}})\|\right\}}, \tag{7}$$

where $\mathbf{M} = \mathbb{E}\{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}\dot{\boldsymbol{g}}^{\otimes 2}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{t}})\}$.

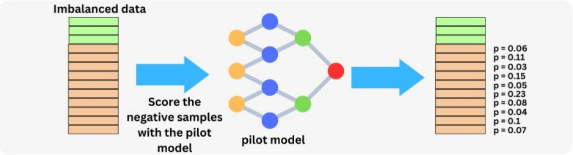
How to optimally sample Imbalanced Data?

Step 1: Train a "pilot" model

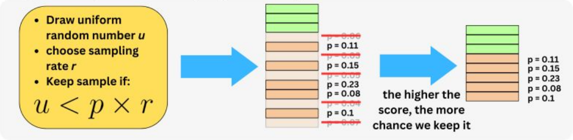
TheAiEdge.io



Step 2: Score the negative samples with the pilot model



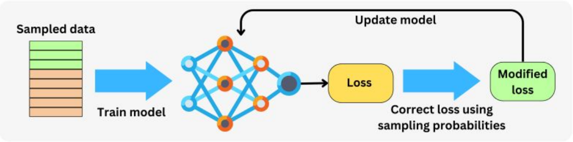
Step 3: Sample the data proportionally to the scores



- Draw uniform random number *u*
- choose sampling rate *r*
- Keep sample if:

$$u < p \times r$$

Step 4: Correct Likelihood function to produce unbiased estimates



Outline

1. Introduction
2. Subsampling with Variable selection
3. Penalized MSCL estimator and its theoretical analysis
4. Numerical experiments

Adaptive lasso [4]

- Not all available features/covariates are useful.
- Let $\mathcal{A} = \{j : \beta_{\mathrm{t}(j)} \neq 0\}$ be the active set, and $\mathcal{A}^c = \{j : \beta_{\mathrm{t}(j)} = 0\}$.
- The full data adpative lasso of rare-events data:

$$\widehat{\boldsymbol{\theta}}^{\mathrm{adp}}_{\mathrm{mle}} = \arg\max_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{N} [y_i g(\boldsymbol{x}_i; \boldsymbol{\theta}) - \log\{1 + e^{g(\boldsymbol{x}_i;\boldsymbol{\theta})}\}] - \lambda_N \sum_{j=1}^{p} \frac{|\beta_{(j)}|}{|\hat{\beta}_{\mathrm{pl}(j)}|^{\gamma}} \right\}, \quad (8)$$

  where $\hat{\boldsymbol{\beta}}_{\mathrm{pl}}$ is a consistent **pilot estimate**, and λ_N and γ are tuning parameters.

- **The penalty term** is critical for oracle properties.
- Optimal subsampling also requires a **pilot estimate**, so it is natural to combine adaptive lasso with optimal subsampling.

[4] Zou (2006) and Zhang and Lu (2007)

Oracle properties

Theorem 2.1 (Oralce properties of full data adaptive lasso)

Let $\hat{\mathcal{A}}_{\mathrm{adp}} := \{j : \hat{\beta}^{\mathrm{adp}}_{\mathrm{mle}(j)} \neq 0\}$. *Under some regularity conditions:*

1. *Consistency in variable selection:* $\lim_{N\to\infty} \mathbb{P}(\hat{\mathcal{A}}_{\mathrm{adp}} = \mathcal{A}) = 1$.
2. *Asymptotic normality:*

$$\sqrt{N_1}(\widehat{\boldsymbol{\theta}}^{\mathrm{adp}}_{\mathrm{mle}(\mathcal{A})} - \boldsymbol{\theta}_{\mathrm{t}(\mathcal{A})}) \xrightarrow{D} \mathbb{N}(\mathbf{0}, \mathbf{V}_{\mathrm{mle}(\mathcal{A})}), \tag{9}$$

where $\mathbf{V}_{\mathrm{mle}(\mathcal{A})} = \mathbb{E}\{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}\}\mathbf{M}^{-1}_{(\mathcal{A})}$ *and*
$\mathbf{M}_{(\mathcal{A})} = \mathbb{E}\{e^{f(\boldsymbol{x}_i;\boldsymbol{\beta}_{\mathrm{t}})}\dot{\boldsymbol{g}}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{t}})\}$.

