View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by IUPUIScholarWorks

A-OPTIMAL SUBSAMPLING FOR BIG DATA GENERAL ESTIMATING EQUATIONS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Chung Ching Cheung

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Hanxiang Peng, co-Chair
    Department of Mathematical Sciences, IUPUI
Dr. Leonid Rubchinsky, co-Chair
    Department of Mathematical Sciences, IUPUI
Dr. Benzion Boukai
    Department of Mathematical Sciences, IUPUI
Dr. Guang Lin
    Department of Mathematics, Purdue University
Dr. Mohammad AL Hasan
    Department of Computer and Information Science, IUPUI

Approved by:

Dr. Eugene Mukhin
Head of the Graduate Program
Department of Mathematical Sciences, IUPUI

This thesis is dedicated to my dear parents and my wife Haley Yu.

ACKNOWLEDGMENTS

I want to express my deepest gratitude to Prof. Hanxiang Peng and Prof. Leonid Rubchinsky for their contribution of time, energy and patience which make my journey for PhD pursuit fruitful and enjoyable. During my first two years of PhD study in IUPUI, Prof. Rubchinsky has broadened my horizon on neuroscience, he spent lots of time teaching me about neuroscience. I also feel grateful to know Leonid's post-doctoral researcher Dr. Shiva Ratnadurai who, together with Leonid, guide me on the research patiently and wholeheartedly. It is my honor to produce some good results with them. In my last three years of PhD study, Prof. Peng has brought me to the world of statistics. I have learned a lot from Prof. Peng, in particular on optimal subsampling methods of big data which is a very hot topic in statistics nowadays. His endless encouragement and guidance has motivated me to produce some interesting results in my research.

Besides, I would like to thank Prof. Benzion Boukai for his help on my internship, Prof. Jyoti Sarkar for introducing me the interesting project that we work together, and Prof. Honglang Wang for his time and effort spent on discussing with me about high dimensional statistics. My sincere thanks also go to Prof. Guang Lin and Prof. Hasan Mohammad AL who generously become my thesis committee.

Last but not the least, I would like to appreciate those who have supported my study and research in IUPUI. I would also want to thank my wife Haley Yu for her encouragement and support throughout these years of my PhD study.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

Cheung, Chung Ching Ph.D., Purdue University, August 2019. A-Optimal Subsampling for Big Data General Estimating Equations. Major Professor: Hanxiang Peng.

A significant hurdle for analyzing big data is the lack of effective technology and statistical inference methods. A popular approach for analyzing data with large sample is subsampling. Many subsampling probabilities have been introduced in literature (Ma, *et al.*, 2015) for linear model. In this dissertation, we focus on generalized estimating equations (GEE) with big data and derive the asymptotic normality for the estimator without resampling and estimator with resampling. We also give the asymptotic representation of the bias of estimator without resampling and estimator with resampling. we show that bias becomes significant when the data is of high-dimensional. We also present a novel subsampling method called A-optimal which is derived by minimizing the trace of some dispersion matrices (Peng and Tan, 2018). We derive the asymptotic normality of the estimator based on A-optimal subsampling methods. We conduct extensive simulations on large sample data with high dimension to evaluate the performance of our proposed methods using MSE as a criterion. High dimensional data are further investigated and we show through simulations that minimizing the asymptotic variance does not imply minimizing the MSE as bias not negligible. We apply our proposed subsampling method to analyze a real data set, gas sensor data which has more than four millions data points. In both simulations and real data analysis, our A-optimal method outperform the traditional uniform subsampling method.

1. INTRODUCTION

This dissertation introduces the A-optimal subsampling method for estimators obtained by solving general estimating equations (GEE). We focus on the cases where the sample size n is large and the dimension p is high. Asymptotic distribution of the A-optimal subsampling is dervied and the magnitude of the corresponding bias is computed. The simulation results indicated that the A-optimal sampling outperformed the uniform sampling by the criterion of mean square errors.

1.1 Big Data Analysis

Thanks to the advancement in computing technology, big data has become a hot topic in statistics nowadays. Big data are data that is massive on scale where traditional computers cannot handle. A significant hurdle for analyzing big data is the lack of effective statistical computing and inference. Divide-and-Conquer is one common approach to tackle the big data problems. An important example is the MapReduce paradigm (Dean and Ghemawat, 2004) which processes large data sets in parallel fashion. On the other hand, subsampling method is popular in statistics to handle big data. In this method, one takes a small data set from the original large sample and uses it as a surrogate to perform statistical analysis. For example, uniform sampling is often used in subsampling for computational intensive problem. However, uniform sampling is not effective in extracting relevant information and the performance of the statistical inference can be very poor. See the simulations of uniform sampling by Peng and Tan (2018). Motivated by this, we seek non-uniform data-dependent sampling methods for big data in the framework of general estimating equations (GEE) by the criterion of A-optimality, that is, minimizing the traces of certain variance-covariance matrices.

A lot of work on non-uniform data-dependent sampling algorithms for data analysis problems can be found in literature. For instance, Ma and Sun (2014) and Ma, *et al.* (2015) has studied the leverage scores and used it as a non-uniform sampling distribution for linear regression. Ma, *et al.* (2015) proposed the shrinkage leveraging estimator (SLEV) and unweighted leveraging estimator (LEVUNW). They derived the bias and variance formulas for the weighted subsampling estimators. Drineas, *et al.* (2008) and Mahoney and Drineas (2009) has studied the matrix-based problems that are related to least squares approximation, where normalized statistical leverage scores are used as the non-uniform sampling distribution. Zhu, *et al.* (2015) obtained the optimal subsampling distribution for large sample linear regression, and proposed the predictor-length subsampling method. Wang, *et al.* (2018) constructed optimal subsampling for large sample in logistic regression model. Wang, *et al.* (2019) developed the information-based optimal subdata selection for big data linear regression where the subsampling method is based on D-optimality criterion. On the other hand, fast computational methods on subsampling algorithms have been well-studied by statisticians and computer scientists. Drineas, *et al.* (2006) constructed fast Monte-Carlo algorithms for approximating matrix multiplications. Drineas, *et al.* (2010) has studied randomized algorithms for least squares approximation, where the leverage scores are computed approximately. The monograph of Mahoney (2011) has a detailed discussion on this method. In this dissertation, we present non-uniform subsampling methods in GEE using A-optimality criterion. Peng and Tan (2018) proposed the A-optimal probability distribution for linear regression model. They derived the asymptotic expansions for the subsmapling estimator and the asymptotic normality under certain conditions. Also, they have proposed data truncation for fast computation.

1.2 Challenges of Big Data Analysis

Fan, *et al.* (2014) discussed that big data has the following salient features: (1) massive sample size and high dimension whereas traditional data set has sample size larger than the dimension, (2) heterogeneity of data, (3) noise accumulation, (4) spurious correlation and (5) incidental endogeneity. We will elaborate each of them in details and discuss the impacts of big data on statistical thinking.

1.2.1 Heterogeneity of data

Big data is always created by aggregating data from different data sources corresponding to different subpopulations. Each subpopulations may be characterized by its unique features which are distinct from others. Data from small subpopulations are considered as outliers. One way to deal with outliers is to remove them before performing statistical analysis. However, under big data era, even the smallest subpopulation will have significant size since the full sample size is massive. This helps us to better understand the heterogeneity of subpopulations. For example, the large amount of genome sequencing data enables us to discover the relationship between certain genes (covariates) and rare diseases (outcomes) (Worthey, *et al.*, 2010). This discovery is infeasible if the sample size is not large enough. Besides the benefits brought from big data, the heterogeneity of big data also comes with statistical and computational challenges. For instances, we have to impose some regularizations to avoid overfitting in finite mixture of regression models (Khalili and Chen, 2007).

1.2.2 Noise accumulation

Big data analysis often requires us to simultaneously estimate many parameters (high dimension). The noise from the data will be accumulated when estimating a large number of parameters. One way to tackle this problem is to assume sparsity on the model (Hastie, *et al.*, 2009). L_1- regularization method (Lasso) is used to select a subset of the variables which best describe the model. In fact, with more parameters included in the model, it will not only increase the noise which may even dominate the true signal, but also make the interpretation of the model more difficult as more parameters are considered. In our simulation studies, we will demonstrate how the noise accumulation in high dimensional data poses a challenge to statistical inference under our A-optimal subsampling.

1.2.3 Spurious correlation

Spurious correlation refers to the situation that many uncorrelated random variables may have correlation in high dimensions. For example, Fan, *et al.* (2008) demonstrated that when the dimensionality of the data is very high, variable selection becomes challenging since there could be high correlation between those significant variables and spurious variables. This may lead to unreliable statistical conclusion. In particular, Fan, *et al.* (2008) considered the case when the dimension p (800, 6400) is larger than the sample size n (60). They assume $x_1, \ldots, x_n$ be n independent observations of a p-dimensional Gaussian vector $X \sim N_p(0, I_p)$. The maximum absolute sample correlations between the first variable and the other variables, $r = \max_{i \geq 2} |corr(X_1, X_i)|$ with 1000 repetitions are then computed. The simulation results show that the maximum absolute correlation increases when the dimension increases.

1.2.4 Incidental endogeneity

Endogeneity refers to the fact that correlation exists between the variables X and the residual noise ε in a regression. This contradicts with the exogenous assumption that the predictors should be uncorrelated with the noise in regression model. Big data is more prone to have endogeneity problem because big data is usually an aggregation of data from multiple sources, this implies more measurement errors and thus the endogeneity problem. Also, Big data usually comes with high dimensionality, that is, more predictors are included in the model. This increases the possibility that some predictors are correlated with the residual noise. The existence of incidental endogeneity will make traditional statistical methods invalid, and the impact of it on high dimensional statistics is still not well understood.

1.3 Count Data regression

Zhao (2018) studied A-optimal subsampling theory with emphasis on count data regression model, for instance, Poisson regression model, zero-inflated Poisson regression model and negative binomial regression model. However, the sample size of the data is of ten of thousands which is not the usual size for Big Data. In this dissertation, we study Big Data with sample size n equals millions and dimension p equals 50. We develop A-optimal subsampling theory for general estimating equation (GEE) with arbitrary data structure, that is, data can be random or deterministic, dependent or independent. These results are parallel to those obtained in linear regression model in Peng and Tan (2018). We will also derive the Taylor expansions of the bias of regression parameter estimators with resampling and without resampling. Both Taylor expansions show that the magnitude of the remainder terms of the biases are significant when the data is of high-dimensional. We will show in the simulations that under big data with massive n and large p, the remainder term of the bias is not negligible. Also, we will demonstrate through simulations that the bias is not negligible when the random variable of the dataset does not have finite high-order moments.

This dissertation is organized as follows. We briefly review the nonuniform subsampling methods for linear regression model in Chapter 2. In Chapter 3, we give some classical examples of generalized estimating equations (GEE). In Chapter 4, we prove the asymptotic normality of the bias of estimators with resampling and without resampling. We also derive the general expression of the biases with specific order for the remainder terms. We discuss the theoretical results related to A-optimal distributions under GEE framework. We shall focus on asymptotic normality, asymptotic behaviors under A-optimal sampling for fixed dimension and growing dimension. Simulation studies of big data is presented in Chapter 5. A real data example will be given in Chapter 6 to demonstrate our methods.

2. LEAST-SQUARES AND LINEAR MODELS

In this chapter, we provide an overview of subsampling methods for the linear regression problems.

Consider a linear regression model

$$y_i = \mathbf{x}_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.0.1}$$

where $y_i \in \mathbb{R}$ is a response variable, $\mathbf{x}_i \in \mathbb{R}^p$ is a p-dimensional design vector, $\beta \in \mathbb{R}^p$ is a p-dimensional regression parameter and $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identical distributed (i.i.d.) random errors with zero mean $\mathbf{E}(\varepsilon) = 0$ and finite variance $\sigma^2 = Var(\varepsilon) < \infty$. We shall assume that $\mathbf{x}_i$'s are nonrandom although the results will hold also for random $\mathbf{x}_i$'s, and the true regression parameter satisfying the linear model is β_0.

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ be the $n \times p$ design matrix, $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ be the response vector and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$ be the error vector. Assume throughout that $\mathbf{X}$ has full rank. Then the linear model can be expressed as

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon} \tag{2.0.2}$$

A common estimator of β is the ordinary least squares (OLS) estimator $\hat{\beta}_{\text{ols}}$ given by

$$\hat{\beta}_{\text{ols}} = \text{argmin}_{\beta \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\beta\|^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{2.0.3}$$

where $\|\cdot\|$ represents the Euclidean norm on $\mathbb{R}^n$. The predicted value is given by $\hat{\mathbf{y}} = H\mathbf{y}$ where $H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the so-called hat matrix. The i^{th} diagonal element of the hat matrix $h_{ii} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$ is called the statistical leverage score of the i^{th} observation in literature, and we shall use the leverage score also to refer to the distribution. Note that $\sum_{i=1}^n h_{ii} = \text{tr}(H) = \text{tr}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) = \text{tr}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}) = p$. Statistical leverage scores have been used to quantify influential observation. It is clear that $h_{ii} = \dfrac{d\hat{y}_i}{dy_i}$, when h_{ii} tends to one, then $\hat{y}_i$ tends to y_i which implies the i^{th} observation is leveraged.

2.1 Subsmapling Methods

When the sample size n is extremely large, it becomes infeasible to compute the full sample OLS using conventional computer. An alternative way is to draw a subsample of size $r \ll n$ using certain sampling probability $\{\pi_i\}_{i=1}^n$, i.e., draw r rows from the original data according to $\{\pi_i\}_{i=1}^n$, and construct a weighted estimator $\hat{\beta}_r^*$ on the subsample. We summarize the weighted estimation algorithm as follows (Ma, *et al.* (2015), Zhu, *et al.* (2015)).

Weighted Estimation Algorithm (subsmapling with replacement)

- Step 1. Construct a sampling probability $\{\pi_i\}_{i=1}^n$ for all the data points $(\mathbf{x}_i, y_i)$. Use the distribution to draw a subsample of size $r \ll n$ and denoted it as $(\mathbf{X}^*, \mathbf{y}^*)$ with the corresponding probabilities $\boldsymbol{\pi}^*$.
- Step 2. Construct the weighted matrix $\mathbf{W}^* = \text{diag}\left\{\dfrac{1}{r\pi_j^*}\right\}_{j=1}^r$.
- Step 3. Compute the weighted least squares estimator as follows.

$$\hat{\beta}_r^* = (\mathbf{X}^{*T}\mathbf{W}^*\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\mathbf{W}^*\mathbf{y}^*.$$

Note that we can also use the weight matrix $\mathbf{W} = \text{diag}\left\{\dfrac{k_i}{r\pi_i}\right\}_{i=1}^n$ where k_i is the number of times the i^{th} data point has been selected. Then the weighted least squares estimator is computed as

$$\hat{\beta}_r = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}.$$

Denote $\mathbf{w} = (w_1, w_2, \ldots, w_n)^T$ as a random vector where w_i's are the diagonal entries of $\mathbf{W}$. Then $\mathbf{w}$ follows as a scaled multinomial distribution,

$$\mathbf{P}\left(w_1 = \frac{k_1}{r\pi_1}, w_2 = \frac{k_2}{r\pi_2}, \ldots, w_n = \frac{k_n}{r\pi_n}\right) = \frac{r!}{k_1!k_2!\cdots k_n!}\pi_1^{k_1}\pi_2^{k_2}\cdots\pi_n^{k_n}, \qquad (2.1.1)$$

The following lemma is useful in later chapters.

Lemma 2.1.1 *Consider the scaled multinomial distribution* $\mathbf{w}$ *as described in (2.1.1). We have the following results.*

1. $\mathbf{E}(w_i) = 1$, $\mathbf{E}(w_i^2) = \frac{1}{r}\left(\frac{1}{\pi_i} - 1\right) + 1$, *for* $i = 1, \ldots, n$,

2. $\mathbf{E}(w_i w_j) = 1 - \frac{1}{r}$, *for* $i \neq j$,

3. $\mathbf{E}[(w_i - 1)(w_j - 1)] = \begin{cases} \frac{1}{r}\left(\frac{1}{\pi_i} - 1\right), & \textit{for } i = j \\ -\frac{1}{r}, & \textit{for } i \neq j \end{cases}$

Rewrite the above results in matrix form, we have $\mathbf{V}(\mathbf{w}) = \mathbf{E}[(\mathbf{w} - \mathbf{1})(\mathbf{w} - \mathbf{1})^T] = diag\left\{\frac{1}{r\boldsymbol{\pi}}\right\} - \frac{1}{r}\mathbf{J}_n$, *where* $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^T$ *and* $\mathbf{J}_n$ *is a* $n \times n$ *matrix of ones.*

Proof of Lemma 2.1.1. To prove 1, note that the expected value of w_i is given $\mathbf{E}(w_i) = \frac{1}{r}\sum_k \frac{r!}{k_1!k_2!\cdots k_n!}\pi_1^{k_1}\pi_2^{k_2}\cdots\pi_n^{k_n}\left(\frac{k_i}{\pi_i}\right) = \frac{1}{r}\sum_k \frac{r!}{k_1!\cdots k_n!}\pi_1^{k_1}\cdots\pi_i^{k_i-1}\cdots\pi_n^{k_n}k_i$. On the other hand, we have

$$(\pi_1 + \cdots + \pi_n)^r = \sum_k \frac{r!}{k_1!\cdots k_n!}\pi_1^{k_1}\cdots\pi_n^{k_n}. \qquad (2.1.2)$$

Differentiate the left hand side of (2.1.2) with respect to π_i, we have $r(\pi_1 + \cdots + \pi_n)^{r-1} = r$, and differentiate the right hand side of (2.1.2) with respect to π_i gives $r\mathbf{E}(w_i)$. Hence, $\mathbf{E}(w_i) = 1$. Note that $\mathbf{E}(w_i^2) = \frac{1}{r^2}\sum_k \frac{r!}{k_1!\cdots k_n!}\pi_1^{k_1}\cdots\pi_i^{k_i-2}\cdots\pi_n^{k_n}k_i^2$. Differentiate (2.1.2) with respect to π_i twice gives $r(r-1) = r^2\mathbf{E}(w_i^2) - \frac{r}{\pi_i}\mathbf{E}(w_i)$, solving it gives $\mathbf{E}(w_i^2) = \frac{1}{r}\left(\frac{1}{\pi_i} - 1\right) + 1$.

To prove 2, for $i \neq j$, we have $\mathbf{E}(w_i w_j) = \frac{1}{r^2}\sum_k \frac{r!}{k_1!\cdots k_n!}\pi_1^{k_1}\cdots\pi_i^{k_i-1}\pi_j^{k_j-1}\cdots\pi_n^{k_n}k_ik_j$. Differentiate (2.1.2) with respect to π_i and π_j gives $r(r-1) = r^2\mathbf{E}(w_iw_j)$. Thus, $\mathbf{E}(w_iw_j) = 1 - \frac{1}{r}$.

To prove 3, note that when $i = j$, $\mathbf{E}[(w_i - 1)^2] = \mathbf{E}(w_i^2) - \mathbf{E}(w_i)^2 = \frac{1}{r}(\frac{1}{\pi_i} - 1)$, and for $i \neq j$, $\mathbf{E}[(w_i - 1)(w_j - 1)] = \mathbf{E}(w_i w_j) - 1 = -\frac{1}{r}$. ■

The weighted least squares estimator is determined by the probability distribution $\{\pi_i\}_{i=1}^n$. The followings are several subsampling distribution that have been discussed in literature (Ma, *et la.* (2014, 2015), Zhu, *et al.* (2015)).

- **Uniform Sampling Estimator (UNIF).** Draw the subsample according to the uniform sampling probability, $\pi_i = 1/n$. The corresponding weighted LS estimator is given by $\hat{\beta}_r^* = (\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\mathbf{y}^*$.

- **Basic Leveraging Estimator (LEV).** Draw the subsample according to the leverage scores, $\pi_i = \frac{h_{ii}}{p}$.

- **Shrinkage Leveraging Estimator (SLEV).** Consider the convex combination of the leverage scores distribution and the uniform distribution, $\pi_i = \alpha\frac{h_{ii}}{p} + (1 - \alpha)\frac{1}{n}$, where $\alpha \in (0, 1)$. This subsampling method was originally proposed by Ma, *et al.* (2014, 2015).

- **Unweighted Leveraging Estimator (LEVUNW).** The subsampling probability is the leverage scores with the computation of the unweighted LS estimator:

$$\hat{\beta}_r^{*u} = \text{argmin}_{\beta \in \mathbb{R}^p} \|y^* - \mathbf{X}^*\beta\|^2.$$

  This subsampling algorithm is originially proposed by Ma *et al.* (2015).

- **Optimal Subsampling Estimator (OPT).** Let $\pi_i = \frac{\sqrt{1-h_{ii}}\|\mathbf{x}_i\|}{\sum_{i=1}^n \sqrt{1-h_{ii}}\|\mathbf{x}_i\|}$. This is the subsampling probability which minimizes the expectation of the trace of $\mathbf{V}_c$, where $\mathbf{V}_c = \frac{1}{r}\sum_{i=1}^n \frac{e_i^2}{\pi_i}\mathbf{x}_i\mathbf{x}_i^T$ with $e_i = y_i - \mathbf{x}_i^T\hat{\beta}_{\text{ols}}$. This subsampling method is proposed by Zhu, *et al.* (2015). The computational cost of OPT is $O(np^2)$.

- **Predictor-length Subsampling Estimator (PL).** Let $\pi_i = \frac{\|\mathbf{x}_i\|}{\sum_{i=1}^n \|\mathbf{x}_i\|}$. This subsampling distribution is proposed by Zhu *et al.* (2015) as an improvement of (OPT) on the computation cost which is $O(np)$.

UNIF is simple to implement but the performance is usually poor. LEV is the first nonuniform data dependent subsampling method based on leverage scores. However, LEV can cause inflation of MSE due to the leverage scores being too small (Ma *et al* (2015)). SLEV take advantages of both leverage scores and uniform, it makes sure that the sampling probabilities is not too small (hence, avoid inflated variance) and sample respect to probabilities near the leverage scores so the influential data points can be selected. LEVUNW is different than LEV since they have different distributions of sampling and reweighting. Ma, *et al* (2015) showed that both SLEV and LEVUNW have better empirical performance than UNIF and LEV in terms of mean square errors. Zhu *et al.* (2015) showed empirically that both OPT and PL have better performance than LEV and UNIF, while OPT and PL have similar performance.

2.2 Asymptotic Theory for Weighted Subsampling Estimators

We now give some theoretical results of bias and variance of weighted LS estimators. These results are based on a series expansion of $\hat{\beta}_r^*$ around the expected value $\hat{\beta}_{\text{ols}}$ (Peng and Tan (2018), Ma *et al.* (2014, 2015), Zhu *et al.* (2015)).

Lemma 2.2.1 *Let $\hat{\beta}_r^*$ be the weighted least squares estimator obtained from the weighted estimation algorithm. Then we have the following expansion of $\hat{\beta}_r^*$ around $\hat{\beta}_{ols}$*

$$\hat{\beta}_r^* = \hat{\beta}_{ols} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T diag\{\hat{\mathbf{e}}\}(\mathbf{w} - \mathbf{1}) + R_W \tag{2.2.1}$$

where $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\beta}_{ols}$ is the residual, and R_W is the remainder.

Note that the randomness of the estimators is of two-fold, the randomness originated from the model as well as the randomness inherent from the subsampling distribution. Given Lemma 2.2.1, we can derive the conditional (concern with the randomness of the subsampling, and conditioned on the data $\mathbf{y}$) and unconditional (concern with the randomness of the model) expectations and variances of the weighted subsampling estimators. The following result is from Ma *et al.* (2015). Better results can also be found in Peng and Tan (2018).

Theorem 2.2.1 *Let $\hat{\beta}_r^*$ be the weighted least squares estimator obtained from the weighted estimation algorithm. The conditional expectation and variance are given by*

$$\begin{aligned} &\mathbf{E}^*(\hat{\beta}_r^*) = \hat{\beta}_{ols} + \mathbf{E}^*(R_W), \\ &\mathbf{V}^*(\hat{\beta}_r^*) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \left[diag\{\hat{\mathbf{e}}\} diag\left\{\frac{1}{r\boldsymbol{\pi}}\right\} diag\{\hat{\mathbf{e}}\} \right] \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} + \mathbf{V}^*(R_W). \end{aligned} \tag{2.2.2}$$

The unconditional expectation and variance are given by

$$\begin{aligned} &\mathbf{E}(\hat{\beta}_r^*) = \beta_0, \\ &\mathbf{V}(\hat{\beta}_r^*) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} + \frac{\sigma^2}{r}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T diag\left\{\frac{(1-h_{ii})^2}{\pi_i}\right\} \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} + \mathbf{V}(R_W). \end{aligned} \tag{2.2.3}$$

Remark 2.2.1 The conditional expectation means that when we compute the weighted subsampling estimators from a data set N times, the average of the N estimators will be centered roughly at the LS estimator given the remainder term is negligible. While the unconditional expectation means that from the true model, we first generate lots of data sets from it. Then we compute the weighted subsampling estimator from each of the data set, and the average of these estimators is centered at the true parameter.

Remark 2.2.2 The conditional and unconditional variances of $\hat{\beta}_r^*$ is inversely proportional to the subsample size r. Under LEV procedure, the second term of the variance could be inflated by very small leverage scores as $\pi_i = h_{ii}/p$. Under UNIF procedure, the second term of the variance depends on n/r which is larger than p/r from LEV when $p \ll n$. Thus, UNIF also has variance inflation problem.

3. GENERALIZED ESTIMATING EQUATIONS

One way to obtain estimators for regression parameters in statistics is by solving the some estimating equations. In this chapter, we give some examples of likelihood-based models and derive the estimation equations. We will also review the estimating equations for generalized linear model (GLM).

3.1 Linear regression with Normal distribution

Let Y has a normal (Gaussian) distribution $N(\mu, \sigma^2)$ where the mean is $\mathbf{E}(y) = \mu$ and $\mathbf{V}(y) = \sigma^2$. The probability density of Y is given by

$$f(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}. \tag{3.1.1}$$

The likelihood of n independent random variables $y_i \sim N(\mu_i, \sigma)$ for $i = 1, \ldots, n$ is given by

$$\begin{aligned} L(\boldsymbol{\mu}, \sigma^2|y_1, \ldots, y_n) &= \prod_{i=1}^{n} \exp\left\{-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(y_i-\mu_i)^2}{2\sigma^2}\right\} && (3.1.2) \\ &= \exp\left\{\sum_{i=1}^{n}\left(-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(y_i-\mu_i)^2}{2\sigma^2}\right)\right\} && (3.1.3) \end{aligned}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$. Consider the linear model where $g(\mu_i) = \mu_i = \mathbf{E}(y_i) = \mathbf{x}_i^T\beta$, where $\mathbf{x}_i \in \mathbb{R}^p$ is the covariates vector for the i^{th} observation, and $\beta \in \mathbb{R}^p$ is the regression parameters, and g is the link function is identity. The log-likelihood model for the linear regression is

$$L(\beta, \sigma^2|\mathbf{X}, y_1, \ldots, y_n) = \sum_{i=1}^{n}\left(-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(y_i - \mathbf{x}_i^T\beta)^2}{2\sigma^2}\right) \tag{3.1.4}$$

If the response variable Y is always positive, we could use the log-link $g(\mu_i) = \ln(\mu_i) = \mathbf{x}_i^T\beta$ which gives the log-linear regression model. To estimate the regression parameters, we use the maximum likelihood estimator which is obtained by solving the first derivative of the log-likelihood, $\frac{\partial L}{\partial \beta}$,

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} \mathbf{x}_i (y_i - \mathbf{x}_i^T \beta) = \mathbf{0} \tag{3.1.5}$$

This is the estimating equation for linear model. Solving it gives the ordinary least squares estimator $\hat{\beta}_{\text{ols}}$.

3.2 Poisson regression

Poisson distribution is a natural choice when the response variable is non-negative counts. Let Y has a Poisson distribution $Poi(\mu)$ where $\mathbf{E}(y) = \mathbf{V}(y) = \mu > 0$. The probability density of Poisson is given by

$$f(y|\mu) = \frac{e^{-\mu}\mu^y}{y!}, \quad y = 0, 1, 2, \ldots, \tag{3.2.1}$$

The likelihood of n independent random variables $y_i \sim Poi(\mu_i)$ for $i = 1, \ldots, n$ is given by

$$\begin{aligned} L(\boldsymbol{\mu}|y_1, \ldots, y_n) &= \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} && (3.2.2) \\ &= \prod_{i=1}^{n} \exp\{-\mu_i + y_i \ln(\mu_i) - \ln(y_i!)\} && (3.2.3) \end{aligned}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$. Consider the linear model with log link, $\ln(\mu_i) = \mathbf{x}_i^T \beta$. The log-likelihood model is

$$L(\beta, \boldsymbol{\mu}|\mathbf{X}, y_1, \ldots, y_n) = \sum_{i=1}^{n} \{\exp(\mathbf{x}_i^T \beta) + y_i \mathbf{x}_i^T \beta - \ln(y_i!)\} \tag{3.2.4}$$

Differentiate the log-likelihood with respect to β gives the estimating equation,

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} (y_i - \exp(\mathbf{x}_i^T \beta))\mathbf{x}_i = \mathbf{0} \tag{3.2.5}$$

The estimating equation can be solved by Newton-Raphson and iteratively reweighted least squares (IRLS) algorithms.

Note that in real data, the condition $\mathbf{E}(y) = \mathbf{V}(y)$ usually fails to hold. This is known as overdispersion problem. Negative binomial regression is commonly used to handle this problem.

3.3 Negative Binomial regression

Let Y follows a negative binomial distribution $NB(r,p)$ with the probability distribution given by

$$f(y|r,p) = \frac{\Gamma(y+r)}{y!\Gamma(r)} p^y (1-p)^r, \quad y = 0,1,2,\dots \tag{3.3.1}$$

where the mean $\mathbf{E}(y) = \frac{pr}{1-p}$, and variance $\mathbf{V}(y) = \frac{pr}{(1-p)^2}$. By making a transformation of the parameter

$$\mu := \mathbf{E}(y) = \frac{pr}{1-p} \Rightarrow p = \frac{\mu}{\mu + r} \tag{3.3.2}$$

and set $\alpha := \frac{1}{r}$, we have the density as

$$f(y|\mu,\alpha) = \frac{\Gamma(y+1/\alpha)}{y!\Gamma(1/\alpha)} \left(\frac{\mu}{\mu+1/\alpha}\right)^y \left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha}, \quad y = 0,1,2,\dots \tag{3.3.3}$$

Then, $\mathbf{V}(y) = \mu + \alpha\mu^2 > \mathbf{E}(y)$, and $\mathbf{E}(y) = \mathbf{V}(y)$ if and only if $\alpha = 0$. The term α is known as dispersion parameter or shape parameter. The negative binomial model can be used to handle the overdispersed Poisson. Consider a log-linear model with link function: $\mu_i = \mathbf{E}(y_i) = \exp(\mathbf{x}_i^T\beta)$, where $y_i \sim NB(\mu_i,\alpha), i = 1\dots,n$. Then the log likelihood model is

$$L(\boldsymbol{\mu},\alpha|\mathbf{X},y_1,\dots,y_n) = \sum_{i=1}^{n}\left(y_i \ln\left(\frac{\mu_i}{\mu_i+1/\alpha}\right) + \frac{1}{\alpha}\ln\left(\frac{1}{1+\alpha\mu_i}\right)\right) + C \tag{3.3.4}$$

$$= \sum_{i=1}^{n} y_i[\mathbf{x}_i^T\beta - \ln(\exp(\mathbf{x}_i^T\beta) + 1/\alpha)] \tag{3.3.5}$$

$$- \sum_{i=1}^{n} 1/\alpha \ln(1 + \alpha\exp(\mathbf{x}_i^T\beta)) + C \tag{3.3.6}$$

The estimating equation is given by

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} y_i \mathbf{x}_i - \frac{(y_i + 1/\alpha)\exp(\mathbf{x}_i^T \beta)\mathbf{x}_i}{\exp(\mathbf{x}_i^T \beta) + 1/\alpha} \tag{3.3.7}$$

$$= \sum_{i=1}^{n} \frac{(y_i - \exp(\mathbf{x}_i^T \beta))\mathbf{x}_i}{1 + \alpha \exp(\mathbf{x}_i^T \beta)} = \mathbf{0} \tag{3.3.8}$$

Note that we get back the estimating equation of Poisson when $\alpha = 0$. In fact, $Poi(\mu) = \lim_{r \to \infty} NB(r, \frac{\mu}{\mu + r})$ provided that $\frac{rp}{1-p} \to \mu$ as $r \to \infty$.

