Information-based Optimal Subdata Selection for Big Data Logistic Regression

Qianshun Cheng, Haiying Wang and Min Yang[1]

Monsanto(Bayer) Company, University of Connecticut and University of Illinois at Chicago

Abstract

Technological advances have enabled an exponential growth in data volumes, and proven statistical methods are no longer applicable for extraordinary large data sets due to computational limitations. Subdata selection is an effective strategy to address this issue. In this study, we investigate existing sampling approaches and propose a novel framework of selecting subsets of data for logistic regression models. We show that, while the information contained in the subdata based on random sampling approaches is limited by the size of the subset, the information contained in the subdata based on the new framework increases as the size of the full data set increases. Performances of the proposed approach and that of other existing methods are compared under various criteria via extensive simulation studies.

Keywords: D-Optimality, Information matrix, Subsampling

1. Introduction

Technological advances have enabled an exponential growth in data collection and the size of data sets. For example, the cross-continental Square Kilometre Array, the next generation of astronomical telescopes, will generate 700 TB of data per second [1]. While the extraordinary sizes of data sets provide researchers golden opportunities for scientific discoveries, they also bring tremendous challenges when attempting to analyze these large data sets. Proven statistical methods are no longer applicable due to computational

*Corresponding author

Email address: chengqianshun1@gmail.com \& haiying.wang@uconn.edu \& myang2@uic.edu (Qianshun Cheng, Haiying Wang and Min Yang)

limitations. Recent advances in statistical analysis to deal with these challenges are arguably on two major different strategies: the divide-and-conquer approach and the subdata selection approach.

The divide-and-conquer approach takes advantage of the parallel computing technology. A large data set is split into chunks of reasonable sizes, and analysis is implemented separately on each chunk of data and a specified aggregation method is implemented to merge pieces of information from chunks to produce final analysis. The analysis and aggregation methods depend on the structure of the data set and model assumptions. For the linear regression model, the least squares estimate can be directly decomposed into a weighted average of the least squares estimate based on each chunk. This has become the standard aggregation method for merging solutions from blocks with linear models. For nonlinear models, several aggregation methods are proposed. [2] proposed an approach for approximating the estimating equation estimator using a first order Taylor expansion. Under certain conditions, accuracy of the final estimator from aggregation is proved to be close to the direct estimator from the full data. [3] considered a divide-and-conquer approach for generalized linear models (GLM) where both the number of observations n and the number of covariates p are large. They incorporated variable selection via penalized regression into the subset processing step, and showed that, under certain regularity conditions, the aggregated estimator in model selection is consistent and asymptotically equivalent to the penalized estimator based on the full dataset. In [4], an approach similar to the divide-and-conquer approach is proposed, where accumulated parameter estimators based on data chunks arrived can be updated using future coming data. The divide-and-conquer approach gains efficiency mainly from the implementation of parallel computing, and it may not reduce computational time if implemented with a single core.

The subdata approach reduces the computation burden by downsizing the data volume. The key question here is how to select an informative subdata such that it maintains as much information as possible. As noted in a recent NSF program guideline, "Tradeoffs between computational costs and statistical efficiency" is one of six research directions need to be addressed for theoretical foundation of data science [5].

Existing subdata approaches are mainly based on random subsampling. Combining the methods of subsampling [6] and bootstrapping [7, 8], [9] proposed a novel approach

called bags of little bootstraps (BLB) to achieve computational efficiency. [10] proposed a mean log-likelihood approach using Monte Carlo averages of estimates from subsamples to approximate the quantities needed in the analysis. The BLB and mean log-likelihood methods select subsamples using simple random sampling. Another line of the subsampling method is based on leverage sampling algorithms. In this approach, a sampling probability is assigned to each dataline according to its leverage score. [11] reviewed existing subsampling methods in the context of linear regression and termed the methods leveraging algorithms, considered the statistical properties of leveraging algorithms, and proposed a shrinkage algorithmic leveraging method.

A major limitation of random subsampling methods is that the amount of information in a resulting subdata is proportional to the size of the subdata, which is often significantly smaller than the full data size. [12] proved that, in linear regression, the variance of an estimator based on the random subsampling method converge to zero at a rate proportional to the inverse of the subdata size. Is it possible that the information contained in a subdata is related to the size of the full data rather than that of the subdata only? Ideally we would want to choose the subdata with the maximum amount of information among all possible subdata sets. However this is infeasible in practice since there are $\binom{n}{r}$ subsets of data with size r from a full data set of size n. This combination number is quickly out of reach even for moderate n and r, so an alternative approach has to be employed. Under linear models, [12] proposed a novel approach called Information-Based Optimal Subdata Selection (IBOSS) to select a subdata. Unlike random subsampling methods, IBOSS is a deterministic approach. It selects a subdata based on the characterization of the D-optimal design. Under certain conditions, [12] showed that the variance of the resultant estimator converge to zero at a rate corresponding to the size of the full data. The simulation studies demonstrated that the IBOSS approach significantly outperformed random subsampling approaches.

While the IBOSS approach effectively addresses the trade off between the computational complexity and statistical efficiency, it is under the linear model context. Does this strategy also work under nonlinear models? Unlike linear models, where the corresponding information matrices are relatively simple with an explicit form, the problem for nonlinear models is remarkably different, where the information matrices are much

more complicated and depend on unknown parameters. Consequently, the problem under nonlinear models is considerably harder than that under linear models.

