OPTIMAL SUBSAMPLING FOR BIG DATA GENERALIZED LINEAR MODELS

A PREPRINT

Fei Tan
Department of Mathematical Sciences, IUPUI
ftan@iupui.edu

Hanxiang Peng
Department of Mathematical Sciences, IUPUI
402 N Blackford St., LD 270, Indianapolis, IN 46202, USA
hanxpeng@iupui.edu

September 10, 2023

ABSTRACT

Motivated by the computational burden in fitting a single index model caused by high parameter dimensionality and possibly compounded by data of massive size, we propose an optimal subsampling approach based on the criterion of maximum information. We derive the optimal sampling distributions and prove asymptotic normality of the subsampling estimator for either/both fixed or/and growing number of the index parameters p or/and spline knots d as the subsample size r tends to infinity with the sample size n. We construct a fast algorithm with running time $O(r^2(p+d))$ with r far less than n, and study the behaviors of the proposed approach using both simulated and real data.

Keywords A-optimality · Asymptotic normality · Big data · Infinite dimension · Penalized spline · Single index model

1 Introduction

2 SIMULATION STUDY

In this section, we use simulation studies to evaluate the A-optimal subsampling approach proposed in previous sections. The design matrix $\mathbf{X}$ is generated from one of the four following multivariate distributions. (1) Gaussian distribution $N(0,\Sigma)$, $\Sigma_{i,j} = 0.3^{|i-j|}$. (2) Mixture Gaussian distribution with $\frac{1}{2}N(0,\Sigma) + \frac{1}{2}N(0,3\Sigma)$. (3) Log-normal distribution $LN(0,\frac{1}{2}\Sigma)$. (4) The t distribution with 5 degree of freedom $\mathbf{T}_5(0,\frac{1}{2}\Sigma)$. We choose $n = 50,000$, $p = 50$, $\boldsymbol{\beta} = (0.1, -0.1 \times \mathbf{1}^{\top}_{(p/2)}, 0.1 \times \mathbf{1}^{\top}_{(p/2)})$. We consider the response y_i from Poisson distribution and Negative Binomial distribution with variance structure $V(y_i) = \mu_i + 5\mu_i^2$. We use logarithm link in the above two situations:

$\log(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}, \quad i = 1, \cdots, n$. The subsampling probabilities are calculated from the following formulas:

$$\begin{aligned}
\hat{\pi}_i^{(2)} &= \frac{\|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-1}\mathbf{x}_i\||\hat{e}_i|}{\sum_{i=1}^n \|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-1}\mathbf{x}_i\||\hat{e}_i|}, \\
\hat{\pi}_i^{(1)} &= \frac{\|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-\frac{1}{2}}\mathbf{x}_i\||\hat{e}_i|}{\sum_{i=1}^n \|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-\frac{1}{2}}\mathbf{x}_i\||\hat{e}_i|}, \\
\hat{\pi}_i^{(0)} &= \frac{\|\mathbf{x}_i\||\hat{e}_i|}{\sum_{i=1}^n \|\mathbf{x}_i\||\hat{e}_i|}, \\
\bar{\pi}_i^{(2)} &= \frac{\|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-1}\mathbf{x}_i\|\hat{g}_i}{\sum_{i=1}^n \|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-1}\mathbf{x}_i\|\hat{g}_i}, \\
\bar{\pi}_i^{(1)} &= \frac{\|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-\frac{1}{2}}\mathbf{x}_i\|\hat{g}_i}{\sum_{i=1}^n \|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-\frac{1}{2}}\mathbf{x}_i\|\hat{g}_i}, \\
\bar{\pi}_i^{(0)} &= \frac{\|\mathbf{x}_i\|\hat{g}_i}{\sum_{i=1}^n \|\mathbf{x}_i\|\hat{g}_i}, \quad i = 1, \cdots, n.
\end{aligned} \tag{2.1}$$

For Poisson regression,

$$\begin{aligned}
W(\hat{\boldsymbol{\beta}}) &= \mathrm{Diag}(\hat{\mu}_i), \quad \hat{\mu}_i = \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}}), \\
\hat{e}_i = y_i - \hat{\mu}_i, &\quad \hat{g}_i = \sqrt{\hat{\mu}_i} \quad i = 1, \cdots, n.
\end{aligned} \tag{2.2}$$