3.4 Generalized Linear Models (GLMs)

The theory of generalized linear models (GLMs) was introduced by Nelder and Wedderburn (1972). In GLM, the response variableis a member of the exponential family. Examples of exponential family member include Gaussian, Poisson, Bernoulli, binomial, negative binomial, etc. The exponential family with a canonical location parameter θ, nuisance parameter ϕ, and known function a, b, c has the following density

$$f(y|\theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\} \tag{3.4.1}$$

where $a(\phi)$ is a scale parameter and $c(y, \phi)$ is a normalizing term which ensures integration of the density is one. The expected value and variance are

$$\mathbf{E}(y) = b'(\theta) = \mu \tag{3.4.2}$$

$$\mathbf{V}(y) = \mathbf{V}(\mu) = a(\phi)\frac{\partial \mu}{\partial \theta} = a(\phi)b''(\theta) \tag{3.4.3}$$

where the variance is a function of the mean μ. The likelihood of n independent random variables $y_1, \ldots, y_n$ of exponential family is given by

$$L(\boldsymbol{\theta}, \phi | y_1, \ldots, y_n) = \prod_{i=1}^{n} \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\} \tag{3.4.4}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. The log-likelihood is

$$L(\boldsymbol{\theta}, \phi | y_1, \ldots, y_n) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\} \tag{3.4.5}$$

We consider the linear model $g(\mu_i) = g(\mu(\theta_i)) = \mathbf{x}_i^T\beta$, which implies $b'(\theta_i) = \mu_i = h(\mathbf{x}_i^T\beta)$ where $h = g^{-1}$. Note that

$$\frac{\partial\theta}{\partial\mu} = \left(\frac{\partial\mu}{\partial\theta}\right)^{-1} = \frac{1}{b''(\theta)} \tag{3.4.6}$$

Hence the estimating equations of GLMs with this linear model is

$$\begin{aligned}\frac{\partial L}{\partial\beta} &= \left(\frac{\partial L}{\partial\theta}\right)\left(\frac{\partial\theta}{\partial\mu}\right)\left(\frac{\partial\mu}{\partial\beta}\right) && (3.4.7)\\ &= \sum_{i=1}^{n}\left(\frac{y_i - b'(\theta_i)}{a(\phi)}\right)\left(\frac{1}{b''(\theta_i)}\right)h'(\mathbf{x}_i^T\beta)\mathbf{x}_i && (3.4.8)\\ &= \sum_{i=1}^{n}\left(\frac{y_i - h(\mathbf{x}_i^T\beta)}{\mathbf{V}(y_i)}\right)h'(\mathbf{x}_i^T\beta)\mathbf{x}_i = \mathbf{0} && (3.4.9)\end{aligned}$$

For canonical link $g(\mu(\theta)) = \theta$, we have

$$1 = g'(\mu(\theta))\mu'(\theta) = g'(\mu(\theta))b''(\theta) \tag{3.4.10}$$

and $h(g(\mu(\theta)) = \mu(\theta)$ implies $h'(\mathbf{x}^T\beta) = \frac{1}{g'(\mu(\theta))}$. Thus, we can further simplify the estimating equations of GLMs as

$$\sum_{i=1}^{n}(y_i - h(\mathbf{x}_i^T\beta))\mathbf{x}_i = \mathbf{0} \tag{3.4.11}$$

The estimating equations of GLM are highly non-linear. But it can be solved by IWLS through application of Newton's method.

4. A-OPTIMAL SUBSAMPLING AND ASYMPTOTIC THEORY

In this chapter, we give the asymptotic normality of the subampling generalized bootstrap estimate under fixed and growing dimension. We will derive the order of the remainder term of the bias of this subsampling estimate. The A-optimal subsampling distribution will be dervied and the A-optimal scoring algorithm will be discussed.

We follow the framework of Chatterjee and Bose (2005): Let $\{Z_{ni} : 1 \leq i \leq n,\, n \geq 1\}$ be a sequence of random variables defined on some probability space $(\Omega, \mathbb{P})$ and $\beta \in \mathcal{B} \subset \mathbb{R}^p$ be a parameter vector. Consider a triangular array of smooth functions $\{\psi_{ni}(Z_{ni}; \beta)\colon\ 1 \leq i \leq n,\ n \geq 1\}$ taking values in $\mathbb{R}^p$ and mapping to $\mathbb{R}^p$, with each $\mathbb{E}(\psi_{ni}(Z_{ni}; \beta_0)) = 0$ for some unique $\beta_0 \in \mathcal{B}$. The parameter β_0 is unknown and we estimate β_0 by $\hat{\beta}_n$ which solves the estimating equations,

$$\Psi_n(\beta) = \sum_{i=1}^{n} \psi_{ni}(Z_{ni}; \beta) = 0. \tag{4.0.1}$$

We consider the case when the sample size n is *extremely large* and the dimension p is also large, in which case conventional methods suffer from large sample size challenge and also high-dimension challenge. To deal with large p, which is a typical case for big data, it is common to assume the sparsity principle, that is the response variable only depends on a subset of predictors. See Tibshirani (1996) and Bühlmann and van de Geer. S. (2011) for the variables selection methods in high-dimensional statistics. When n is large, due to the limit of computing technologies, computation of the full sample estimate $\hat{\beta}_n$ is infeasible or time-consuming to obtain. One pupular method for handling large sample is subsampling. In this dissertation, we seek the A-optimal sampling distribution on the data points, and use it to take a small proportion of the data as a surrogate of the whole sample for model fitting and statistical inference. We shall also look at the case of growing $p = p_n$ with sample size n. Asymptotic properties of the subsampling estimate when p_n is growing with n is investigated.

Let $\pi_n = (\pi_{ni}, i = 1, \ldots, n)$ be a sampling distribution on the n data points Z_{ni}. Assume π_n is known for now. A subsample $Z^* = \{Z_j^* : j = 1, \ldots, r\}$ with the subsample size $r << n$ is selected based on this sampling distribution. Let $\pi^* = (\pi_j^* : j = 1, \ldots, r)$ be the corresponding sampling probabilities. The full sample estimator $\hat{\beta}$ is then approximated by the subsampling generalized bootstrap estimate $\hat{\beta}_r^*$ which solves the estimating equations

$$\Psi_r^*(\beta) =: \sum_{j=1}^{r} \frac{\psi_{nj}(Z_{nj}^*; \beta)}{\pi_j^*} = 0. \tag{4.0.2}$$

An important feature of (4.0.2) is that it uses the sampling probability as the weights of the estimating equations. This is analogous to the Hansen-Hurwitz estimate in classical sampling (Hansen and Hurwitz (1943)). In fact, the conditional expectation of $r^{-1}\Psi_r^*(\beta)$ is

$$\begin{aligned}
\mathbb{E}^*(r^{-1}\Psi_r^*(\beta)) &= \frac{1}{r}\sum_{j=1}^{r} \mathbb{E}^*(\frac{\psi_{nj}(Z_{nj}^*; \beta)}{\pi_j^*}) \\
&= \sum_{i=1}^{n} \pi_i \frac{\psi_{ni}(Z_{ni}; \beta)}{\pi_i} \\
&= \sum_{i=1}^{n} \psi_{ni}(Z_{ni}; \beta)
\end{aligned}$$

which is original estimating equations $\Psi_n(\beta)$.

The theory of weighted (generalized) bootstrap has been extensively studied in the literature, see e.g. Efron (1979), Mammen (1993) and Bose and Chatterjee (2002). However, in order to makes the proposed bootstrap computational friendly, most of the existing weights are exchangeable non-negative random variable and independent of data. Only some of these weights can improve Efron's bootstrap by using Edgeworth expansions. See Chapter II of the monograph by Barbe and Bertail (1995) and the references therein. Unlike existing weights, we pursue weights which are data dependent and not exchangeable. In fact, our weights are derived by minimizing the trace of certain variance covariance matrix. They are referred to as the A-optimal weights which are different from existing weights.

4.1 Notation and some elementary results

NOTATION.

- Let $\{a_n\}_{n=1}^{\infty}, \{b_n\}_{n=1}^{\infty}$ be two sequences of real numbers. $a_n = o(1)$ means that $a_n \to 0$ as $n \to \infty$. $a_n = o(b_n)$ means that $a_n/b_n = o(1)$.

- $a_n = O(1)$ means that for all large n, $|a_n| \leq C$ for some C. $a_n = O(b_n)$ means that $a_n/b_n = O(1)$.

- For any sequences of random variables $\{X_n\}_{n=1}^{\infty}$ and $\{Y_n\}_{n=1}^{\infty}$, $X_n = o_p(1)$, if for every $\varepsilon > 0$, $\mathbb{P}(|X_n| > \varepsilon) \to 0$ as $n \to \infty$. $X_n = o_p(Y_n)$ if $X_n/Y_n = o_p(1)$.

- $X_n = O_p(1)$ if for every $\varepsilon > 0$, there is a $C > 0$ and N such that if $n \geq N$ then $\mathbb{P}(|X_n| > C) \leq \varepsilon$. $X_n = O_p(Y_n)$ if $X_n/Y_n = O_p(1)$.

- Abbreviate $\psi_{ni}(\beta) = \psi_{ni}(Z_{ni}; \beta)$, its d-th component $\psi_{ni,d}(\beta)$, and $\psi_{ni} = \psi_{ni}(\beta_0)$.

- For $\psi : \mathbb{R}^p \to \mathbb{R}$, define $\dot{\psi} : \mathbb{R}^p \to \mathbb{R}^p$ by

$$\dot{\psi}(\beta) = \frac{\partial \psi}{\partial \beta} = \left(\frac{\partial \psi}{\partial \beta_1}, \cdots, \frac{\partial \psi}{\partial \beta_p} \right).$$

  For $\psi : \mathbb{R}^p \to \mathbb{R}^p$, define $\dot{\psi} : \mathbb{R}^{p \times p} \to \mathbb{R}^{p \times p}$ to be the $p \times p$ matrix $\dot{\psi}(B)$ whose d-th row is $\dot{\psi}(\beta_d)^T$ for $B = (\beta_1, \cdots, \beta_p)$ with $\beta_d \in \mathbb{R}^p$. Similarly, define $\ddot{\psi}(B)$ to be the $p^2 \times p$ matrix resulted from stacking $\ddot{\psi}_d(\beta_d) = \frac{\partial^2}{\partial \beta^T \partial \beta} \psi_d(\beta_d)$. We write $\dot{\psi}(\beta)$ for $\dot{\psi}(B)$ if $B = (\beta, \cdots, \beta)$ and similarly $\ddot{\psi}(\beta)$ for $\ddot{\psi}(B)$.

- For any vector $v, w \in \mathbb{R}^p$, we define the $\circ$ notation as follows

$$v^T \circ \ddot{\psi}_{ni}(\beta) \circ w := [v^T \ddot{\psi}_{ni,1}(\beta) w, v^T \ddot{\psi}_{ni,2}(\beta) w, \ldots, v^T \ddot{\psi}_{ni,p}(\beta) w]^T.$$

  Hence, $v^T \circ \ddot{\psi}_{ni}(\beta) \circ w$ is a p-dimensional vector.

- Define the norm $\|\cdot\|_{oe}$ as follows: for $\psi = (\psi_1, \cdots, \psi_p)^T$,

$$\|\ddot{\psi}\|_{oe} = \sqrt{\sum_{d=1}^{p} \|\ddot{\psi}_d\|_o^2}.$$

- For any random variable X, we define the centered version of X as

$$\bar{X} = X - \mathbb{E}(X)$$

- For matrix A, denote $A^\top$ the transpose of A, $A^{\otimes 2} = AA^\top$, $A^{-\top} = (A^{-1})^\top$, $\mathbb{E}^{-1}(A) = (\mathbb{E}(A))^{-1}$, and $A^{(s)} = \frac{1}{2}(A + A^\top)$.

- Write $\lambda_{\max}(A)$ ($\lambda_{\text{amax}}(A)$) the maximum (maximum absolute) eigenvalue of A, etc.

- We write $\|A\|$ for the euclidean norm and $\|A\|_o$ for the operator (spectral) norm of matrix A which are defined by

$$\|A\|^2 := \text{Tr}(A^T A) = \sum_{i,j} A_{ij}^2,$$
$$\|A\|_o := \sup_{\|u\|=1} \|Au\| = \sup_{\|u\|=1} (u^T A^T A u)^{1/2} = \lambda_{\max}^{1/2}(AA^T)$$

  where A_{ij} denotes the (i, j)-th entry of A.

In other words, $\|A\|^2$ equals the sum of the eigenvalues of $A^T A$, while $\|A\|_o^2$ is the largest eigenvalue of $A^T A$. Consequently,

$$\|A\|_o \leq \|A\|$$

If A is $p \times p$, then $\|A\| \leq \sqrt{p}\|A\|_o$. Thus we have

$$\|Ax\| \leq \|A\|_o \|x\| \leq \|A\|\|x\|$$

for compatible vector x. Also,

$$\|A\|_o = \sup_{\|u\|=1} \sup_{\|v\|=1} u^T A v$$

and it simplifies to

$$\|A\|_o = \sup_{\|u\|=1} u^T A u$$

if A is symmetric. Using this and the Cauchy-Schwartz inequality,

$$\left\| \int f g^T du \right\|_o^2 \leq \left\| \int f f^T du \right\|_o \left\| \int g g^T du \right\|_o,$$

$$(\|\mathbb{E}(XY^T)\|_o \leq \|\mathbb{E}(XX^T)\|_o^{1/2} \|\mathbb{E}(YY^T)\|_o^{1/2}).$$

In particular,

$$\left\| \int f f^T du \right\|_o \leq \int \|f\|^2 du,$$

$$(\|\mathbb{E}(XX^T)\|_o \leq \mathbb{E}(\|X\|^2)).$$

4.2 Asymptotic behaviors of the subsampling M-estimators for fixed/growing dimension

We need the following assumptions. Let

$$\hat{J}_n = J_n(\hat{\beta}_n) = \sum_{i=1}^n \pi_{ni}^{-1} \psi_{ni}(\hat{\beta}_n)^{\otimes 2},\ \hat{\lambda}_n = \lambda_{\max}^{1/2}(\hat{J}_n),\ \Sigma_n = \dot{\Psi}_n^{-1} J_n \dot{\Psi}_n^{-\top}\big|_{\hat{\beta}_n}.$$

$$\tilde{J}_n = \tilde{J}_n(\hat{\beta}_n) = \sum_{i=1}^n \psi_{ni}(\hat{\beta}_n)^{\otimes 2},\ \tilde{\lambda}_n = \lambda_{\max}(\tilde{J}_n),\ \tilde{\Sigma}_n = \dot{\Psi}_n^{-1} \tilde{J}_n \dot{\Psi}_n^{-\top}\big|_{\hat{\beta}_n}.$$

Note that $\hat{\lambda}_n \approx n$, $\tilde{\lambda}_n \approx \sqrt{n}$, and $\frac{\tilde{\lambda}_n^2}{\max \pi_i} \leq \hat{\lambda}_n^2 \leq \frac{\tilde{\lambda}_n^2}{\min \pi_i}$.

Let $\sigma_n^2 > 0$ be an arbitrary sequence. Typically, $\sigma_n^2 = \frac{1}{n \max \pi_i}$.

(R1) There is a constant $c_0 > 0$ such that $\hat{\lambda}_n \xrightarrow{p} \infty$ and

$$\mathbb{P}(\hat{\lambda}_n^{-1} \lambda_{\text{amin}}(\dot{\Psi}_n^{(s)}(\hat{\beta}_n)) > c_0) \to 1.$$

(R2) Each component $\psi_{ni,d}(\beta)$ admits the second order expansion

$$\psi_{ni,d}(\beta_0 + t) = \psi_{ni,d}(\beta_0) + \dot{\psi}_{ni,d}^\top(\beta_0) t + 1/2 t^\top \ddot{\psi}_{ni,d}(\tilde{\beta}_{ni,d}) t, \quad d = 1, \ldots, p,$$

for $\|t\| \leq t_0$ with some $t_0 > 0$, where $\tilde{\beta}_{ni,d}$ lies in between β_0 and $\beta_0 + t$.

(R3) The sampling probabilities π_{ni} and subsample size r_n satisfy

$$\sum_{i=1}^{n} \pi_{ni}^{-1} \|\dot{\psi}_{ni}(\hat{\beta}_n)\|^2 = o_p(p^2 \hat{\lambda}_n^2).$$

(R4) There exists a neighborhood $\mathbb{N}_0$ of β_0 such that $\ddot{\Psi}_{n,d}(\beta)$ is either positive or negative definite in $\mathbb{N}_0$, and that there is a random variable $\eta_{ni,d}$

$$\sup_{\beta \in \mathbb{N}_0} \lambda_{\text{amax}}(\ddot{\Psi}_{n,d}(\beta)) \le \eta_{ni,d}, \quad d = 1, \ldots, p,$$

where the random vector $\eta_{ni} = (\eta_{ni,1}, \ldots, \eta_{ni,p})^\top$ satisfies

$$\sum_{i=1}^{n} \|\eta_{ni}\|^2 = o_p(\tilde{\lambda}_n^2 p).$$

(R5) $$\lambda_{\max}(J_n(\hat{\beta}_n))/\lambda_{\min}(J_n(\hat{\beta}_n)) = O_p(1).$$

(R6) Fix $u \in \mathbb{R}^{p_n}$ with $\|u\| = 1$. The double array $z_{nj}^* = s_n^{-1} u^\top \dot{\Psi}_n^{-\top}(\hat{\beta}_n) \psi_{nj}^*(\hat{\beta}_n)/\pi_{nj}^*$, $j = 1, 2, \ldots, r,\ r \ge 1$ satisfies the Lindeberg condition: for every $t > 0$,

$$\sum_{i=1}^{n} \pi_{ni} \|z_{n,i}\|^2 \mathbf{1}[\|z_{ni}\| \ge \sqrt{r}t] = o_p(1), \quad as \quad r \to \infty,$$

where $s_n^2 = u^\top \Sigma_n u$.

To prove asymptotic normality of the subsampling estimator, we need (R3) and (R4) replaced by the followings assumptions

(R3") The sampling probabilities π_{ni} and subsample size r_n satisfy

$$\sum_{i=1}^{n} \pi_{ni}^{-1} \|\dot{\psi}_{ni}(\hat{\beta}_n)\|^2 = o_p(p^{-1} r_n \hat{\lambda}_n^2 \sigma_n^{-2}).$$

(R4") There exists a neighborhood $\mathbb{N}_0$ of β_0 such that $\ddot{\Psi}_{n,d}(\beta)$ is either positive or negative definite in $\mathbb{N}_0$ and that there is a random variable $\eta_{ni,d}$

$$\sup_{\beta \in \mathbb{N}_0} \lambda_{\text{amax}}(\ddot{\Psi}_{n,d}(\beta)) \le \eta_{ni,d}, \quad d = 1, \ldots, p,$$

where the random vector $\eta_{ni} = (\eta_{ni,1}, \ldots, \eta_{ni,p})^\top$ satisfies

$$\Big\| \sum_{i=1}^{n} \eta_{ni} \Big\|^2 + \sum_{i=1}^{n} (n + (r_n \pi_{ni})^{-1}) \|\eta_{ni}\|^2 = o_p(p^{-2} r_n \sigma_n^{-4} \hat{\lambda}_n^2).$$

Theorem 4.2.1 *Suppose (R1)–(R4) hold. Assume $\hat{\beta}_n$ is a solution of (4.0.1) such that $\hat{\beta}_n = \beta_0 + o_p(1)$. Assume*

$$\sum_{i=1}^{n} \pi_{ni}^{-1} \|\psi_{ni}(\hat{\beta}_n)\|^2 = O_p(p\hat{\lambda}_n^2 \sigma_n^2). \tag{4.2.1}$$

Then these exists a sequence of solutions $\hat{\beta}_r^*$ of (4.2.18) such that if $p\sigma_n^2/r = o_p(1)$, then it holds in probability that

$$\hat{\beta}_r^* - \hat{\beta} = O_{p^*}(p^{1/2} r^{-1/2} \sigma_n), \tag{4.2.2}$$

$$\dot{\Psi}_n(\hat{\beta}_n)\sqrt{r_n}(\hat{\beta}_r^* - \hat{\beta}_n) = -\frac{1}{\sqrt{r_n}} \sum_{j=1}^{r_n} \frac{\psi_{nj}^*(\hat{\beta}_n)}{\pi_j^*} + o_{p^*}(\hat{\lambda}_n \sqrt{p}). \tag{4.2.3}$$

If, further, (R5)–(R6) are satisfied for $u \in \mathbb{R}^{p_n}$ with $\|u\| = 1$, and (R3) and (R4) are replaced by (R3') and (R4'), then it holds in probability that

$$s_n^{-1}\sqrt{r_n} u^\top (\hat{\beta}_r^* - \hat{\beta}_n) \Rightarrow \mathscr{N}(0,1), \quad \textit{in probability}, \quad r_n \to \infty. \tag{4.2.4}$$

Proof of Theorem 4.2.1. For $t \in \mathbb{R}^p$, let $t_n = p^{1/2} r^{-1/2} \sigma_n \tilde{\lambda}_n^{-1} t$, and

$$\Delta_n^*(t) = p^{-1/2} r_n^{1/2} \sigma_n^{-1} \hat{\lambda}_n^{-1} \sum_{i=1}^{n} w_i \big(\psi_{ni}(\hat{\beta}_n + t_n) - \psi_{ni}(\hat{\beta}_n)\big) - \hat{\lambda}_n^{-1} \dot{\Psi}_n(\hat{\beta}_n) t. \tag{4.2.5}$$

For arbitrary $C > 0$, fix $\|t\| \leq C$. Then $t_n \to 0$ by assumption. For notational brevity, we now drop the subscript n when there appears to be no ambiguity. It follows from $\hat{\beta}_n = \beta_0 + o_p(1)$, (R2) and the inequality in (R4) that

$$\|\Delta_n^*(t)\|^2 \leq 2C^2 \hat{\lambda}^{-2} \Big\| \sum_{i=1}^{n} \bar{w}_i \dot{\psi}_{ni}(\hat{\beta}_n) \Big\|^2 + \frac{1}{2} C^4 p r^{-1} \sigma_n^2 \hat{\lambda}^{-2} \Big\| \sum_{i=1}^{n} w_i \eta_{ni} \Big\|^2,$$

for large subsample size r and with large probability (meaning it holds on an event whose probability converges to one as the subsample size tends to infinity). Using some algebra, one easily derive

$$r\mathbb{E}^*(\| \sum_i \bar{w}_i \dot{\psi}_{ni}(\hat{\beta}_n)\|^2) \leq \sum_i {\pi_i}^{-1} \|\dot{\psi}_{ni}(\hat{\beta}_n)\|^2 =: A_n,$$

$$\mathbb{E}^*\big(\| \sum_{i=1}^{n} w_i \eta_{ni}\|^2\big) \leq \mathbb{E}^*(\sum_{i=1}^{n} w_i \|\eta_{ni}\|^2) = \sum_{i=1}^{n} \|\eta_{ni}\|^2 =: B_n.$$

Hence it follows from (R3) and the equality in (R4) that

$$\mathbb{E}^*(\sup_{\|t\|\leq C}\|\Delta_n^*(t)\|^2) \leq 2C^2r^{-1}\hat{\lambda}_n^{-2}A_n + C^4pr^{-1}\sigma_n^2\hat{\lambda}_n^{-2}B_n = o_p(\sigma_n^{-2}). \tag{4.2.6}$$

Recall $\dot{\Psi}_n^{(s)}(\hat{\beta}_n) = 1/2(\dot{\Psi}_n(\hat{\beta}_n) + \dot{\Psi}_n^\top(\hat{\beta}_n))$. Assume without loss of generality that $\lambda_{\text{amin}}(\dot{\Psi}_n^{(s)}(\hat{\beta}_n)) = \lambda_{\min}(\dot{\Psi}_n^{(s)}(\hat{\beta}_n)) > 0$ for large n. By (4.2.5),

$$\begin{aligned}\ell_n^*(C) =: &\inf_{\|t\|=C}\{p^{-1/2}r^{1/2}\sigma_n^{-1}\hat{\lambda}_n^{-1}t^\top\sum_{i=1}^n w_i\psi_{ni}(\hat{\beta}_n + t_n)\}\\ &\geq C^2\hat{\lambda}_n^{-1}\lambda_{\text{amin}}(\dot{\Psi}_n^{(s)}(\hat{\beta}_n)) - C\sup_{\|t\|=C}\|\Delta_n^*(t)\| \qquad (4.2.7)\\ &- Cp^{-1/2}r^{1/2}\sigma_n^{-1}\hat{\lambda}_n^{-1}\|\sum_i w_i\psi_{ni}(\hat{\beta}_n)\|.\end{aligned}$$

By (R1), $\hat{\lambda}_n^{-1}\lambda_{\min}(\dot{\Psi}_n^{(s)}(\hat{\beta}_n)) \geq c_0$ for some $c_0 > 0$ and with large probability for large n. For large $K > 0$, as $\Psi_n(\hat{\beta}_n) = 0$ and by (4.2.1), we have

$$\begin{aligned}&\mathbb{P}^*(p^{-1/2}r^{1/2}\sigma_n^{-1}\hat{\lambda}_n^{-1}\|\sum_i w_i\psi_{ni}(\hat{\beta}_n)\| > K)\\ &= \mathbb{P}^*(p^{-1/2}r^{1/2}\sigma_n^{-1}\hat{\lambda}_n^{-1}\|\sum_i \bar{w}_i\psi_{ni}(\hat{\beta}_n)\| > K)\\ &\leq K^{-2}p^{-1}\sigma_n^{-2}\hat{\lambda}_n^{-2}\sum_i \pi_i^{-1}\|\psi_{ni}(\hat{\beta}_n)\|^2\\ &= K^{-2}O_p(1) = o_p(1).\end{aligned}$$

This and (4.2.6)–(4.2.7) yield that for large C,

$$\begin{aligned}\mathbb{P}^*(\ell_n^*(C) > 0) &\geq 1 - \mathbb{P}^*(\sup_{\|t\|=C}\|\Delta_n^*(t)\| > c_0C)\\ &- \mathbb{P}^*(p^{-1/2}r^{1/2}\hat{\lambda}_n^{-1}\|\sum_i w_i\psi_{ni}(\hat{\beta}_n)\| > c_0C) = 1 - o_p(1).\end{aligned}$$

Using the same argument of Chatterjee and Bose (2005), on the set $\ell_n^*(C) > 0$ the continuity of $\Psi_r^*(\beta)$ implies that there is a root $t = T_n$ of $\Psi_r^*(\hat{\beta}_n + p^{1/2}r^{-1/2}\sigma_n t) = 0$ with $|T_n| \leq C$. This holds on an event with probability approaching one as C tends to infinity. Thus $\hat{\beta}_r^* = \hat{\beta}_n + p^{1/2}r^{-1/2}\sigma_n T_n$ is a root of (4.2.18) and satisfies $\mathbb{P}^*(p^{-1/2}r^{1/2}\sigma_n^{-1}\|\hat{\beta}_r^* - \beta_0\| \leq C) \geq 1 - o_p(1)$. By (4.2.6), $\Delta_n^*(T_n) = o_p(\sigma_n^{-1})$, which is, in view of (4.2.5), amount to

$$\hat{\lambda}_n^{-1}\dot{\Psi}_n(\hat{\beta}_n)\sqrt{r}(\hat{\beta}_r^* - \beta_0) = -\hat{\lambda}_n^{-1}\sqrt{r}\sum_i w_i\psi_{ni}(\hat{\beta}_n) + o_{p^*}(\sqrt{p}).$$

This shows (4.2.3) by the stochastic equivalence.

Using (R3") and the equality in (R4"), (4.2.6) becomes

$$\mathbb{E}^*(\sup_{\|t\|\leq C} \|\Delta_n^*(t)\|^2) \leq 2C^2 r^{-1}\hat{\lambda}_n^{-2}A_n + C^4 p r^{-1}\sigma_n^2\hat{\lambda}_n^{-2}B_n = o_p(p^{-1}\sigma_n^{-2}). \tag{4.2.8}$$

Following the same argument as above with $\Delta_n^*(T_n) = o_p(p^{-1/2}\sigma_n^{-1})$, we have the expression

$$\hat{\lambda}_n^{-1}\dot{\Psi}_n(\hat{\beta}_n)\sqrt{r}(\hat{\beta}_r^* - \beta_0) = -\hat{\lambda}_n^{-1}\sqrt{r}\sum_i w_i\psi_{ni}(\hat{\beta}_n) + o_{p^*}(1). \tag{4.2.9}$$

The asymptotic normality (4.2.4) follows from the established relation (4.2.9) and the Lindeberg-Feller theorem (e.g. Theorem 7.2.1 of Chung, 2001). More specifically, the Lindeberg condition (R6) implies that the main term on the left side of (4.2.9) has an asymptotic standard normal in conditional probability given the data, while the remainder term is negligible,

$$s_n^{-1}\hat{\lambda}_n u^\top \dot{\Psi}_n^{-1}(\hat{\beta}_n)\alpha_n^* = o_p(1),$$

where $\alpha_n^* = o_{p^*}(1)$. This follows from

$$\frac{\hat{\lambda}_n}{s_n} \leq \frac{\lambda_{\max}(J_n(\hat{\beta}_n))}{\lambda_{\min}((J_n(\hat{\beta}_n))\|u^\top\dot{\Psi}_n^{-1}(\hat{\beta}_n)\|} \leq \frac{B}{\|u^\top\dot{\Psi}_n^{-1}(\hat{\beta}_n)\|},$$

where B is a constant implied by (R6). The proof is now complete. ■

Note that when conditions (R3") and (R4") are used, the remainder term of the expansion of $\hat{\beta}_r^*$ is of order $o(\hat{\lambda}_n/\sqrt{r_n})$. This is used for deriving the asymptotic normality of $\hat{\beta}_r^*$. To derive the asymptotic bias and variance of the subsampling estimator, we will use the Taylor expansion (4.2.3) because (R3") and (R4") are implied by (R3) and (R4). (R5) and (R6) are used for asymptotic normality.

A result from multivariable calculus and a related inequality which will be used later is worth to note here. Here we use the notation introduced in Section 4.1.

Lemma 4.2.1 *Let $f : \mathbb{R}^p \to \mathbb{R}^p$ be a continuous function which is twice differentiable. Denote $f = (f_1, \dots, f_p)$. Then the second order Taylor expansion of f about $x_0 \in \mathbb{R}^p$ is*

$$\begin{pmatrix} f_1 \\ \vdots \\ f_p \end{pmatrix}(x_0 + t) = \begin{pmatrix} f_1 \\ \vdots \\ f_p \end{pmatrix}(x_0) + \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_p} \\ \vdots & & \\ \frac{\partial f_p}{\partial x_1} & \cdots & \frac{\partial f_p}{\partial x_p} \end{pmatrix}(x_0)t + \frac{1}{2}t^T \circ \ddot{f}(\tilde{X}) \circ t \qquad (4.2.10)$$

where $\tilde{X} = (\tilde{x}, \cdots, \tilde{x})$ and $\tilde{x}$ lies in between x_0 and $x_0 + t$. $\ddot{f}$ is a stack of $p \times p$ matrix $\ddot{f}_1, \dots, \ddot{f}_p$ where $\ddot{f}_i = \begin{pmatrix} \frac{\partial^2 f_i}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f_i}{\partial x_1 \partial x_p} \\ \vdots & & \\ \frac{\partial^2 f_i}{\partial x_p \partial x_1} & \cdots & \frac{\partial^2 f_i}{\partial x_p \partial x_p} \end{pmatrix}$ and $\ddot{f} = [\ddot{f}_1, \dots \ddot{f}_p]^T \in \mathbb{R}^{p^2 \times p}$.

Lemma 4.2.2 *Let $f : \mathbb{R}^p \to \mathbb{R}^p$ be a continuous function which is twice differentiable. Denote $f = (f_1, \dots, f_p)^T$. Consider the second derivative matrix $\ddot{f} = [\ddot{f}_1, \dots \ddot{f}_p]^T \in \mathbb{R}^{p^2 \times p}$, and vectors $a, b \in \mathbb{R}^p$, we have*

$$\|a \circ \ddot{f} \circ b\| \le \|a\| \|b\| \|\ddot{f}\|_{oe} \qquad (4.2.11)$$

where by definition $\|\ddot{f}\|_{oe} = \sqrt{\sum_{i=1}^p \|\ddot{f}_i\|_o^2}$.