Adaptive lasso with subsample IPW estimators

- Consider a penalized IPW estimator

$$\widehat{\boldsymbol{\theta}}_{\mathrm{w}}^{\mathrm{adp}} := \arg\max_{\boldsymbol{\theta}} \left\{ \ell_{\mathrm{w}}^{\mathrm{sub}}(\boldsymbol{\theta}) - \lambda_{\mathrm{w}} \sum_{j=1}^{p} \frac{|\beta_{(j)}|}{|\hat{\beta}_{\mathrm{pl}(j)}|^{\gamma}} \right\}, \tag{10}$$

where

$$\ell_{\mathrm{w}}^{\mathrm{sub}}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \delta_i \frac{y_i g(\boldsymbol{x}_i; \boldsymbol{\theta}) - \log\{1 + e^{g(\boldsymbol{x}_i;\boldsymbol{\theta})}\}}{\pi(\boldsymbol{x}_i, y_i)}. \tag{11}$$

- $\widehat{\boldsymbol{\theta}}_{\mathrm{w}}^{\mathrm{adp}}$ also has **oracle properties**.

Oracle properties

Theorem 2.2 (Subsample IPW adaptive lasso)

Under some regularity conditions:

1. *Consistency in variable selection:* $\lim_{N\to\infty} \mathbb{P}(\hat{\mathcal{A}}_{\mathrm{w}} = \mathcal{A}) = 1$.
2. *Asymptotic normaility:*

$$\sqrt{N_1}(\widehat{\boldsymbol{\theta}}^{\mathrm{adp}}_{\mathrm{w}(\mathcal{A})} - \boldsymbol{\theta}_{\mathrm{t}(\mathcal{A})}) \xrightarrow{D} \mathbb{N}(\mathbf{0}, \mathbf{V}_{\mathrm{w}(\mathcal{A})}), \tag{12}$$

where $\mathbf{V}_{\mathrm{w}(\mathcal{A})} = \mathbf{V}_{\mathrm{mle}(\mathcal{A})} + \mathbf{V}_{\mathrm{sub}(\mathcal{A})}$, *and*

$$\mathbf{V}_{\mathrm{sub}(\mathcal{A})} = c\mathbb{E}\{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}\}\mathbf{M}^{-1}_{(\mathcal{A})}\mathbb{E}\left[\varphi^{-1}(\boldsymbol{x}_{(\mathcal{A})})e^{2f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}\dot{\boldsymbol{g}}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right]\mathbf{M}^{-1}_{(\mathcal{A})}.$$

Here $c\mathbb{E}\{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}\} = \lim_{P} \frac{N_1}{N_0\rho}$, the positive/negative ratio in the subsample.

Optimal subsampling for adaptive lasso

Proposition 1 (Optimal subsampling functions)

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

$$\varphi_{\mathrm{A-OS}}^{\mathrm{adp}}(\boldsymbol{x}) = \frac{p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\mathbf{M}_{(\mathcal{A})}^{-1}\dot{\boldsymbol{g}}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|}{\mathbb{E}\left\{p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\mathbf{M}_{(\mathcal{A})}^{-1}\dot{\boldsymbol{g}}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\right\}}. \tag{13}$$

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

$$\varphi_{\mathrm{L-OS}}^{\mathrm{adp}}(\boldsymbol{x}) = \frac{p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\dot{\boldsymbol{g}}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|}{\mathbb{E}\left\{p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\dot{\boldsymbol{g}}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\right\}}. \tag{14}$$

Inactive variables may affect optimal function

- Optimal functions only dependent on **active variables**.
- We need to estimate $\mathcal{A}$ using pilot estimates.
- It is unavoidable to include **inactive variables** in pilot estimates.
- Existing optimal functions affected by **inactive variables** in practice.

Illustration of the issue of scale dependence

Consider a logistic regression with independent features.

$$\varphi^{\mathrm{mle}}_{\mathrm{A-OS}}(\boldsymbol{x}) \propto p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\sqrt{K_{\mathcal{A}} + \mathbb{E}\{e^{\boldsymbol{x}_{(\mathcal{A})}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{t}}}\}^{-2}\sum_{j\in\mathcal{A}^c}\frac{x_{(j)}^2}{\mathbb{V}(x_{(j)})^2}}, \tag{15}$$

$$\varphi^{\mathrm{mle}}_{\mathrm{L-OS}}(\boldsymbol{x}) \propto p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\sqrt{1 + \sum_{j\in\mathcal{A}} x_{(j)}^2 + \sum_{j\in\mathcal{A}^c} x_{(j)}^2}. \tag{16}$$