Nonlinear models, however, are widely applied in practice. Specifically, logistic regression models have played important roles in categorical data analysis. They have been used in various fields, like finance, medicine, and social sciences. Unlike linear models, where the estimators have closed form solutions, the estimators for logistic regression models have no closed form solutions in general. We have to utilize iterative approaches to calculate the estimates numerically. Compared with linear regression models, the computation cost for logistic regression models is much higher for big data sets. There is limited research on how to choose a subdata from a full data set for a logistic regression model, perhaps due to the complexity of the nonlinearity feature. [13] proposed the optimal subsampling method under the A-optimality criterion (OSMAC) algorithm, where the probability weights are specified according to the A-optimality in optimal design theory [14]. However, like many other random subsampling approaches for linear models, we shall show in the next section that the information extracted from the OSMAC approach is limited by the subsample size.

In this paper, we study subdata selection under logistic regression models utilizing the IBOSS strategy. A new algorithm of selecting subdata is proposed. Compared with existing subsampling approaches, the new algorithm has the following two advantages: the estimation efficiency of the algorithm is significantly higher and the computational cost is competitive.

The key contribution of this paper is that, under logistic regression models, it (i) proves that the information from random subsampling based subdata selection method is limited by the size of the subdata, (ii) proposes a new approach for the trade off between the computational complexity and statistical efficiency, and (iii) proves that the information from the new algorithm increases along with the size of full data. Theses results give a theoretical justification for the information based subdata selection under nonlinear models. Since "data reduction is perhaps the most critical component in retrieving information in big data" [15], this is a significant step in big data analysis under nonlinear models.

The rest of the paper is organized as follows: Section 2 introduces notations, provides

a summary of existing methods, and presents lower-bounds of the variance covariance matrices for subsampling-based estimators. Section 3 introduces a new algorithm and discusses its asymptotic properties. Section 4 compares the performance of the new algorithm, the OSMAC algorithm, and the simple random sampling method using various simulation settings. Section 5 provides a brief summary of this paper and its possible extensions. All technical details are provided in the supplemental material.

2. Notations and Existing Methods

We present the model setup and existing methods in this section. Let $\mathcal{F}_n = \{(Y_i, Z_i), i = 1, \ldots, n\}$ denote the full data, where Y_i is a binary response variable and $Z_i = (z_{i1}, \cdots, z_{im})^T$ is a m dimensional explanatory variable. Assume the logistic regression model:

$$\text{Prob}(Y_i = 1|X_i) = p_i(\boldsymbol{\beta}) = \frac{e^{X_i^T\boldsymbol{\beta}}}{1 + e^{X_i^T\boldsymbol{\beta}}}, \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_m)^T$ and $X_i = (1, Z_i^T)^T = (1, z_{i1}, \cdots, z_{im})^T$. Here, β_0 is the intercept parameter and $(\beta_1, \cdots, \beta_m)^T$ is the m dimensional slope parameter. Like in linear models, $\boldsymbol{\beta}$ is frequently estimated by the maximum likelihood estimator (MLE),

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} [Y_i \log p_i(\boldsymbol{\beta}) + (1 - Y_i) \log\{1 - p_i(\boldsymbol{\beta})\}].$$

However, for logistic regression, there is no general closed-form solution to the MLE and iterative algorithms such as the Newton-Raphson algorithm [16] are often used to find it numerically. The computational cost of calculating $\hat{\boldsymbol{\beta}}$ based on the full data is at the order of $O(\Delta nm^2)$, where Δ is the number of iterations in the optimization algorithm.

For extraordinary large n, the computational cost could be beyond the available computation capacity. We may have to consider analyzing a subdata instead of the full data. Here, we focus on the scenario that we can only analyze a subdata of size k, and the question is how to choose a subdata that contains the most amount of information about unknown parameters. In literature, many subsampling strategies are developed for this purpose. Most of these strategies, however, are under linear models and usually cannot be easily extended to logistic regression due to the nonlinearity of the logistic regression

models. The few strategies suitable for logistic models are the uniform sampling and the OSMAC algorithm [13].

2.1. Existing Subsampling Approaches and Their Limitations

In a random subsampling approach, a subsample is taken randomly according to some sampling distribution and the sampling procedure is often with replacement. We use $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)$ with $\sum_{i=1}^{n} \eta_i = k$ to denote a subsample of size k taken randomly from the full sample, where η_i denotes the number of times that the ith data point is included in a subsample. Let π_i, $i = 1, ..., n$ be subsampling probabilities such that $\sum_{i=1}^{n} \pi_i = 1$. A subsampling-based estimator, say $\hat{\boldsymbol{\beta}}^{\boldsymbol{\eta}}$, has the general form of

$$\hat{\boldsymbol{\beta}}^{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \frac{\eta_i}{\pi_i} [Y_i \log p_i(\boldsymbol{\beta}) + (1 - Y_i) \log\{1 - p_i(\boldsymbol{\beta})\}].$$

When the uniform sampling is implemented, all datalines have equal chances to be selected, i.e., $\pi_i = \frac{1}{n}$. It is a widely-used technique to downsize data due to its simplicity and low cost on computational resources. However, in terms of the information retrieved from a big dataset, uniform sampling may not be the best choice.

