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

Jing Wang

University of Connecticut

[1]Collaborates: Haiying Wang[1], Hao Helen Zhang[2]
[2]Institutions: University of Connecticut[1], University of Arizona[2]

Motivation

Rare-events data with sparse models

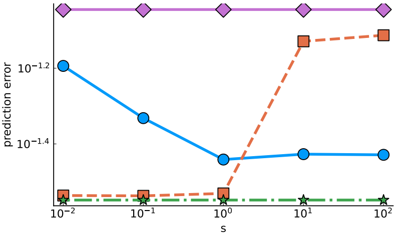
- Rare-events data are **highly imbalanced** binary response data.
  - Rare diseases, click-data on recommendation system.
  - Massive, highly imbalanced.
  - Invovle sparse models, e.g., limited number of key genes related to rare-diseases.
  - Variable selection is not studied.

- **Subsampling** is a popular approach for rare-events data analysis.
  - Data balancing, reducing computational burdens.
  - Usually done with strategy:
    1. Keeping all ones.
    2. Subsampling zeros according to an important function $\varphi(\boldsymbol{x})$.
  - Non-uniform subsampling reduces information loss.

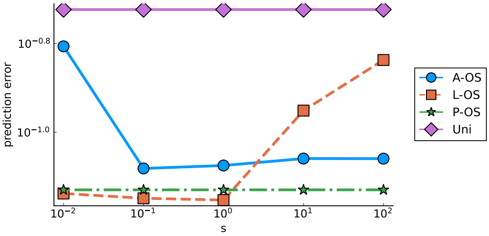
Motivation

Scale-dependent issue

- Existing optimal subsampling functions are scale-dependent.
- May lead to inefficient results.
- A wide concern in literature for various data types and models.



(a) Non-sparse parameter



(b) Sparse parameter

Problem setup

- Rare-events Model:

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

  Then, $\alpha_{\mathrm{t}} \to -\infty$ as $N \to \infty$ implies that $\frac{N_1}{N_0} \to 0$.

- IPW Adaptive Lasso for Variable selection

$$\hat{\boldsymbol{\theta}}_{\mathrm{w}}^{\mathrm{adp}} := \arg\max_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{N_{\mathrm{sub}}^{*}} \frac{\ell_i^{\mathrm{sub}}}{\pi(\boldsymbol{x}_i^{\mathrm{sub}}, y_i^{\mathrm{sub}})} - \lambda_N \sum_{j=1}^{p} \frac{|\beta_{(j)}|}{|\hat{\beta}_{\mathrm{pl}(j)}|^{\gamma}} \right\}, \quad (1)$$

  where $\ell_i^{\mathrm{sub}} = y_i^{\mathrm{sub}} g(\boldsymbol{x}_i^{\mathrm{sub}};\boldsymbol{\theta}) - \log\{1+e^{g(\boldsymbol{x}_i^{\mathrm{sub}};\boldsymbol{\theta})}\}$.

- Both optimal probabilities and adaptive lasso requires a pilot estimator. It is natural to combine them into one unified framework.

Theoretical analysis

Asymptotic properties of $\hat{\boldsymbol{\theta}}_{\mathrm{w}}^{\mathrm{adp}}$

1. Consistency in variable selection: $\lim_{N\to\infty} \mathbb{P}(\hat{\mathcal{A}}_{\mathrm{w}} = \mathcal{A}) = 1$
2. Asymptotic normality: $\sqrt{N_1}\,\boldsymbol{V}_{\mathrm{w}(\mathcal{A})}^{-1/2}(\hat{\boldsymbol{\theta}}_{\mathrm{w}(\mathcal{A})}^{\mathrm{adp}} - \boldsymbol{\theta}_{\mathrm{t}(\mathcal{A})}) \rightsquigarrow \mathbb{N}(0, \boldsymbol{I})$,

$$\boldsymbol{V}_{\mathrm{w}(\mathcal{A})} = \underbrace{\boldsymbol{V}_{\mathrm{mle}(\mathcal{A})}^{-1}}_{\text{Full data}} + \underbrace{c\,\boldsymbol{V}_{\mathrm{sub}(\mathcal{A})}}_{\text{Information loss}}, \text{ where}$$

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

- $c = \lim_{N\to\infty} \frac{e^{\alpha_{\mathrm{t}}}}{\rho}$ is the imbalance rate in the subsample.