If we rescale $\boldsymbol{x}_{(\mathcal{A}^c)}$ to $\tau\boldsymbol{x}_{(\mathcal{A}^c)}$, we have

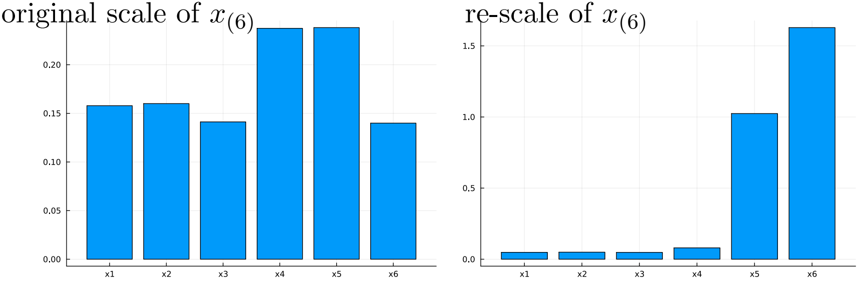
$$\varphi^{\mathrm{mle}}_{\mathrm{A-OS}}(\boldsymbol{x}) \propto p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\sqrt{K_{\mathcal{A}} + \frac{1}{\tau^2}\mathbb{E}\{e^{\boldsymbol{x}_{(\mathcal{A})}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{t}}}\}^{-2}\sum_{j\in\mathcal{A}^c}\frac{x_{(j)}^2}{\mathbb{V}(x_{(j)})^2}}, \tag{17}$$

$$\varphi^{\mathrm{mle}}_{\mathrm{L-OS}}(\boldsymbol{x}) \propto p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\sqrt{1 + \sum_{j\in\mathcal{A}} \boldsymbol{x}_{(j)}^2 + \tau^2\sum_{j\in\mathcal{A}^c} x_{(j)}^2}. \tag{18}$$

Numerical illustration

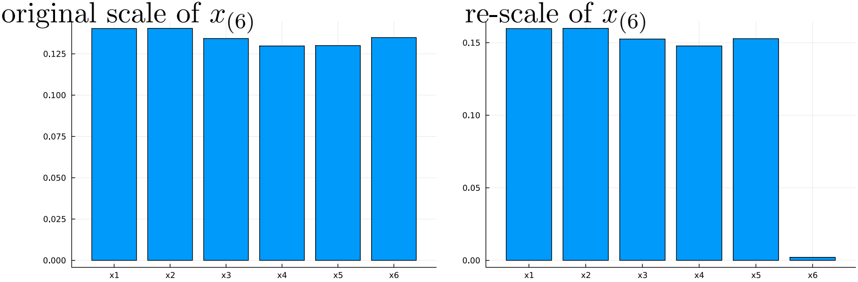
- Consider a logistic model with $\boldsymbol{\beta}_{\mathrm{t}} = (1, 1, 0, 0, 0, 0)$.
- Components of $\boldsymbol{x}_{(\mathcal{A}^c)}$ are not independent.
- Variance of each variable is 1 in the original scale.
- We draw barcharts of contributions of each variable to optimal probabilities.
- We multiply 0.1 to $x_{(6)}$ (an inactive variable) to show the impact of scaling on sampling probabilities.

Contribution of variables to optimal probabilities


original scale of $x_{(6)}$

re-scale of $x_{(6)}$

(a) A-optimality probabilities


original scale of $x_{(6)}$

re-scale of $x_{(6)}$

(b) L-optimality probabilities

Scale invariant optimal function

- Consider the **prediction error** of an estimator $\widehat{\boldsymbol{\theta}}$.

$$\mathrm{MSPE}(\widehat{\boldsymbol{\theta}}) = \int \left\{p(\boldsymbol{x};\widehat{\boldsymbol{\theta}}) - p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\right\}^2 \mathrm{d}\mathbb{P}_{\boldsymbol{x}},$$

Theorem 2.3 (Asymptotic distribution of predition error)

