Unweighted Esitmation based on Optimal sample under Measurement Constraints

Jing Wang, HaiYing Wang, Shifeng Xiong

Department of Statistics, University of Connecticut and AMSS, Chinese Academy of Sciences

{jing.7.wang}@uconn.edu

Introduction

Big data bring new challenges to data storage and processing, especially when computational resources are limited. Researchers have developed many subsampling methods for various models, such as linear, logistic and generalized linear models(GLMs) (see Ma *et al.* (2015), Wang *et al.* (2018), Ai *et al.* (2021)). Most algorithms developed for GLMs rely on all responses of the full data, which limits the application scope of subsampling when responeses are difficult to acuqire. To handle this problem, Zhang *et al.* (2021) proposed a response-free optimal sampling shceme. However, they use a reweighted estimator which assigns smaller weights for more informative data points. Thus, their approach is not efficient. We introduce an unweighted estimator to improve the estimating efficiency and investigate the theoretical properties of both estimators. Asymptotic nomality is established using martingale techniques without conditioning on pilot estimation, which has been less investigated in existing subsampling literature. Both theoretical analysis and numerical experiments show that our estimator is more efficient and has a better performance without increasing computational complexity.

Background and model setup

We consider GLMs:$f(y|x,\beta_0,\sigma) \propto \exp\{yx^T\beta_0 - b(x^T\beta_0)/c(\sigma)\}$, where β_0 is the unknown parameter, $b(\cdot)$ and $c(\cdot)$ are known functions, and σ is the dispersion parameter. The maximum likelihood estimator (MLE) of β_0 is as following:

$$\hat{\beta}_{\text{MLE}} := \arg\max_{\beta} \frac{1}{n}\sum_{i=1}^{n}\{Y_iX_i^T\beta - b(X_i^T\beta)\},$$

The computational burden of computing MLE is usually intensive facing massive data. In some situation, it is costy to measure responses, which makes existing subsampling methods, such as OSMAC (Ai *et al.* (2021)), hard to implement. To handle these difficulties, Zhang *et al.* (2021) proposed optimal subsampling under measurement constraints. Considering $\{\pi_i\}_{i=1}^n$ as sampling probabilities, we use a reweighted estimator to obtain subsampling estimator:

$$\hat{\beta}_{\text{w}} := \arg\max_{\beta} \frac{1}{r}\sum_{i=1}^{r}\frac{Y_i^*X_i^{*T}\beta - b(X_i^{*T}\beta)}{n\pi_i^*}, \tag{1}$$

where * denote values obtained from sampling. Using this model setup, Zhang *et al.* (2021) developed optimal sampling probability under measurement constraints (OSUMC) through A-optimality criterion:

$$\pi_i^{\text{OS}}(\beta_0,\Phi) = \frac{\sqrt{b''(X_i^T\beta_0)}\|\Phi^{-1}X_i\|}{\sum_{j=1}^{n}\sqrt{b''(X_j^T\beta_0)}\|\Phi^{-1}X_j\|} \tag{2}$$

Unweighted Algorithm

- **Problem of OSUMC** We can notice that in (1), an inverse probability weight is used to estimate β_0. As pointed in Wang (2019), the weighting scheme does not bring us the most efficient estimator because intuitively, if a data point (X_i, Y_i) has a larger sampling probability, it contains more information about β_0. However, in (1), data points with higher sampling proabilities have smaller weights, which reduce the estimation efficiency. Thus, we propose a **more efficient estimator** based on the unweighted target function. We define our estimator as following:

$$\hat{\beta}_{\text{uw}} := \arg\max_{\beta} \frac{1}{r}\sum_{i=1}^{r}\{Y_i^*X_i^{*T}\beta - b(X_i^{*T}\beta)\}, \tag{3}$$

- **Unweighted algorithm** We propose the following two-step unweighted estimating procedure

Algorithm Unweighted estimation for GLM under measurement constraints

1. Take a pilot subsample of size r_{p}: $\{(X_i^{*\text{p}}, Y_i^{*\text{p}})\}_{i=1}^{r_{\text{p}}}$ with simple random sampling from the full data set $\{(X_i,Y_i)\}_{i=1}^n$. Calculate the pilot estimate of β_0, $\hat{\beta}_{\text{p}}$, and the pilot estimate of Φ, $\hat{\Phi}_{\text{p}}$
2. Use $\hat{\beta}_{\text{p}}$ and $\hat{\Phi}_{\text{p}}$ to replace β_0 and Φ in (2) and caculate the sampling probabilities $\{\pi_i^{\text{OS}}(\hat{\beta}_{\text{p}},\hat{\Phi}_{\text{p}})\}_{i=1}^n$.
3. Obtain a subsample $\{(X_i^*, Y_i^*)\}_{i=1}^r$ of size r according to the sampling probabilities $\{\pi_i^{\text{OS}}(\hat{\beta}_{\text{p}},\hat{\Phi}_{\text{p}})\}_{i=1}^n$ using sampling with replacement, and solve the estimation equation:

$$\Psi_{\text{uw}}^*(\beta) := \frac{1}{r}\sum_{i=1}^{r}\{b'(X_i^{*T}\beta) - Y_i^*\}X_i^* = 0,$$

to obtain the unweighted estimator defined in (3).