Theorem 4.2.2 *Suppose (R1)–(R4) hold. Assume $\hat{\beta}_n$ is a solution of the GEE such that $\hat{\beta}_n = \beta_0 + o_p(1)$. Let $\hat{\beta}_r^*$ be the subsampling estimator according to the sampling probability $\pi = (\pi_1, \dots, \pi_n)$. Then the bias is given by*

$$\mathbb{E}^*(\hat{\beta}_r^*) - \hat{\beta}_n = \frac{1}{r_n} \dot{\Psi}_n^{-1}(\hat{\beta}_n) \sum_{i=1}^n \frac{1}{\pi_i} \left[\dot{\psi}_{ni}(\hat{\beta}_n) \bar{a}_{ni} - \frac{1}{2} b_{nii} \right] + o_{p^*}(\frac{p^{3/2}}{r}). \qquad (4.2.12)$$

where $a_{ni} = \dot{\Psi}_n^{-1} \psi_{ni}|_{\hat{\beta}_n}$ and $b_{nij} = \bar{a}_{ni}^T \circ \ddot{\Psi}_n(\hat{\beta}_n) \circ \bar{a}_{nj}$.

Proof of Theorem 4.2.2. Define

$$\Psi_n^*(\beta) = \sum_{j=1}^{r_n} \frac{\psi_{nj}^*(\beta)}{r_n \pi_j^*} =: \sum_{i=1}^n w_i \psi_{ni}(\beta). \qquad (4.2.13)$$

where the equality follows the stochastic equivalence, see Peng and Tan (2018). Consider the Taylor expansion

$$\Psi_n^*(\hat{\beta}_r^*) = \Psi_n^*(\hat{\beta}_n) + \dot{\Psi}_n^*(\hat{\beta}_n)(\hat{\beta}_r^* - \hat{\beta}_n) + \frac{1}{2}(\hat{\beta}_r^* - \hat{\beta}_n)^T \circ \ddot{\Psi}_n^*(\tilde{B}^*) \circ (\hat{\beta}_r^* - \hat{\beta}_n) \tag{4.2.14}$$

where $\tilde{B}^*$ lies in between $\hat{\beta}_r^*$ and $\hat{\beta}_n$. Note that $\Psi_n^*(\hat{\beta}_r^*) = 0$. Apply expectation to get

$$\begin{aligned} 0 &= \mathbb{E}^*(\Psi_n^*(\hat{\beta}_n)) + \mathbb{E}^*[\dot{\Psi}_n^*(\hat{\beta}_n)(\hat{\beta}_r^* - \hat{\beta}_n)] + \frac{1}{2}\mathbb{E}^*[(\hat{\beta}_r^* - \hat{\beta}_n)^T \circ \ddot{\Psi}_n^*(\tilde{B}^*) \circ (\hat{\beta}_r^* - \hat{\beta}_n)] \\ &= \mathbb{E}^*[(\dot{\Psi}_n^*(\hat{\beta}_n) - \mathbb{E}^*(\dot{\Psi}_n^*(\hat{\beta}_n)))(\hat{\beta}_r^* - \hat{\beta}_n)] + \mathbb{E}^*(\dot{\Psi}_n^*(\hat{\beta}_n))\mathbb{E}^*(\hat{\beta}_r^* - \hat{\beta}_n) \\ &\quad + \frac{1}{2}\mathbb{E}^*[(\hat{\beta}_r^* - \hat{\beta}_n)^T \circ \ddot{\Psi}_n(\hat{\beta}_n) \circ (\hat{\beta}_r^* - \hat{\beta}_n)] \\ &\quad + \frac{1}{2}\mathbb{E}^*[(\hat{\beta}_r^* - \hat{\beta}_n)^T \circ [\ddot{\Psi}_n^*(\tilde{\beta}^*) - \ddot{\Psi}_n(\hat{\beta}_n)] \circ (\hat{\beta}_r^* - \hat{\beta}_n)] \\ &= \mathbb{E}^*\left(\sum_{i=1}^n (w_i - 1)\dot{\psi}_{ni}(\hat{\beta}_n)(\hat{\beta}_r^* - \hat{\beta}_n)\right) + \dot{\Psi}_n(\hat{\beta}_n)(\mathbb{E}^*(\hat{\beta}_r^*) - \hat{\beta}_n) + \Delta_1^* + \Delta_2^*, \end{aligned} \tag{4.2.15}$$

where

$$\Delta_1^* := \frac{1}{2}\mathbb{E}^*[(\hat{\beta}_r^* - \hat{\beta}_n)^T \circ \ddot{\Psi}_n(\hat{\beta}_n) \circ (\hat{\beta}_r^* - \hat{\beta}_n)], \tag{4.2.16}$$

$$\Delta_2^* := \frac{1}{2}\mathbb{E}^*[(\hat{\beta}_r^* - \hat{\beta}_n)^T \circ [\ddot{\Psi}_n^*(\tilde{\beta}^*) - \ddot{\Psi}_n(\hat{\beta}_n)] \circ (\hat{\beta}_r^* - \hat{\beta}_n)]. \tag{4.2.17}$$

The second equality holds since $\mathbb{E}^*(\Psi_n^*(\hat{\beta}_n)) = \sum_{i=1}^n \mathbb{E}(w_i)\psi_{ni}(\hat{\beta}_n) = \sum_{i=1}^n \psi_{ni}(\hat{\beta}_n) = 0$, and the last equality follows from the result that $\dot{\Psi}_n^*(\hat{\beta}_n) - \mathbb{E}^*(\dot{\Psi}_n^*(\hat{\beta}_n)) = \sum_{i=1}^n (w_i - 1)\dot{\psi}_{ni}(\hat{\beta}_n)$.

It follows from (4.2.3) and the stochastic equivalence that

$$\begin{aligned} \hat{\beta}_r^* - \hat{\beta}_n &= -\sum_{i=1}^n w_i \dot{\Psi}_n^{-1}(\hat{\beta}_n)\bar{\psi}_{ni}(\hat{\beta}_n) + \dot{\Psi}_n^{-1}(\hat{\beta}_n)o_p(\hat{\lambda}_n\sqrt{p}/\sqrt{r_n}) \\ &= -\sum_{i=1}^n \bar{w}_i \dot{\Psi}_n^{-1}(\hat{\beta}_n)\bar{\psi}_{ni}(\hat{\beta}_n) + \dot{\Psi}_n^{-1}(\hat{\beta}_n)\alpha_n^* \\ &= -\sum_{i=1}^n \bar{w}_i \bar{a}_{ni} + \tilde{\alpha}_n^* \end{aligned} \tag{4.2.18}$$

where $\bar{a}_{ni} = \dot{\Psi}_n^{-1}\bar{\psi}_{ni}|_{\hat{\beta}_n}$ and $\tilde{\alpha}_n^* = \dot{\Psi}_n^{-1}(\hat{\beta}_n)\alpha_n^*$ and $\alpha_n^* = o_p(\hat{\lambda}_n\sqrt{p}/\sqrt{r_n})$.

Note that the expected value of $\tilde{\alpha}_n^*$ is

$$\begin{aligned}\mathbb{E}^*\|\tilde{\alpha}_n^*\|^2 &= \mathbb{E}^*\|\dot{\Psi}_n^{-1}(\hat{\beta}_n)\alpha_n^*\|^2 \\ &\leq \|\dot{\Psi}_n^{-1}(\hat{\beta}_n)\|_o^2\mathbb{E}^*\|\alpha_n^*\|^2 \\ &\leq \|\dot{\Psi}_n^{-1}(\hat{\beta}_n)\|^2 o_p\left(\frac{\hat{\lambda}_n^2 p}{r}\right) \\ &= O\left(\frac{1}{\hat{\lambda}_n^2}\right) o_p\left(\frac{\hat{\lambda}_n^2 p}{r}\right) \\ &= o_p\left(\frac{p}{r}\right)\end{aligned} \tag{4.2.19}$$

Substituting (4.2.18) to the first term of (4.2.15), we have

$$\begin{aligned}\mathbb{E}^*\left(\sum_{i=1}^n \bar{w}_i\dot{\psi}_{ni}(\hat{\beta}_n)(\hat{\beta}_r^* - \hat{\beta}_n)\right) &= -\mathbb{E}^*\left(\sum_{i=1}^n \bar{w}_i\dot{\psi}_{ni}(\hat{\beta}_n)\left[\sum_{j=1}^n \bar{w}_j\bar{a}_{nj} + \tilde{\alpha}_n^*\right]\right) \\ &= -\mathbb{E}^*\left(\sum_{i=1}^n\sum_{j=1}^n \bar{w}_i\bar{w}_j\dot{\psi}_{ni}(\hat{\beta}_n)\bar{a}_{nj}\right) - \delta_{1n} \\ &= -\sum_{i=1}^n \frac{1}{r_n}\left(\frac{1}{\pi_i} - 1\right)\dot{\psi}_{ni}(\hat{\beta}_n)\bar{a}_{ni} \\ &\quad + \frac{1}{r_n}\sum\sum_{i\neq j}\dot{\psi}_{ni}(\hat{\beta}_n)\bar{a}_{nj} - \delta_{1n} \\ &= -\frac{1}{r_n}\sum_{i=1}^n \frac{1}{\pi_i}\dot{\psi}_{ni}(\hat{\beta}_n)\bar{a}_{ni} - \delta_{1n}\end{aligned} \tag{4.2.20}$$

where

$$\delta_{1n} := \mathbb{E}^*\left(\sum_{i=1}^n \bar{w}_i\dot{\psi}_{ni}(\hat{\beta}_n)\tilde{\alpha}_n^*\right) \tag{4.2.21}$$

The third equality follows from $\mathbb{E}[(w_i - 1)(w_j - 1)] = \frac{1}{r_n\pi_i} - \frac{1}{r_n}$ for $i = j$ and $\mathbb{E}[(w_i - 1)(w_j - 1)] = -\frac{1}{r_n}$ for $i \neq j$, and $\Psi_n(\hat{\beta}_n) = \sum_{i=1}^n \psi_{ni}(\hat{\beta}_n) = 0$.

Substituting (4.2.18) in Δ_1^*, we have

$$\begin{aligned}
\Delta_1^* &= \frac{1}{2}\mathbb{E}^*[(\hat{\beta}_r^* - \hat{\beta}_n)^T \circ \ddot{\Psi}_n(\hat{\beta}_n) \circ (\hat{\beta}_r^* - \hat{\beta}_n)] \\
&= \frac{1}{2}\mathbb{E}^* \left((\sum_{i=1}^n \bar{w}_i \bar{a}_{ni} + \tilde{\alpha}_n^*)^T \circ \ddot{\Psi}_n(\hat{\beta}_n) \circ (\sum_{i=1}^n \bar{w}_i \bar{a}_{ni} + \tilde{\alpha}_n^*) \right) \\
&= \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \mathbb{E}^*(\bar{w}_i \bar{w}_j) \bar{a}_{ni}^T \circ \ddot{\Psi}_n(\hat{\beta}_n) \circ \bar{a}_{nj} + \frac{1}{2}\mathbb{E}^*(\tilde{\alpha}_n^{*T} \circ \ddot{\Psi}_n(\hat{\beta}_n) \circ \tilde{\alpha}_n^*) \qquad (4.2.22) \\
&= \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \mathbb{E}^*(\bar{w}_i \bar{w}_j) b_{nij} + \delta_{2n} \\
&= \frac{1}{2r_n}\sum_{i=1}^n \frac{1}{\pi_i} b_{nii} + \delta_{2n}
\end{aligned}$$

where $b_{nij} := \bar{a}_{ni}^T \circ \ddot{\Psi}_n(\hat{\beta}_n) \circ \bar{a}_{nj}$, and

$$\delta_{2n} := \frac{1}{2}\mathbb{E}^*(\tilde{\alpha}_n^{*T} \circ \ddot{\Psi}_n(\hat{\beta}_n) \circ \tilde{\alpha}_n^*) \qquad (4.2.23)$$

Substituting (4.2.20) and (4.2.22) to (4.2.15), we have

$$\dot{\Psi}_n(\hat{\beta}_n)(\mathbb{E}^*(\hat{\beta}_r^*) - \hat{\beta}_n) = \frac{1}{r_n}\sum_{i=1}^n \frac{1}{\pi_i}\left[\dot{\psi}_{ni}(\hat{\beta}_n)\bar{a}_{ni} - \frac{1}{2}b_{nii}\right] + \delta_{1n} - \delta_{2n} - \Delta_2^* \qquad (4.2.24)$$

By (R3) and (4.2.19), the order of δ_{1n} is

$$\begin{aligned}
\|\delta_{1n}\|^2 &= \left\| \mathbb{E}^* \left(\sum_{i=1}^n \bar{w}_i \dot{\psi}_{ni}(\hat{\beta}_n) \tilde{\alpha}_n^* \right) \right\|^2 \\
&\leq \mathbb{E}^* \left\| \sum_{i=1}^n \bar{w}_i \dot{\psi}_{ni}(\hat{\beta}_n) \right\|_o^2 \mathbb{E}^* \|\tilde{\alpha}_n^*\|^2 \\
&\leq \mathbb{E}^* \left\| \sum_{i=1}^n \bar{w}_i \dot{\psi}_{ni}(\hat{\beta}_n) \right\|^2 o_p\left(\frac{p}{r}\right) \\
&= \mathrm{tr}\left[\sum_{i=1}^n \sum_{j=1}^n \mathbb{E}^*(\bar{w}_i \bar{w}_j) \dot{\psi}_{ni}(\hat{\beta}_n) \dot{\psi}_{nj}^T(\hat{\beta}_n) \right] o_p\left(\frac{p}{r}\right) \\
&\leq \frac{1}{r}\sum_{i=1}^n \frac{1}{\pi_i} \|\dot{\psi}_{ni}(\hat{\beta}_n)\|^2 o_p\left(\frac{p}{r}\right) \\
&= \frac{1}{r} o_p(p^2 \hat{\lambda}_n^2) o_p\left(\frac{p}{r}\right) \\
&= o_p\left(\frac{p^3 \hat{\lambda}_n^2}{r^2}\right)
\end{aligned}$$

Hence, $\|\delta_{1n}\| = o_p(p^{3/2}\hat{\lambda}/r)$. Multiply it with the inverse of $\dot{\Psi}_n(\hat{\beta}_n)$, the order of this remainder term of bias equals $o_p(p^{3/2}/r)$.

By (R4), we have the order of δ_{2n} as

$$\begin{aligned}
4\|\delta_{2n}\|^2 &\leq \mathbb{E}^*\|\tilde{\alpha}_n^{*T} \circ \ddot{\Psi}_n(\hat{\beta}_n) \circ \tilde{\alpha}_n^*\|^2 \\
&= \mathbb{E}^* \left(\sum_{d=1}^{p} (\tilde{\alpha}_n^{*T} \sum_{i=1}^{n} \ddot{\Psi}_{ni,d}(\hat{\beta}_n) \tilde{\alpha}_n^*)^2 \right) \\
&\leq \mathbb{E}^* \left(\sum_{d=1}^{p} \|\tilde{\alpha}_n^*\|^4 \sum_{i=1}^{n} \eta_{ni,d}^2 \right) \\
&= \mathbb{E}^*(\|\tilde{\alpha}_n^*\|^4) \sum_{i=1}^{n} \|\eta_{ni}\|^2 \\
&\leq o_p\left(\frac{p^2}{r^2}\right) O_p(p\tilde{\lambda}_n^2) \\
&= o_p\left(\frac{p^3\tilde{\lambda}_n^2}{r^2}\right)
\end{aligned}$$

Hence, $\|\delta_{2n}\| = o_p(p^{3/2}\tilde{\lambda}_n/r)$. Multiply it with the inverse of $\dot{\Psi}_n(\hat{\beta}_n)$, the order of this remainder term of bias equals $o_p(\frac{p^{3/2}}{r}\frac{\tilde{\lambda}_n}{\hat{\lambda}_n})$ which is faster than the above.

By (R4) and $\|\hat{\beta}_r^* - \hat{\beta}_n\| = O_p(p^{1/2}r^{-1/2}\sigma_n)$ from Theorem 4.2.1, we have the order of Δ_2^* as

$$\begin{aligned}
4\|\Delta_2^*\|^2 &\leq \mathbb{E}^*\|(\hat{\beta}_r^* - \hat{\beta}_n)^T \circ [\ddot{\Psi}_n^*(\tilde{\beta}^*) - \ddot{\Psi}_n(\hat{\beta}_n)] \circ (\hat{\beta}_r^* - \hat{\beta}_n)\|^2 \\
&= \mathbb{E}^* \left(\sum_{d=1}^{p} \{(\hat{\beta}_r^* - \hat{\beta}_n)^T \sum_{i=1}^{n} [\ddot{\Psi}_{ni,d}^*(\tilde{\beta}_d^*) - \ddot{\Psi}_{ni,d}(\hat{\beta}_n)](\hat{\beta}_r^* - \hat{\beta}_n)\}^2 \right) \\
&\leq \mathbb{E}^* \left(\sum_{d=1}^{p} \|\hat{\beta}_r^* - \hat{\beta}_n\|^4 (\sum_{i=1}^{n} 2\eta_{ni,d})^2 \right) \\
&\leq \mathbb{E}^* \left(\|\hat{\beta}_r^* - \hat{\beta}_n\|^4 \sum_{d=1}^{p} 4n \sum_{i=1}^{n} \eta_{ni,d}^2 \right) \\
&= 4\mathbb{E}^*\|\hat{\beta}_r^* - \hat{\beta}_n\|^4 \sum_{i=1}^{n} n\|\eta_{ni}\|^2 \\
&= O_p\left(\frac{p^2\sigma_n^4}{r^2}\right) o_p\left(\frac{r\hat{\lambda}_n^2}{p\sigma_n^4}\right) \\
&= o_p\left(\frac{pn}{r}\right)
\end{aligned}$$

Hence, $\|\Delta_2^*\| = o_p\left(\frac{\sqrt{pn}}{\sqrt{r}}\right)$. Multiply it with the inverse of $\dot{\Psi}_n(\hat{\beta}_n)$, the order of this remainder term of bias equals $o_p(\frac{\sqrt{p}}{\sqrt{rn}})$.

Comparing the rate of the three remainders, we get the largest rate being $o_p(p^{3/2}/r)$. This completes the proof. ■

Remark 4.2.1 From expression (4.2.12), we have the remainder equals $o_p(p^{3/2}/r)$. This means the bias decreases with the sample size r in fixed dimension. However, when p is large, the remainder term may not be negligible even when r increases. Thus, the bias for high dimension data may be significant even when the sample size increases.

Remark 4.2.2 Note that in the expression of the bias, we have first derivative of the estimating functions $\dot{\psi}$ and the second derivative $\ddot{\psi}$ involved. In the context of GLM, the original estimating function is known as the score function of a likelihood. The first derivative of the score function is the Hessian matrix. Our result shows that to compute the bias asymptotically, we need to consider the third derivative of the likelihood.

Remark 4.2.3 Theorem 4.2.2 shows that the order of the remainder term of bias equals $o_p(p^{3/2}/r)$ in GEE. Comparing this result with the result of bias in Peng and Tan, 2018 for linear model. We get the remainder converges in probability with rate $p^{3/2}/r$ which is different than the stochastic bound of order equals $O_p({r_n}^{-3/2})$ in their paper.

4.3 Examples of biases of some subsampling estimators

Example 1 (Linear Regression) The normal equation for the least squares estimator (LSE) in linear regression is

$$\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)\mathbf{x}_i = 0 \tag{4.3.1}$$

From the equation, we have $\psi_{ni}(\beta) = (y_i - \mathbf{x}_i^T\beta)\mathbf{x}_i$, $\dot{\psi}_{ni}(\beta) = -\mathbf{x}_i\mathbf{x}_i^T$ and $\dot{\Psi}(\beta) = -\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T = -\mathbf{X}^T\mathbf{X}$. Since the second derivative $\ddot{\psi}_{ni} = 0$, by Theorem 4.2.2, the bias is given by

$$\begin{aligned}\text{Bias}^*(\hat{\beta}_r^*) &= \frac{-1}{r}(\mathbf{X}^T\mathbf{X})^{-1}\sum_{i=1}^n \frac{1}{\pi_i}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i(y_i - \mathbf{x}_i^T\hat{\beta}_n) + o_p(\sqrt{\frac{p}{rn}}) \\ &= \frac{-1}{r}(\mathbf{X}^T\mathbf{X})^{-1}\sum_{i=1}^n \frac{h_{ii}\mathbf{x}_i\hat{e}_i}{\pi_i} + o_p(\sqrt{\frac{p}{rn}})\end{aligned} \tag{4.3.2}$$

where h_{ii} is the i-th diagonal element of the hat matrix $H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ and $\hat{e}_i = y_i - \mathbf{x}_i^T\hat{\beta}_n$ is the residual of the i-th observation. Note that the main term becomes zero when the subsampling probability is proportional to the leverage scores, i.e. $\pi_i = h_{ii}/p, \quad i = 1, \ldots, n$.

Example 2 (Poisson Regression) Let Y follows a Poisson distribution with mean parameter μ, Poi(μ). Then the probability mass function of Y is given by

$$f_{\text{poi}}(y;\mu) = \exp(-\mu)\frac{\mu^y}{y!}, \quad y = 0, 1, 2, \ldots \tag{4.3.3}$$

The mean of Y and covariate vector $\mathbf{x}_i$ satisfy $E(Y_i) = \mu_i = \exp(\mathbf{x}_i^T\beta)$ where $\beta \in \mathbb{R}^p$ is the regression coefficient and the inverse link function is $h(t) = \exp(t)$. The normal equation for Poisson regression is

$$\sum_{i=1}^n (y_i - \exp(\mathbf{x}_i^T\beta))\mathbf{x}_i = 0 \tag{4.3.4}$$

where $\sum_{i=1}^n \psi_{ni}(\hat{\beta}_n) = 0$. Hence, we have $\psi_{ni}(\beta) = \mathbf{x}_i(y_i - \exp(\mathbf{x}_i^T\beta)$, $\dot{\psi}_{ni}(\beta) = -\mathbf{x}_i\mathbf{x}_i^T\exp(\mathbf{x}_i^T\beta)$, and $\dot{\Psi}(\beta) = -\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\exp(\mathbf{x}_i^T\beta) = -\mathbf{X}^T\Phi\mathbf{X}$, where $\Phi = \text{Diag}(\exp(\mathbf{x}_i^T\beta))$. Moreover,

$$\ddot{\Psi}_n(\beta) = \begin{pmatrix} -\sum_{j=1}^n x_{j1}\exp(\mathbf{x}_j^T\beta)\mathbf{x}_j\mathbf{x}_j^T \\ \vdots \\ -\sum_{j=1}^n x_{jp}\exp(\mathbf{x}_j^T\beta)\mathbf{x}_j\mathbf{x}_j^T \end{pmatrix}$$

Hence, the first term of main term of (4.2.12) is

$$\begin{aligned}\dot{\psi}_{ni}(\hat{\beta}_n)a_{ni} &= \mathbf{x}_i\mathbf{x}_i^T\exp(\mathbf{x}_i^T\hat{\beta}_n)(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\exp(\mathbf{x}_i^T\hat{\beta}_n))^{-1}\mathbf{x}_i(y_i - \exp(\mathbf{x}_i^T\hat{\beta}_n) \\ &= \tilde{h}_{ii}\hat{e}_i\mathbf{x}_i\end{aligned} \tag{4.3.5}$$

where $\tilde{h}_{ii} = \mathbf{x}_i^T \exp(\mathbf{x}_i^T\hat{\beta}_n)(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T \exp(\mathbf{x}_i^T\hat{\beta}_n))^{-1}\mathbf{x}_i$ is the diagonal of the generalized hat matrix $\tilde{H} = \Phi^{1/2}\mathbf{X}(\mathbf{X}^T\Phi\mathbf{X})^{-1}\mathbf{X}^T\Phi^{1/2}$ evaluated at $\hat{\beta}_n$, and $\hat{e}_i = y_i - \exp(\mathbf{x}_i^T\hat{\beta}_n)$ is the residual.

The second term of the main term of (4.2.12) is

$$\begin{aligned}
c_{nii} &= a_{ni}^T \circ \ddot{\Psi}_n(\beta_n) \circ a_{ni} \\
&= \mathbf{x}_i^T\hat{e}_i(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T \exp(\mathbf{x}_i^T\hat{\beta}_n))^{-1} \circ \ddot{\Psi}_n(\hat{\beta}_n) \circ (\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T \exp(\mathbf{x}_i^T\hat{\beta}_n))^{-1}\mathbf{x}_i\hat{e}_i
\end{aligned}$$

The k-th componenet of c_{nii} is

$$\begin{aligned}
c_{nii,k} &= -\hat{e}_i^2\mathbf{x}_i^T(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T \exp(\mathbf{x}_i^T\hat{\beta}_n))^{-1} \\
&\qquad (\sum_{j=1}^n x_{jk}\exp(\mathbf{x}_j^T\hat{\beta}_n)\mathbf{x}_j\mathbf{x}_j^T)(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T \exp(\mathbf{x}_i^T\hat{\beta}_n))^{-1}\mathbf{x}_i \qquad (4.3.6) \\
&= -\hat{e}_i^2 \exp(-\mathbf{x}_i^T\hat{\beta}_n)\sum_{j=1}^n x_{jk}\tilde{h}_{ij}^2
\end{aligned}$$

Thus, the bias of $\hat{\beta}_r^*$ under Poisson regression model is

$$\text{Bias}^*(\hat{\beta}_r^*) = \frac{-(\mathbf{X}^T\Phi\mathbf{X}|_{\hat{\beta}_n})^{-1}}{r}\sum_{i=1}^n \frac{\hat{e}_i}{\pi_i}(\tilde{h}_{ii}\mathbf{x}_i + \frac{1}{2}\hat{e}_i\exp(-\mathbf{x}_i^T\hat{\beta}_n)\mathbf{X}^T\tilde{h}_i^2) + o_p(\sqrt{\frac{p}{rn}}) \qquad (4.3.7)$$

where we define $\tilde{h}_i^2 := [\tilde{h}_{i1}^2, \dots, \tilde{h}_{in}^2]^T$ for $i = 1, \dots, n$.

Example 3 (Negative Binomial regression) Let Y follows a negative binomial distribution with mean μ and overdispersion parameter $\alpha > 0$, $\text{Nb}(\mu, \alpha)$. Then the probability mass function of y is given by

$$f_{\text{nb}}(y; \mu, \alpha) = \frac{\Gamma(y + 1/\alpha)}{\Gamma(1/\alpha)y!}(1 + \alpha\mu)^{-1/\alpha}(\mu/(\mu + 1/\alpha))^{-y}, \ y = 0, 1, 2, \dots \qquad (4.3.8)$$

Similar to Poisson regression, the mean of Y and covariate vector $\mathbf{x}_i$ satisfy $E(Y_i) = \mu_i = \exp(\mathbf{x}_i^T\beta)$ and the normal equation for negative binomial regression is

$$\sum_{i=1}^n \frac{y_i - \exp(\mathbf{x}_i^T\beta)}{1 + \alpha\exp(\mathbf{x}_i^T\beta)}\mathbf{x}_i = 0 \qquad (4.3.9)$$

Thus, we have $\psi_{ni} = \frac{y_i - \exp(\mathbf{x}_i^T\beta)}{1+\alpha\exp(\mathbf{x}_i^T\beta)}\mathbf{x}_i$, $\dot{\psi}_{ni} = -\frac{(1+\alpha y_i)\exp(\mathbf{x}_i^T\beta)}{(1+\alpha\exp(\mathbf{x}_i^T\beta))^2}\mathbf{x}_i\mathbf{x}_i^T$, and $\dot{\Psi}(\beta) = -\mathbf{X}^T\Phi\mathbf{X}$, where $\Phi = \mathrm{Diag}\left(\frac{(1+\alpha y_i)\exp(\mathbf{x}_i^T\beta)}{(1+\alpha\exp(\mathbf{x}_i^T\beta))^2}\right)$. In addition, let

$$s_i = \frac{(1+\alpha y_i)(1-\alpha\exp(\mathbf{x}_i^T\beta))\exp(\mathbf{x}_i^T\beta)}{(1+\alpha\exp(\mathbf{x}_i^T\beta))^3} = \phi_{ii}u_i$$

Then the second derivative is

$$\ddot{\Psi}_n(\beta) = \begin{pmatrix} -\sum_{j=1}^n x_{j1}s_j\mathbf{x}_j\mathbf{x}_j^T \\ \vdots \\ -\sum_{j=1}^n x_{jp}s_j\mathbf{x}_j\mathbf{x}_j^T \end{pmatrix}$$

The first term of the bias in (4.2.12) is

$$\begin{aligned} \dot{\psi}_{ni}(\hat{\beta}_n)a_{ni} &= \frac{(1+\alpha y_i)\exp(\mathbf{x}_i^T\hat{\beta}_n)}{(1+\alpha\exp(\mathbf{x}_i^T\hat{\beta}_n))^2}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\Phi\mathbf{X})^{-1}|_{\hat{\beta}_n}\frac{y_i-\exp(\mathbf{x}_i^T\hat{\beta}_n)}{1+\alpha\exp(\mathbf{x}_i^T\hat{\beta}_n)}\mathbf{x}_i \\ &= \tilde{h}_{ii}\hat{e}_i\mathbf{x}_i \end{aligned} \tag{4.3.10}$$

where the generalized hat matrix is $\tilde{H} = \Phi^{1/2}\mathbf{X}(\mathbf{X}^T\Phi\mathbf{X})^{-1}\mathbf{X}^T\Phi^{1/2}$ and $\tilde{h}_{ii}$ is the diagonal entries of $\tilde{H}$ evaluated at $\hat{\beta}_n$. And $\hat{e}_i = \psi_{ni}(\hat{\beta}_n)$. The second term of the bias in (4.2.12) is

$$\begin{aligned} c_{nii} &= a_{ni}^T \circ \ddot{\Psi}_n(\hat{\beta}_n) \circ a_{ni} \\ &= \mathbf{x}_i^T\hat{e}_i(\mathbf{X}^T\Phi\mathbf{X})^{-1} \circ \ddot{\Psi}_n(\hat{\beta}_n) \circ (\mathbf{X}^T\Phi\mathbf{X})^{-1}\mathbf{x}_i\hat{e}_i \end{aligned}$$

The k-th componenet of c_{nii} is

$$\begin{aligned} c_{nii,k} &= -\hat{e}_i^2\mathbf{x}_i^T(\mathbf{X}^T\Phi\mathbf{X})^{-1}(\sum_{j=1}^n x_{jk}s_j\mathbf{x}_j\mathbf{x}_j^T)(\mathbf{X}^T\Phi\mathbf{X})^{-1}\mathbf{x}_i \\ &= -\frac{\hat{e}_i^2}{\phi_{ii}}\sum_{j=1}^n x_{jk}u_j\tilde{h}_{ij}^2 \end{aligned} \tag{4.3.11}$$

Thus, the bias of $\hat{\beta}_r^*$ under negative binomial regression model is

$$\mathrm{Bias}^*(\hat{\beta}_r^*) = \frac{-(\mathbf{X}^T\Phi\mathbf{X}|_{\hat{\beta}_n})^{-1}}{r}\sum_{i=1}^n \frac{\hat{e}_i}{\pi_i}(\tilde{h}_{ii}\mathbf{x}_i + \frac{\hat{e}_i}{2\phi_{ii}}\mathbf{X}^T\tilde{\tilde{h}}_i^2) + o_p(\sqrt{\frac{p}{rn}}) \tag{4.3.12}$$

where we define $\tilde{\tilde{h}}_i^2 := [\tilde{h}_{i1}^2u_1, \ldots, \tilde{h}_{in}^2u_n]^T$ for $i = 1, \ldots, n$.