Under some regularity conditions, the prediction error satisfies

$$N_1 e^{-2\alpha_{\mathrm{t}}} \mathrm{MSPE}(\widehat{\boldsymbol{\theta}}^{\mathrm{adp}}_{\mathrm{w}(\mathcal{A})}) \tag{19}$$

$$\xrightarrow{D} \mathbb{E}^{-1}\left\{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}\right\} \mathbf{Z}^{\mathrm{T}}_{(\mathcal{A})} \mathbf{M}^{1/2}_{\mathrm{w}(\mathcal{A})} \mathbf{M}^{-1}_{(\mathcal{A})} \boldsymbol{\Omega}_{(\mathcal{A})} \mathbf{M}^{-1}_{(\mathcal{A})} \mathbf{M}^{1/2}_{\mathrm{w}(\mathcal{A})} \mathbf{Z}_{(\mathcal{A})}. \tag{20}$$

where $\mathbf{Z}_{(\mathcal{A})} \sim \mathbb{N}(\mathbf{0}, \mathbf{I})$ *and* $\boldsymbol{\Omega}_{(\mathcal{A})} = \mathbb{E}\left\{e^{2f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})} \dot{\boldsymbol{g}}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x}, \boldsymbol{\theta}_{\mathrm{t}})\right\}$.

Scale invariant optimal function

Theorem 2.4 (Scale invariant optimal function)

Minimizing ***the asymptotic mean of the prediction error*** *gives*

$$\varphi^{\mathrm{adp}}_{\mathrm{P-OS}}(\boldsymbol{x}) = \frac{p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\boldsymbol{\Omega}^{\frac{1}{2}}_{(\mathcal{A})}\mathbf{M}^{-1}_{(\mathcal{A})}\dot{\boldsymbol{g}}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|}{\mathbb{E}\left\{p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\boldsymbol{\Omega}^{\frac{1}{2}}_{(\mathcal{A})}\mathbf{M}^{-1}_{(\mathcal{A})}\dot{\boldsymbol{g}}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\right\}}. \tag{21}$$

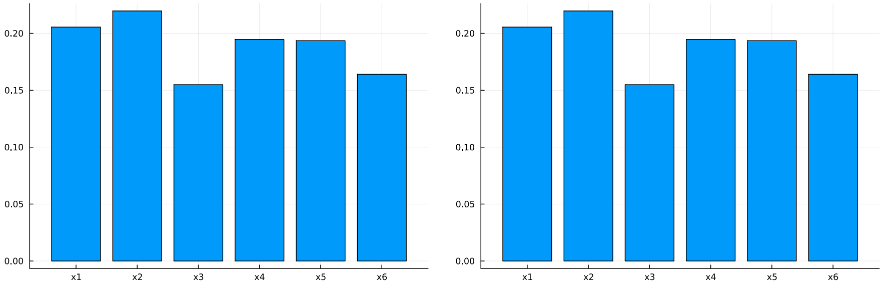
Proposition 2 (Scale invariant property)

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

$$g(\mathbf{A}\boldsymbol{x};\mathbf{B}^{\mathrm{T}}\boldsymbol{\theta}) = g(\boldsymbol{x};\boldsymbol{\theta}), \tag{22}$$

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

Contribution of variables to scale invariant optimal probabilities



(a) Original scale of $\boldsymbol{x}$.

(b) Re-scale of $\boldsymbol{x}$.

Outline

1 Introduction

2 Subsampling with Variable selection

3 Penalized MSCL estimator and its theoretical analysis

4 Numerical experiments

The limitation of the IPW estimator

- Remember the penalized IPW estimator

$$\widehat{\boldsymbol{\theta}}_{\mathrm{w}}^{\mathrm{adp}} := \arg\max_{\boldsymbol{\theta}} \left\{ \ell_{\mathrm{w}}^{\mathrm{sub}}(\boldsymbol{\theta}) - \lambda_{\mathrm{w}} \sum_{j=1}^{p} \frac{|\beta_{(j)}|}{|\hat{\beta}_{\mathrm{pl}(j)}|^{\gamma}} \right\}, \tag{23}$$

where

$$\ell_{\mathrm{w}}^{\mathrm{sub}}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \delta_i \frac{y_i g(\boldsymbol{x}_i;\boldsymbol{\theta}) - \log\{1 + e^{g(\boldsymbol{x}_i;\boldsymbol{\theta})}\}}{\pi(\boldsymbol{x}_i, y_i)}. \tag{24}$$

- Data points with **higher** $\pi(\boldsymbol{x}_i, y_i)$ have **lower** weights.
- The estimation effciency can be further improved.