Inspired by the A-optimality criteria from optimal design theories, [13] proposed the novel OSMAC algorithm, in which the sampling probabilities are assigned to each dataline in a way to optimize the A-optimality criteria of the asymptotic covariance matrix of subsample parameter estimators, which is equivalent to minimizing the asymptotic mean squared error (MSE) of some parameter. They recommended result that minimizes the asymptotic MSE of $M_X \hat{\boldsymbol{\beta}}^{\boldsymbol{\eta}}$, where $M_X = \frac{1}{n} \sum_{i=1}^{n} \Psi(\hat{c}_i) X_i X_i^T$, $\hat{c}_i = X_i^T \hat{\boldsymbol{\beta}}$, and $\Psi(c_i) = \frac{e^{c_i}}{(1+e^{c_i})^2}$, and call the corresponding strategy mVc strategy as in [13]. The optimal sampling probabilities under the mVc strategy are

$$\pi_i^{mVc} = \frac{|Y_i - p_i(\hat{\boldsymbol{\beta}})| \|X_i\|}{\sum_{j=1}^{n} |Y_j - p_j(\hat{\boldsymbol{\beta}})| \|X_j\|}, i = 1, ..., n.$$

Further details and interpretations of these subsampling probabilities can be found in [13]. They showed that the mVc strategy has high estimation accuracy and low computational cost through simulation studies and real data analysis. They also derived the asymptotic property of $\hat{\boldsymbol{\beta}}^{\boldsymbol{\eta}}$ in approximating $\hat{\boldsymbol{\beta}}$. Specifically, they proved that under

certain conditions, conditional on the full data, for large n and k,

$$\hat{\boldsymbol{\beta}}^{\boldsymbol{\eta}} - \hat{\boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{V}), \tag{2.2}$$

where $\mathbf{V} = M_X^{-1}\mathbf{V_c}M_X^{-1}$, $\mathbf{V_c} = \frac{1}{kn^2}\sum_{i=1}^{n}\frac{(Y_i - p_i(\hat{\boldsymbol{\beta}}))^2 X_i X_i^T}{\pi_i}$, and $\overset{a}{\sim}$ means that the two quantities have the same asymptotic distribution.

The result in (2.2) builds the bridge between the full data MLE and the subdata estimator based on random subsampling strategies. However, the following Theorem 2.1 shows that $\mathbf{V}$ is bounded from below by a term that is at the order $1/k$ and does not converge to zero as n goes to infinity if k is fixed. Therefore, the information that we can extract from the random sampling is limited by the subsampling size k, which is a limitation of the random sampling strategy.

Theorem 2.1. *For the random subsampling-based estimator $\hat{\boldsymbol{\beta}}^{\boldsymbol{\eta}}$, the asymptotic variance covariance V in (2.2) is larger than, in Lowering ordering, a matrix proportional to the inverse of subdata size, namely,*

$$\mathbf{V} \geq \frac{4\{1 + o_P(1)\}}{k}\left(\sum_{i=1}^{n}\pi_i X_i X_i^T\right)^{-1}.$$

Applying Theorem 2.1 to existing sampling-based methods, we see some limitations of these methods. For uniform sampling, $\pi_i = \frac{1}{n}$, $i = 1, \cdots, n$, we have the following theorem.

Theorem 2.2. *For uniform sampling, if $X_i, i = 1, \ldots, n$, are generated independently from the same distribution X which has finite second moment, then we have*

$$\mathbf{V} \geq \frac{4\{E(XX^T)\}^{-1} + o_P(1)}{k}.$$

For the mVc sampling strategy, just like that for uniform subsampling, we have the following Theorem showing its limitation.

Theorem 2.3. *If $X_i, i = 1, \ldots, n$, are generated independently from the same distribution X, which satisfies that $E(\|X\|^3) < \infty$, then for a subdata obtained according to the mVc subsampling probabilities,*

$$\mathbf{V} \geq \frac{a(E(\|X\|XX^T))^{-1} + o_P(\mathbf{1})}{k}, \tag{2.3}$$

where $a > 0$ is a constant that does not depend on n or k.

The aforementioned theorems show that the variance covariance matrix of the parameter estimator from a subdata sampled through existing subsampling strategies for logistic regression is bounded in probability from below by a term that is related to the size of the subdata only.

3. IBOSS Algorithm for Logistic Regression Models

Recently, [12] proposed a novel IBOSS subsampling approach for linear models. Unlike random subsampling approach which selects a subdata according to some sampling distribution, the IBOSS approach directly utilizes the structure of D-optimal design under linear models and deterministically selects informative subsets. Based on both simulated and real data, [12] showed that the resultant estimator by implementing this procedure has significantly higher estimation efficiency. They also showed theoretically that the information matrix of the subdata from the IBOSS approach with a fixed k is not bounded as long as $n \to \infty$ and the covariate distribution is not bounded. These results have built the theoretical foundation for the IBOSS strategy, which pave the way for applying IBOSS strategy for more complexity analysis, for example, LASSO. On the other hand, these results are under linear models. Can such strategy apply for nonlinear models, specifically, logistic models? This paper gives a positive answer to this question.