Message from theoretical analysis

- The asymptotic variances $\boldsymbol{V}_{\mathrm{mle}(\mathcal{A})}$ and $\boldsymbol{V}_{\mathrm{w}(\mathcal{A})}$ are of order $\frac{1}{N_1}$.
- If **remain enough 0's**, e.g., $c = 0$, there will be **no information loss**.
- In case there is information loss $c > 0$, we can choose $\varphi(\boldsymbol{x})$ to minimize the information loss.

Optimal subsampling function

Traditional optimal subsampling function and limitations

1. **A-optimality**:
   $\min \operatorname{tr}(\boldsymbol{V}_{\mathrm{w}(\mathcal{A})}) \Rightarrow \varphi^{\mathrm{adp}}_{\mathrm{A-OS}}(\boldsymbol{x}) \propto p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\boldsymbol{M}^{-1}_{(\mathcal{A})}\dot{g}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|$.
2. **L-optimality**:
   $\min \operatorname{tr}(\boldsymbol{M}_{\mathrm{w}(\mathcal{A})}) \Rightarrow \varphi^{\mathrm{adp}}_{\mathrm{L-OS}}(\boldsymbol{x}) \propto p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|\dot{g}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|$.

- If $g(\boldsymbol{x},\boldsymbol{\theta}) = \alpha + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$, then $\varphi^{\mathrm{adp}}_{\mathrm{L-OS}}(\boldsymbol{x}) \propto p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})(1+\|\boldsymbol{x}_{(\mathcal{A})}\|)$.
  - Due to inaccurate pilot, scale of $\boldsymbol{x}_{(\mathcal{A}^c)}$ will affect $\hat{\varphi}^{\mathrm{adp}}_{\mathrm{L-OS}}(\boldsymbol{x})$.
- Construct optimal function by focusing on **prediction error**:

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

Optimal subsampling function

Scale-invariant optimal subsampling function

1. We prove that

$$N_1 e^{-2\alpha_{\mathrm{t}}} \mathrm{MSPE}(\hat{\boldsymbol{\theta}}^{\mathrm{adp}}_{\mathrm{w}(\mathcal{A})}) \rightsquigarrow \mathbb{E}^{-1}\left\{e^{f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})}\right\} \boldsymbol{Z}^{\mathrm{T}}_{(\mathcal{A})} \boldsymbol{L}^{\mathrm{T}}_{(\mathcal{A})} \boldsymbol{\Omega}_{(\mathcal{A})} \boldsymbol{L}_{(\mathcal{A})} \boldsymbol{Z}_{(\mathcal{A})}.$$

where $\boldsymbol{Z}_{(\mathcal{A})} \sim \mathbb{N}(0, \boldsymbol{I})$, $\boldsymbol{\Omega}_{(\mathcal{A})} = \mathbb{E}\left[e^{2f(\boldsymbol{x};\boldsymbol{\beta}_{\mathrm{t}})} \dot{g}^{\otimes 2}_{(\mathcal{A})}(\boldsymbol{x}, \boldsymbol{\theta}_{\mathrm{t}})\right]$, and $\boldsymbol{L}_{(\mathcal{A})} = \boldsymbol{M}^{-1}_{(\mathcal{A})} \boldsymbol{M}^{1/2}_{\mathrm{w}(\mathcal{A})}$.

2. The optimal function that minimizes the asymptotic mean is

$$\varphi^{\mathrm{adp}}_{\mathrm{P-OS}}(\boldsymbol{x}) \propto p(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}}) \|\boldsymbol{\Omega}^{\frac{1}{2}}_{(\mathcal{A})} \boldsymbol{M}^{-1}_{(\mathcal{A})} \dot{g}_{(\mathcal{A})}(\boldsymbol{x};\boldsymbol{\theta}_{\mathrm{t}})\|,$$

which is scale-invariant for a class of g including neural networks.