For Negative Binomial distribution,

$$\begin{aligned}
W(\hat{\boldsymbol{\beta}}) &= \mathrm{Diag}(\frac{\hat{\mu}_i}{1+\alpha\hat{\mu}_i}), \quad \hat{\mu}_i = \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}}), \\
\hat{e}_i = \frac{y_i - \hat{\mu}_i}{1+\alpha\hat{\mu}_i}, &\quad \hat{g}_i = \sqrt{\frac{\hat{\mu}_i}{1+\alpha\hat{\mu}_i}} \quad i = 1, \cdots, n.
\end{aligned} \tag{2.3}$$

We take subsamples of size r according to the sampling distributions calculated from the above step, and obtain the estimate $\hat{\boldsymbol{\beta}}_r^*$. We calculate the empirical mean square errors at different subsample sizes r for each of $B = 1000$ subsamples using the following formula:

$$MSE = \frac{1}{B}\sum_{b=1}^{B} \|\hat{\boldsymbol{\beta}}_{r,b}^* - \hat{\boldsymbol{\beta}}\|^2,$$

where $\hat{\boldsymbol{\beta}}_{r,b}^*$ is the estimate from the b^{th} subsample with subsample size r.

3 Full SAMPLE REAL DATA ANALYSIS: BIKE SHARING DATA

3.1 Introduction of the Real Data

This data set is available from the UCI Machine Learning Repository website. Bike sharing systems can be considered new generation of old-fashioned bike rentals. The use of this system is not restricted to rentals and returns at the same docking station, bikes can be returned to any docking station after usage. Predicting the hourly bike request will help in planing, expanding and maintaining adequate number of bikes. In United States, the bike sharing system has been proved to be very successful in major cities, including Washington, DC, New York, Chicago, Los Angles, where bikes sharing has become a popular transportation option. Our goal in this example is to build statistical models to predict hourly request of bikes in Washington DC area. There are totally 17,389 observations in this data set, the response variables are the counts of casual rentals, registered rentals, and total rented bikes including both casual and registered. We split the data into two sets, using the 2011 year data to build the models and the 2012 data to calculate the prediction error. The predictor variables are season, workingday, daytime, weathersit, temp, hum and windspeed. The season variables include spring indicator, summer indicator and fall indicator, winter is the reference level. Variable workingday indicates whether a day is a working day, the reference level is weekend of holiday. Variable daytime indicates if the time is between 7am to 22pm, with the referencing time range from 0am to 5am representing the reference level night time. Weathersit varaible has 4 categories, the first category represents clear, few clouds, and partly cloudy; the second category represents mist plus cloudy, mist plus broken clouds, and mist; the third category represents light snow, light rain; and the fourth category represents heavy rain, and snow plus frog. Since the fourth category only has three observations, we combine the third and fourth categories. We choose the first category as the reference. Temp is normalized temperature in Celsius. Hum is normalized humidity. Windspeed is normalized wind speed. There are 11 regression coefficients for the predictor variables including the intercept.

4 A-OPTIMAL SUBSAMPLING FOR REAL DATA ANALYSIS: BIKE SHARING DATA

We now apply our proposed subsampling methods to conduct Quasipoisson regression and Negative Binomial regression. The subsampling probabilities are caculated as follows:

$$
\begin{aligned}
\hat{\pi}_i^{(2)} &= \frac{\|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-1}\mathbf{x}_i\||\hat{e}_i|}{\sum_{i=1}^n \|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-1}\mathbf{x}_i\||\hat{e}_i|},\\
\hat{\pi}_i^{(1)} &= \frac{\|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-\frac{1}{2}}\mathbf{x}_i\||\hat{e}_i|}{\sum_{i=1}^n \|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-\frac{1}{2}}\mathbf{x}_i\||\hat{e}_i|},\\
\hat{\pi}_i^{(0)} &= \frac{\|\mathbf{x}_i\||\hat{e}_i|}{\sum_{i=1}^n \|\mathbf{x}_i\||\hat{e}_i|},\\
\bar{\pi}_i^{(2)} &= \frac{\|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-1}\mathbf{x}_i\|\hat{g}_i}{\sum_{i=1}^n \|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-1}\mathbf{x}_i\|\hat{g}_i},\\
\bar{\pi}_i^{(1)} &= \frac{\|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-\frac{1}{2}}\mathbf{x}_i\|\hat{g}_i}{\sum_{i=1}^n \|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-\frac{1}{2}}\mathbf{x}_i\|\hat{g}_i},\\
\bar{\pi}_i^{(0)} &= \frac{\|\mathbf{x}_i\|\hat{g}_i}{\sum_{i=1}^n \|\mathbf{x}_i\|\hat{g}_i}, \quad i = 1, \cdots, n.
\end{aligned}
\tag{4.1}
$$