4.4 Asymptotic behaviors of the M-estimators for fixed/growing dimension

We study the case of the estimator based on a general estimating equation for finite dimension $p = p_n < \infty$. We need a similar version of conditions (R1)-(R6), which can be formally obtained by setting all $\pi_{ni} = 1$ (the uniform sampling). Let

$$J_{1n}(\beta) = \sum_{i=1}^{n} \mathbb{E}(\psi_{ni}(\beta)^{\otimes 2}), \quad \lambda_{1n} = \lambda_{\max}^{1/2}(J_{1n}).$$

(R1') $$\lambda_{1n} \to \infty, \quad \inf_{n \geq n_0}\{\lambda_{1n}^{-2}\lambda_{\text{amin}}(\mathbb{E}(\dot{\Psi}_n^{(s)}))\} > 0.$$

(R2') Assume $\exists \tilde{B} \in \mathbb{R}^{p\times p}$ where $\tilde{B} = [\tilde{\beta}_1, \ldots, \tilde{\beta}_p]$ with $\tilde{\beta}_i$ lies in a neighborhood of β_0 such that

$$\Psi_n(\beta) = \Psi_n(\beta_0) + \dot{\Psi}_n(\beta - \beta_0) + \frac{1}{2}(\beta - \beta_0)^T \circ \ddot{\Psi}_n(\tilde{B}) \circ (\beta - \beta_0)$$

(R3') $$\textstyle\sum_{i=1}^{n} \mathbb{E}\big(\|\dot{\psi}_{ni} - \mathbb{E}(\dot{\psi}_{ni})\|^2\big) = O(p^2\lambda_{1n}^2).$$

(R4') Same as (R4) except that η_{ni} are replaced with η_{1ni} which satisfy

$$\sum_{i=1}^{n} \|\eta_{1ni}\|^2 = O_P(n^{-1}p\lambda_{1n}^2).$$

(R5') $$\lambda_{\max}(J_{1n})/\lambda_{\min}(J_{1n}) = O(1).$$

(R6') Fix $u \in \mathbb{R}^{p_n}$ with $\|u\| = 1$. Let $s_{1n}^2 = u^\top \mathbb{E}^{-1}(\dot{\Psi}_n) J_{1n} \mathbb{E}^{-\top}(\dot{\Psi}_n)u$. The double array $z_{1ni} = s_{1n}^{-1}u^\top \mathbb{E}^{-1}(\dot{\Psi}_n)\psi_{ni}$, $i = 1, 2, \ldots, n$, $n \geq 1$ satisfies

$$\sum_{i=1}^{n} \|z_{1ni}\|^2 = o_P(1), \quad \mathbb{E}(\max_i \|z_{1ni}\|) = o(1).$$

for every $t > 0$.

(R7') There exists $\eta_{n,d}$ and a neighborhood $\mathbb{N}_0$ of β_0 such that $\forall B = [\beta, \dots, \beta]^T \in \mathbb{R}^{p\times p}$, where β lies in $\mathbb{N}_0$,

$$\lambda_{\text{amax}}(\ddot{\Psi}_{n,d}(\beta)) \le \eta_{nd}, \quad d = 1, \dots, p,$$

$$\mathbb{E}(\|\eta_n\|^{2k}) = O(p^k \lambda_{1n}^{4k}), \quad k = 1, 2, \dots$$

where $\eta_n = (\eta_{n1}, \dots, \eta_{np})^T$.

(R8') $$\|\mathbb{E}(\bar{\dot{\Psi}}_n \bar{\dot{\Psi}}_n^T)\|_o = O(\lambda_{1n}^2).$$

The following theorem describes the asymptotic behaviors of the M-estimator for both fixed and growing parameter dimension.

Theorem 4.4.1 *Suppose (R1'), (R2), (R3')–(R5') hold. Then there exists a sequence of solutions $\hat{\beta}_n$ of (4.0.1) such that if $p/\lambda_{1n}^2 = o(1)$, then*

$$p^{-1/2}\lambda_{1n}(\hat{\beta}_n - \beta_0) = O_P(1), \tag{4.4.1}$$

$$\lambda_{1n}^{-1}\mathbb{E}(\dot{\Psi}_n)(\hat{\beta}_n - \beta_0) = -\lambda_{1n}^{-1}\sum_{i=1}^n \psi_{ni} + o_P(1). \tag{4.4.2}$$

If, further, (R5')–(R6') are satisfied for $u \in \mathbb{R}^p$ with $\|u\| = 1$, then

$$s_{1n}^{-1}u^\top(\hat{\beta}_n - \beta_0) \Rightarrow \mathscr{N}(0,1), \quad \textit{in probability.} \tag{4.4.3}$$

Proof of Theorem 4.4.1. For $t \in \mathbb{R}^p$, let

$$\Delta_n(t) = p^{-1/2}\lambda_{1n}^{-1}\big(\Psi_n(\beta_0 + p^{1/2}\lambda_{1n}^{-1}t) - \Psi_n(\beta_0)\big) - \lambda_{1n}^{-2}\mathbb{E}(\dot{\Psi}_n)t. \tag{4.4.4}$$

For arbitrary $C > 0$, fix $\|t\| \le C$. Then $p^{1/2}\lambda_{1n}^{-1}t \to 0$. By (R2) and the inequality in (R4'),

$$\|\Delta_n(t)\|^2 \le 2C^2\lambda_{1n}^{-4}\big\|\dot{\Psi}_n - \mathbb{E}(\dot{\Psi}_n)\big\|_o^2 + 1/2C^4 p\lambda_{1n}^{-6}\big\|\sum_{i=1}^n \eta_{1ni}\big\|^2.$$

Clearly, $\mathbb{E}(\big\|\dot{\Psi}_n - \mathbb{E}(\dot{\Psi}_n)\big\|_o^2) \le \mathbb{E}(\big\|\dot{\Psi}_n - \mathbb{E}(\dot{\Psi}_n)\big\|^2) = \sum_i \mathbb{E}\big(\|\dot{\psi}_{ni} - \mathbb{E}(\dot{\psi}_{ni})\|^2\big)$. This, (R3') and the equality in (R4') imply that

$$\mathbb{E}(\sup_{\|t\|\le C} \|\Delta_n(t)\|^2) = o(1/p). \tag{4.4.5}$$

Recall $\dot{\Psi}_n^{(s)} = 1/2(\dot{\Psi}_n + \dot{\Psi}_n^\top)$. Assume without loss of generality that $\lambda_{\text{amin}}(\mathbb{E}(\dot{\Psi}_n^{(s)})) = \lambda_{\min}(\mathbb{E}(\dot{\Psi}_n^{(s)})) > 0$ for large n. By (4.4.4),

$$\begin{aligned} \ell_n(C) =&: \inf_{\|t\|=C} \left\{ p^{-1/2}\lambda_{1n}^{-1} t^\top \Psi_n(\beta_0 + p^{1/2}\lambda_{1n}^{-1} t) \right\} \\ &\geq C^2 \lambda_{1n}^{-2} \lambda_{\text{amin}}(\mathbb{E}(\dot{\Psi}_n^{(s)})) - C \sup_{\|t\|=C} \big\|\Delta_n(t)\big\| - Cp^{-1/2}\lambda_{1n}^{-1}\big\|\Psi_n\big\|. \end{aligned} \tag{4.4.6}$$

By (R1'), $\lambda_{1n}^{-2}\lambda_{\min}(\mathbb{E}(\dot{\Psi}_n^{(s)})) \geq 2c_0$ for some $c_0 > 0$. Hence it follows from $\mathbb{E}(\|\Psi_n\|^2) \leq p\lambda_{1n}^2$ and (4.4.5) that for large C,

$$\begin{aligned} \mathbb{P}(\ell_n(C) > 0) &\geq 1 - \mathbb{P}(\sup_{\|t\|=C} \|\Delta_n(t)\| > c_0 C) - \mathbb{P}(p^{-1/2}\lambda_{1n}^{-1}\|\Psi_n\| > c_0 C) \\ &= 1 - o(1). \end{aligned}$$

Following Chatterjee and Bose (2005), on the set $\ell_n(C) > 0$ the continuity of $\Psi_n(\beta)$ implies that there is a root $t = t_n$ of $\Psi_n(\beta_0 + p^{1/2}\lambda_{1n}^{-1}t) = 0$ with $\|t_n\| \leq C$. This holds on an event with probability approaching one and for large C. Thus $\hat{\beta}_n = \beta_0 + p^{1/2}\lambda_{1n}^{-1}t_n$ is a root of (4.0.1) and satisfies $\mathbb{P}(p^{-1/2}\lambda_{1n}\|\hat{\beta}_n - \beta_0\| \leq C) \geq 1 - o(1)$, which shows (4.4.1). By (4.4.5), $\Delta_n(t_n) = o_p(p^{-1/2})$, which shows (4.4.2).

The asymptotic normality (4.4.3) follows from the established relation (4.4.2) and Theorem 5.4.2 of Borovskikh and Korolyuk (1997). More specifically, (R6') implies that the main term on the left side of (4.4.2) has an asymptotic standard normal, while the remainder term is negligible, that is,

$$s_{1n}^{-1}\lambda_{1n} u^\top \mathbb{E}^{-1}(\dot{\Psi}_n)\alpha_n = o_p(1),$$

where $\alpha_n = o_p(1)$. This follows from

$$\frac{\lambda_{1n}}{s_{1n}} \leq \frac{\lambda_{\max}(J_{1n})}{\lambda_{\min}(J_{1n})\|u^\top \mathbb{E}^{-1}(\dot{\Psi}_n)\|} \leq \frac{c}{\|u^\top \mathbb{E}^{-1}(\dot{\Psi}_n)\|},$$

where c is a constant implied by (R5'). The proof is now complete. ■

4.5 Asymptotic bias of the full sample estimator

We first derive the general expression of the bias of the full sample estimator $\hat{\beta}_n$ for GEE. We will show below that the bias is not negligible when the dimension is high.

Under generalized linear model (GLM), it is well known that the maximum likelihood estimator (MLEs) is biased when the sample size n is small or the Fisher information is small. Bias of MLEs is well studied in literature. Cordeiro and McCullagh (1991) has derived the general formulae for first-order biases of MLEs in GLM. Cook *et at.* (1986) has derived the biases of MLEs for normal non-linear regression models. Young and Bakir (1987) has presented the MLEs in the generalized log-gamma regression model. Again, Cordeiro and Botter (2001) has found the second-order biases of MLEs in overdispersed generalized linear models.

Note that the existence of estimator $\hat{\beta}_n$ does not guarantee that the expected value of $\hat{\beta}_n$ also exists. Below we give an example to illustrate.

Example 4 Consider $X_1, \ldots, X_n \overset{iid}{\sim} N(\frac{1}{\beta}, 1)$. Then the estimating equations for β is

$$\sum_{i=1}^{n}(-X_i + \frac{1}{\beta}) = 0$$

Hence, the MLE is $\hat{\beta}_n = \dfrac{1}{\bar{X}}$. By the asymptotic theory, with true value $\beta_0 \neq 0$, $\sqrt{n}(\hat{\beta}_n - \beta_0) \Rightarrow N(0, \beta^4)$. Unfortunately, the expected value $\mathbb{E}(\hat{\beta}_n)$ does not exist. In fact, the distribution of $\bar{X}$ is $N(\frac{1}{\beta}, \frac{1}{n})$. That is, $\sqrt{n}\bar{X} \sim N(\frac{1}{\beta}, 1)$. The expected value of $\hat{\beta}_n$ is then

$$\mathbb{E}(\hat{\beta}_n) = \sqrt{n}\mathbb{E}(\frac{1}{\sqrt{n}\bar{X}}) = \frac{\sqrt{n}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{t} \exp^{-\frac{1}{2}(t-\frac{1}{\beta})^2} dt \qquad (4.5.1)$$

where the integral does not exist, and thus the expected value of the MLE does not exist.

This example provokes us to impose some further conditions for the existence of the expected value of estimators of GEE. Let $\mathcal{B} = \mathbb{N}_0$ where $\mathbb{N}_0$ is a neighborhood of β_0,

(U_0) $\quad \sup_{B\in\mathcal{B}^p} |\mathbb{E}[\dot{\Psi}_n^T(B)\dot{\Psi}_n(B)]^{-1}|_o = O(\lambda_{1n}^{-4})$,

(U_1) $\quad \mathbb{E}(\|\Psi_n\|^{2k}) = O(p^k\lambda_{1n}^{2k}), \quad$ for $k = 0, 1, 2, \ldots$,

(U_2) $\quad \sup_{B\in\mathcal{B}^p} \mathbb{E}(|\dot{\Psi}_n^{-1}(B)|_o^{2k}) = O(\lambda_{1n}^{-4k}), \quad$ for $k = 0, 1, 2, \ldots$.

Theorem 4.5.1 *Assume* $\dot{\Psi}_n^{-1}(B)$ *exists for all* $B \in \mathcal{B}^p$. *Assume* (U_0) *and* (U_1) *hold. Then* $\|\mathbb{E}(\hat{\beta}_n - \beta_0)\| = o(\sqrt{p}\lambda_{1n}^{-1})$.

Proof of Theorem 4.5.1. Consider the first-order Taylor expansion of $\Psi_n(\hat{\beta}_n)$ about β_0

$$0 = \Psi_n(\hat{\beta}_n) = \Psi_n + \dot{\Psi}_n(\tilde{B})(\hat{\beta}_n - \beta_0) \tag{4.5.2}$$

where $\tilde{\beta}_d$ lies in between β_0 and $\hat{\beta}_n$. By (U_0) and (U_1) with $k = 1$,

$$\begin{aligned} \|\mathbb{E}(\hat{\beta}_n - \beta_0)\|^2 &= \|\mathbb{E}(\dot{\Psi}_n^{-1}(\tilde{B})\Psi_n)\|^2 \\ &\leq |\mathbb{E}(\dot{\Psi}_n^{-1}(\tilde{B})\dot{\Psi}_n^{-T}(\tilde{B}))|_o \mathbb{E}(\|\Psi_n\|^2) \qquad (4.5.3) \\ &\leq \sup_{B \in \mathcal{B}^p} |\mathbb{E}[\dot{\Psi}_n^T(B)\dot{\Psi}_n(B)]^{-1}|_o \mathbb{E}(\|\Psi_n\|^2) = O(p\lambda_{1n}^{-2}). \end{aligned}$$

Hence, $\|\mathbb{E}(\hat{\beta}_n - \beta_0)\| = O(\sqrt{p}\lambda_{1n}^{-1})$. ■

Typically, $\lambda_{1n} = O(\sqrt{n})$. Theorem 4.5.1 shows that the bias of $\hat{\beta}_n$ exists and the main component of the bias is of order $O(\sqrt{p/n})$. For fixed dimension, the bias is negligible however for growing dimension the bias is not negligible unless $p/n \to 0$. Even if $p/n \to 0$, the bias is asymptotically zero, but the rate can be very slow. We will establish this result through getting the Taylor expansion of the bias. We need some more lemmas for the proof of our main result.

Lemma 4.5.1 *Assume* (U_1) *and* (U_2) *hold. Then for each positive integer* k,

$$\mathbb{E}(\|\hat{\beta}_n - \beta_0\|^k) = O(p^{k/2}\lambda_{1n}^{-k}). \tag{4.5.4}$$

Proof of Lemma 4.5.1. Using the Taylor expansion of Ψ_n as in Theorem 4.5.1 and Cauchy-Schwarz,

$$\begin{aligned} \mathbb{E}(\|\hat{\beta}_n - \beta_0\|^k) &\leq \mathbb{E}(|\dot{\Psi}_n^{-1}(\tilde{B})|_o^k \|\Psi_n\|^k) \\ &\leq \sqrt{\mathbb{E}(|\dot{\Psi}_n^{-1}(\tilde{B})|_o^{2k})}\sqrt{\mathbb{E}(\|\Psi_n\|^{2k})} \\ &\leq \sqrt{\sup_{B \in \mathcal{B}^p} \mathbb{E}(|\dot{\Psi}_n^{-1}(B)|_o^{2k})}\sqrt{\mathbb{E}(\|\Psi_n\|^{2k})} \\ &= O(\lambda_{1n}^{-2k})O(p^{k/2}\lambda_{1n}^{k}) \\ &= O(p^{k/2}\lambda_{1n}^{-k}) \end{aligned}$$

■

Lemma 4.5.2 *Assume (R7') holds. Then for each positive integer k,*

$$\mathbb{E}\left(\sup_{\tilde{B}\in\mathbb{N}_0^p}\|\ddot{\Psi}_n(\tilde{B})\|_{oe}^{2k}\right) = O(p^k\lambda_{1n}^{4k}). \tag{4.5.5}$$

Proof of Lemma 4.5.2. By (R7'),

$$\begin{aligned}\mathbb{E}(\sup_{\tilde{B}\in\mathbb{N}_0^p}\|\ddot{\Psi}_n(\tilde{B})\|_{oe}^{2k}) &= \mathbb{E}\left(\sup_{\tilde{B}\in\mathbb{N}_0^p}\sum_{d=1}^{p}\|\ddot{\Psi}_{n,d}(\tilde{\beta}_d)\|_o^2\right)^k \\ &\leq \mathbb{E}(\sum_{d=1}^{p}\eta_{nd}^2)^k \\ &= \mathbb{E}(\|\eta_n\|^{2k}) = O(p^k\lambda_{1n}^{4k})\end{aligned}$$

■

Throughout the proof, if there is no parameter indicated, we assume the function is evaluated at β_0.

Theorem 4.5.2 *Suppose (R1')–(R5'), (R7') and (U_2) hold. Assume $\hat{\beta}_n$ is a solution to (4.0.1) from Theorem 4.4.1. Then we have,*

$$\hat{\beta}_n - \beta_0 = -\dot{\Psi}_n^{-1}\Psi_n - \frac{1}{2}\dot{\Psi}_n^{-1}(\hat{\beta}_n-\beta_0)^T\circ\ddot{\Psi}_n(\tilde{B})\circ(\hat{\beta}_n-\beta_0) =: l_n - \alpha_n \tag{4.5.6}$$

where $\mathbb{E}(\|l_n\|^{2k}) = O(p^k\lambda_{1n}^{-2k})$, *and* $\mathbb{E}(\|\alpha_n\|^{2k}) = O(p^{3k}\lambda_{1n}^{-4k})$.

Proof of Theorem 4.5.2. By (R2'),

$$0 = \Psi_n(\hat{\beta}_n) = \Psi_n + \dot{\Psi}_n(\hat{\beta}_n-\beta_0) + \frac{1}{2}(\hat{\beta}_n-\beta_0)^T\circ\ddot{\Psi}_n(\tilde{B})\circ(\hat{\beta}_n-\beta_0), \tag{4.5.7}$$

Rearranging the terms, we get

$$\hat{\beta}_n - \beta_0 = -\dot{\Psi}_n^{-1}\Psi_n - \frac{1}{2}\dot{\Psi}_n^{-1}(\hat{\beta}_n-\beta_0)^T\circ\ddot{\Psi}_n(\tilde{B})\circ(\hat{\beta}_n-\beta_0). \tag{4.5.8}$$

Define l_n as the first term and α_n as the second term of (4.5.8). Define $a_{ni} = \dot{\Psi}_n^{-1}\psi_{ni}$. It is easy to see $l_n = -\dot{\Psi}_n^{-1}\mathbb{E}(\dot{\Psi}_n)\mathbb{E}^{-1}(\dot{\Psi}_n)\bar{\Psi}_n = \dot{\Psi}_n^{-1}\mathbb{E}(\dot{\Psi}_n)\tilde{l}_n$, and $\|l_n\| \leq |\dot{\Psi}_n^{-1}|_o|\mathbb{E}(\dot{\Psi}_n)|_o\|\tilde{l}_n\|$, where $\tilde{l}_n = -\mathbb{E}^{-1}(\dot{\Psi}_n)\bar{\Psi}_n = -\sum_{i=1}^n \mathbb{E}^{-1}(\dot{\Psi}_n)\bar{\psi}_{ni}$. Then,

$$\begin{aligned}
\mathbb{E}(\|l_n\|^{2k}) &\leq |\mathbb{E}(\dot{\Psi}_n)|_o^{2k}\mathbb{E}(|\dot{\Psi}_n^{-1}|_o^{2k}\|\tilde{l}\|^{2k}) \\
&\leq |\mathbb{E}(\dot{\Psi}_n)|_o^{2k}\sqrt{\mathbb{E}(|\dot{\Psi}_n^{-1}|_o^{4k})}\sqrt{\mathbb{E}(\|\tilde{l}\|^{4k})} \\
&\leq \lambda_{1n}^{4k}\sqrt{O(\lambda_{1n}^{-8k})}\sqrt{O(p^{2k}\lambda_{1n}^{-4k})} \\
&= O(p^k\lambda_{1n}^{-2k})
\end{aligned}$$

By (R7'), (U$_2$), Lemma 4.5.1 and Lemma 4.5.2, for each positive integer k,

$$\begin{aligned}
4\mathbb{E}(\|\alpha_n\|^{2k}) &= \mathbb{E}(\|\dot{\Psi}_n^{-1}(\hat{\beta}_n - \beta_0)^T \circ \ddot{\Psi}_n(\tilde{B}) \circ (\hat{\beta}_n - \beta_0)\|^{2k}) \\
&\leq \mathbb{E}(|\dot{\Psi}_n^{-1}|_o^{2k}\|(\hat{\beta}_n - \beta_0)^T \circ \ddot{\Psi}_n(\tilde{B}) \circ (\hat{\beta}_n - \beta_0)\|^{2k}) \\
&\leq \mathbb{E}(|\dot{\Psi}_n^{-1}|_o^{2k}\|\hat{\beta}_n - \beta_0\|^{4k}\|\ddot{\Psi}_n(\tilde{B})\|_{oe}^{2k}) \\
&\leq \sqrt[4]{\mathbb{E}(|\dot{\Psi}_n^{-1}|_o^{8k})}\sqrt[4]{\mathbb{E}(\|\hat{\beta}_n - \beta_0\|^{16k})}\sqrt{\mathbb{E}(\|\ddot{\Psi}_n(\tilde{B})\|_{oe}^{4k})} \\
&= O(\lambda_{1n}^{-4k})O(p^{2k}\lambda_{1n}^{-4k})O(p^k\lambda_{1n}^{4k}) \\
&= O(p^{3k}\lambda_{1n}^{-4k})
\end{aligned}$$

■

Theorem 4.5.3 *Suppose (R1')-(R5'), (R7') and (R8') hold. Assume (U$_0$), (U$_1$), (U$_2$) hold for $k = 4$ and $\hat{\beta}_n$ is a solution to (4.0.1) from Theorem 4.4.1. Then the bias of $\hat{\beta}_n$ is*

$$\mathbb{E}(\hat{\beta}_n) - \beta_0 = -\mathbb{E}^{-1}(\dot{\Psi}_n)\sum_{i=1}^n \mathbb{E}[-\bar{\dot{\psi}}_{ni}\bar{a}_{ni} + \frac{1}{2}(\bar{a}_{ni}^T \circ \ddot{\Psi}_n \circ \bar{a}_{ni})] + O(\frac{p^{7/2}}{\lambda_{1n}^3}) \qquad (4.5.9)$$

where $\bar{\dot{\psi}}_{ni} = \dot{\psi}_{ni} - \mathbb{E}(\dot{\psi}_{ni})$, $\bar{a}_{ni} = \dot{\Psi}_n^{-1}\bar{\psi}_{ni}$.

Proof of Theorem 4.5.3. Consider the second-order Taylor expansion of $\Psi_n(\hat{\beta}_n)$ about β_0

$$0 = \Psi_n(\hat{\beta}_n) = \Psi_n + \dot{\Psi}_n(\hat{\beta}_n - \beta_0) + \frac{1}{2}(\hat{\beta}_n - \beta_0)^T \circ \ddot{\Psi}_n(\tilde{B}) \circ (\hat{\beta}_n - \beta_0),$$

where $\tilde{B} = [\tilde{\beta}_1, \ldots, \tilde{\beta}_p]^T$, and each $\tilde{\beta}_i$ lies in between $\hat{\beta}_n$ and β_0. Simple algebra yields,

$$\begin{aligned} 0 = \Psi_n + \mathbb{E}(\dot{\Psi}_n)(\hat{\beta}_n - \beta_0) + \bar{\dot{\Psi}}_n(\hat{\beta}_n - \beta_0) + \frac{1}{2}(\hat{\beta}_n - \beta_0)^T \circ \ddot{\Psi}_n \circ (\hat{\beta}_n - \beta_0) + \\ \frac{1}{2}(\hat{\beta}_n - \beta_0)^T \circ [\ddot{\Psi}_n(\tilde{B}) - \ddot{\Psi}_n] \circ (\hat{\beta}_n - \beta_0). \end{aligned}$$

Denote by $W_n \in \mathbb{R}^p$ the last term. Taking expectation on both sides of the above equality, we get

$$0 = \mathbb{E}(\dot{\Psi}_n)(\mathbb{E}(\hat{\beta}_n) - \beta_0) + \mathbb{E}(\bar{\dot{\Psi}}_n(\hat{\beta}_n - \beta_0)) + \frac{1}{2}\mathbb{E}[(\hat{\beta}_n - \beta_0)^T \circ \ddot{\Psi}_n \circ (\hat{\beta}_n - \beta_0)] + \mathbb{E}(W_n),$$

Rearranging the terms, we have

$$-\mathbb{E}(\dot{\Psi}_n)(\mathbb{E}(\hat{\beta}_n) - \beta_0) = \mathbb{E}(\bar{\dot{\Psi}}_n(\hat{\beta}_n - \beta_0)) + \frac{1}{2}\mathbb{E}[(\hat{\beta}_n - \beta_0)^T \circ \ddot{\Psi}_n \circ (\hat{\beta}_n - \beta_0)] + \mathbb{E}(W_n). \tag{4.5.10}$$

Substituting $\hat{\beta}_n - \beta_0 = l_n - \alpha_n$ from Theorem 4.5.2 to (4.5.10),

$$\begin{aligned} -\mathbb{E}(\dot{\Psi}_n)(\mathbb{E}(\hat{\beta}_n) - \beta_0) = \mathbb{E}(\bar{\dot{\Psi}}_n l_n + \frac{1}{2} l_n^T \circ \ddot{\Psi}_n \circ l_n) - \mathbb{E}(\bar{\dot{\Psi}}_n \alpha_n) - \mathbb{E}(l_n^T \circ \ddot{\Psi}_n \circ \alpha_n) + \\ \frac{1}{2}\mathbb{E}(\alpha_n^T \circ \ddot{\Psi}_n \circ \alpha_n) + \mathbb{E}(W_n). \end{aligned} \tag{4.5.11}$$

We now show that the first term on the right has the slowest rate and thus is the main term of bias. Write $\bar{l}_n = -\sum_{i=1}^n \bar{a}_{ni}$ where $\bar{a}_{ni} = \dot{\Psi}_n^{-1}\bar{\psi}_{ni}$. By (R8'), the order is

$$\begin{aligned} \|\mathbb{E}(\bar{\dot{\Psi}}_n l_n)\|^2 &\le |\mathbb{E}(\bar{\dot{\Psi}}_n \bar{\dot{\Psi}}_n^T)|_o \mathbb{E}(\|l_n\|^2) \\ &= O(\lambda_{1n}^2) O(p\lambda_{1n}^{-2}) \\ &= O(p). \end{aligned}$$

Hence, $\|\mathbb{E}(\bar{\dot{\Psi}}_n l_n)\| = O(\sqrt{p})$. Note that the d-th component of $\mathbb{E}(l_n^T \circ \ddot{\Psi}_n \circ l_n)$ is

$$\begin{aligned} \mathbb{E}(l_n^T \circ \ddot{\Psi}_{n,d} \circ l_n) &= \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}(a_{ni}^T \sum_{l=1}^n \ddot{\psi}_{nl,d}\; a_{nj}) \\ &= \sum_{i=1}^n \sum_{l=1}^n \mathbb{E}(a_{ni}^T \ddot{\psi}_{nl,d} a_{ni}) + \sum\sum_{i \ne j} \sum_{l=1}^n \mathbb{E}(a_{ni}^T \ddot{\psi}_{nl,d} a_{nj}) \\ &= \sum_{i=1}^n \sum_{l=1}^n \mathbb{E}(a_{ni}^T \ddot{\psi}_{nl,d} a_{ni}). \end{aligned}$$

Stacking them together, we have

$$\mathbb{E}(l_n^T \circ \ddot{\Psi}_n \circ l_n) = \sum_{i=1}^{n} \mathbb{E}(a_{ni}^T \circ \ddot{\Psi}_n \circ a_{ni}).$$

By Lemma 4.5.2 with $k = 4$ and Theorem 4.5.2, the order of this term is

$$\begin{aligned}
\|\mathbb{E}(l_n^T \circ \ddot{\Psi}_n \circ l_n)\|^2 &\leq \mathbb{E}(\|l_n\|^4 \|\ddot{\Psi}_n\|_{oe}^2) \\
&\leq \sqrt{\mathbb{E}(\|l_n\|^8)} \sqrt{\mathbb{E}(\|\ddot{\Psi}_n\|_{oe}^4)} \\
&= O(p^2 \lambda_{1n}^{-4}) O(p \lambda_{1n}^4) \\
&= O(p^3)
\end{aligned}$$

Hence, $\|\mathbb{E}(l_n^T \circ \ddot{\Psi}_n \circ l_n)\| = O(p^{3/2})$. Consider the second term of (4.5.11), by (R8')

$$\begin{aligned}
\|\mathbb{E}(\bar{\dot{\Psi}}_n \alpha_n)\|^2 &\leq \|\mathbb{E}(\bar{\dot{\Psi}}_n \bar{\dot{\Psi}}_n^T)\|_o \mathbb{E}(\|\alpha_n\|^2) \\
&\leq O(\lambda_{1n}^2) O(p^3 \lambda_{1n}^{-4}) = O(p^3 \lambda_{1n}^{-2}).
\end{aligned}$$

Hence, $\|\mathbb{E}(\bar{\dot{\Psi}}_n \alpha_n)\| = O(p^{3/2} \lambda_{1n}^{-1})$. Consider the third term,

$$\begin{aligned}
\|\mathbb{E}(l_n^T \circ \ddot{\Psi}_n \circ \alpha_n)\|^2 &\leq \mathbb{E}(\|l_n\|^2 \|\alpha_n\|^2 \|\ddot{\Psi}_n\|_{oe}^2) \\
&\leq \sqrt[4]{\mathbb{E}(\|l_n\|^8)} \sqrt[4]{\mathbb{E}(\|\alpha_n\|^8)} \sqrt{\mathbb{E}(\|\ddot{\Psi}_n\|_{oe}^4)} \\
&= O(p \lambda_{1n}^{-2}) O(p^3 \lambda_{1n}^{-4}) O(p \lambda_{1n}^4) \\
&= O(p^5 \lambda_{1n}^{-2})
\end{aligned}$$

Hence, $\|\mathbb{E}(l_n^T \circ \ddot{\Psi}_n \circ \alpha_n)\| = O(p^{5/2} \lambda_{1n}^{-1})$. Consider the fourth term,

$$\begin{aligned}
\|\mathbb{E}(\alpha_n^T \circ \ddot{\Psi}_n \circ \alpha_n)\|^2 &\leq \mathbb{E}(\|\alpha_n\|^4 \|\ddot{\Psi}_n\|_{oe}^2) \\
&\leq \sqrt{\mathbb{E}(\|\alpha_n\|^8)} \sqrt{\mathbb{E}(\|\ddot{\Psi}_n\|_{oe}^4)} \\
&= O(p^6 \lambda_{1n}^{-8}) O(p \lambda_{1n}^4) \\
&= O(p^7 \lambda_{1n}^{-4})
\end{aligned}$$

Hence, $\|\mathbb{E}(\alpha_n^T \circ \ddot{\Psi}_n \circ \alpha_n)\| = O(p^{7/2}\lambda_{1n}^{-2})$. Let's establish an inequality for the difference,

$$\begin{aligned}
\|\ddot{\Psi}_n(\tilde{B}) - \ddot{\Psi}_n\|_{oe}^2 &= \sum_{d=1}^{p} |\ddot{\psi}_{nd}(\tilde{\beta}_d) - \ddot{\psi}_{nd}|_o^2 \\
&\le \sum_{d=1}^{p}\sum_{s=1}^{p}\sum_{t=1}^{p} |\ddot{\psi}_{nd,(s,t)}(\tilde{\beta}_d) - \ddot{\psi}_{nd,(s,t)}|^2 \\
&\le \|\hat{\beta}_n - \beta_0\|^2 \sum_{d=1}^{p}\sum_{s=1}^{p}\sum_{t=1}^{p} (\sqrt{p}\delta_{d,(s,t)})^2 \\
&\le p\|\hat{\beta}_n - \beta_0\|^2 \sum_{d=1}^{p} \|\Delta_d\|^2.
\end{aligned}$$

where we assume $\|\Delta_d\| = O(p\lambda_{1n}^2)$ for $d = 1, \dots, p$. Apply the above result on the last term,

$$\begin{aligned}
4\|\mathbb{E}(W_n)\|^2 &= \|\mathbb{E}[(\hat{\beta}_n - \beta_0)^T \circ [\ddot{\Psi}_n(\tilde{B}) - \ddot{\Psi}_n] \circ (\hat{\beta}_n - \beta_0)]\|^2 \\
&\le \mathbb{E}(\|\hat{\beta}_n - \beta_0\|^4 \|\ddot{\Psi}_n(\tilde{B}) - \ddot{\Psi}_n\|_{oe}^2) \\
&\le p\sum_{d=1}^{p} \mathbb{E}(\|\Delta_d\|^2 \|\hat{\beta}_n - \beta_0\|^6) \\
&\le p\sum_{d=1}^{p} \sqrt{\mathbb{E}(\|\Delta_d\|^4)}\sqrt{\mathbb{E}(\|\hat{\beta}_n - \beta_0\|^{12})} \\
&\le p^2 O(p^2\lambda_{1n}^4) O(p^3\lambda_{1n}^{-6}) \\
&= O(p^7\lambda_{1n}^{-2})
\end{aligned}$$

Hence, $\|\mathbb{E}(W_n)\| = O(p^{7/2}\lambda_{1n}^{-1})$. Note that this is the slowest rate except the first term. Thus, dividing $\mathbb{E}(\dot{\Psi}_n)$ onto the other side in (4.5.11), and assume $\mathbb{E}(\dot{\Psi}_n) = O(\lambda_{1n}^2)$, we have the remainder equals $O(p^{7/2}\lambda_{1n}^{-3})$. ■

Remark 4.5.1 For fixed $p_n = p$, the bias is $\mathbb{E}(\hat{\beta}_n) - \beta_0 = G_{1n}^{-1}\sum_{i=1}^{n} b_{ni} + O(p^{7/2}/n^{3/2})$. We construct the bias-corrected estimator $\hat{\beta}_{bc} = \hat{\beta}_n - G_{1n}^{-1}\sum_{i=1}^{n} b_{ni}$. Then the bias is $\mathbb{E}(\hat{\beta}_{bc}) - \beta_0 = O(p^{7/2}/n^{3/2})$. Hence, using the bias-corrected estimator, we can improve the bias by $\frac{p^{7/2}n^{-3/2}}{p^{3/2}n^{-1}} = \frac{p^2}{\sqrt{n}}$.