The critical step in IBOSS strategy is to characterize optimal subdata utilizing the information matrix. Under some regularity conditions [17, 18], for a large n, the full data MLE $\hat{\boldsymbol{\beta}}$ satisfies that, asymptotically,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \overset{a}{\sim} N\bigg(0, \Big(\sum_{i=1}^{n} \Psi(c_i^0) X_i X_i^T\Big)^{-1}\bigg),$$

where $\Psi(c_i^0) = \frac{e^{c_i^0}}{(1+e^{c_i^0})^2}$, $c^0 = X_i^T \boldsymbol{\beta}^0$, and $\boldsymbol{\beta}^0$ is the true parameter. Here, the term $I = \sum_{i=1}^{n} \Psi(c_i^0) X_i X_i^T$ is called the Fisher information matrix and an optimal design is to optimize some meaningful functions of I. The nonlinearity of its elements within the information matrix as well as its dependency on unknown parameters complicate the characterization of an optimal subdata. Consequently, the picture of IBOSS strategy under logistic regression models is remarkably different from that of linear models. It is

arguably much harder.

3.1. Characterization and the proposed algorithm

Let $\alpha_1, \ldots, \alpha_n$ be the indicators showing whether the corresponding data points are selected or not, i.e., $\alpha_i = 1$ if (y_i, X_i) is selected in the subdata and $\alpha_i = 0$ otherwise. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ and $\hat{\boldsymbol{\beta}}^{\boldsymbol{\alpha}}$ be the resulting estimator from the subdata selected through $\boldsymbol{\alpha}$. We first study the characterization of an optimal subdata under the D-optimality criterion. If a subdata is selected deterministically, the corresponding information matrix can be written as

$$I_{\boldsymbol{\alpha}} = \sum_{i=1}^{n} \alpha_i \Psi(c_i^0) X_i X_i^T. \tag{3.1}$$

The following result gives an upper bound for the determinant of the subdata information matrix (3.1).

Theorem 3.1. *For subdata of size k from full data (X_i, Y_i), $i = 1, \ldots, n$,*

$$|I_{\boldsymbol{\alpha}}| \le \min_{1 \le \ell \le m} \left\{ \frac{k^{m+1}}{4^{m-1}\beta_\ell^2} (c^*)^2 \Psi(c^*)^{m+1} \prod_{j=1,\ j \ne \ell}^{m} (Z_{(n)j} - Z_{(1)j})^2 \right\}, \tag{3.2}$$

where c^* maximizes $c^2\Psi(c)^{m+1}$, and $Z_{(n)j}$ and $Z_{(1)j}$ are the extreme order statistics for the jth covariate. If the minimum on the right hand side of (3.2) occurs at $\ell = m$, say, then the equality in (3.2) holds for subdata, if it exists, with $k/2^m$ data points that have $Z_i^* = (z_{i1}, \ldots, z_{i,m-1}, z_{i,m})$, $i = 1, 2, \ldots, 2^m$, where $z_{ij} = Z_{(n)j}$ or $Z_{(1)j}$, $j = 1, \ldots, m-1$ and z_{im} is chosen to make $(X_i^*)^T\boldsymbol{\beta}$ equal to c^* or $-c^*$, with each of the 2^m possible combinations for Z_i^* appearing equally often.

Often, k is much smaller than 2^m, and the subdata with the equality in Theorem 3.1 does not exist. Nonetheless the characterization of the subdata can guide us to select a more informative subdata. Notice that the IBOSS algorithm for linear models in [12] cannot be directly applied for the models considered in this paper due to the different characterizations of the subdata. We will propose a novel algorithm motivated by Theorem 3.1.

The characterization in Theorem 3.1 requires $c = X^T\boldsymbol{\beta}$ to be fixed at constant $\pm c^*$, which depends on the unknown parameter. To address this issue, we first draw a small

subdata of size k_0 through uniform sampling and obtain a rough estimator, say $\hat{\boldsymbol{\beta}}_{k_0}$, of the parameter $\boldsymbol{\beta}$. To best meet the characterization, a two-stage subdata selection strategy is proposed. In the first stage, a relatively large portion of the full data with their c values falling into a pre-specified neighbor of $\pm c^*$ is selected. For example, we can choose some $\delta > 0$ and then collect all the data lines (X_i, Y_i) with $\{i \mid min(|c_i - c^*|, |c_i + c^*|) \leq \delta\}$, where $c_i = X_i^T \hat{\boldsymbol{\beta}}_{k_0}$. These selected datalines will be treated as the new database for the second stage data selection procedure. The second stage procedure is similar to the IBOSS procedure proposed for linear regression, i.e., to select data lines according to extreme values of all the m covariates. The proposed IBOSS procedure to select a subdata of size k is described in details below.

Stage 1:

1. Prefix a constant δ as maximum tolerance on the c values.

2. Given data set $\mathcal{F}_n = \{(Y_i, Z_i),\ i = 1, \ldots, n\}$, use random sampling to take a subdata of size k_0, and derive an estimate $\hat{\boldsymbol{\beta}}_{k_0}$ based on the selected subdata.

3. Compute $c_i = X_i^T \hat{\boldsymbol{\beta}}_{k_0}$, where $X_i^T = (1, Z_i^T)$, for $i = 1, \cdots, n$, and construct $B = \{i \mid \min(|c_i - c^*|, |c_i + c^*|) \leq \delta\}$.