For Quasioisson regression,

$$
\begin{aligned}
W(\hat{\boldsymbol{\beta}}) &= \mathrm{Diag}(\frac{\hat{\mu}_i}{\hat{\phi}}), \quad \hat{\mu}_i = \exp(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}),\\
\hat{e}_i &= \frac{y_i - \hat{\mu}_i}{\hat{\phi}}, \quad \hat{g}_i = \sqrt{\frac{\hat{\mu}_i}{\hat{\phi}}} \quad i = 1, \cdots, n.
\end{aligned}
\tag{4.2}
$$

For Negative Binomial regression,

$$
\begin{aligned}
W(\hat{\boldsymbol{\beta}}) &= \mathrm{Diag}(\frac{\hat{\mu}_i}{1+\hat{\alpha}\hat{\mu}_i}), \quad \hat{\mu}_i = \exp(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}),\\
\hat{e}_i &= \frac{y_i - \hat{\mu}_i}{1+\hat{\alpha}\hat{\mu}_i}, \quad \hat{g}_i = \sqrt{\frac{\hat{\mu}_i}{1+\hat{\alpha}\hat{\mu}_i}} \quad i = 1, \cdots, n,
\end{aligned}
\tag{4.3}
$$

where $\hat{\phi}$ and $\hat{\alpha}$ are the moment estimators based on full sample. We take a uniform subsample of size $r_0 = 200$ to get an initial estimates $\hat{\boldsymbol{\beta}}^*_{r_0}$; then plug in to the formulas to approximate the A-optimal subsampling distributions; take subsample of size r according to the approximate distributions to get $\hat{\boldsymbol{\beta}}^*_r$; We calculate the empirical mean square errors for different subsample sizes r for each of $B = 1000$ subsamples using the formula:

$$
MSE = \frac{1}{B}\sum_{b=1}^{B} \|\hat{\boldsymbol{\beta}}^*_{r,b} - \hat{\boldsymbol{\beta}}\|^2,
$$

where $\hat{\boldsymbol{\beta}}^*_{r,b}$ is the estimate from the b^{th} subsample with subsample size r.

5 Casual Bike Rentals

5.1 Quasipoisson Regression Model for Casual Data

5.2 Negative Binomial Regression Model for Casual Data

6 A-OPTIMAL SUBSAMPLING FOR REAL DATA ANALYSIS: BLOG FEEDBACK DATA

In this chapter, we apply the proposed subsampling method to analyze the Blog Feedback data set, available from the UCI machine learning repository. This data was collected and processed from raw html of the blog posts. The goal is to predict the number of comments in the upcoming 24 hours relative to the base time. The base time was chosen from the past, and the blog posts selected were published within 72 hours before base time. The features were recorded at the base time based on the selected blog posts.

There are 52,397 observations in the training data set, and 7,624 observations in the test data set. We use the training data set to build the model, and the test data set to calculate the prediction errors. The 23 features are total number of comments before base time (Tc), number of comments in the 24 hours right before the base time (Cl24), number of comments in the time period between T1 and T2 (Ct1t2), where T1 denotes the date time 48 hours before base time, T2 denotes the date time 24 hours before base time, number of comments in 24 hours immediately after publication of the post but before base time (Cf24), total number of trackbacks before base time (Tt), number of trackbacks in the last 24 hours before the base time (Tl24), number of trackbacks between T1 and T2 (Tt1t2), where T1 is the time point 48 hours before basetime and T2 the time point 24 hours before basetime, number of trackbacks within 24 hours immediately after publication of the post but before basetime (Tf24), the length of time between the publication of the blog post and base time (Ltime), the length of the blog post (Lbp), indicators (0 or 1) for whether Monday to whether Saturday of the base time (Mbt, Tbt, Wbt, THbt, Fbt, Sbt), indicators (0 or 1) for whether Monday to whether Saturday of the blog publication date (Mpb, Tpb, Wpb, THpb, Fpb, Spb), number of parent pages(Ppage).

Poisson regression model is not appropriate for this data set because of observed overdispersion in data and inflated number of zeros. Quasipoisson regression model has the same parameter estimates as the Poisson regression model and does not accommodate zero-inflation, so it is not a good choice either. Zero-inflated Poisson regression model allows inflated zeros hence is an appropriate choice.