Next we compute the variance of the full sample estimator. We will make use of the results from Theorem 4.5.2 and Theorem 4.5.3.

Theorem 4.5.4 *Suppose (R1'), (R2), (R3')–(R5'), (R7') and (R8') hold. Assume $\hat{\beta}_n$ is a solution to (4.0.1) from Theorem 4.4.1. Then the variance of $\hat{\beta}_n$ is*

$$Var(\hat{\beta}_n) = \mathbb{E}(l_n^{\otimes 2}) + O(p^2\lambda_{1n}^{-3}) \tag{4.5.12}$$

Proof of Theorem 4.5.4. By (4.5.6) and (4.5.9), we have

$$\hat{\beta}_n - \beta_0 = l_n - \alpha_n \tag{4.5.13}$$

$$\mathbb{E}(\hat{\beta}_n) - \beta_0 = G_{1n}^{-1}(\sum_{i=1}^n b_{ni} + r_n). \tag{4.5.14}$$

where $r_n = O(p^{7/2}/n^{1/2})$. Substracting the first equation from the second, we have

$$\hat{\beta}_n - \mathbb{E}(\hat{\beta}_n) = l_n - \alpha_n - G_{1n}^{-1}(\sum_{i=1}^n b_{ni} + r_n). \tag{4.5.15}$$

Hence, the variance of $\hat{\beta}_n$ is

$$\begin{aligned}\mathbb{E}[(\hat{\beta}_n - \mathbb{E}(\hat{\beta}_n))^{\otimes 2}] = {} & \mathbb{E}(l_n^{\otimes 2}) + \mathbb{E}(\alpha_n^{\otimes 2}) - \mathbb{E}(l_n\alpha_n^T) - \mathbb{E}(\alpha_n l_n^T) \\ & - G_{1n}^{-1}(\sum_{i=1}^n b_{ni} + r_n)\mathbb{E}(l_n^T - \alpha_n^T) \\ & - \mathbb{E}(l_n - \alpha_n)(\sum_{i=1}^n b_{ni} + r_n)^T G_{1n}^{-T} \\ & + G_{1n}^{-1}(\sum_{i=1}^n b_{ni} + r_n)^{\otimes 2} G_{1n}^{-T}\end{aligned} \tag{4.5.16}$$

Note that by Theorem 4.5.2,

$$\begin{aligned}\|\mathbb{E}(\alpha_n^{\otimes 2})\| &\le \mathbb{E}(\|\alpha_n^{\otimes 2}\|) \\ &= \mathbb{E}(\|\alpha_n\|^2) \\ &= O(p^3\lambda_{1n}^{-4}).\end{aligned}$$

and

$$\begin{aligned}\|\mathbb{E}(l_n\alpha_n^T)\|^2 &\le \mathbb{E}(\|l_n\|^2)\mathbb{E}(\|\alpha_n\|^2) \\ &= O(p\lambda_{1n}^{-2})O(p^3\lambda_{1n}^{-4}) \\ &= O(p^4\lambda_{1n}^{-6})\end{aligned}$$

Thus, $\|\mathbb{E}(l_n \alpha_n^T)\| = O(p^2\lambda_{1n}^{-3})$. Also,

$$\begin{aligned}\|\mathbb{E}(l_n - \alpha_n)\|^2 &\leq \mathbb{E}(\|l_n\|^2) + \mathbb{E}(\|\alpha_n\|^2)\\ &= O(p\lambda_{1n}^{-2}) + O(p^3\lambda_{1n}^{-4})\\ &= O(p\lambda_{1n}^{-2}(1 + p^2\lambda_{1n}^{-2}))\end{aligned}$$

By Theorem 4.5.3, we have the order of magnitude $G_{1n}^{-1}(\sum_{i=1}^n b_{ni} + r_n) = O(p^{3/2}\lambda_{1n}^{-2})$. Hence, $G_{1n}^{-1}(\sum_{i=1}^n b_{ni} + r_n)\mathbb{E}(l_n^T - \alpha_n^T) = O(p^2\lambda_{1n}^{-3}\sqrt{1 + p^2\lambda_{1n}^{-2}})$, and the last term of (4.5.16) has magnitude $O(p^3\lambda_{1n}^{-4})$. Therefore, the slowest rate among the remainders is $O(p^2\lambda_{1n}^{-3})$ and the proof is completed. ■

4.6 Examples of biases of some full sample estimators

We will derive the asymptotic biases of full sample estimators of GLM with canonical link and noncanonical links using Theorem 4.5.3. The results are in agreement with that in Cordeiro and McCullagh (1991) section four.

Example 5 (GLM with canonical link) Consider the estimating equations for GLM with canonical link

$$\sum_{i=1}^{n}(y_i - h(\mathbf{x}_i^T\beta))\mathbf{x}_i = 0 \tag{4.6.1}$$

where h is the inverse link function, $\mathbb{E}(y_i) = \mu_i = h(\mathbf{x}_i^T\beta)$. Thus we have $\psi_{ni}(\beta) = (y_i - h(\mathbf{x}_i^T\beta))\mathbf{x}_i$ and $\dot{\psi}_{ni}(\beta) = -h'(\mathbf{x}_i^T\beta)\mathbf{x}_i\mathbf{x}_i^T$. Note that $\bar{\ddot{\Psi}}_n(\beta_0) = 0$ which makes the first term of the main term in (4.5.9) equals to zero. Let $\mathbf{W} = \text{Diag}(h'(\mathbf{x}_i^T\beta_0))$. The other variables are

$$\begin{aligned}a_{ni} &= -(\sum_{i=1}^{n} h'(\mathbf{x}_i^T\beta_0)\mathbf{x}_i\mathbf{x}_i^T)^{-1}(y_i - h(\mathbf{x}_i^T\beta_0))\mathbf{x}_i\\ &= -(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\varepsilon_i\mathbf{x}_i\end{aligned} \tag{4.6.2}$$

where $\dot{\Psi}_n^{-1} = -(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$, and the $\varepsilon_i = y_i - h(\mathbf{x}_i^T\beta_0)$ is the error. The second-order derivatives of the estimating functions is

$$\ddot{\Psi}_n(\beta) = -\begin{pmatrix} \sum_{j=1}^n h''(\mathbf{x}_j^T\beta_0)x_{j1}\mathbf{x}_j\mathbf{x}_j^T \\ \vdots \\ \sum_{j=1}^n h''(\mathbf{x}_j^T\beta_0)x_{jp}\mathbf{x}_j\mathbf{x}_j^T \end{pmatrix} \tag{4.6.3}$$

Write $\mathbf{Z} = \{z_{ij}\} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{1/2}$, and $\mathbf{F} = \text{Diag}(h''(\mathbf{x}_i^T\beta_0)/h'(\mathbf{x}_i^T\beta_0))$. Then the k-th componenet of the second main term in (4.5.9) is

$$\begin{aligned} C_{n,k} &= \sum_{i=1}^n \mathbb{E}(a_{ni}^T \circ \ddot{\Psi}_n(B_0) \circ a_{ni})_k \\ &= -\mathbb{E}(\sum_{j=1}^n h''(\mathbf{x}_j^T\beta_0)x_{jk}\sum_{i=1}^n \frac{z_{ij}^2}{h'(\mathbf{x}_i^T\beta_0)h'(\mathbf{x}_j^T\beta_0)}\varepsilon_i^2) \\ &= -\sum_{j=1}^n \frac{h''(\mathbf{x}_j^T\beta_0)}{h'(\mathbf{x}_j^T\beta_0)}x_{jk}\sum_{i=1}^n \frac{z_{ij}^2}{h'(\mathbf{x}_i^T\beta_0)}V(y_i) \\ &= -\sum_{j=1}^n \frac{h''(\mathbf{x}_j^T\beta_0)}{h'(\mathbf{x}_j^T\beta_0)}x_{jk}\sum_{i=1}^n z_{ij}^2 \end{aligned}$$

Thus, the second term can be expressed as

$$C_n = -\mathbf{X}^T\text{Diag}(\mathbf{Z}\mathbf{Z}^T)\mathbf{F}\mathbf{1}. \tag{4.6.4}$$

Hence, the bias of $\hat{\beta}_n$ is

$$\text{Bias}(\hat{\beta}_n) = -\frac{1}{2}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\text{Diag}(\mathbf{Z}\mathbf{Z}^T)\mathbf{F}\mathbf{1} + +O(\frac{p^{7/2}}{n^{3/2}}). \tag{4.6.5}$$

Example 6 (GLM with noncanonical link) Consider the estimating equations of GLM with noncanonical link

$$\sum_{i=1}^n \frac{y_i - h(\mathbf{x}_i^T\beta)}{V(y_i)}h'(\mathbf{x}_i^T\beta)\mathbf{x}_i = 0 \tag{4.6.6}$$

Hence, we have $\psi_{ni}(\beta_0) = s_i h'(\mathbf{x}_i^T\beta_0)\mathbf{x}_i$, where we denote s_i be the fraction $(y_i - h(\mathbf{x}_i^T\beta_0))/V(y_i)$. The first derivative is $\dot{\psi}_{ni} = (s_i'h_i' + s_ih_i'')\mathbf{x}_i\mathbf{x}_i^T = u_i\mathbf{x}_i\mathbf{x}_i^T$ with

$$u_i := s_i'h_i' + s_ih_i'' = \frac{-h'(\mathbf{x}_i^T\beta_0) + (y_i - h(\mathbf{x}_i^T\beta_0))h''(\mathbf{x}_i^T\beta_0)}{V(y_i)}$$

Note that this time the centered version $\bar{\dot{\Psi}}(\beta_0)$ may not be zero since there is y_i inside the term. Also,

$$
\begin{aligned}
\mathbb{E}(\dot{\Psi}_n) &= \sum_{i=1}^{n} \mathbb{E}(\dot{\psi}_{ni}(\beta_0)) \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\frac{-h'(\mathbf{x}_i^T\beta_0) + (y_i - h(\mathbf{x}_i^T\beta_0))h''(\mathbf{x}_i^T\beta_0)}{V(y_i)}\right]\mathbf{x}_i\mathbf{x}_i^T \\
&= -\sum_{i=1}^{n} \frac{h'(\mathbf{x}_i^T\beta_0)}{V(y_i)}\mathbf{x}_i\mathbf{x}_i^T \\
&= -\mathbf{X}^T\mathbf{W}\mathbf{X}
\end{aligned}
$$

where $W := \text{Diag}(h'(\mathbf{x}_i^T\beta_0)/V(y_i))$. The variable a_{ni} is approximated by

$$
a_{ni} = (\mathbb{E}(\dot{\Psi}_n))^{-1}\psi_{ni} = -(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}s_i h_i' \mathbf{x}_i
$$

and the centred version $\bar{\dot{\Psi}}_n$ is

$$
\begin{aligned}
\bar{\dot{\Psi}}_n = \dot{\Psi}_n - \mathbb{E}(\dot{\Psi}_n) &= \sum_{i=1}^{n} \frac{(y_i - h(\mathbf{x}_i^T\beta_0))h''(\mathbf{x}_i^T\beta_0)}{V(y_i)}\mathbf{x}_i\mathbf{x}_i^T \\
&= \mathbf{X}^T\mathbf{G}\mathbf{X}
\end{aligned}
$$

The first term of the main term of (4.5.9) is

$$
\begin{align*}
\mathbb{E}(\bar{\dot{\Psi}}_n \sum_{i=1}^{n} a_{ni}) &= -\mathbb{E}[(\mathbf{X}^T\mathbf{G}\mathbf{X})(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\sum_{j=1}^{n} s_j h_j' \mathbf{x}_j] \\
&= \mathbb{E}(\sum_{i=1}^{n} \frac{\varepsilon_i h_i''}{V_i}\mathbf{x}_i\mathbf{x}_i^T \sum_{j=1}^{n} \frac{\varepsilon_j h_j'}{V_j}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{x}_j) \\
&= \mathbb{E}(\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\varepsilon_i\varepsilon_j h_i''\sqrt{h_j'}}{\sqrt{V_iV_j}\sqrt{h_i'}} z_{ij}\mathbf{x}_i) \tag{4.6.7} \\
&= \mathbb{E}(\sum_{i=1}^{n} \frac{\varepsilon_i^2 h_i''}{V_i} z_{ii}\mathbf{x}_i + \sum_{i\neq j}\sum \frac{\varepsilon_i\varepsilon_j h_i''\sqrt{h_j'}}{\sqrt{V_iV_j}\sqrt{h_i'}} z_{ij}\mathbf{x}_i) \\
&= \sum_{i=1}^{n} h_i'' z_{ii}\mathbf{x}_i = \mathbf{X}^T\mathbf{Z}_d\mathbf{F}\mathbf{1}
\end{align*}
$$

where $\mathbf{Z} = \{z_{ij}\} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{1/2}$ and $\mathbf{F} = \text{Diag}(h''(\mathbf{x}_i^T\beta_0))$. The second-order derivatives of the estimating functions is

$$\ddot{\Psi}_n(\beta) = -\begin{pmatrix} \sum_{j=1}^n u'_j x_{j1}\mathbf{x}_j\mathbf{x}_j^T \\ \vdots \\ \sum_{j=1}^n u'_j x_{jp}\mathbf{x}_j\mathbf{x}_j^T \end{pmatrix} \tag{4.6.8}$$

The k-th component of the second term of the main term of the bias is

$$\begin{aligned} C_{n,k} &= \sum_{i=1}^n \mathbb{E}(a_{ni}^T \circ \ddot{\Psi}_n(B_0) \circ a_{ni})_k \\ &= -\mathbb{E}(\sum_{i=1}^n \frac{\varepsilon_i h'_i}{V_i}\mathbf{x}_i^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\sum_{j=1}^n u'_j x_{jk}\mathbf{x}_j\mathbf{x}_j^T)\frac{\varepsilon_i h'_i}{V_i}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{x}_i) \\ &= -\mathbb{E}(\sum_{j=1}^n \frac{u'_j V_j}{h'_j} x_{jk} \sum_{i=1}^n \frac{\varepsilon_i^2 h'_i}{V_i} z_{ij}^2) \end{aligned} \tag{4.6.9}$$

Note that $u'_j = \frac{1}{V_j}(\varepsilon_j h'''_j - 3h'_j h''_j)$. Substitute it to (4.6.9), we have

$$\begin{aligned} C_{n,k} &= -\sum_{j=1}^n \frac{E(\varepsilon_j^3)h'''_j}{V_j} z_{jj}^2 x_{jk} + 3\sum_{j=1}^n\sum_{i=1}^n h''_j h'_i z_{ij}^2 x_{jk} \\ &= -\sum_{j=1}^n \left[\frac{E(\varepsilon_j^3)h'''_j}{V_j} z_{jj}^2 - 3h''_j(\sum_{i=1}^n h'_i z_{ij}^2)\right] x_{jk} = -\sum_{j=1}^n q_j x_{jk} \end{aligned}$$

where $\mathbf{Q} = \text{Diag}\{q_j\}$ and $q_j = \frac{E(\varepsilon_j^3)h'''_j}{V_j} z_{jj}^2 - 3h''_j(\sum_{i=1}^n h'_i z_{ij}^2)$. Then the second term of the bias can be expressed as

$$C_n = -\mathbf{X}^T\mathbf{Q}\mathbf{1} \tag{4.6.10}$$

and the bias for $\hat{\beta}_n$ is

$$\text{Bias}(\hat{\beta}_n) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}_d\mathbf{F}\mathbf{1} - \frac{1}{2}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Q}\mathbf{1} + O(\frac{p^{7/2}}{n^{3/2}}) \tag{4.6.11}$$

Note that the GLM with noncanonical link has higher bias since the first term is non-zero, while the GLM with canonical link has the first term being zero. Also, for GLM with noncanonical link, the first term of bias given in (4.6.11) is the same as that in Cordeiro and McCullagh (1991) equation (4.2). In their paper, they apply the general expression for biases of the MLEs given by Cox and Snell (1968) and McCullagh (1987) to derive what they called the crucial quantity for the bias.

4.7 The A-optimal sampling distribution

In view of Theorem 4.2.1, we have

$$\mathrm{Var}^*(\hat{\beta}_r^*) = \frac{1}{r}\Sigma_n + o_p(1) = \frac{1}{r}\sum_{i=1}^n \frac{1}{\pi_i}\dot{\Psi}_n^{-1}\psi_{ni}\psi_{ni}^\top \dot{\Psi}_n^{-\top}|_{\hat{\beta}_n} + o_p(1). \tag{4.7.1}$$

Since Σ_n is a function of the sampling distribution $\pi = (\pi_1, \ldots, \pi_n)$ on the data points, we seek a sampling distribution which minimizes the trace of the matrix Σ_n. Following Peng and Tan (2018), we have

$$\tau(\pi) =: \mathrm{Tr}\big(\Sigma_n\big) = \sum_{i=1}^n \frac{\|a_{ni}\|^2}{\pi_i}, \quad \pi \in \mathscr{P}_n,$$

where $a_{ni} = \dot{\Psi}_n^{-1}\psi_{ni}|_{\hat{\beta}_n}$, and $\mathscr{P}_n$ is the probability simplex $\mathscr{P}_n = \{\pi : \pi_i \geq 0, \sum_i \pi_i = 1\}$ in $\mathbb{R}^n$. By using Lagrange multipliers, one can obtain the minimizer of the trace of Σ_n which is stated in the following theorem. From the perspective of design theory, this minimizer is referred to as *A-optimal* sampling distribution. Equivalently, an A-optimal distribution of the subsampling estimator $\hat{\beta}_r^*$ is the distribution which minimizes the trace of the main term of the conditional variance of $\hat{\beta}_r^*$. Let

$$\hat{H}_k = A_n(\dot{\Psi}_n^\top \dot{\Psi}_n)^{-k/2} A_n^\top|_{\hat{\beta}_n}, \quad k = 0, 1, 2. \tag{4.7.2}$$

where $A_n(\beta) = (\psi_{n1}(\beta), \ldots, \psi_{nn}(\beta))^\top$.

Theorem 4.7.1 *Suppose $\dot{\Psi}_n(\hat{\beta}_n)$ is invertible. Then the square roots of the diagonal entries of $\hat{H}_2$ gives an (asymptotically) A-optimal distribution $\hat{\pi}$ on the data points for $\hat{\beta}_r^*$ to approximate $\hat{\beta}_n$. Suppose, further, $\psi_{ni}(\hat{\beta}_n) \neq 0$ for $i = 1, \ldots, n$. Then $\hat{\pi}$ is unique.*

Specifically, the sampling probabilities are given by

$$\hat{\pi}_i \propto \|a_{ni}\| = (\psi_{ni}^\top(\dot{\Psi}_n^\top \dot{\Psi}_n)^{-1}\psi_{ni})^{1/2}|_{\hat{\beta}_n}, \quad i = 1, \ldots, n, \tag{4.7.3}$$

where $p_i \propto a_i$ denotes $p_i = a_i / \sum_{i=1}^n a_i$ for $a_i \geq 0, i = 1, \ldots, n$.

Following Peng and Tan (2018), $\hat{H}_2$ is referred to as the *A-optimal score matrix* with its diagonal entries inducing the unique *A-optimal* distribution, while $\hat{H}_1$ is the “hat” matrix for the GEE model in an obvious way to mimic the hat matrix $H_1 = X(X^\top X)^{-1}X^\top$ for the linear model. Clearly, we have $\hat{H}_0$ being the identity matrix of size $p \times p$.

In general, we can further generalize the sampling probabilities according to the $\hat{H}_k$ as follows

$$\hat{\pi}_i^{(k)} \propto \|a_{ni}^{(k)}\| = (\psi_{ni}^\top(\dot{\Psi}_n^\top\dot{\Psi}_n)^{-k/2}\psi_{ni})^{1/2}|_{\hat{\beta}_n}, \quad i = 1, \dots, n, \tag{4.7.4}$$

for $k = 0, 1, 2$. Then the *A-optimal* distribution is $\hat{\pi}_i = \hat{\pi}_i^{(2)}$.

Remark 4.7.1 $\pi_i^{(1)}$ and $\pi_i^{(0)}$ are another two cases of interest because both have computational ease. While $\hat{H}_2$ and $\hat{H}_1$ have the same running time as the full data estimator $\hat{\beta}_n$, $\hat{H}_0$ has the advantage of less computational burden as only the lengths $\|\psi_{ni}\|$ needed to be computed. In our simulations study, we will compare the effectiveness and of the sampling estimator based on these three probability distributions, and also the time for computing these probabilities.

4.8 The A-optimal sampling distribution by conditioning

Suppose $n^{-1}\dot{\Psi}_n(\hat{\beta}_n) = \dot{\Psi}_0 + o_p(1)$. Then by (4.7.1) we have the approximation,

$$n^{-2}\mathrm{Var}^*(\hat{\beta}_r^*) \approx \frac{1}{r}\sum_{i=1}^{n}\frac{1}{\pi_i}\dot{\Psi}_0^{-1}\hat{\psi}_{ni}^{\otimes 2}\dot{\Psi}_0^{-\top}. \tag{4.8.1}$$

where $\hat{\psi}_{ni} = \psi_{ni}(\hat{\beta}_n)$. Consider $Z = (Y, X)$, where Y is a response and X a covariate. Given $\{X_i\}$, the conditional expectation is given by

$$n^{-2}\mathrm{Var}(\hat{\beta}_r^*|\mathbf{X}) \approx \frac{1}{r}\sum_{i=1}^{n}\frac{1}{\pi_i}\dot{\Psi}_0^{-1}\mathbb{E}(\hat{\psi}_{ni}^{\otimes 2}|\{X_i\})\dot{\Psi}_0^{-\top}. \tag{4.8.2}$$

Similar to the A-optimal sampling distribution, we derive the conditional $\bar{A}$-optimal sampling distribution $\bar{\pi}$ given by the following theorem. Let

$$\bar{H}_k = \mathbb{E}(\hat{A}_n(\dot{\Psi}_0^\top\dot{\Psi}_0)^{-k/2}\hat{A}_n^\top|\{X_i\}), \quad k = 0, 1, 2. \tag{4.8.3}$$

where $\hat{A}_n = (\hat{\psi}_{n1}, \dots, \hat{\psi}_{nn})^\top$.

Theorem 4.8.1 *Suppose $\dot{\Psi}_0$ is invertible. Assume (R12) holds. Then the square roots of the diagonal entries of $\bar{H}_2$ gives an (asymptotically) $\bar{A}$-optimal distribution $\bar{\pi}$ on the data points for $\hat{\beta}_r^*$ to approximate $\hat{\beta}_n$.*

Specifically, the sampling probabilities are given by

$$\bar{\pi}_i \propto \left(\mathbb{E}(\hat{\psi}_{ni}^\top(\dot{\Psi}_0^\top\dot{\Psi}_0)^{-1}\hat{\psi}_{ni})|\{X_i\}\right)^{1/2}, \quad i = 1, \ldots, n. \tag{4.8.4}$$

Again, we generalize the sampling probabilities with respect to $\bar{H}_k$ as follows

$$\bar{\pi}_i^{(k)} \propto \left(\mathbb{E}(\hat{\psi}_{ni}^\top(\dot{\Psi}_0^\top\dot{\Psi}_0)^{-k/2}\hat{\psi}_{ni})|\{X_i\}\right)^{1/2}, \quad i = 1, \ldots, n. \tag{4.8.5}$$

for $k = 0, 1, 2$, where the $\bar{A}$-optimal sampling probability is $\bar{\pi}_i = \bar{\pi}_i^{(2)}$.

Remark 4.8.1 Similar to the previous A-optimal distribution, we have three versions of $\bar{A}$-optimal disbributions. The $\bar{\pi}_i^{(0)}$ and $\bar{\pi}_i^{(1)}$ have the advantage of computational ease. On the other hand, in order to have conditions (R3)–(R4), (R3')–(R4') and (R6) hold, the sampling probabilities π_i's must be bounded away from zero. This is not required for the other conditions.

The A-optimal Scoring Algorithm. The bottleneck for computing the sampling probabilities is to compute the inverse matrix $\dot{\Psi}_n^\top\dot{\Psi}_n|_{\hat{\beta}_n}$. In order to overcome this computational hurdle, we apply the *A-optimal Scoring Method* for the linear model proposed in Peng and Tan (2018) in our GEE framework. We select a uniform pre-subsample $Z_{nk,0}^*$ of size r_0 from the data set Z_{ni} and approximate $\hat{\beta}_n$ by $\hat{\beta}_{r_n,0}^*$ which is the solution to the equation,

$$\Psi_{n,0}^*(\beta) = \sum_{k=1}^{r_0} \psi_{nk}(Z_{nk,0}^*; \beta) = 0.$$

The A-optimal score matrix $\hat{H}_2$ is then approximated by

$$\hat{H}_{2,0}^* = A_n(\dot{\Psi}_{n,0}^{*\top}\dot{\Psi}_{n,0}^*)^{-1}A_n^\top|_{\hat{\beta}_{r_n,0}^*}.$$

That is, the A-optimal sampling probabilities $\hat{\pi}_i$ are approximated by $\hat{\pi}_i^*$, the normalized square roots of the diagonal entries of $\hat{H}_{2,0}^*$. Specifically,

$$\hat{\pi}_i^* \propto (\psi_{ni}^\top(\dot{\Psi}_{n,0}^{*\top}\dot{\Psi}_{n,0}^*)^{-1}\psi_{ni})^{1/2}|_{\hat{\beta}_{r_n,0}^*}, \quad i = 1, \ldots, n. \tag{4.8.6}$$

Now we use $\hat{\pi}^*$ to take a subsample Z^*_{nj} of size r from the remaining data $\{Z_{ni}\} \setminus \{Z^*_{nk,0}\}$, and compute the subsampling estimate $\hat{\beta}^*_r$ as the solution to the GEE (4.0.2) based on the subsample Z^*_{nj} and the corresponding sampling probabilities $\hat{\pi}^*_j$. This procedure shall also be referred to as *the A-optimal Scoring Method (for the GEE model).*

Remark 4.8.2 (1) Note that in the original formulation of the A-optimal distribution, the probabilities are dependent on the $\hat{\beta}_n$ which is also the target we want to approximate. Hence, it make more sense to approximate $\hat{\beta}_n$ by $\hat{\beta}^*_{r_n,0}$ in the calculation as in A-optimal Scoring Method. (2) The A-optimal Scoring Method gives an algorithm which has faster running time than the algorithm given by the full data GEE model since instead of computing the whole sample size n, we only need to focus on a much smaller size r_0+r. (3) Three visits of the dataset suffices to compute the subsampling estimate $\hat{\beta}^*_r$. (4) *Parallel computing* can be used to calculate the approximate sampling distribution $\hat{\pi}^*$ in (4.8.6). This shortens the time for getting the estimate $\hat{\beta}^*_r$. (5) $\hat{\pi}^*$ is sequentially updatable for stream data.

4.9 Asymptotic behaviors under A-optimal sampling for fixed p

Let $l_{ni}, i = 1, 2, \dots, n, n \geq 1$ be a double array of positive numbers. Like in Peng and Tan (2018), we truncate $\hat{\pi}$ from below by $l_n = (l_{ni}/n)$ as follows:

$$\hat{\pi}^{(l_n)}_{ni} \propto \hat{\pi}_{ni}\mathbf{1}[\hat{\pi}_{ni} \geq l_{ni}/n] + l_{ni}\mathbf{1}[\hat{\pi}_{ni} < l_{ni}/n], \quad i = 1, \dots, n. \tag{4.9.1}$$

Though, typically, we require $l_{ni} \geq l_0 > 0$ for some l_0, we shall investigate conditions to allow for $l_{ni} \to 0$ as n tends to infinity.