Stage 2:

4. For $l = 1$, from $\{(Y_i, Z_i^T),\ i \in B\}$, pick $\lceil \frac{k}{2m} \rceil$ data lines with largest values of z_{il} and $\lceil \frac{k}{2m} \rceil$ data lines with smallest values of of z_{il}. Include these datalines in the subsample, and remove their index from the set B.

5. Repeat Step 4 for $l = 2, \ldots, m$.

3.2. Asymptotic properties

As we have shown in Section 2.1, one limitation of a subsampling-based procedure is that the asymptotic variance covariance matrix of the resultant estimator is bounded from below by a term proportional to the inverse of subdata size k. In other words, the information matrix of the subdata is bounded even if n goes to infinity. We will show that the proposed new algorithm is not restricted by this limitation.

The two-stage procedure, however, makes it extremely challenge to investigate the asymptotic property for the general case. We need to study the asymptotic distribution of order statistics conditional on the event that $min(|c_i - c^*|, |c_i + c^*|) \leq \delta$. In addition, c_i depends on $\hat{\boldsymbol{\beta}}_{k_0}$ and no closed form solution is available. Fortunately, focusing on the case of $m = 2$, we manage to prove the following theorem.

Theorem 3.2. *For logistic regression with $X_i = (1, z_{i1}, z_{i2})^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$, assume that the two dimension covariate $Z_i = (z_{i1}, z_{i2})$, $i = 1, \cdots, n$, are generated independently from a bivariate normal distribution $\mathcal{Z}$ with mean vector u and variance covariance matrix $\boldsymbol{\Sigma}$. Let I^{IBOSS} be the information matrix of $\boldsymbol{\beta}$ based on two-stage procedure and n_1 represent the number of remaining datalines after the first stage. Suppose that $\beta_1 \neq 0$, $\beta_2 \neq 0$ and Σ is nonsingular, then at least one eigenvalue of I^{IBOSS} goes to ∞ when n_1 goes to ∞.*

The assumptions about β_1, β_2, and Σ are not restricted, and they are reasonable in practice. In practice, δ is usually specified to keep a certain percentage of the full data after the first stage, i.e., n_1 is proportional to n and thus goes to infinity.

From Theorem 3.2, as long as the size of the remaining data after first stage goes to infinity, at least one eigenvalue of the information matrix for the final subdata picked using the extended IBOSS subdata selection procedure goes to infinity even with a fixed k. While we could not directly prove that the variance of the slope parameters goes to zero as $n \to \infty$ due to mathematical complicity, Theorem 3.2 is a significant step towards the ideal result. Despite the case studied here is simple, it shows the great potential on estimation efficiency of the proposed subdata selection procedure. The performance of the IBOSS strategy will be demonstrated numerically in various simulated scenarios in next section.

4. Simulation settings and result

In this section, the IBOSS procedure is evaluated in various distributions of Z_i. The distributions used to generate Z_i's are listed below.

- MzNormal: Multivariate-normal distribution with mean vector $u = (0, \cdots, 0)^T$

and variance covariance matrix $\mathbf{\Sigma}$, where $\Sigma_{ij} = 0.5$ if $i \neq j$ and $\Sigma_{ij} = 1$ if $i = j$.

- NzNormal: Multivariate-normal distribution with mean vector $u = (1, \cdots, 1)^T$ and variance covariance matrix $\mathbf{\Sigma}$ as defined above.

- MixNormal: Mixed normal distribution $\frac{1}{2}\mathcal{N}(u, \mathbf{\Sigma}) + \frac{1}{2}\mathcal{N}(-u, \mathbf{\Sigma})$, where $u = (1, \cdots, 1)^T$ and $\mathbf{\Sigma}$ is the same as defined above.

- T3: Multivariate T distribution with 3 degrees of freedom with location parameter $u = (0, \cdots, 0)^T$ and shape parameter matrix $\mathbf{\Sigma}/10$, where $\mathbf{\Sigma}$ is the same as defined above.

These distributions were used in [13] to demonstrate the efficiency of the mVc subsampling strategy. The performance of the new IBOSS strategy is compared with the uniform sampling strategy and the mVc sampling strategy in all scenarios. To be consistent with the simulation settings in [13], we assume that there is no intercept parameter β_0 unless otherwise specified and set the of dimension of $\boldsymbol{\beta}$ to be the same as in [13].Since we know the true value $\boldsymbol{\beta}$ in these simulation studies, the mean squared error is used to evaluate the deviation as well as the biasness of the estimated $\hat{\boldsymbol{\beta}}$ from $\boldsymbol{\beta}$. All the log(MSE) shown in figures in this section is based on a $\log_{10}$ scale. For all simulations, subsamples of size k are used for the mVc algorithm and the new IBOSS algorithm, while subsamples of size $k + k_0$ are used for uniform sampling to account for the sampling cost in the first stage. For most of the scenarios, the performance of the full data estimator is not presented in figures as we mainly concern about the relative performance among different subsampling algorithms. In subsection 4.1 and subsection 4.2, for simulation cases with MzNormal, MixNormal and T3 distributions, the δ criterion for first stage selection of the new algorithm is fixed at 0.5. For simulation scenarios with NzNormal distribution, the δ is fixed at 2.5 as the distribution of $c = X^T\boldsymbol{\beta}$ is shifted to the right of zero under this type of distribution. In subsection 4.3, the full data sample size n is fixed at 500000 with initial $k_0 = 1000$ and subsample size $k = 5000$. Multiple δ criterions are tested according to the proportion of data kept in the first stage when implementing the new algorithm. Some insights on the selection of a proper δ value are drawn upon simulation results. In subsection 4.4, the computational costs of different algorithms are compared under various simulation scenarios.