Maximum sampled conditional likelihood (MSCL)

Instead of penalized IPW, we proposed a penalized MSCL estimator

$$\widehat{\boldsymbol{\theta}}_{\text{lik}}^{\text{adp}} := \arg\max_{\boldsymbol{\theta}} \left\{ \ell_{\text{lik}}^{\text{sub}}(\boldsymbol{\theta}) - \lambda_{\text{lik}} \sum_{j=1}^{p} \frac{|\beta_{(j)}|}{|\hat{\beta}_{\text{pl}(j)}|^{\gamma}} \right\}, \tag{25}$$

where

$$\ell_{\text{lik}}^{\text{sub}}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \delta_i [y_i g(\boldsymbol{x}_i; \boldsymbol{\theta}) - \log\{1 + e^{g(\boldsymbol{x}_i;\boldsymbol{\theta}) + l_i}\}], \tag{26}$$

and $l_i = -\log\{\rho\varphi(\boldsymbol{x}_i)\}$.

- MSCL estimator is **more effcient** than IPW (Wang and Kim, 2022; Wang *et al.*, 2021).

Practical implementation

1:
- 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 pilot estimator with **lasso penality**. We call this **first stage screening**.
- Record $\hat{\mathcal{A}}_{\mathrm{pl}} = \{j : \hat{\beta}_{\mathrm{pl}(j)} \neq 0\}$ and calculate approximate optimal sampling probabilities.

2: Use the estimated optimal sampling probabilities to obtain a subsample and the adaptive lasso estimator:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{lik}}^{\mathrm{adp}} := \arg\max_{\boldsymbol{\theta}} \left\{ \ell_{\mathrm{lik}}^{\mathrm{sub}}(\boldsymbol{\theta}) - \lambda_{\mathrm{lik}} \sum_{j \in \hat{\mathcal{A}}_{\mathrm{pl}}} \frac{|\beta_{(j)}|}{|\hat{\beta}_{\mathrm{pl}(j)}|^{\gamma}} \right\}, \tag{27}$$

We call this step the **second stage screening**.

Oracle properties

Theorem 3.1 (MSCL adaptive lasso estimator)

Under some regularity conditions:

1. *Consistency in variable selection:* $\lim_{N\to\infty} \mathbb{P}(\hat{\mathcal{A}}_{\text{lik}} = \mathcal{A}) = 1$
2. *Asymptotic normality:*

$$\sqrt{N_1}\mathbf{V}_{\text{lik}(\mathcal{A})}^{-1/2}(\widehat{\boldsymbol{\theta}}_{\text{lik}(\mathcal{A})}^{\text{adp}} - \boldsymbol{\theta}_{\text{t}(\mathcal{A})}) \xrightarrow{D} \mathbb{N}(\mathbf{0}, \mathbf{I}), \tag{28}$$

where $\mathbf{V}_{\text{lik}(\mathcal{A})} = \mathbb{E}\left\{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\text{t}})}\right\}\boldsymbol{\Lambda}_{\text{lik}(\mathcal{A})}^{-1}$ *and*

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

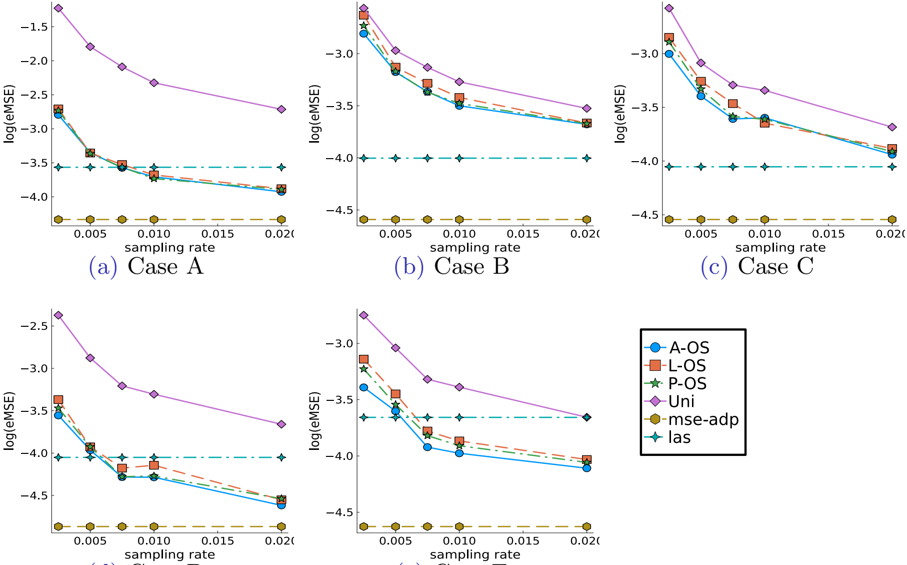
Outline

1. Introduction
2. Subsampling with Variable selection
3. Penalized MSCL estimator and its theoretical analysis
4. Numerical experiments