As 64.05% of the values in the response variable are 0, we shall consider fitting the zero-inflated Poisson regression model for this data set. The estimating equation of zero-inflated Poisson regression contains the parameter $0 \leq \rho \leq 1$, which accounts for the amount of positive structural zeros beyond the sampling zeros explained by the Poisson distribution f_{poi}. In the literature, ρ can be modeled as a function of the predictor variables, for example, via the logistic link. Here for simplifying the estimating process, we shall estimate ρ first. Specifically, based on the interpretation of ρ and noting that 64.05% is the proportion of zeros in the response variable while $\exp(\mu)$ is the probability of taking zero value in the Poisson distribution, we estimate ρ by

$$\hat{\rho} = 0.6405 - \exp(-\hat{\mu}). \tag{6.1}$$

where $\hat{\mu}$ is an estimator of μ. As Y follows the zero inflated model **??**, we have

$$P(Y = 0) = \rho + (1 - \rho)\exp(-\mu).$$

On the other hand, $E(Y) = (1 - \rho)\mu$. Thus $\mu = E(Y)/(1 - \rho)$ and we get

$$P(Y = 0) = \rho + (1 - \rho)\exp(-E(Y)/(1 - \rho).$$

The moment equation of this is

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[Y_i = 0] = \rho + (1 - \rho)\exp(-\bar{Y}/(1 - \rho)).$$

Since $\bar{y} = 6.765$, $\hat{p} = 0.6405$, we solve the equtions to get $\hat{\rho} \approx \hat{p} = 0.6405$. Alternatively, we can use 6.1 to get $\hat{\rho}$ by plugging in $\hat{\mu} = 214.9628$, yielding the same value.

To compare Poisson, Quasipoisson, with zero-inflated Poisson regression models, we report the full sample estimates, standard errors, P-values for these three models in Table (7.1). Many parameters in the Quasipoisson model are not significant, while these parameters in zero-inflated Poisson model are significant. Also, we perform proposed subsampling method in the zero-inflated Poisson regression model.

References

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

[2] Bickel, P. J. and Freedman, D. A. (1983). Bootstrapping regression models with many parameters. In A Festschrift for Erich L. Lehmann (P. J. Bickel, K. A. Doksum and J. L. Hodges, Jr., eds.) 28-48. Wadsworth, Belmont, CA.

[3] Chung, K. L. (2001) A Course in Probability Theory. Academic Press.

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

[5] Djalil, C., Didier, C. (2007). On the strong consistency of asymptotic M Estimators. J. Statist. Plan. Infer. **137**:2774-2783.

[6] Drineas P., Kannan R. and Mahoney M.W. (2006a). Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. SIAM Journal on Computing, **36**: 132-157.

[7] Drineas P., Mahoney M.W. and Muthukrishnan S. (2006b). Sampling algorithms for l2 regression and applications. Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms, pages 1127-1136.

[8] Drineas, P., Mahoney, M., Muthukrishnan, S. & Sarlos, T. (2011). Faster least squares approximation. Numerische Mathematik, **117**: 219-249.

[9] Efron, B. (1979), Bootstrap methods: another look at the jackknife. Annals of Statistics, **7**: 1-26.

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

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

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

[13] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, **96**(456): 1348-1360.

[14] Hardle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. Annals of Statistics, **21**(1): 157-178.

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

[16] Hoerl, E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. Journal of the American Statistical Association, **12**(1): 55-67.

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

[18] Mammen, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. Annals of Statistics, **17**: 382-400.

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

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

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

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

[23] Peng, Hanxiang and Tan, Fei (2018). Big data linear regression via A-optimal subsampling. Submitted to the Annals of Statistics, 2018.

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

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

[26] Ruppert, D. (2002). Selecting the number of knots for penalized splines. Journal of the American Statistical Association, **11**(4): 735-757.

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

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

[29] Ruppert D., Wand M., Carroll R. Semiparametric Regression. New York: Cambridge University Press, 2003.

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

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

[32] Wang, G. and Wang, L. (2015). Spline estimation and variable selection for single index prediction models with diverging number of index parameters. Journal of Statistical Planning and Inference, **162**: 1–19.

[33] 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.

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

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

[36] Xia, Y. C. and Hardle, W. (2006). Semi-parametric estimation of partially linear single-index models. Journal of Multivariate Analysis. **97**: 1162-1184.

[37] Xia, Y. C., Tong, H., LI, W. K. and Zhu, L. (2002). An adaptive estimation of dimension reduction space (with discussion). Journal of the Royal Statistical Society. Series B, Statistical methodology, **64**: 363-410.

[38] Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. Journal of the American Statistical Association, **97**(460): 1042-1054.

[39] Yurinskii, V. (1976). Exponential Inequalities for Sums of Random Vectors. Journal of Multivariate Analysis, **6**: 473-499.

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

[41] Y. le Cun, Modeles Connexionnistes de l'Apprentissage. PhD thesis, Universite Pierre et Marie Curie, Paris, France, 1987.

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