4.1. A fixed n with varying k

In this scenario, the full data sample size is $n = 500000$ and the X_i's are 7×1 vectors generated from the distribution settings mentioned above. The true parameter $\boldsymbol{\beta}_0 = (0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)^T$. In the first stage, a subdata of size $k_0 = 1000$ will be randomly drawn. In the second stage, the subsample size k is chosen to be 1000, 2000, 5000, and 8000. We repeat the simulations 1000 times to calculate empirical MSEs.

Results for this scenario are shown in Figure 1. The x-axis represents the size of the subdata in the second stage and the y-axis represents $\log_{10}$ of empirical MSEs. Under all the distribution settings, the performance of the new subsampling strategy is compared with the uniform sampling strategy and the mVc subsampling strategy. Under MzNormal and MixNormal settings, the new IBOSS strategy performs better than the mVc strategy, which performs better than the uniform sampling strategy. Under NzNormal setting, the new IBOSS strategy and the mVc strategy perform similarly, and they are both better than the uniform sampling strategy. For the T3 distribution, the performance of the new IBOSS strategy is significant better. The empirical MSE from the extended IBOSS can be less than $\frac{1}{6}$ of that from the mVc strategy or the uniform sampling strategy.

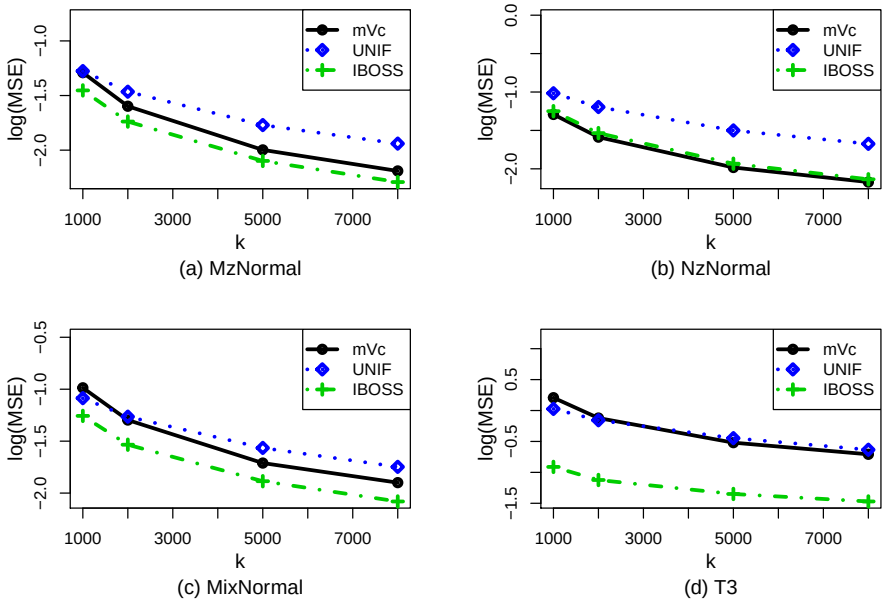

Figure 1: MSEs for different strategies with a fixed n and varying k.

4.2. A fixed k with varying n

Intuitively, the new IBOSS approach tries to pick data points close to the D-optimal data points for the logistic regression model. As we have showed in theory, the subdata should become more informative as n grows larger. This indicates that the new algorithm should experience an improvement in estimation accuracy as n increases, even when the size of subdata k is fixed. In this subsection, simulations are conducted to see whether we can observe this trend. Here we use the same distribution and parameter settings from the previous scenario except that we increase the dimension of X_i's to be 9. We set the true $\boldsymbol{\beta}_0$ to be a 9 dimensional vector of 0.5 and choose n to be 50000, 200000, 800000, and 3200000. The subsample size is fixed at k =5000. Log_{10} of empirical MSEs are shown in Figure 2. We also provide the results from using the full data as comparisons. In this figure, for T3 distributions, a clear trend of increasing estimation efficiency can be detected for the new IBOSS strategy as data size grows larger. The trend is more clear if we use the original MSE as y-axis. Here we keep the log transformation for the sake

of consistence.