Simulations

- Consider a logistic regression, i.e. $g(\boldsymbol{x};\boldsymbol{\theta}) = \alpha + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$.
- Imbalance rate: 0.5%.
- Five cases of parameters $\boldsymbol{\beta}$:
  - A $\boldsymbol{\beta}_{\mathrm{t}} = (3, 1.5, 0, 0, 2, 0, 1, 0, 1, \underbrace{0, , ..., 0}_{41})^{\mathrm{T}}$.
  - B $\boldsymbol{\beta}_{\mathrm{t}} = (0.65, 0.65, \underbrace{0, , ..., 0}_{7}, 0.65, 0, 0.65, \underbrace{0, , ..., 0}_{7}, 0.65, \underbrace{0, , ..., 0}_{30})^{\mathrm{T}}$.
  - C $\boldsymbol{\beta}_{\mathrm{t}} = (0.75, 0.75, \underbrace{0, ..., 0}_{7}, 0.75, 0, 0.75, 0.75, \underbrace{0, ..., 0}_{37})^{\mathrm{T}}$.
  - D $\boldsymbol{\beta}_{\mathrm{t}} = (3, 2, \underbrace{0, ..., 0}_{7}, 0.85\, \underbrace{0, ..., 0}_{40})^{\mathrm{T}}$.
  - E $\boldsymbol{\beta}_{\mathrm{t}} = (3, -2, \underbrace{0, ..., 0}_{7}, 0.85, 0, -0.75, \underbrace{0, ..., 0}_{18})^{\mathrm{T}}$.
- Two scenarios of distributions of covariates $\boldsymbol{x}$:
  1. Some inactive variables have large variances.
  2. Some inactive variables have small variances.