Theoretical analysis of unweighted algorithm

- **Asymptotic normality** Under some regularity conditions

$$\sqrt{r}(\hat{\beta}_{\text{uw}} - \beta_0) \xrightarrow{d} N(0,\Sigma_{\text{uw}}^{\rho}), \quad \Sigma_{\text{uw}}^{\rho} := m\Gamma^{-1} + \rho\Gamma^{-1}\Omega\Gamma^{-1}$$

- **Efficiency comparison** We can restate the results in Zhang *et al.* (2021) as:

$$\sqrt{r}(\hat{\beta}_{\text{w}} - \beta_0) \xrightarrow{N} (0,\Sigma_{\text{w}}^{\rho}), \quad \Sigma_{\text{w}}^{\rho} := m\Phi^{-1}\Gamma\Phi^{-1} + \rho\Phi^{-1}$$

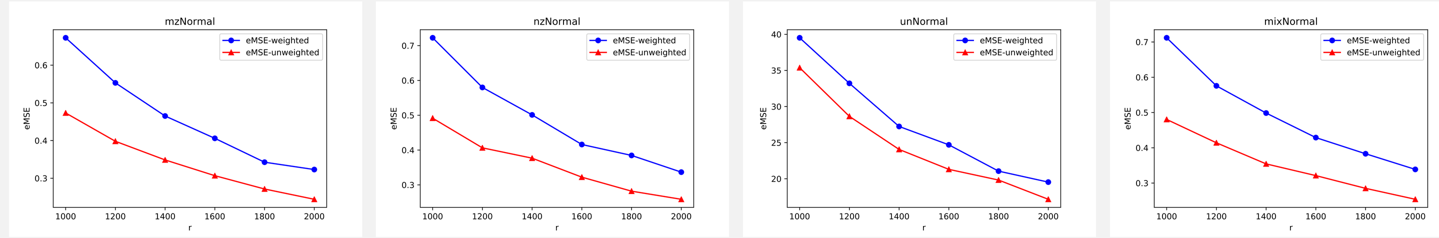
We can prove that

$$\Gamma^{-1} \le \Phi^{-1}\Gamma\Phi^{-1}, \quad \text{and } \Gamma^{-1}\Omega\Gamma^{-1} \ge \Phi^{-1}.$$

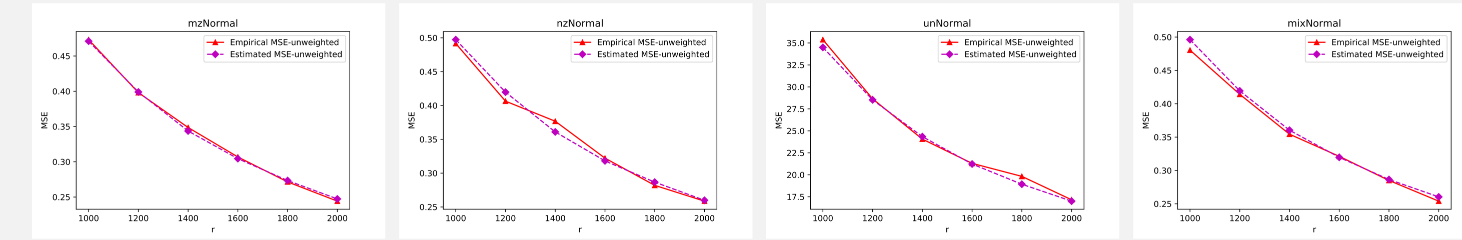
Therefore, under subsampling scenario, usually $r/n \to 0$, we know that unweighted algorithm is more efficient for parameter estimation