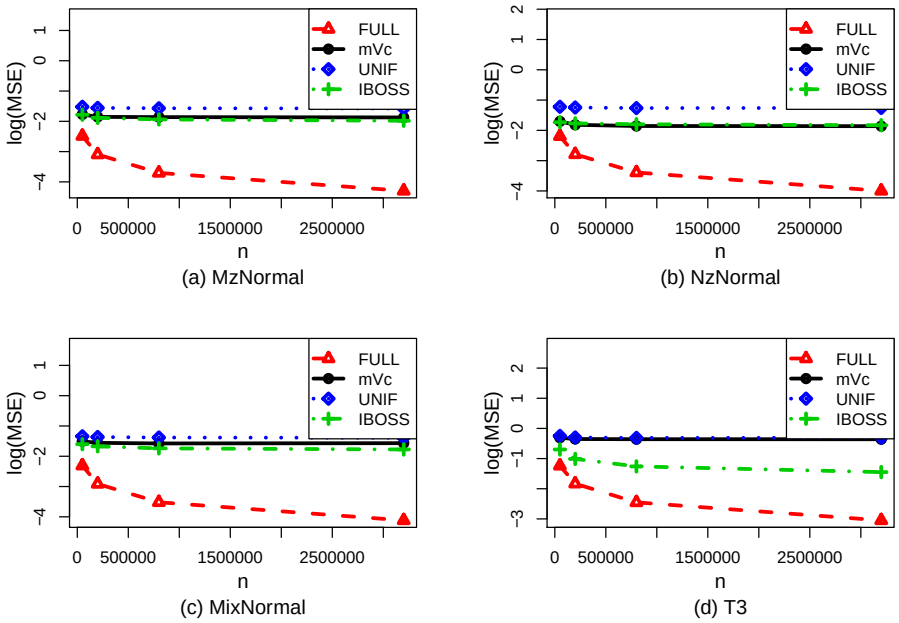

Figure 2: Results of the subsampling algorithms with different full data sizes

4.3. Some insights on determining δ

In all the simulation scenarios above, δ is pre-specified for the first stage filtering of data. The value of δ, along with full data size n and distributions to generate Z_i, affect the estimation accuracy of the new algorithm. The selection of the δ value in the new algorithm need more theoretical investigations on the asymptotic distribution of extreme order statistics $Z_{(1)j}$ and $Z_{(n)j}$, as well as many other quantities. [12] studied the asymptotic property of $Z_{(1)j}$ and $Z_{(n)j}$ when implementing the IBOSS with in linear regression. Unfortunately, there is no theoretical result to help select a proper δ for a specific dataset with logistic regression. However, we can use simulations to get some insights on selecting a δ value. Here, we still use the distribution settings in the beginning of this section. The slope parameter $(\beta_1, \cdots, \beta_9)$ is again set as a 9 dimensional vector of 0.5. We consider 3 cases for the distribution of $c = X^T\boldsymbol{\beta}$ described below.

- Balanced case: Intercept β_0 is not include in the model. In this case, c is centered around 0 for MixNormal, MzNormal and T3 distributions, and it is centered around some positive number depending on the dimension of the covariate for the NzNormal distribution.

- Right shift case: Intercept $\beta_0 = 2$ is include in the model. For this case, the distribution center for c is shifted to the right hand side of the center of the balanced case.

- Left shift case: Intercept $\beta_0 = -2$ is include in the model. For this case, the distribution center for c is shifted to the left hand side of the center of the balanced case.

For all these three cases, we pick certain δ to keep certain percentages of the full data after the first stage of the IBOSS procedure. The percentage tested are $0.25, 0.35, 0.45, 0.55, 0.65$, and 0.75. The full data size is $n = 500000$. For each setting in this scenario, 1000 repetitions of the simulation are used to calculate the empirical MSEs of the slope parameters. The results for all of these cases are shown in the following Figure 3, Figure 4, and Figure 5.

From Figures 3, 4, and 5, one can find that, regardless the distributions of covariates Z_i and the shift of the center of c, the 25% to 35% extraction rate range for first stage seems to have a good and robust performance. This indicates that when we try to pick a proper value for δ, we can pick a δ which filtered out around 65% to 75% percent of the full data and kept around 25% to 35% according to the distance criteria we set up. It is also worth noting that for the T3 distribution, the different quantiles tested generally perform well and a slight gain on accuracy can be obtained as the percentage increases.

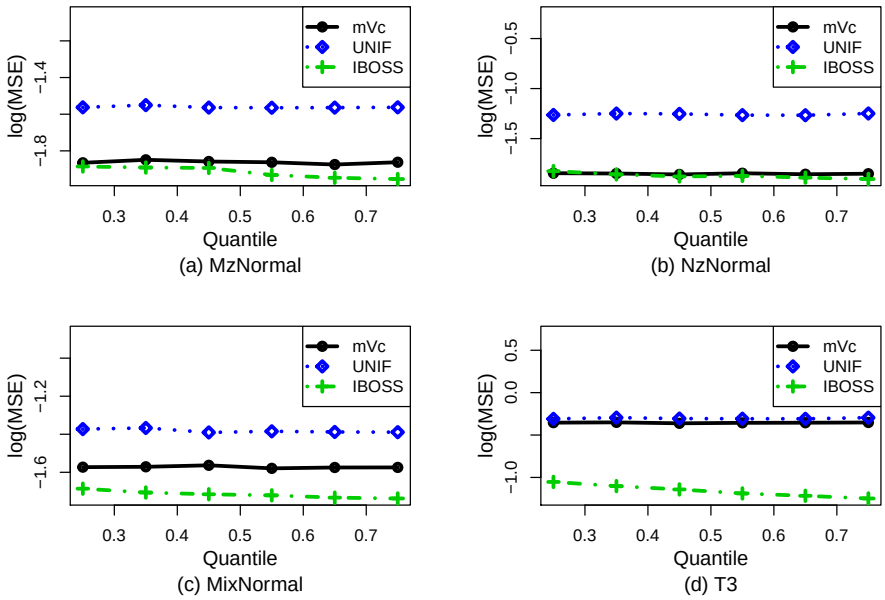

Figure 3: Results of performance under different percentages for the new algorithm: Balanced case