(R11) There is some constant $c_0 > 0$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \|\psi_{ni}(\beta_0)\| = c_0 + o_p(1).$$

(R12) There is a constant matrix $\dot{\Psi}_0$ with $\lambda_{\text{amin}}(\dot{\Psi}_0) > 0$ such that

$$\frac{1}{n}\dot{\Psi}_n = \frac{1}{n}\sum_{i=1}^{n} \dot{\psi}_{ni}(\beta_0) = \dot{\Psi}_0 + o_p(1).$$

(R13) There is a positive definite matrix A_0 such that

$$\delta_n \sum_{i=1}^{n} \frac{\psi_{ni}^{\otimes 2}}{\|\psi_{ni}\|} = A_0 + o_p(1).$$

(R31) There is a positive sequence of $l_n = (l_{ni} : i = 1, \ldots, n)$ such that

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\|\dot{\psi}_{ni}\|^2}{\|\psi_{ni}\|} \mathbf{1}[\|\psi_{ni}\| \geq l_{ni}] = o_p(r_n).$$

(R41) There exists a neighborhood $\mathbb{N}_0$ of β_0 such that $\ddot{\Psi}_{n,d}(\beta)$ is either positive or negative definite in $\mathbb{N}_0$ and that there is a rv $\eta_{ni,d}$

$$\sup_{\beta \in \mathbb{N}_0} \lambda_{\text{amax}}(\ddot{\Psi}_{n,d}(\beta)) \leq \eta_{ni,d}, \quad d = 1, \ldots, p,$$

where the random vector $\eta_{ni} = (\eta_{ni,1}, \ldots, \eta_{ni,p})^\top$ satisfies

$$\frac{1}{n} \sum_{i=1}^{n} \big(1 + \frac{\mathbf{1}[\|\psi_{ni}\| \geq l_{ni}]}{r_n \|\psi_{ni}\|}\big) \|\eta_{ni}\|^2 = o_p(r_n n \delta_n).$$

(R61) The double array $z_{nj}^{(l_n)*} = \Sigma_n^{-1/2} \dot{\Psi}_n^{-\top}(\hat{\beta}_n) \psi_{nj}^*(\hat{\beta}_n) / \hat{\pi}_{nj}^{(l_n)*}$, $j = 1, 2, \ldots, r$, $r \geq 1$ satisfies the Lindeberg condition: for every $t > 0$,

$$\sum_{i=1}^{n} \hat{\pi}_{ni}^{(l_n)} \|z_{n,i}^{(l_n)}\|^2 \mathbf{1}[\|z_{ni}\| \geq \sqrt{r} t] = o_p(1), \quad \textit{as} \quad r \to \infty.$$

Theorem 4.9.1 *Suppose (R11)-(R13), (R2), (R31)-(R41) and (R4’) hold. Assume $\hat{\beta}_n$ is a solution of (4.0.1) such that $\hat{\beta}_n = \beta_0 + o_p(1)$. Then these exists a sequence of solutions $\hat{\beta}_r^*$ of (4.2.18) such that*

$$\dot{\Psi}_n(\hat{\beta}_n) \sqrt{r_n} (\hat{\beta}_{r_n}^* - \hat{\beta}_n) = -\frac{1}{\sqrt{r_n}} \sum_{j=1}^{r_n} \frac{\psi_{nj}^*(\hat{\beta}_n)}{\pi_j^*} + o_p(\hat{\lambda}_n). \tag{4.9.2}$$

If, further, (R61) hold for the truncated sampling distribution in (4.9.1), then

$$V_n^{-1/2} \sqrt{r_n} (\hat{\beta}_r^* - \hat{\beta}_n) \Rightarrow \mathscr{N}(0, 1), \quad \textit{in probability}, \quad r_n \to \infty. \tag{4.9.3}$$

*where V_n equals Σ_n in (**??**) under the truncated sampling distribution (4.9.1).*

Proof of Theorem 4.9.1. We shall verify the conditions of Theorem 4.2.1 for the case of fixed dimenion $p_n = p$. In this case, (R31)-(R41) and (R61) imply (R3)-(R4) and (R6), respectively. Let $\hat{\psi}_{ni} = \psi_{ni}(\hat{\beta}_n)$. By (R2), (R12), (R41) and (R4'),

$$\frac{1}{n}\sum_{i=1}^{n}(\|\hat{\psi}_{ni}\| - \|\psi_{ni}\|) = o_p(1). \tag{4.9.4}$$

This and (R11) yield

$$\frac{1}{n}\sum_{i=1}^{n}\|\hat{\psi}_{ni}\| = c_0 + o_p(1). \tag{4.9.5}$$

By (R4') again,

$$\frac{1}{n}\dot{\Psi}_n(\hat{\beta}_n) - \frac{1}{n}\dot{\Psi}_n = \frac{1}{n}\sum_{i=1}^{n}\left(\dot{\psi}_{ni}(\hat{\beta}_n) - \dot{\psi}_{ni}\right) = o_p(1). \tag{4.9.6}$$

This and (R12) give

$$\frac{1}{n^2}\left(\dot{\Psi}_n^\top \dot{\Psi}_n|_{\hat{\beta}_n}\right)^{-1} = \left(\dot{\Psi}_0^\top \dot{\Psi}_0\right)^{-1} + o_p(1). \tag{4.9.7}$$

Thus there exist constants $0 < b_0 \le B_0 < \infty$ such that

$$b_0\|\hat{\psi}_{ni}\|/n \le \hat{\pi}_i \le B_0\|\hat{\psi}_{ni}\|/n, \quad i = 1, \ldots, n. \tag{4.9.8}$$

Let us write $J_n(\beta) = J_n(\beta, \pi)$ and $\hat{J}_n = J_n(\hat{\beta}_n, \hat{\pi})$. Then by (R13) and (4.9.7),

$$\delta_n \hat{J}_n = \sum_{i=1}^{n}\|\hat{\psi}_{ni}\|\delta_n\sum_{i=1}^{n}\frac{\hat{\psi}_{ni}^{\otimes 2}}{\|\hat{\psi}_{ni}\|} = n(A_0c_0 + o_p(1)). \tag{4.9.9}$$

Thus $\delta_n\hat{\lambda}_n^2 = \delta_n\lambda_{\max}(\hat{J}_n) = c_1(n + o_p(1))$ for some constant $c_1 > 0$. Consequently, (R5) holds; (R12) and (4.9.6) imply (R1); (4.9.5) yields (4.2.1). We now apply Theorem 4.2.1 to finish the proof. ∎

5. SIMULATION STUDY

In this chapter, we present some simulation results about A-optimal subsampling approach on generalized count regression. Here, we focus on Poisson and Negative Binomial distributions. We choose the true coefficient $\boldsymbol{\beta} = (0.1, -0.1\times\mathbf{1}_{25}^\top, 0.1\times\mathbf{1}_{25}^\top)^\top$, and generate p $= 51$-dimensional covariate vector $(1, \mathbf{X})$ where $\mathbf{X}$ is from one of the four multivariate distributions:

1. Gaussian (GA) $N(0, \Sigma)$,
2. Mixture Gaussian (MG) $\frac{1}{2}N(0,\Sigma) + \frac{1}{2}N(0, 3\Sigma)$,
3. Log-normal (LN) $LN(0, \frac{1}{2}\Sigma)$,
4. T-distribution (T) with degree of freedom equals 5 $T_5(0, \frac{1}{2}\Sigma)$

all with the same 50×50 covariance matrix Σ with the (i,j) entry equal to $\Sigma_{i,j} = 0.3^{|i-j|}$. For sample size $n = 10^6$, we consider the response y_i follows Poisson distribution with log link $\log(\mu_i) = \mathbf{x}_i^\top \beta$, $i = 1, \ldots, n$, or Negative Binomial distribution with variance $V(y_i) = \mu_i + 5\mu_i^2$ and log link $\log(\mu_i) = \mathbf{x}_i^\top\beta$, $i = 1, \ldots, n$. The data sets are then fitted to either a Poisson regression model or negative binomial regression model. To compare the efficiency of A-optimal subsampling and the uniform subsampling methods, we fit a subsample of size r to the model where the A-optimal subsampling probabilities are computed from the formulas:

$$\hat{\pi}_i^{(k)} = \frac{\|(\mathbf{X}^\top\mathbf{W}(\tilde{\boldsymbol{\beta}})\mathbf{X})^{-k/2}\mathbf{x}_i\||\hat{e}_i|}{\sum_{i=1}^n \|(\mathbf{X}^\top\mathbf{W}(\tilde{\boldsymbol{\beta}})\mathbf{X})^{-1}\mathbf{x}_i\||\hat{e}_i|}, \quad k = 0, 1, 2. \tag{5.0.1}$$

$$\bar{\pi}_i^{(k)} = \frac{\|(\mathbf{X}^\top\mathbf{W}(\tilde{\boldsymbol{\beta}})\mathbf{X})^{-k/2}\mathbf{x}_i\|\hat{g}_i}{\sum_{i=1}^n \|(\mathbf{X}^\top\mathbf{W}(\tilde{\boldsymbol{\beta}})\mathbf{X})^{-1}\mathbf{x}_i\|\hat{g}_i}, \quad k = 0, 1, 2. \tag{5.0.2}$$

where

$$\begin{aligned} W(\tilde{\boldsymbol{\beta}}) &= \mathrm{Diag}(\hat{\mu}_i), \quad \hat{\mu}_i = \exp(\mathbf{x}_i\tilde{\boldsymbol{\beta}}), \\ \hat{e}_i &= y_i - \hat{\mu}_i, \quad \hat{g}_i = \sqrt{\hat{\mu}_i} \end{aligned} \tag{5.0.3}$$

for Poisson distribution, and

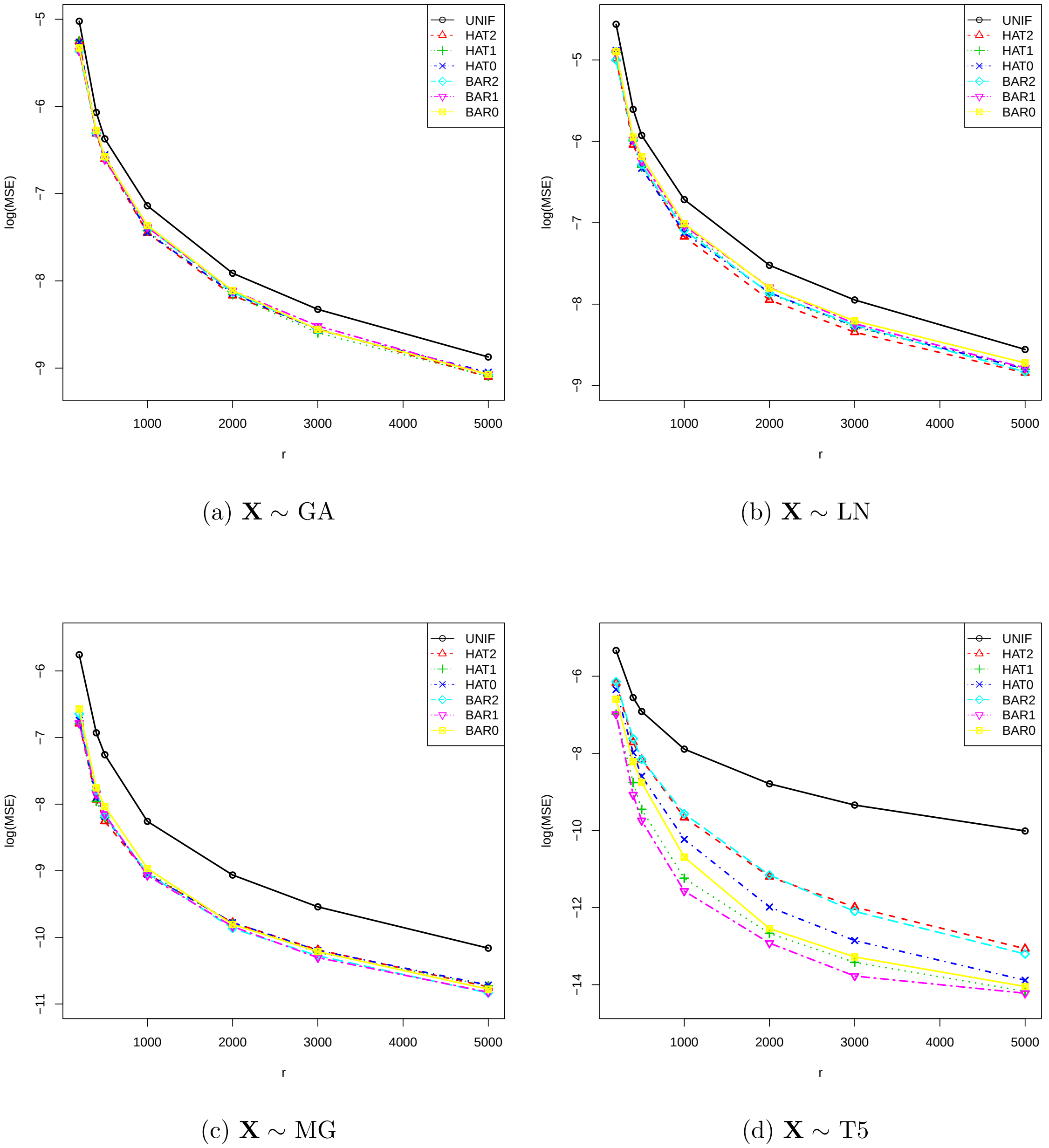
$$
\begin{aligned}
W(\tilde{\boldsymbol{\beta}}) &= \mathrm{Diag}(\frac{\hat{\mu}_i}{1+\alpha\hat{\mu}_i}), \quad \hat{\mu}_i = \exp(\mathbf{x}_i\tilde{\boldsymbol{\beta}}), \\
\hat{e}_i &= \frac{y_i - \hat{\mu}_i}{1+\alpha\hat{\mu}_i}, \quad \hat{g}_i = \sqrt{\frac{\hat{\mu}_i}{1+\alpha\hat{\mu}_i}}
\end{aligned}
\tag{5.0.4}
$$

for Negative Binomial distribution. Here, we use A-optimal scoring method to select a pre-subsample of size $r_0 = 500$ on the full data and fit it to the corresponding model to obtain the estimate $\tilde{\boldsymbol{\beta}}$ which is then used to compute the A-optimal subsampling probabilities. For each subsample size r, we perform the simulations $M = 1000$ times and calculate the empirical mean squared errors (MSE) of the subsampling estimator $\hat{\boldsymbol{\beta}}_r^*$ given by the formula

$$
\mathrm{MSE}(\hat{\boldsymbol{\beta}}_r^*) = \frac{1}{M}\sum_{m=1}^{M} \|\hat{\boldsymbol{\beta}}_{r,m}^* - \hat{\boldsymbol{\beta}}\|^2,
$$

for three subsampling distributions $\boldsymbol{\pi}^{(0)}, \boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}$ with regard to $\hat{A}$- and $\bar{A}$- optimality, and $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ using all the data.

Figure 5.1 presents plots of $\log_{10}$ of the MSEs of $\hat{\boldsymbol{\beta}}_r^*$ using $\hat{\boldsymbol{\pi}}^{(2)}$ (HAT2), $\hat{\boldsymbol{\pi}}^{(1)}$ (HAT1), $\hat{\boldsymbol{\pi}}^{(0)}$ (HAT0), $\bar{\boldsymbol{\pi}}^{(2)}$ (BAR2), $\bar{\boldsymbol{\pi}}^{(1)}$ (BAR1), $\bar{\boldsymbol{\pi}}^{(0)}$ (BAR0) and uniform subsampling (UNIF) probabilities against different subsample sizes $r = 0.02\%n$, $0.04\%n$, $0.05\%n$, $0.1\%n$, $0.2\%n$, $0.3\%n$, $0.5\%n$. To obtain better graphical presentation, we apply logarithm with base 10 on MSEs. The response y follows a Poisson distribution and the data set is fitted to a Poisson regression model. Clearly, all the subsampling methods improve as the subsample size r increases, and all the A-optimal subsampling methods give smaller MSEs than the uniform subsampling method. For example, in $\mathbf{X} \sim$ T5 subplot, A-optimal $\hat{\boldsymbol{\pi}}^{(2)}$ gives a smaller MSE for $r = 2000$ than uniform subsampling for $r = 5000$. Note that A-optimal subsampling $\hat{\boldsymbol{\pi}}^{(2)}$ gives the smallest MSE among all methods in all cases except $\mathbf{X} \sim$ T5. For $\mathbf{X} \sim$ T5, the $\bar{\boldsymbol{\pi}}^{(0)}$ gives the smallest MSE. This seems contradicting to our theory that MSE is dominated by the variance and $\hat{\boldsymbol{\pi}}^{(2)}$ should minimize the variance hence the MSE. We will discuss this abnormality in the following section.



(a) $\mathbf{X} \sim$ GA

(b) $\mathbf{X} \sim$ LN

(c) $\mathbf{X} \sim$ MG

(d) $\mathbf{X} \sim$ T5

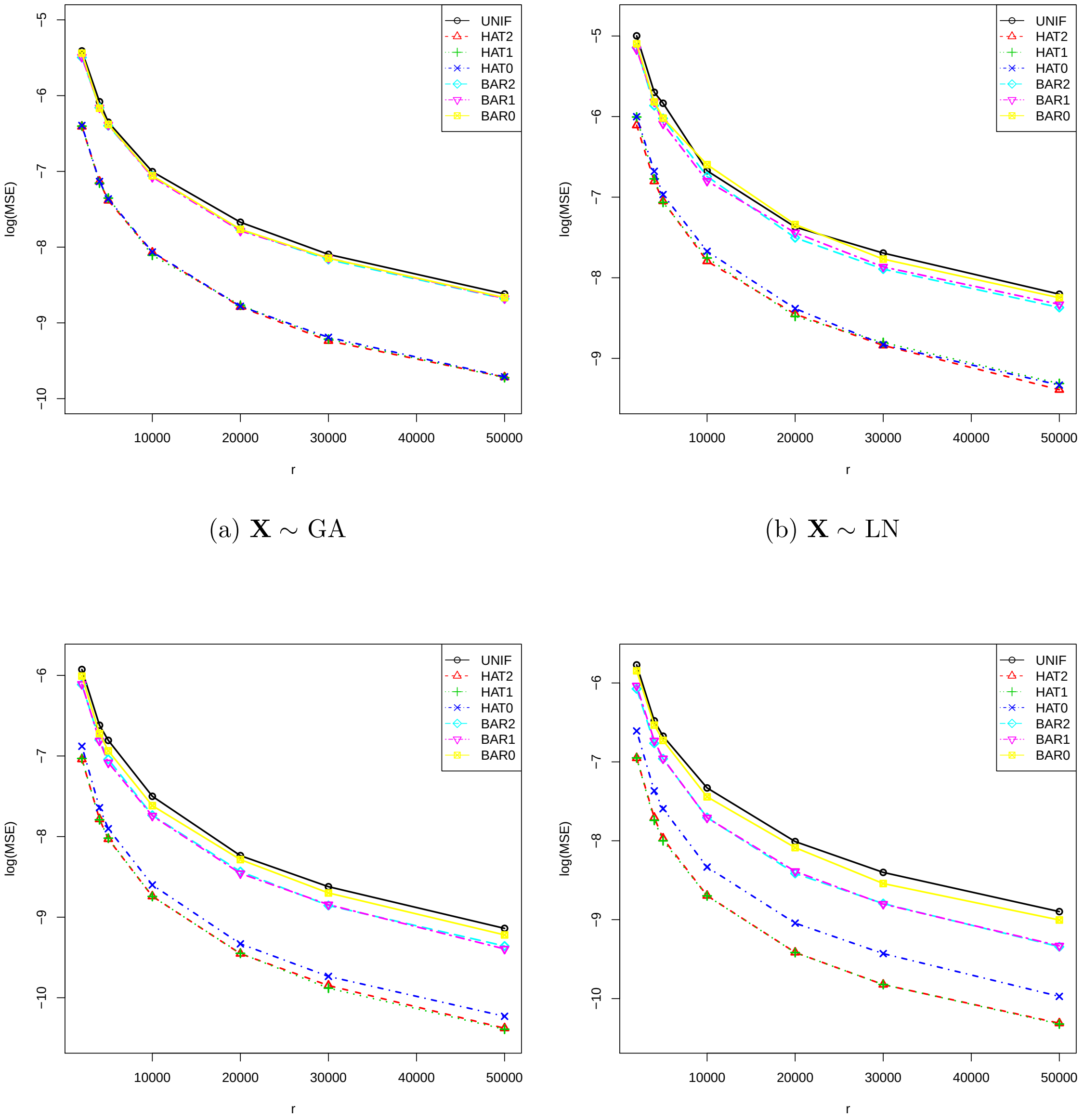
Figure 5.1. MSEs of subsampling estimator for four different distributions of covariate $\mathbf{X}$, with Poisson data and Poisson model. The full data size is $n = 10^6$ and the subsample size r varies.

Figure 5.2 presents plots of $\log_{10}$ of the MSEs against different subsample sizes $r = 0.2\%n, 0.4\%n, 0.5\%n, 1\%n, 2\%n, 3\%n, 5\%n$. The response y follows a negative binomial distribution and the data set is fitted to a negative binomial regression model. We compare all the A-optimal subsampling methods $\hat{\boldsymbol{\pi}}^{(2)}$ (HAT2), $\hat{\boldsymbol{\pi}}^{(1)}$ (HAT1), $\hat{\boldsymbol{\pi}}^{(0)}$ (HAT0), $\bar{\boldsymbol{\pi}}^{(2)}$ (BAR2), $\bar{\boldsymbol{\pi}}^{(1)}$ (BAR1) and $\bar{\boldsymbol{\pi}}^{(0)}$ (BAR0) with uniform subsampling (UNIF). It is clear that the MSEs decrease for all the subsampling methods as the subsample size r increases. The A-optimal subsampling methods within each group, $\hat{A}$- and $\bar{A}$- optimality, have similar MSEs. In particular, the $\hat{A}$ subsampling group performs better than the $\bar{A}$ subsampling group and the uniform subsampling, with $\hat{\boldsymbol{\pi}}^{(2)}$ gives the smallest MSEs for all cases of $\mathbf{X}$.

In Figure 5.3, we fit the data with response y follows a negative binomial distribution to a Poisson regression model, and compare the $\log_{10}$ of the MSEs against subsample size $r = 0.02\%n$, $0.04\%n$, $0.05\%n$, $0.1\%n$, $0.2\%n$, $0.3\%n$, $0.5\%n$ with different subsampling methods. In all the cases, the A-optimal subsampling methods outperform the uniform subsampling. Again, we can see that $\hat{A}$ subsampling methods have similar MSEs, and the same for $\bar{A}$ subsampling methods. For the case of $\mathbf{X} \sim$ T5, the results for $\hat{A}$ group subsampling are significantly better than uniform subsampling. However, $\hat{\boldsymbol{\pi}}^{(2)}$ gives the smallest MSEs for all cases of $\mathbf{X}$ except T5. This is similar to the case in Figure 5.1. We will address this issue in later discussion.

In Figure 5.4, we fit the data with response y follows a Poisson distribution to a negative binomial regression model, and compare the $\log_{10}$ of the MSEs against subsample size $r = 0.2\%n$, $0.4\%n$, $0.5\%n$, $1\%n$, $2\%n$, $3\%n$, $5\%n$ with different subsampling methods. All the A-optimal subsampling methods outperform uniform in all the cases. In particular, $\hat{\boldsymbol{\pi}}^{(2)}$ gives the smallest MSE for all cases.

Consider the case when the Poisson data y is fitted to the Poisson model. We report the 95% coverage probabilities of the first component of $\hat{\boldsymbol{\beta}}_r^*$ with uniform, A-optimal HAT2 and BAR2 subsampling methods in Figure 5.5. The coverage probabilities under these three subsampling methods converge to the the nominal 95% for different cases of $\mathbf{X}$ when r increases.



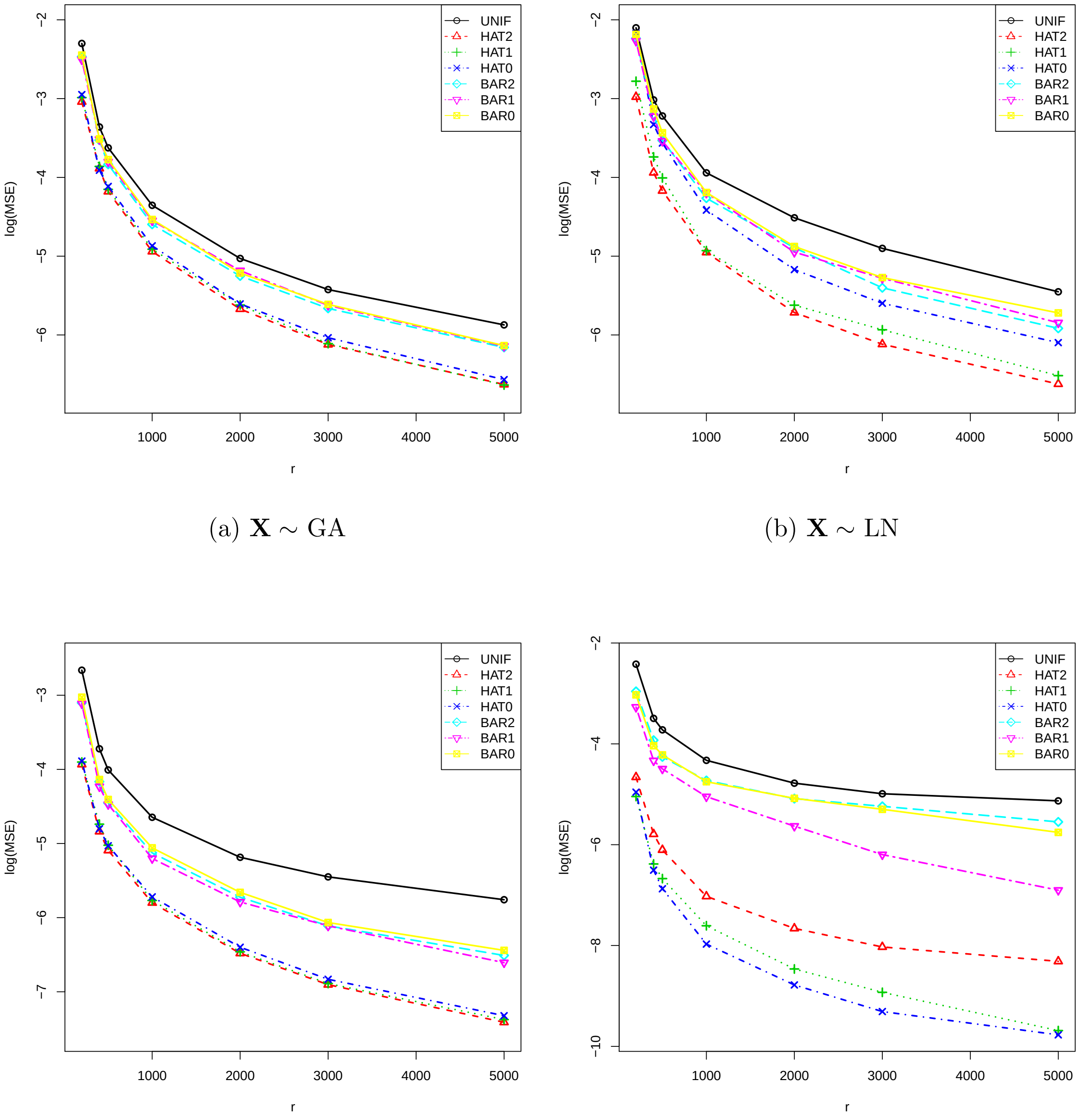
(a) **X** ∼ GA

(b) **X** ∼ LN

(c) **X** ∼ MG

(d) **X** ∼ T5

Figure 5.2. MSEs of subsampling estimator for four different distributions of covariate **X**, with negative binomial data and negative binomial model. The full data size is $n = 10^6$ and the subsample size r varies.



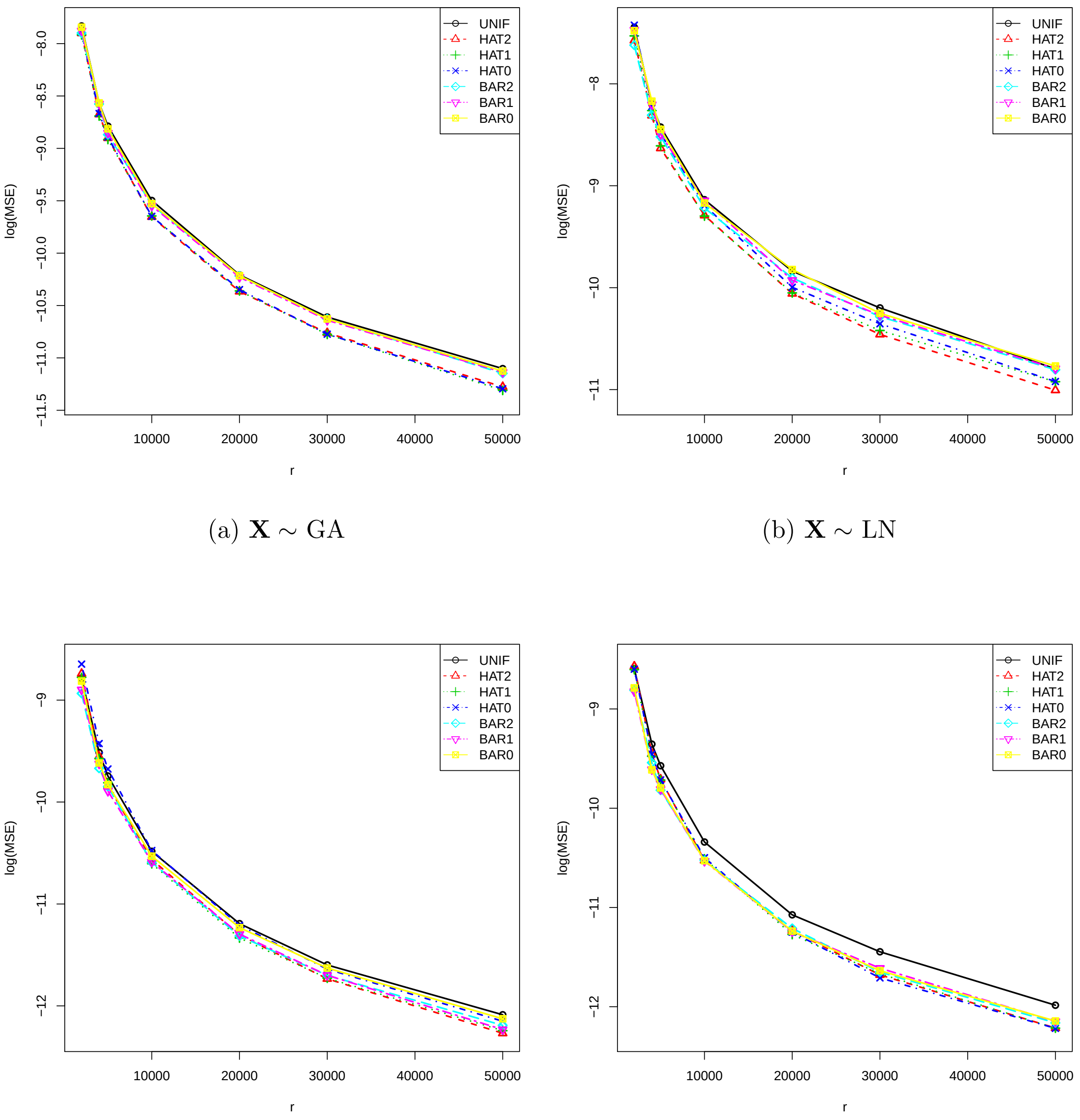
(a) $\mathbf{X} \sim$ GA

(b) $\mathbf{X} \sim$ LN

(c) $\mathbf{X} \sim$ MG

(d) $\mathbf{X} \sim$ T5

Figure 5.3. MSEs of subsampling estimator for four different distributions of covariate $\mathbf{X}$, with negative binomial data and Poisson model. The full data size is $n = 10^6$ and the subsample size r varies.



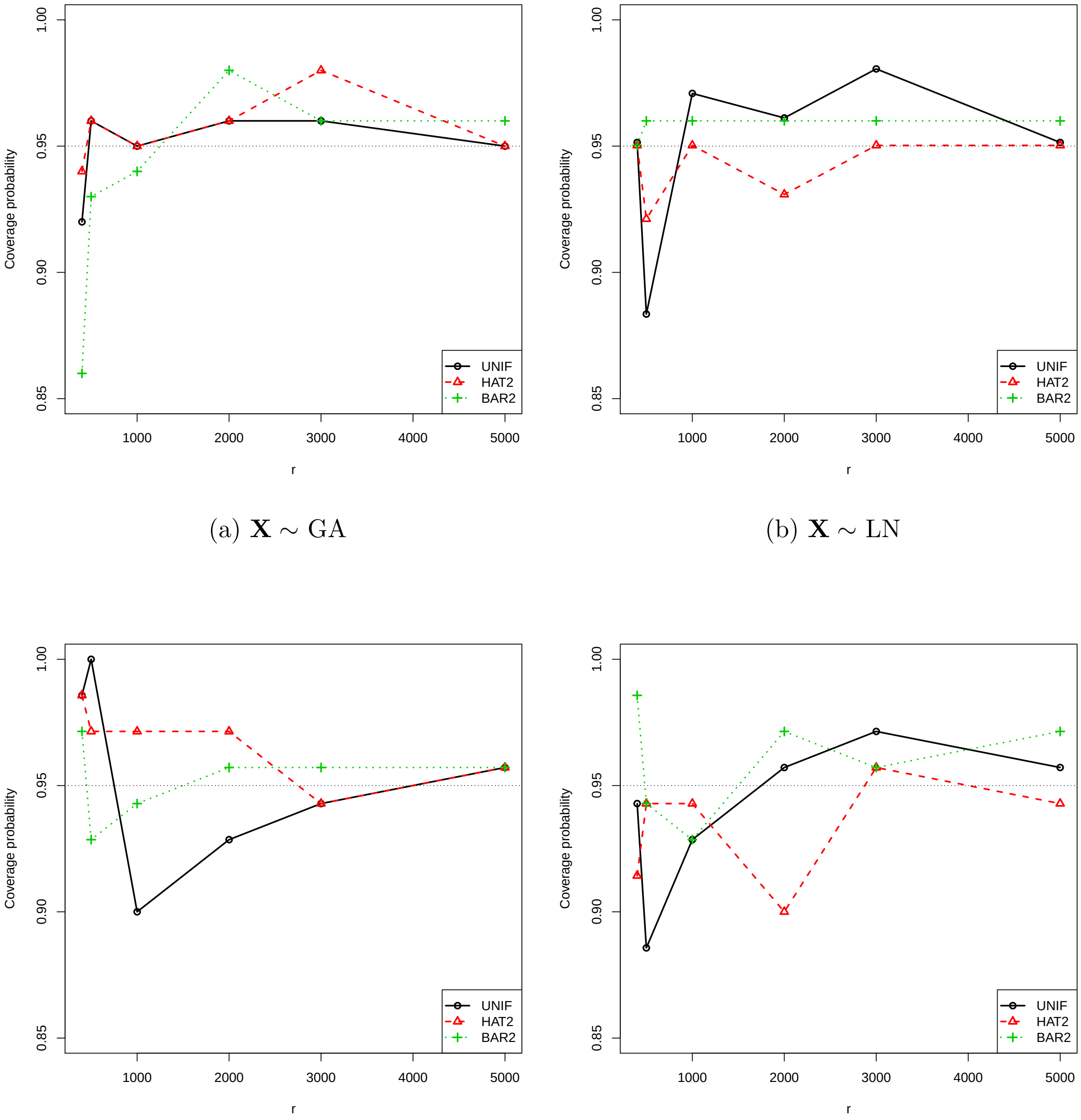
(a) $\mathbf{X} \sim$ GA

(b) $\mathbf{X} \sim$ LN

(c) $\mathbf{X} \sim$ MG

(d) $\mathbf{X} \sim$ T5

Figure 5.4. MSEs of subsampling estimator for four different distributions of covariate $\mathbf{X}$, with Poisson data and negative binomial model. The full data size is $n = 10^6$ and the subsample size r varies.