Numerical experiments

- **Logistic model**

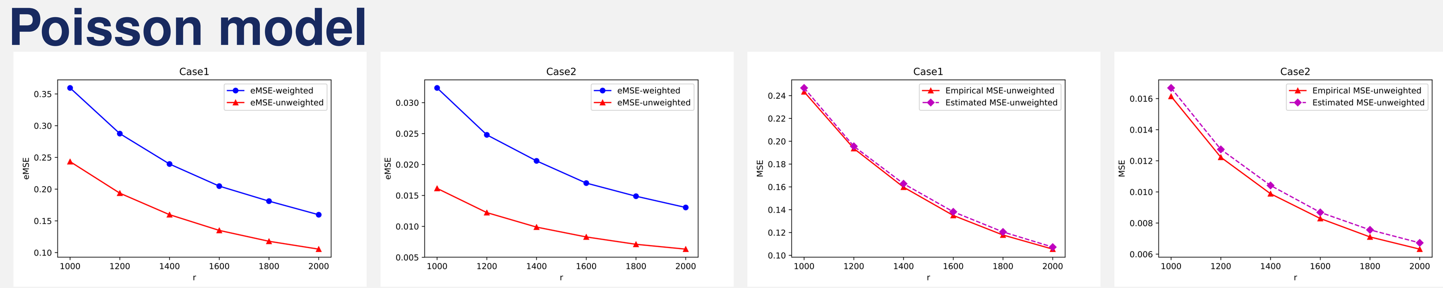


(a) mzNormal (b) nzNormal (c) unNormal (d) mixNormal



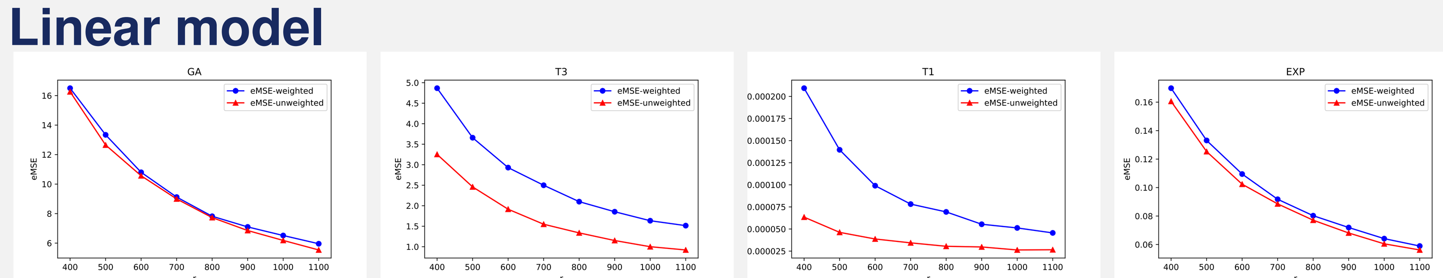
(e) mzNormal (f) nzNormal (g) unNormal (h) mixNormal

- **Poisson model**



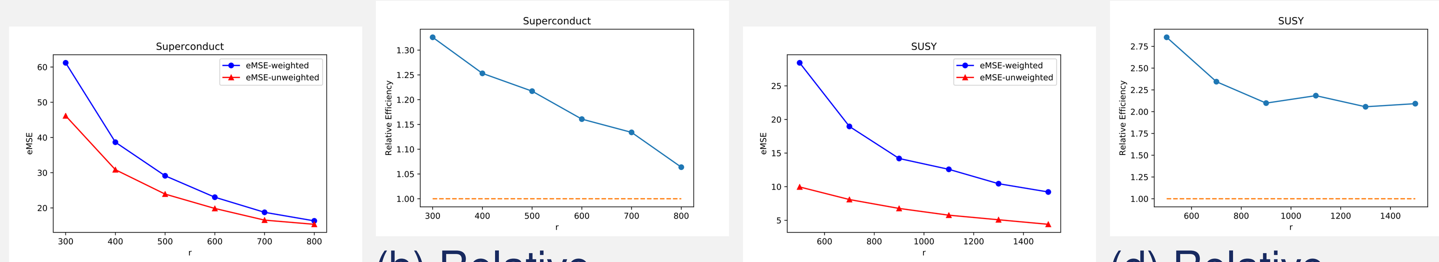
(a) Case1 (b) Case2 (c) Case1 (d) Case2

- **Linear model**



(a) GA (b) T3 (c) T1 (d) EXP

- **Real data**



(a) eMSE (Superconduct) (b) Relative Efficiency (Superconduct) (c) eMSE (SUSY) (d) Relative Efficiency (SUSY)

Conclusion

Both theoretic and numerical results guarantee the better performance of unweighted esitmator.

References

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

Ma, P., Mahoney, M. W., and Yu, B. (2015). A statistical perspective on algorithmic leveraging. *The Journal of Machine Learning Research* **16**, 1, 861–911.

Wang, H. (2019). More efficient estimation for logistic regression with optimal subsamples. *Journal of Machine Learning Research* **20**, 132, 1–59.

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

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