eMSE for Scenarios 1



(a) Case A

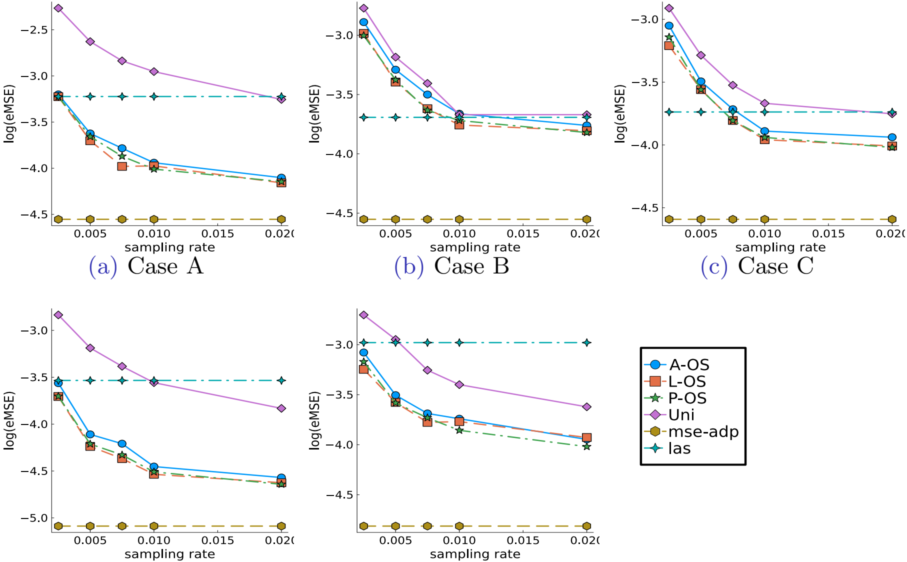
(b) Case B

(c) Case C

(d) Case D

(e) Case E

eMSE for Scenarios 2



(a) Case A

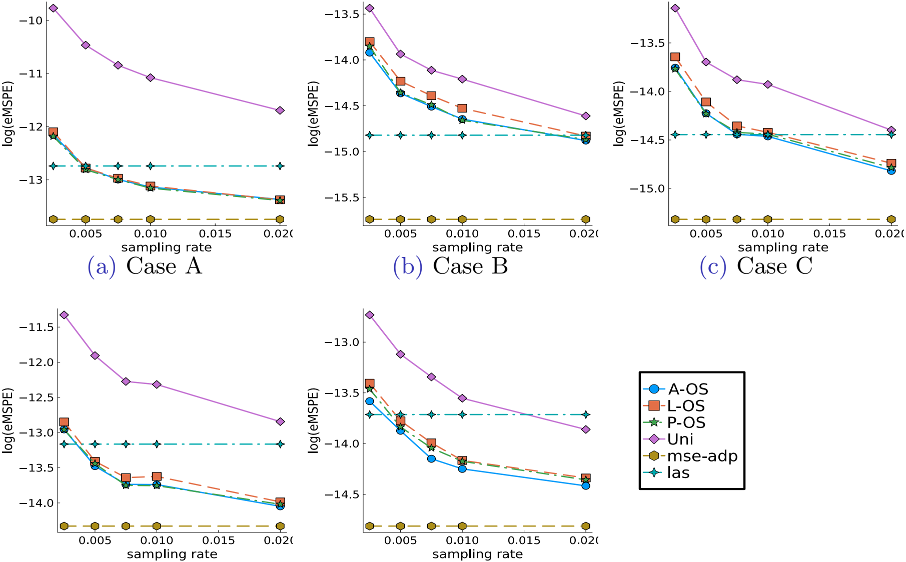
(b) Case B

(c) Case C

(d) Case D

(e) Case E

eMSPE for Scenarios 1



(a) Case A

(b) Case B

(c) Case C

(d) Case D

(e) Case E

Variable selection

Table 2: Mean number of selected variables

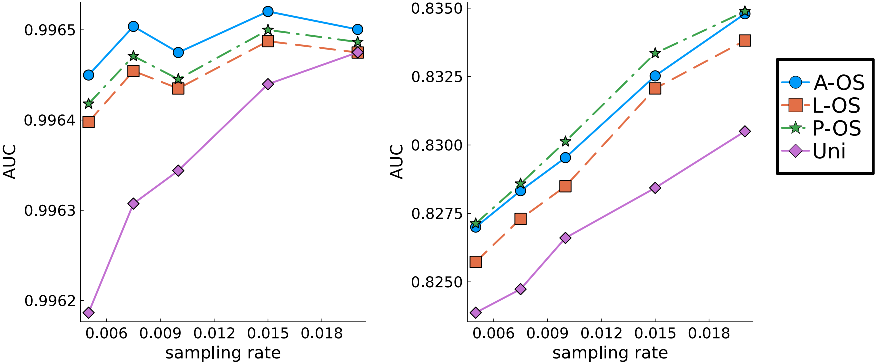
	Case D (three active variables)				
ρ	first-stage	Uni	A-OS	L-OS	P-OS
0.0025	11.12	2.95	3.05	3.06	3.05
0.005	11.55	2.99	3.09	3.10	3.11
0.0075	11.62	3.01	3.08	3.11	3.08

Table 3: Rates of excluding active variables

	Case D			
ρ	Uni	A-OS	L-OS	P-OS
0.0025	0.102	0.050	0.058	0.056
0.005	0.076	0.040	0.040	0.040
0.0075	0.070	0.040	0.040	0.040

Real data performance

- **Covtype:** $N = 581012, p = 52$, Cottonwood/Willow:0.473%
- **Font:** $N = 832670, p = 407$, GADUGI:0.5%



(a) AUC of covtype data set

(b) AUC of font data set

Some take-away

1. No need to use the full data for rare events data; focus on the rare ones.
2. For negative subsampling rare events data, optimal design is not relevant if sufficient zeros can be included.
3. Oracle properties are nice, but we are not the Oracle.
4. Minimizing the prediction error produces scale invariant optimal probabilities.
5. The conditional likelihood is often better than the IPW.

Thank you!

Some references

Wang, H. (2020). Logistic regression for massive data with rare events. In *ICML*, 8264–8271. .

Wang, H. and Kim, J. K. (2022). Maximum sampled conditional likelihood for informative subsampling. *Journal of Machine Learning Research* **23**, 332, 1–50.

Wang, H., Zhang, A., and Wang, C. (2021). Nonuniform negative sampling and log odds correction with rare events data. In *NeurIPS 2021.*

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

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101**, 476, 1418–1429.