(a) $\mathbf{X} \sim$ GA

(b) $\mathbf{X} \sim$ LN

(c) $\mathbf{X} \sim$ MG

(d) $\mathbf{X} \sim$ T5

Figure 5.5. 95% Coverage probabilities of the first component of subsampling estimator for four different distributions of covariate $\mathbf{X}$, with Poisson data and Poisson model. The full data size is $n = 10^6$ and the subsample size r varies.

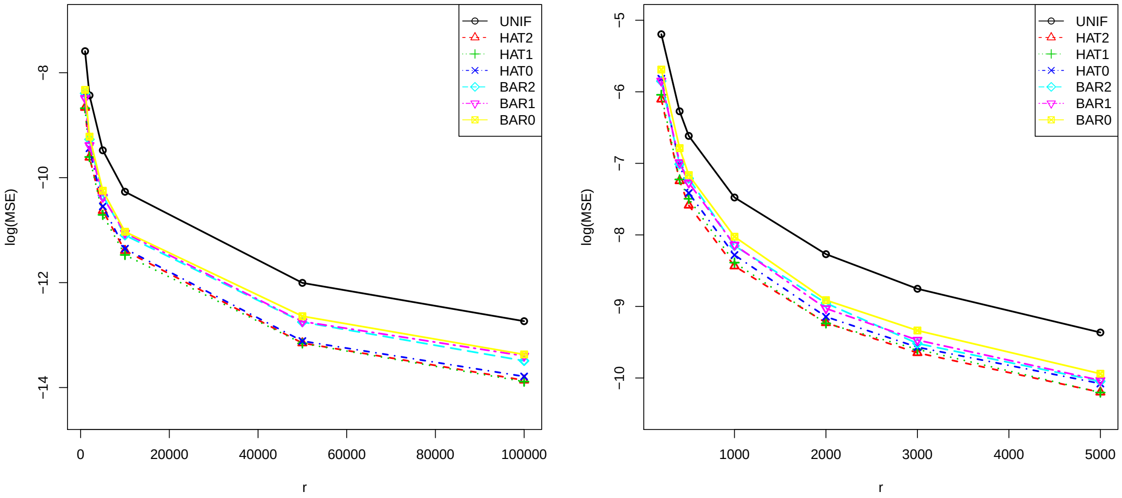
In order to evaluate the computational efficiency of A-optimal subsampling methods, we report the average running time, which includes both the pre-subsampling and subsampling time, for computing $\hat{\beta}_r^*$ using $\hat{\boldsymbol{\pi}}^{(2)}$ and $\hat{\boldsymbol{\pi}}^{(0)}$, and compare it with uniform subsampling. The time for computing MLE $\hat{\boldsymbol{\beta}}$ is also included in the table as FULL for reference. Table 5.1 shows the average CPU time (in seconds) for computing $\hat{\beta}_r^*$ for different full data sizes n and fixed covariate size $p = 200$ and subsample size $r = 10^4$, with repetition of 1000 times. We perform the simulations in R programming language on a laptop with Intel Core i7 processor and 16GB memory. From the table, it is clear that A-optimal subsampling methods use significantly less time than full data approach. In particular, A-optimal subsampling $\hat{\boldsymbol{\pi}}^{(0)}$ compares favorably to $\hat{\boldsymbol{\pi}}^{(2)}$.

Table 5.1.
The average CPU times in seconds for computing $\hat{\beta}_r^*$ for different full data sizes n, and fixed $p = 200$ and $r = 10^4$. The covariate $\mathbf{X}$ is GA and y is Poisson distribution and the data is fitted to a Poisson regression model.

n	HAT2	HAT0	UNIF	FULL
5×10^4	7.95	2.92	1.95	11.01
1×10^5	13.78	3.51	2.22	25.87
5×10^5	55.02	9.52	1.96	116.74
1×10^6	119.90	21.28	1.96	252.78

5.1 Abnormality of A-optimal subsampling on T5 distribution

In Figure 5.1 (d) and Figure 5.3 (d) where $\mathbf{X} \sim$ T5, we can see that although A-optimal subsampling $\hat{\boldsymbol{\pi}}^{(2)}$ gives a smaller MSE than uniform sampling, it is not the smallest among all different variant of A-optimal sampling methods. This contradicts our belief that A-optimal $\hat{\boldsymbol{\pi}}^{(2)}$ should provides the smallest MSE among all subsampling methods as supported in other cases where $\mathbf{X} \sim$ GA, LN and MG. One plausible explanation is that the $\hat{\boldsymbol{\pi}}^{(2)}$ does not minimize the MSE because the squared bias term from the decomposition of MSE is still significant even when the subsample size r is large. For a high dimensional case, the remainder term of the bias as shown in Theorem 4.2.2 is of order $o(\sqrt{p/rn})$ which is not negligible when p is large. In order to investigate the effect of the dimension p and the sample sizes n on the A-optimal subsampling, we consider combinations of sample sizes n: Conventional (100k) and Massive (1M); and dimension p: low (10) and high (50). In the previous cases, we only focus on massive sample size and high dimension case. We want to see if there is any combination of n and p which could makes the bias term negligible so that minimizing the variance will be similar to minimizing the MSE.

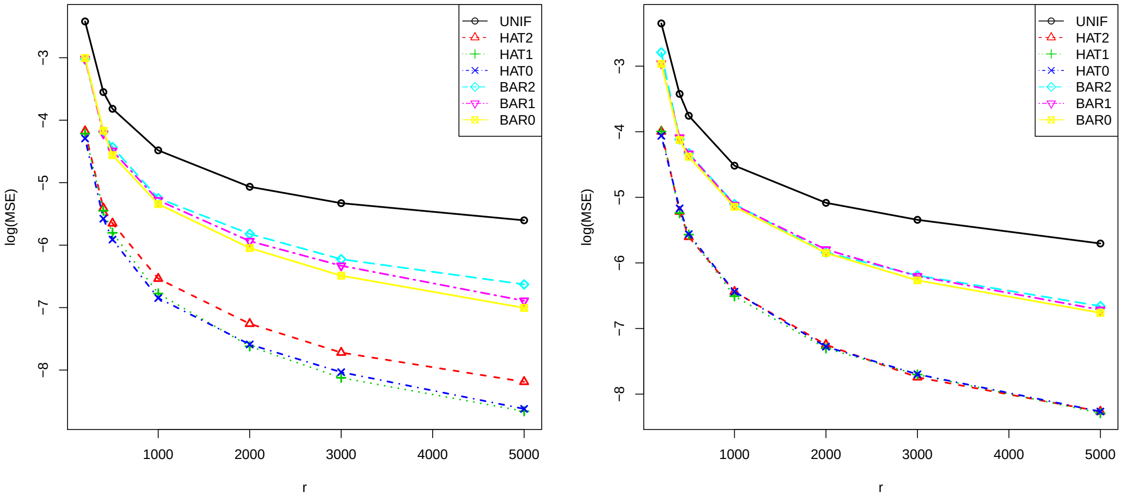


(a) $\mathbf{X} \sim$ T8

(b) $\mathbf{X} \sim$ T10

Figure 5.6. MSEs of subsampling estimator for covariate $\mathbf{X}$ being T8 and T10 distributions, with Poisson data and Poisson model. The full data size is $n = 10^6$ and the subsample size r varies.

On the other hand, we also want to examine if the abnormality is due to the lack of finite moments of higher order in T5 distribution. Note that T5 has up to fourth moment finite. We consider the cases when the degree of freedom of the t-distribution increases to 8 and 10. In Figure 5.6, we show the MSE plots of Poisson data fitted to a Poisson regression model with the same parameters as before (which is our case in Figure 5.1) except the covariate **X** changed to T8 and T10. In both situations, A-optimal subsampling $\hat{\boldsymbol{\pi}}^{(2)}$ gives the smallest MSE among all sampling methods. In Figure 5.7, we show the MSE plots of negative binomial data fitted to a Poisson regression model with the same parameters as before (which is our case in Figure 5.3) except covariate **X** changed to T8 and T10. For T8, $\hat{\boldsymbol{\pi}}^{(2)}$ does not give the smallest MSE, but it does in T10. This suggests that the conditions for A-optimal sampling may require the distribution of the covariates to have finite high-order moments.



(a) $\mathbf{X} \sim$ T8

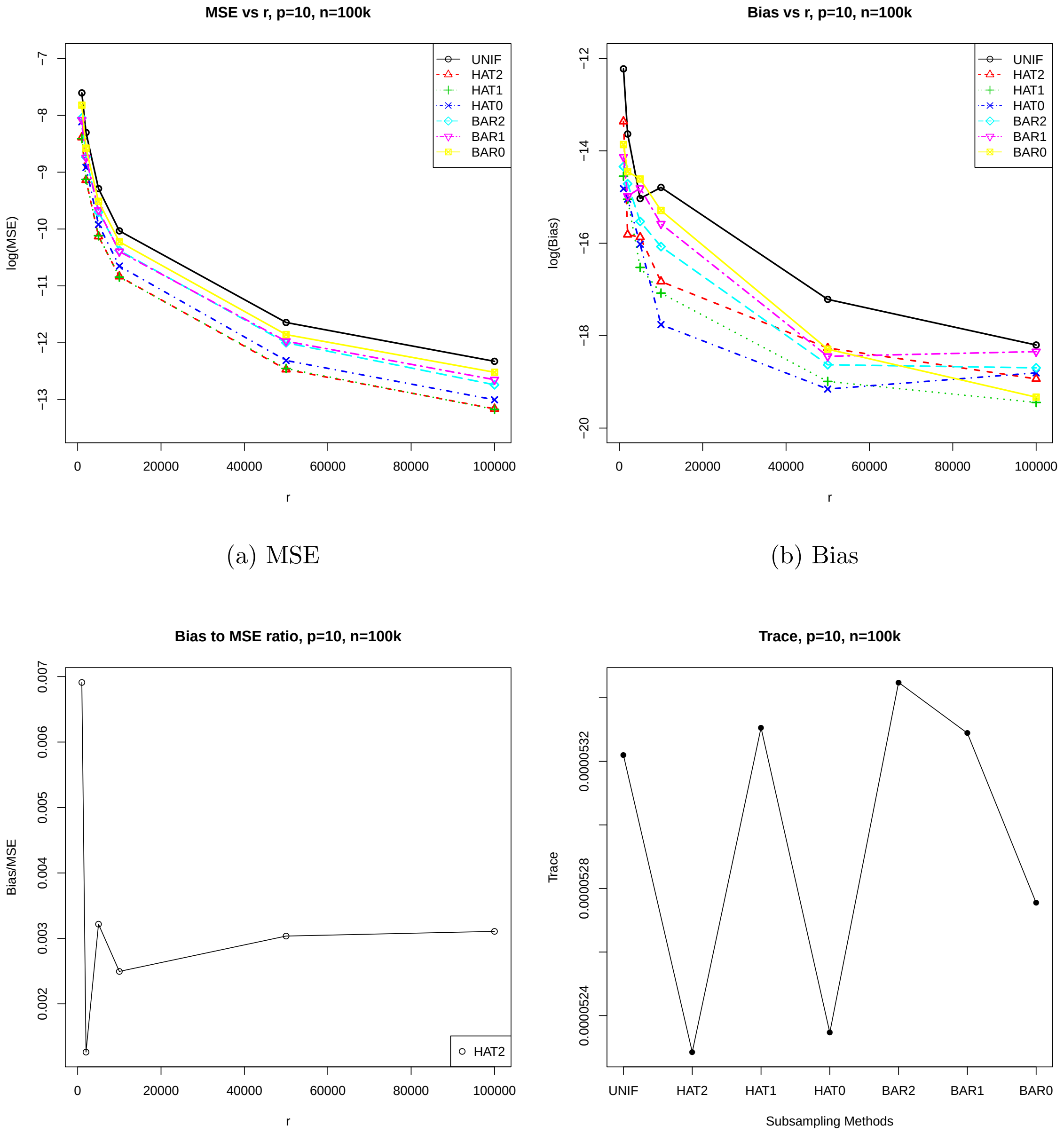
(b) $\mathbf{X} \sim$ T10

Figure 5.7. MSEs of subsampling estimator for covariate **X** being T8 and T10 distributions, with negative binomial data and Poisson model. The full data size is $n = 10^6$ and the subsample size r varies.

Next, we consider different dimensions and sizes of the data with Poisson distribution y and covariate $\mathbf{X} \sim$ T5 fitted to a Poisson regression model. The results are summarized as follows.

1. (**Conventional data size, conventional dimension**) Figure 5.8 shows the log(MSE), the amount of log(bias) , the bias to MSE ratio of $\hat{\boldsymbol{\pi}}^{(2)}$ and the trace of the variance covariance matrix of $\hat{\boldsymbol{\beta}}_r^*$ when the dimension is low $p = 10$ and the data size is conventional $n = 100$k. The A-optimal subsampling $\hat{\boldsymbol{\pi}}^{(2)}$ gives the smallest MSE among all methods, and its bias is comparable with other methods. Note that the bias to MSE ratio is around 0.003 for different r's which implies the bias is negligible and MSE is dominated by variance. Hence, minimizing the variance implies minimizing the MSE and thus $\hat{\boldsymbol{\pi}}^{(2)}$ gives the smallest MSE. The trace of HAT2 method is the smallest among all methods as expected.

2. (**Conventional data size, high dimension**) Figure 5.9 shows the case when the dimension is high $p = 50$ and the data size is conventional $n = 100$k. Note that the MSE corresponds to A-optimal subsampling $\hat{\boldsymbol{\pi}}^{(2)}$ is not the minimum and the bias is significantly higher than the other A-optimal subsampling methods. The percentage of bias to MSE of HAT2 is increasing from around 1% to 6% when r increases. This agrees with Theorem 4.2.2 that the remainder is not negligible when the dimension p is high. Since the bias is now significant, minimizing variance by using HAT2 sampling does not imply minimizing MSE. This explains why the MSE plot of $\hat{\boldsymbol{\pi}}^{(2)}$ is higher than some of the others. Although the trace corresponds to $\hat{\boldsymbol{\pi}}^{(2)}$ is not the minimum, the difference between it and others are insignificant.



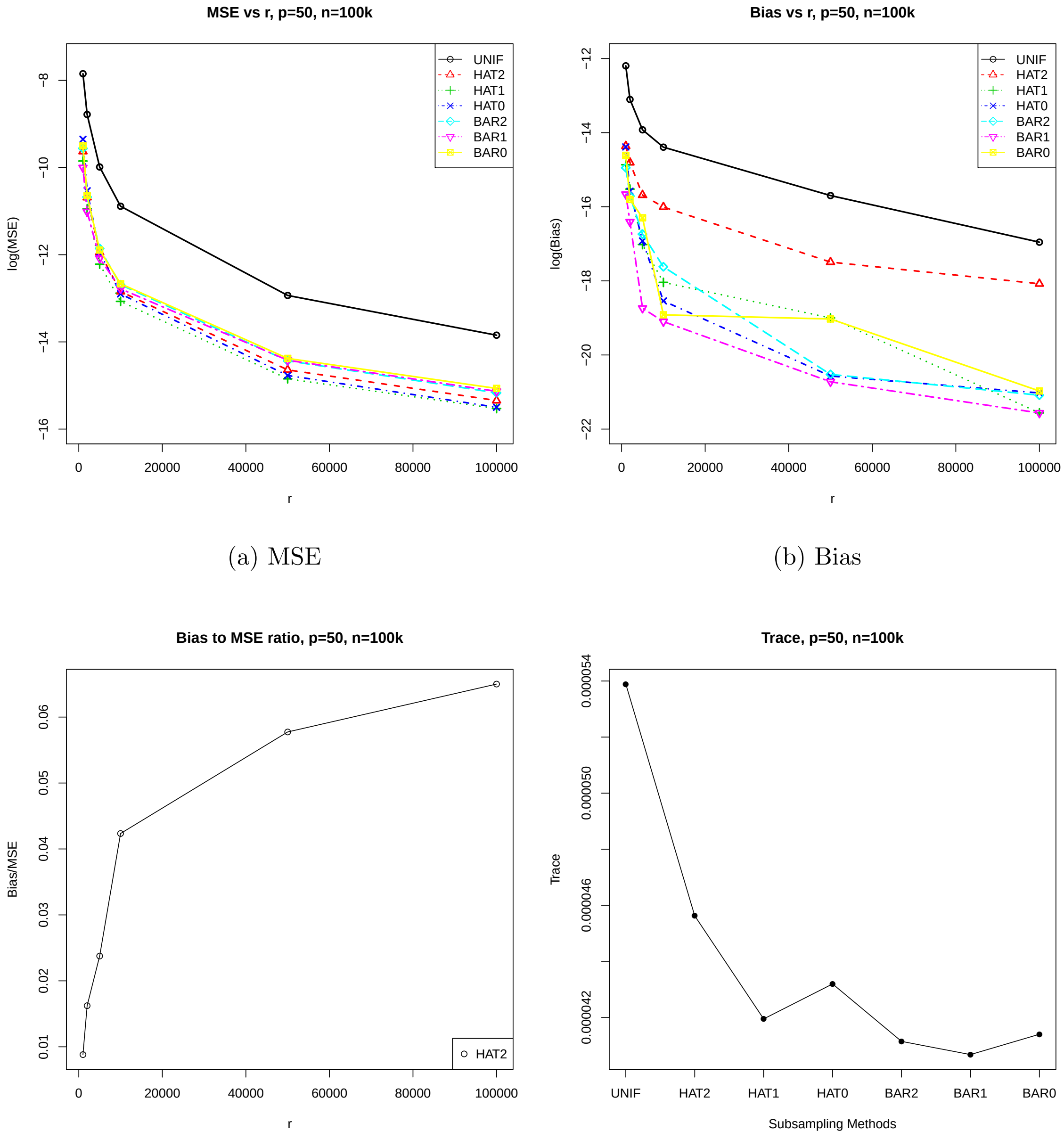
(a) MSE

(b) Bias

(c) Bias/MSE

(d) Trace

Figure 5.8. MSE, Bias, Bias to MSE ratio and trace of subsampling estimator on $\mathbf{X} \sim$ T5, with Poisson data and Poisson model. The full data size is $n = 100$k and the dimension is $p = 10$.



(a) MSE

(b) Bias

(c) Bias/MSE

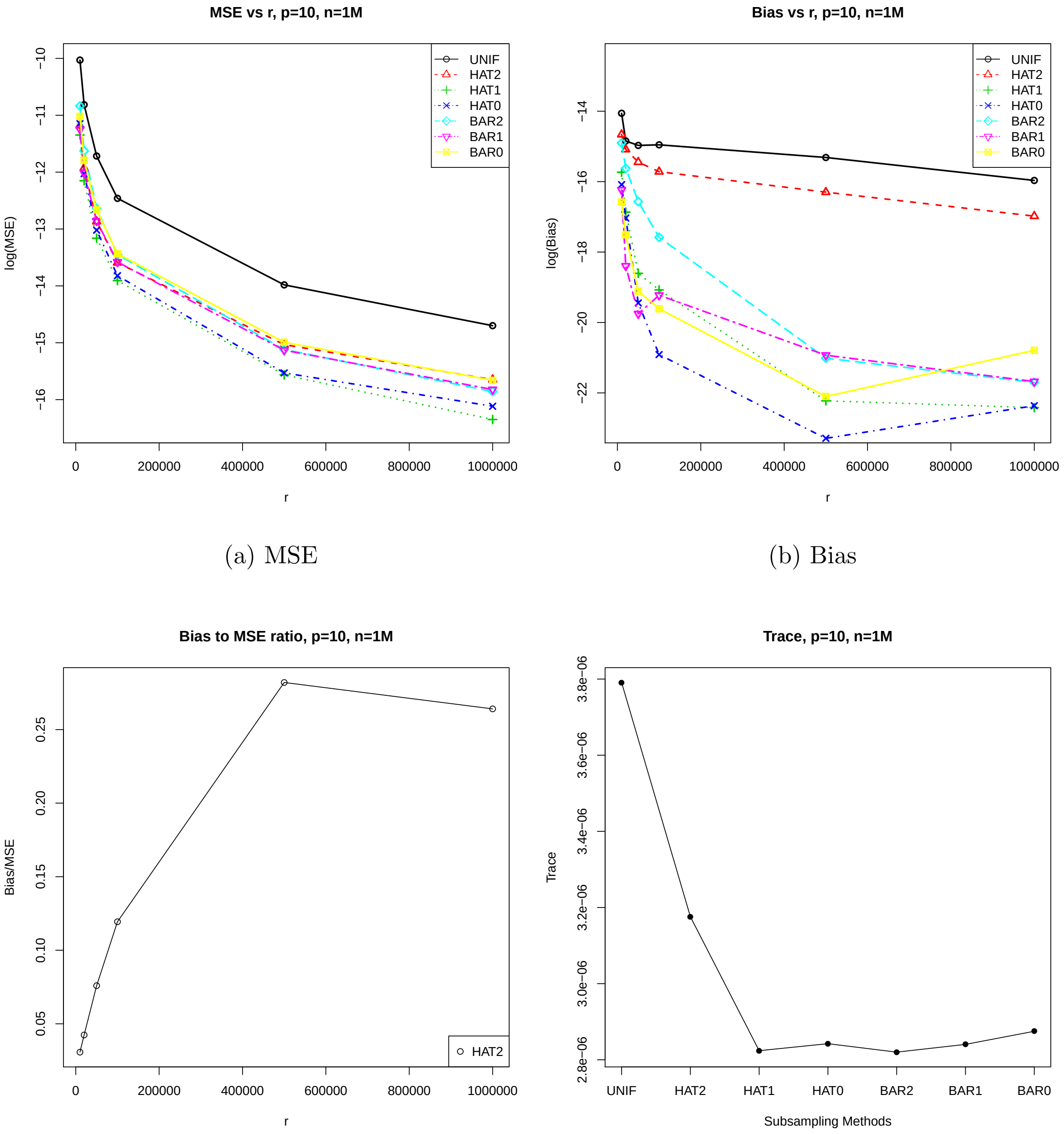
(d) Trace

Figure 5.9. MSE, Bias, Bias to MSE ratio and trace of subsampling estimator on $\mathbf{X} \sim$ T5, with Poisson data and Poisson model. The full data size is $n = 100$k and the dimension is $p = 50$.

3. (**Massive data size, conventional dimension**) Figure 5.10 shows the case when the dimension is low $p = 10$ and the data size is huge $n = 1$M. The MSE corresponds to A-optimal subsampling $\hat{\boldsymbol{\pi}}^{(2)}$ is not the minimum and the bias corresponds to HAT2 is significantly higher than the other A-optimal methods. Note that the percentage of bias to MSE of HAT2 is increasing from around 5% to 25% as r increases. This implies that under massive data, bias is a very significant factor of MSE. Therefore, minimizing the variance would not guarantee a minimum MSE. In addition, the percentage of bias under massive data is much larger than that of conventional data as shown in previous two examples. This indicates strongly that behavior of bias is different when the data is big data.

4. (**Massive data size, high dimension**) Figure 5.11 shows the case when the dimension is high $p = 50$ and the data size is huge $n = 1$M (which is our original setup). A-optimal subsampling $\hat{\boldsymbol{\pi}}^{(2)}$ gives smaller MSE than uniform method but larger than other A-optimal methods. Also, the bias corresponds to $\hat{\boldsymbol{\pi}}^{(2)}$ is the largest among all A-optimal methods. This makes sense since $\hat{\boldsymbol{\pi}}^{(2)}$ is not derived from minimizing the bias. The percentage of bias to MSE is increasing from around 5% to 25% which indicates the bias is significant. Also, the trace of the HAT2 is not the minimium but the difference between it with other traces is minimal.

In conclusion, under $\mathbf{X} \sim$ T5 distribution, A-optimal subsampling $\hat{\boldsymbol{\pi}}^{(2)}$ performs as expected when the dimension is conventional $p = 10$ and the data size is conventional $n = 100$k. Both high dimension p and massive data size n will increase the percentage of bias over MSE. This implies bias is not negligible and minimizing the variance would not necessarily produces the minimum MSE.



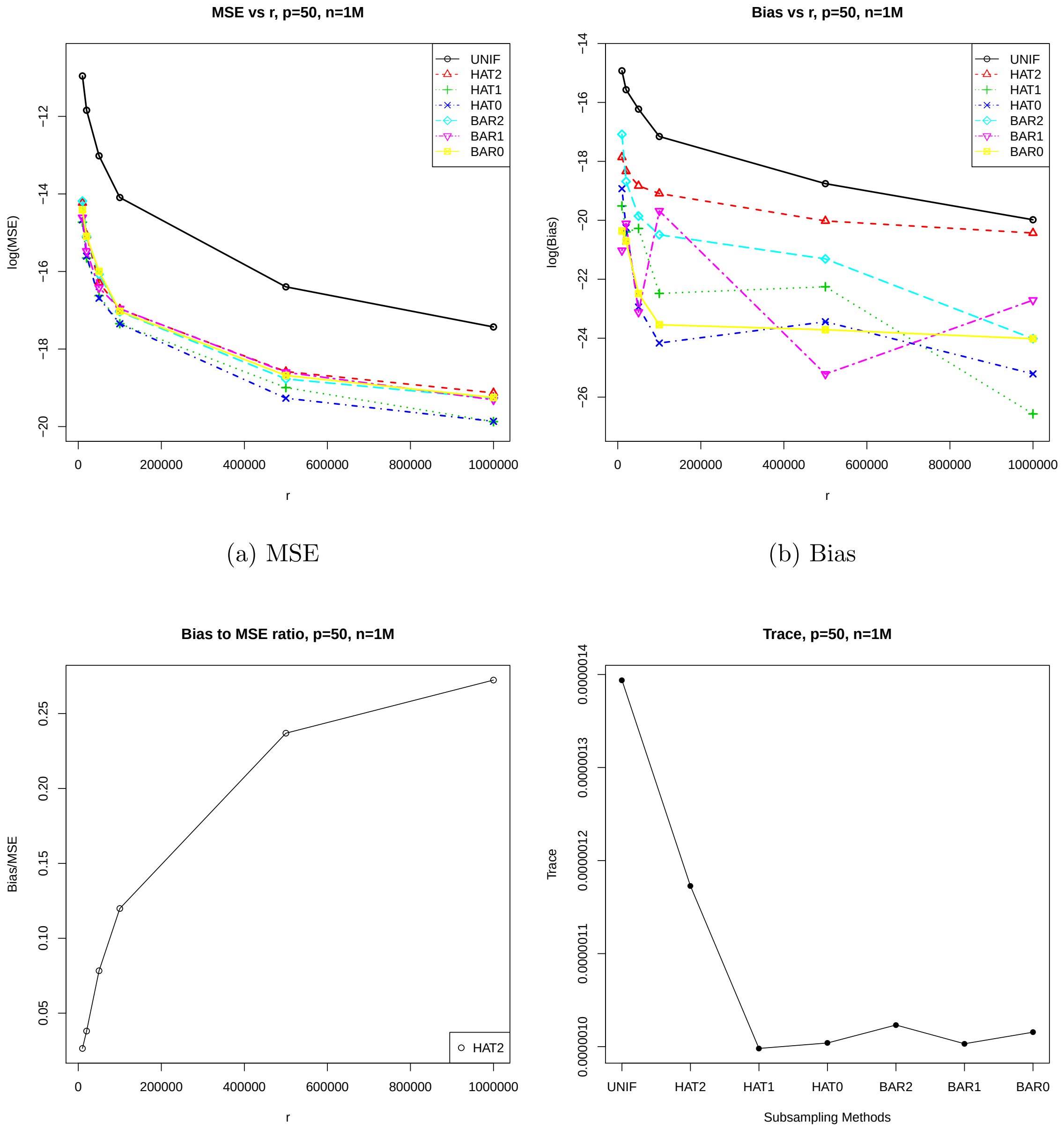
(a) MSE

(b) Bias

(c) Bias/MSE

(d) Trace

Figure 5.10. MSE, Bias, Bias to MSE ratio and trace of subsampling estimator on $\mathbf{X} \sim$ T5, with Poisson data and Poisson model. The full data size is $n = 1$M and the dimension is $p = 10$.



(a) MSE

(b) Bias

(c) Bias/MSE

(d) Trace

Figure 5.11. MSE, Bias, Bias to MSE ratio and trace of subsampling estimator on $\mathbf{X} \sim$ T5, with Poisson data and Poisson model. The full data size is $n = 1$M and the dimension is $p = 50$.

5.2 Numerical error in simulations

In the simulations studies of the previous section, we can conclude that there are two types of errors. First, the statistical error introduced to the models which cannot be avoided. Second, the numerical errors due to rounding off numbers and truncation of numbers during every stage of computations. In this section, we focus on the effect of numerical errors on our simulation studies using A-optimal subsampling methods. To quantify the numerical errors, we use the "Big-O" analysis which measures the time-complexity of the algorithm. The higher the time complexity, the more the numerical errors.

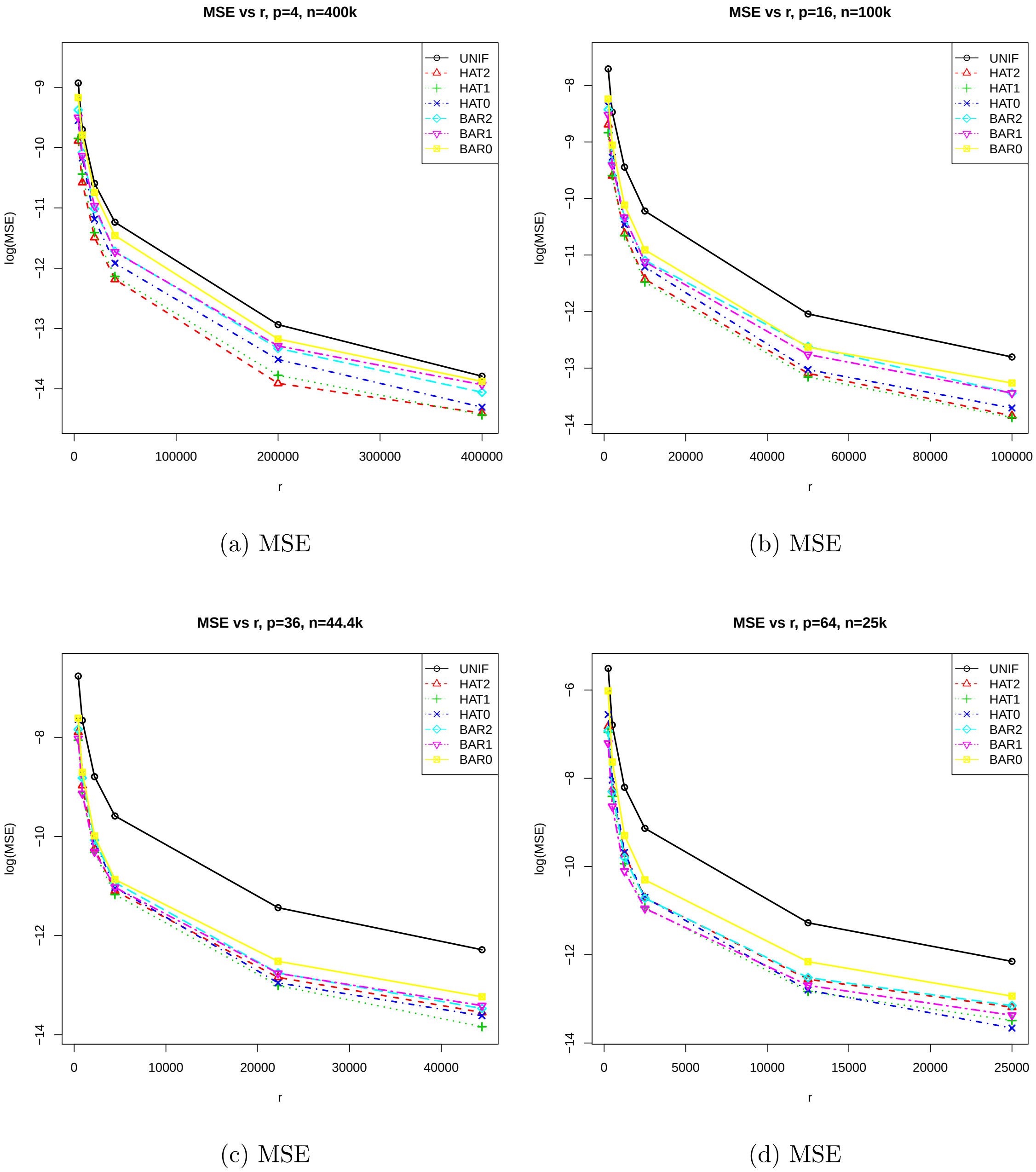
Our $\mathbf{X} \sim$ T5 simulations studies give unexpected results when n is massive or p is of high dimension. Since there are numerical errors involved in the computation which affect our results, we want to get rid of the effect of numerical errors by fixing the error term $O(np)$. We then compute different combinations of n and p to show whether the bias of the estimator increases with the dimension p. The data that we used is Poisson data fitted to a Poisson regression model, and the covariate is $\mathbf{X} \sim$ T5, and the size of the subsample is $r = 0.01n, 0.02n, 0.05n, 0.1n, 0.5n, n$.