Penalized MSCL estimator and practical algorithm

- The IPW assigns smaller weights for more informative data points
- To improve the estimation efficiency, let $l_i^{\mathrm{sub}} = -\log\left\{\rho\varphi(\boldsymbol{x}_i^{\mathrm{sub}})\right\}$,

$$\hat{\boldsymbol{\theta}}_{\mathrm{mscl}}^{\mathrm{adp}} := \arg\max_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{N_{\mathrm{sub}}^*} \ell_{\mathrm{mscl},i}^{\mathrm{sub}} - \lambda_N \sum_{j=1}^{p} \frac{|\beta_{(j)}|}{|\hat{\beta}_{\mathrm{pl}(j)}|^{\gamma}} \right\}, \tag{2}$$

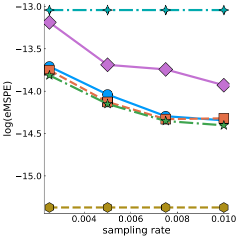
where $\ell_{\mathrm{mscl},i}^{\mathrm{sub}} = y_i^{\mathrm{sub}} g(\boldsymbol{x}_i^{\mathrm{sub}};\boldsymbol{\theta}) - \log\{1 + e^{g(\boldsymbol{x}_i^{\mathrm{sub}};\boldsymbol{\theta}) + l_i^{\mathrm{sub}}}\}$.

- **Efficiency:** $\boldsymbol{V}_{\mathrm{mscl}(\mathcal{A})} \leq \boldsymbol{V}_{\mathrm{w}(\mathcal{A})}$, and $\boldsymbol{V}_{\mathrm{mscl}(\mathcal{A})} = \boldsymbol{V}_{\mathrm{mle}(\mathcal{A})}$ if $c = 0$.

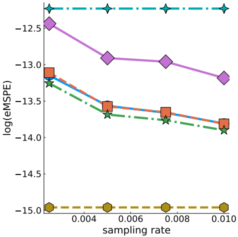
Practical Algorithm

1. **First-stage screening**:
    1. Take a pilot sample, and obtain a lasso estimator.
    2. Estimate $\hat{\varphi}(\boldsymbol{x}_i)$ for $i = 1, ..., N$ and $\hat{\mathcal{A}}$.
2. **Second-stage screening**: Subsampling from 0's with $\hat{\varphi}(\boldsymbol{x}_i)$, and compute adaptive lasso.

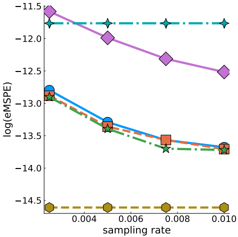
1. **Case A:** Small active effects.
2. **Case B:** Large and small active effects, different signs
3. **Case C:** Large and small active effects, same signs.



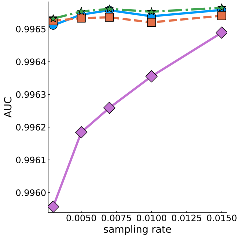
(a) Case A



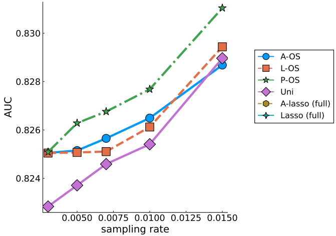
(b) Case B



(c) Case C



(d) covtype



(e) font

Conclusion

Conclusion

1. For rare-events data with sparse models, subsampling estimators can be as efficient as full data estimators under the true model
2. Traditional optimal functions are scale-dependent. The scale-invariant function based on predition error is a better choice.

Limitation and Future work

1. Optimal functions are based on asymptotic normality and asymptotic mean square error.
2. Optimal functions based on the quality of variable selection.
3. Non-asymptotic behaviors.

Thank you!