Three different cases are considered: (1) $np = 1.6$M, (2) $np = 5$M and (3) $np = 10$M which corresponds to those three cases with unexpected results: high dimension p and conventional sample size n, low dimension p and massive sample size n, and high dimension p and massive sample size n. Within each group, we assume that the numerical errors is fixed and thus the error would be due to the different sizes of n and p.

Figure 5.12 shows the MSE of the coefficient estimates based on different cases of n and p: $(n, p) = (400k, 4), (100k, 16), (44.4k, 36), (25k, 64)$, where np is fixed at 1.6M. The MSE corresponds to A-optimal subsampling $\hat{\boldsymbol{\pi}}^{(2)}$ is the minimum when $p = 4$ and $n = 400k$. When p becomes larger, say $p = 64$ and $n = 25k$, we can see that HAT2 does not give the minimum MSE. This implies that even the variance is minimized, the MSE is not, thus the bias is not negligible when p becomes larger.

Figure 5.13 shows the MSE of the coefficient estimates based on different cases of n and p: $(n, p) = (500k, 10), (250k, 20), (166.7k, 30), (125k, 40), (100k, 50), (62.5k, 80),$ $(50k, 100), (33.3k, 150), (25k, 200)$, where np is fixed at 5M. The A-optimal subsampling $\hat{\boldsymbol{\pi}}^{(2)}$ does not give the minimum MSE among all methods except on the case when $n = $ 166.7k and $p = 30$ where $\hat{\boldsymbol{\pi}}^{(2)}$ is the minimum. It is clear that when p gets larger and larger, the HAT2 method does not give the minimum MSE. Again, this implies that bias is a significant factor of MSE when the data is of high-dimensional.

Figure 5.14 shows the MSE of the coefficient estimates based on different cases of n and p: $(n, p) = (1M, 10), (500k, 20), (333.3k, 30), (250k, 40), (200k, 50), (125k, 80),$ $(100k, 100), (66.6k, 150), (50k, 200)$, where np is fixed at 10M. In this large data set, we can see that the MSE of A-optimal subsampling $\hat{\boldsymbol{\pi}}^{(2)}$ is not the minimum for all cases. In particular, when $p = 200$, HAT2 gives the worst MSE among all the A-optimal sampling methods. This agrees with our theory that when p is large, bias is not negligible and minimizing variance does not imply minimizing the MSE.



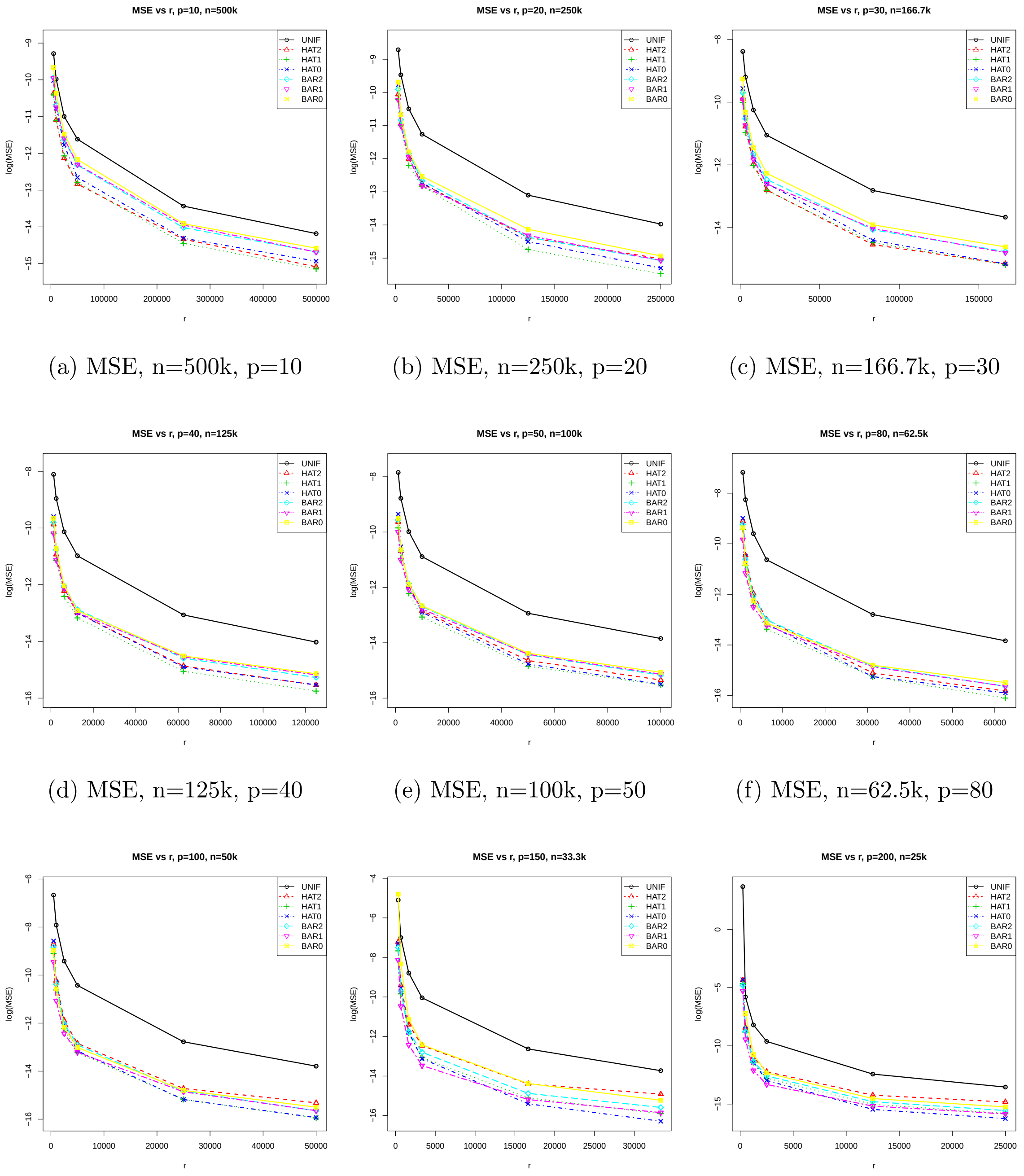
(a) MSE

(b) MSE

(c) MSE

(d) MSE

Figure 5.12. MSE of subsampling estimator on $\mathbf{X} \sim$ T5, Poisson data and Poisson model. np is fixed at 1.6M with n and p assume different values.



(a) MSE, n=500k, p=10

(b) MSE, n=250k, p=20

(c) MSE, n=166.7k, p=30

(d) MSE, n=125k, p=40

(e) MSE, n=100k, p=50

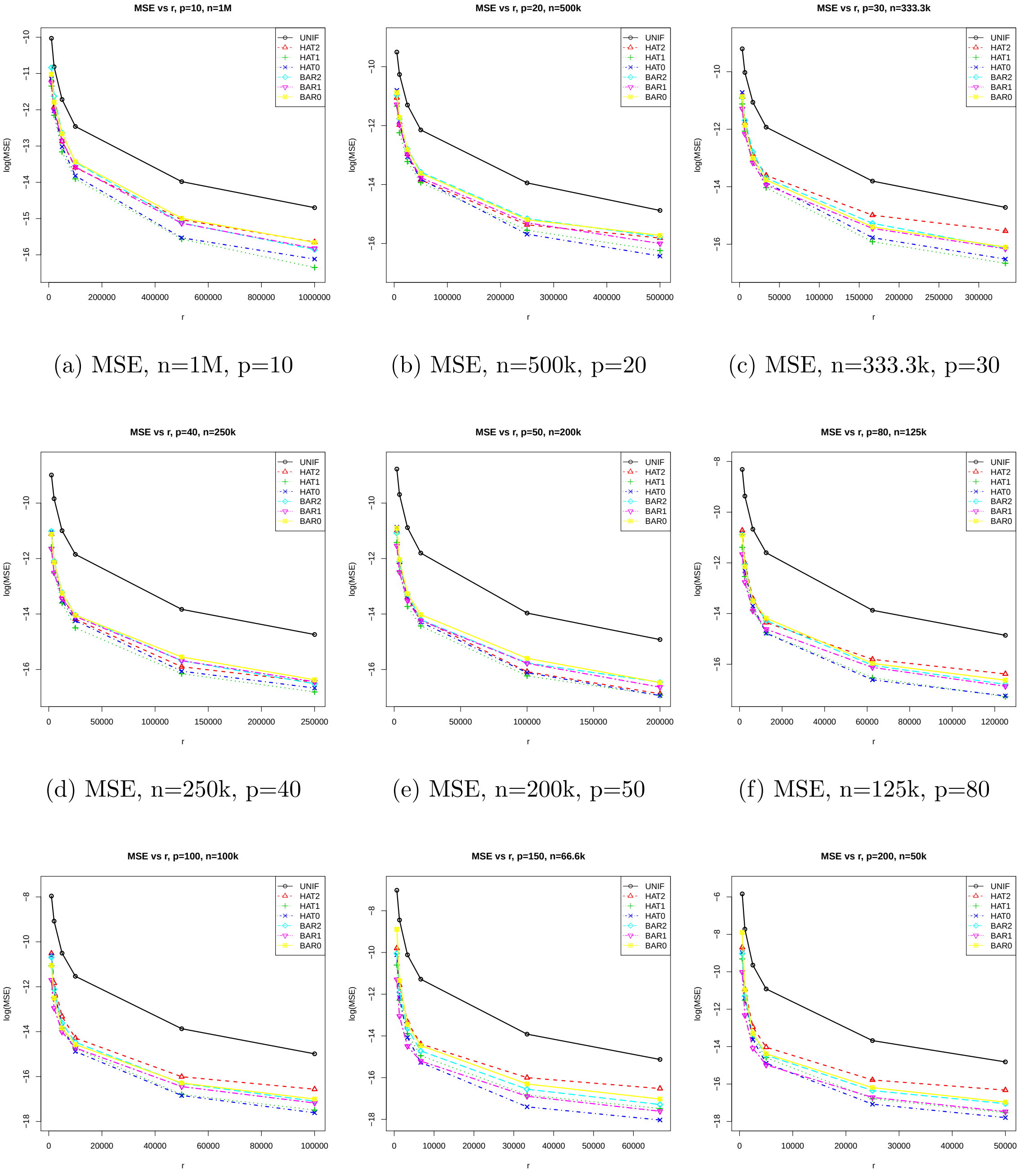
(f) MSE, n=62.5k, p=80

(g) MSE, n=50k, p=100

(h) MSE, n=33.3k, p=150

(i) MSE, n=25k, p=200

Figure 5.13. MSE of subsampling estimator on $\mathbf{X} \sim$ T5, Poisson data and Poisson model. np is fixed at 5M with n and p assume different values.



(a) MSE, n=1M, p=10

(b) MSE, n=500k, p=20

(c) MSE, n=333.3k, p=30

(d) MSE, n=250k, p=40

(e) MSE, n=200k, p=50

(f) MSE, n=125k, p=80

(g) MSE, n=100k, p=100

(h) MSE, n=66.6k, p=150

(i) MSE, n=50k, p=200

Figure 5.14. MSE of subsampling estimator on $\mathbf{X} \sim$ T5, Poisson data and Poisson model. np is fixed at 10M with n and p assume different values.

6. REAL DATA EXAMPLE

6.1 Gas sensors data

We consider the chemical sensors data set collected from the ChemoSignals Laboratory in the BioCircuits Institute, University of California San Diego (Fonollosa *et al.*, 2015). The data set contains the readings of 16 chemical sensors exposed to the gas mixtures of Ethylene and CO at different concentration levels. Each reading was obtained by the continuous acquisition of the 16-sensor array signals with the concentration levels of the mixtures change randomly for about 12 hours without interruption. The concentration ranges of Ethylene and CO were selected such that the induced magnitudes of the 16 sensors measurements were similar. The main purpose of this chemical data collection is to develop a better algorithms for improving the response time of the sensory systems. Further information on this experiment and the detailed explanation of the sensors data can be found in Fonollosa *et al.* (2015).

To illustrate our A-optimal sampling methods, we assume the response variable as the reading from the last chemical sensor, and the readings form the remaining chemical sensors are covariates. Since the readings from the second sensor have 20% of them being negative for some unknown reasons, we exclude it from the covariates. Hence, we have $p = 14$ covariates in total. In order to give a better presentation, log-transformation is applied on the sensors readings. Moreover, we drop the first 20,000 data since they correspond to the run-in time of the first few minutes. Thus, there are $n = 4,188,261$ data points for the resulting full data set.

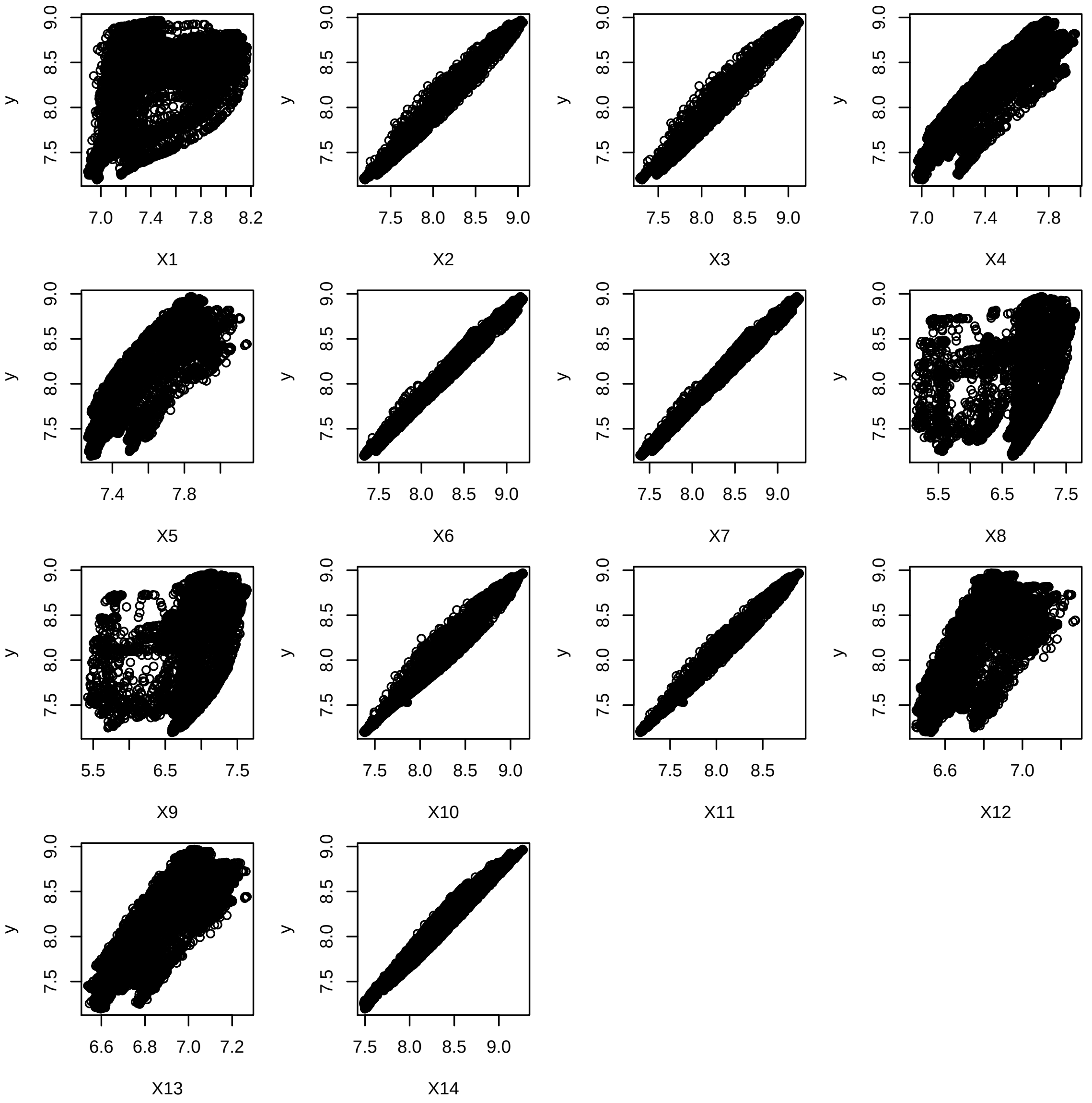


Figure 6.1. Scatter plots of a random sample of 10000 sensor data. The response y is the reading of the last sensor and the predictors X_i's are the readings of the remaining fifteen sensors except the second sensor.

Figure 6.1 gives the scatter plots of the predictors X_i's for $i = 1, \ldots, 15$ with the responses variable y using a simple random sample of 10,000 data points from the full data. We can see that the a linear model is appropriate for fitting the data:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{14} x_{i14} + \varepsilon_i, \quad i = 1, \ldots, n \tag{6.1.1}$$

Hence, the estimating equation is

$$\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)\mathbf{x}_i = 0 \tag{6.1.2}$$

with $\psi_{ni}(\beta) = (y_i - \mathbf{x}_i^T\beta)\mathbf{x}_i$, $\dot{\psi}_{ni}(\beta) = -\mathbf{x}_i\mathbf{x}_i^T$, $\dot{\Psi}_n = -\mathbf{X}^T\mathbf{X}$, and $a_{ni} = \dot{\Psi}_n^{-1}\psi_{ni}|_{\hat{\beta}_n} = -(\mathbf{X}^T\mathbf{X})^{-1}(y_i - \mathbf{x}_i^T\hat{\beta}_n)\mathbf{x}_i$. Hence the A-optimal sampling probabilities distributions $\{\hat{\pi}_i^{(\alpha)}\}_{i=1}^n$ and $\{\bar{\pi}_i^{(\alpha)}\}_{i=1}^n$ where $\alpha = 0, 1, 2$ are given by

$$\hat{\pi}_i^{(\alpha)} = \frac{\sqrt{\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-\alpha}\mathbf{x}_i}|e_i|}{\sum_{i=1}^n\sqrt{\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-\alpha}\mathbf{x}_i}|e_i|}, \quad i = 1, \ldots, n \tag{6.1.3}$$

$$\bar{\pi}_i^{(\alpha)} = \frac{\sqrt{\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-\alpha}\mathbf{x}_i(1-h_{ii})}}{\sum_{i=1}^n\sqrt{\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-\alpha}\mathbf{x}_i(1-h_{ii})}}, \quad i = 1, \ldots, n \tag{6.1.4}$$

To approximate $\hat{\beta}_n$ in the computation of a_{ni}, We will first taking a pre-subsample of size r_1 from the full sample and computing the OLS $\tilde{\beta}$. Then we will use it to replace the $\hat{\beta}_n$ in computing a_{ni}. Thus, the e_i in (6.1.3) is $y_i - \mathbf{x}_i^T\tilde{\beta}$. This is known as the A-optimal scoring method.

We consider subsamples with sizes $r = 0.001n, 0.002n, 0.005n, 0.01n, 0.1n$ with each numbers rounding to decimal place, i.e. $r = \{4188, 8377, 20941, 41883, 418826\}$. The pre-subsample size that we use for computing $\tilde{\beta}$ is 1000. Log-transformatin of MSEs and bias of the regression parameters $\hat{\beta}_r^*$ is computed for different subsampling probability distributions and different subsample sizes. Results from 100 bootstrap samples are plotted in the following figures.

In Figure 6.2, we can see that the A-optimal subsampling distribution $\hat{\pi}^{(2)}$ gives smaller MSEs than other subsampling distributions for all different subsample sizes r. Again, the uniform performs the worst among all distributions as expected. This suggests that under massive data set, A-optimal methods give better statistical inference than uniform, and among all different A-optimal subsampling, the original one dervied from minimizing the trace of the variance-covariance matrix of $\hat{\beta}_r^*$ gives the best MSE.

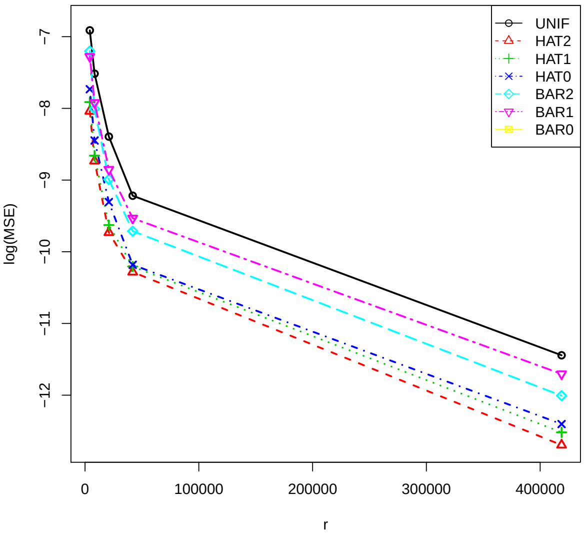


Figure 6.2. log(MSE) for estimating regression parameters for gas sensors data with 100 bootstrap samples.

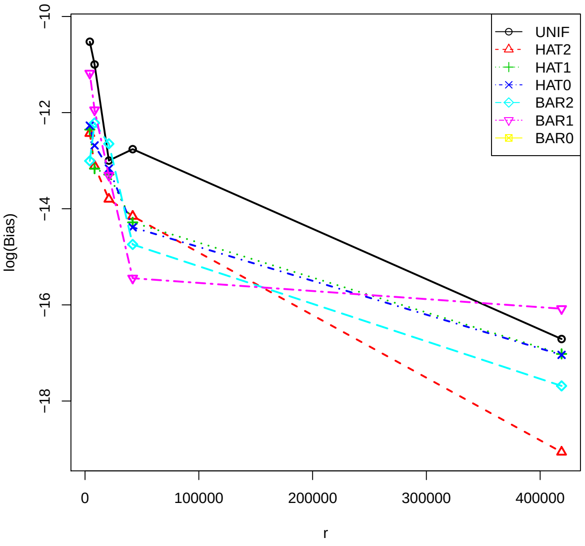


Figure 6.3. log(Bias) for estimating regression parameters for gas sensors data with 100 bootstrap samples.

In Figure 6.3, we compare the empirical bias of the regression parameters of different subsampling distributions for different subsample sizes r's. It can be seen that when r becomes larger, the bias from A-optimal $\hat{\pi}^{(2)}$ is significantly smaller than the other probabilities distributions. Also, uniform incurred the largest bias amount all the subsampling methods.

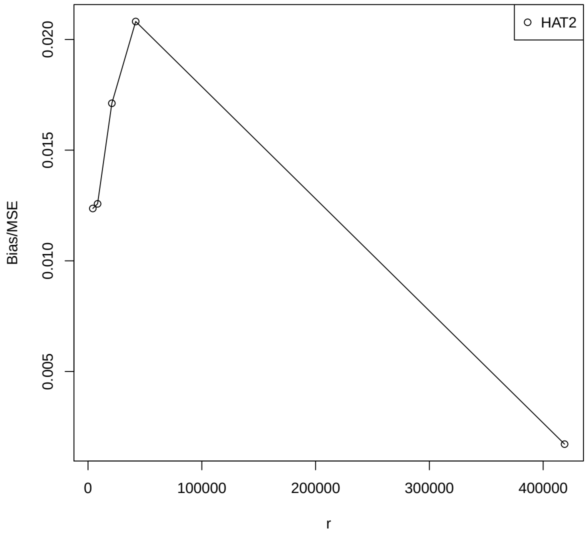


Figure 6.4. Bias/MSE ratios for estimating regression parameters for gas sensors data with 100 bootstrap samples.

Figure 6.4 shows the Bias to MSE ratios of A-optimal $\hat{\pi}^{(2)}$ for different subsample sizes. Note that all the ratios are less than 2.5% of the MSE, and it decreases significantly when r increases. This result agrees with Theorem 4.2.2 in the way that the remainder of bias approaches zero asymptotically when r goes to infinity and p is relatively small. Also, the main term of bias decreases when r increases. Thus, the bias is insignificant compared to the mean square errors.

REFERENCES

REFERENCES

[1] Avron, H., Maymounkov, P. and Toledo, S. (2010). Blendenpik: Supercharging LAPACK's least-squares solver. *SIAM Journal on Scientific Computing*, **32**: 1217–1236.

[2] Barbe, P. and Bertail, P. (1995). *The Weighted Bootstrap.* Springer, New York.

[3] Borovskikh, Yu. V. and Korolyuk, V. S. (1997). *Martingale Approximation.* VSP, Utrecht.

[4] Bose, A. and Chatterjee, S. (2002). Comparison of bootstrap and jackknife variance estimators in linear regression: Second order results. *Statist. Sinica*, **12**: 575–598.

[5] Boutsidis, C., Mahoney, M.W. and Drineas. P. (2009). An improved approximation algorithm for the column subset selection problem. *In Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms*: p. 968–977.

[6] Böhlmann, P. and van de Geer. S. (2011). *Statistics for High-Dimensional Data.* Springer, New York.

[7] Chatterjee, S. and Bose, A. (2002). Dimension asymptotics for generalized bootstrap in linear regression. *Ann. Inst. Statist. Math.* **54** (2): 367–381.

[8] Chatterjee, S. and Bose, A. (2005). Generalized bootstrap for estimating equations. *Ann. Statist.*, **33**: 414-436.

[9] Cameron, C. and Trivedi, P. (1998). *Regression analysis of count data.* Cambridge University Press, Cambridge United Kingdom.

[10] Cheung, T., Peng, H. and Tan, F. (2018). A-optimal subsampling for big data general estimating equations. *Manuscript.* Available at `https://www.math.iupui.edu/~hpeng/preprints_hp.html`.

[11] Cook, R. D., Tsai, C.-L. and Wei, B. C. (1986). Bias in nonlinear regression. *Biometrika*, **73**: 615–623.

[12] Cordeiro, G. M. and Botter, D. A. (2001). Second order biases of maximum likelihood estimates in overdispersed generalized linear models. *Statist. Probab. Lett.*, **55**(3): 269–280.

[13] Cordeiro, G. M. and McCullagh, P. (1991). Bias correction in generalized linear models. *J. R. Statist. Soc. B,* **25**: 305–317.

[14] Cox, D. R. and Snell, E. J. (1968). A general definition of residuals (with discussion). *J. R. Statist. Soc. B,* **30**: 248–275.

[15] Dean, J. and Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. *OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA*, pp. 137-150.

[16] Dobson, A. and Barnett, A. (2002). *An Introduction to Generalized Linear Models.* CRC Press, Boca Raton, FL.

[17] Drineas, P., Magdon-Ismail, M., Mahoney, M. W. and Woodruff, D. P. (2012d). Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research*, **13**: 3475–3506.

[18] 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.

[19] Drineas, P., Mahoney, M. W., Muthukrishnan, S. and Sarlós, T. (2010). Faster least squares approximation. *Numerische Mathematik*, **117**(2): 219–249.

[20] Drineas, P., Mahoney, M. W. and Muthukrishnan, S. (2006b). Sampling algorithms for ℓ_2 regression and applications. *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1127–1136.

[21] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, **7**: 1-26.

[22] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc. B,* **70**: 849–911.

[23] Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics,* **147**: 186–97.

[24] Fan, J., Han, F. and Liu, H. (2014). Challenges of big data analysis. *National Science Review,* **1**: 293–314.

[25] Freedman, D. A. (1981). Bootstrapping regression models. *Ann. Statist.* **9**(6): 1218–1228.

[26] Hansen, M. and Hurwitz, W. N. (1943). On the theory of sampling from a finite population. *Ann. Math. Statist.*, **14**(4): 333–362.

[27] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning.* Berlin: Springer.

[28] Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, **102**: 1025–1038.

[29] Lai, T. L. and C. Z. Wei (1982). A law of the iterated logarithm for double arrays of independent random variables with applications to regression and time series models. *Ann. Probab.* **10**(2): 320–335.

[30] Ma, P. and Sun, X. (2014). Leveraging for big data regression. *Computational Statistics,* **7** (1): 70-76.

[31] Mahoney, M. W. (2011). Randomized algorithms for matrices and data. *arXiv:1104.5557v3* [cs.DS]

[32] Mammen, E. (1993). Bootstrap and wild bootstrap for high-dimensional linear models. *Ann. Statist.*, **21**: 255–285.

[33] Mccullagh, P. and Nelder, J. (1984). *Generalized Linear Models.* Springer-Science+Business Media, New York, NY.

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

[35] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc. A*, **135**: 370–384.

[36] Præstgaard, J. and Wellner, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, **21** (4): 2053–2086.

[37] Peng, H. and Tan, F. (2018a). A fast algorithm for computing the A-optimal sampling distributions in big data linear regression. *Preprint.* Available at `https://www.math.iupui.edu/~hpeng/preprints_hp.html`.

[38] Peng, H. and Tan, F. (2018b). Big data linear regression via A-optimal subsampling. Submitted to *Ann. Statist.* Available at `https://www.math.iupui.edu/~hpeng/preprints_hp.html`.

[39] Sarlós, T. (2006). Improved approximation algorithms for large matrices via random projections. *In Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 143–152.

[40] Teicher, H.(1974). On the law of the iterated logarithm. *Ann. Probab.,* **2**: 714–728.

[41] Tibshirani, R. (1996). Regression shrinkage and selection with the lasso. *J. R. Statist. Soc. B,* **58**: 267–288.

[42] Vershynin, R. (2011). Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027v7*[math.PR].

[43] Wang, C., Chen, M.-H., Schifano, E., Wu, J. and Yan, J. (2016). Statistical methods and computing for big data. *Statistics and Its Interface*, **9**(4): 399-414.

[44] Wang, H. Y., Yang, M. and Stufken, J. (2019). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, **114**(525): 393-405.

[45] Wang, H. Y., Zhu, R., and Ma, P. (2018). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, **113**(552): 829–844.

[46] Worthey, E., Mayer, A. and Syverson, G. *et al.* (2010). Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet. Med.*, **13**: 255-262.

[47] Young, D. H. and Bakir, S. T. (1987). Bias correction for a generalized log-gamma regression model. *Technometrics,* **29**: 183–191.

[48] Zhao, X. (2018). Regression analysis of big count data via a-optimal subsampling. Ph.D. dissertation, Purdue University, USA.

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

VITA

VITA

My name is Chung Ching Cheung. I am from Hong Kong. In my undergraduate, I studied mathematics at The University of Hong Kong and graduated in 2008. After finished my bachelor degree, I pursued a master degree in pure mathematics in the same university and graduated in 2011. After that, I have worked for two years in The Hong Kong Polytechnic University and one year in Hong Kong Baptist University as a lecturer in mathematics. I came to IUPUI in 2014 to pursue my PhD.