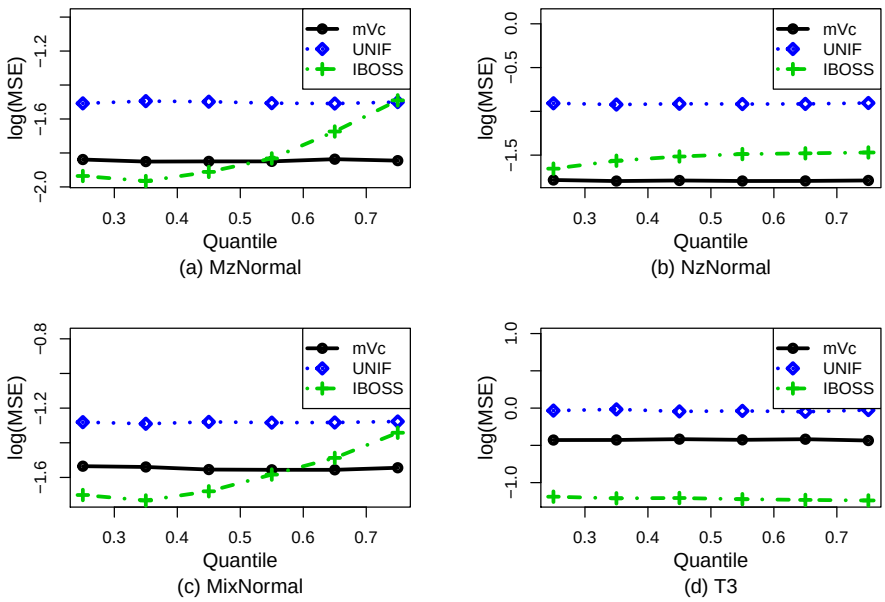

Figure 4: Results of performance under different percentages for the new algorithm: Right skewed case

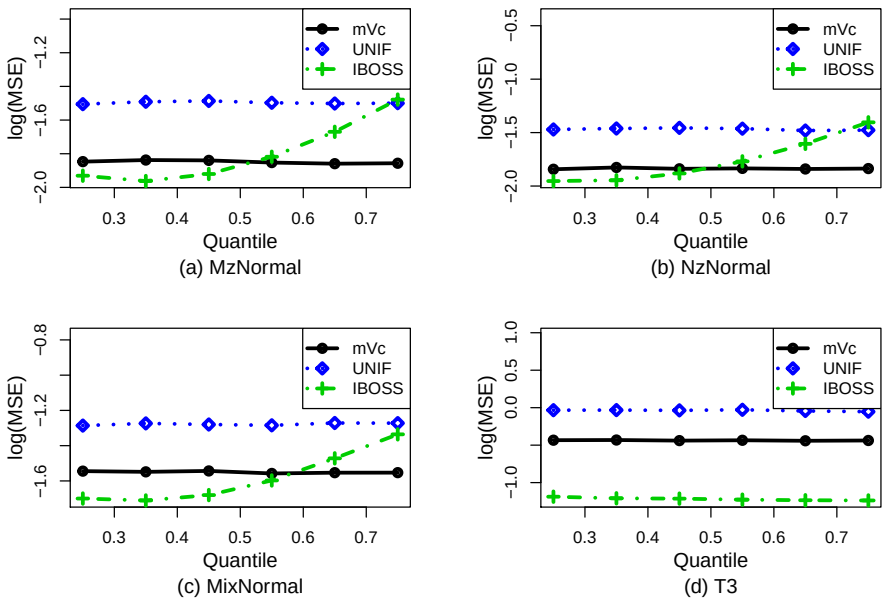

Figure 5: Results of performance under different percentages for the new algorithm: Left skewed case

4.4. Computational costs

The proposed procedure is also competitive compared to the mVc procedure in terms of computational cost. For varieties of combinations of data size n and dimension m, the average computational times (in seconds) with 200 repetitions under the T3 distribution scenario are shown in Table 1. We can easily see that uniform sampling has the smallest computational cost in all the scenarios. The computational cost of the new IBOSS algorithm is slightly better than mVc algorithm and they both use much less time compared with the full data estimator. The reduction on computational cost for mVc and the new algorithm is more and more apparent as size n or dimension m goes larger and larger.

5. Discussion

In this paper, we study subdata selection under logistic regression models. We show that, for random sampling-based strategy, such as the mVc strategy and the uniform sampling,

Table 1: Computational Cost of Subsampling Procedures (Seconds)

n	m	FULL	mVc	UNIF	IBOSS
100000	100	3.9097	0.6454	0.4492	0.5431
200000	100	8.3816	0.9973	0.4553	0.8254
500000	100	22.7350	2.1712	0.4591	1.7737
500000	25	1.9599	0.9151	0.0437	0.5177
500000	50	6.4434	1.3291	0.1294	1.0648
500000	100	22.7350	2.1712	0.4591	1.7737

the information in the subdata is bounded by the size of subdata. A novel information-based optimal subdata selection approach is proposed. For the new approach, we show that at least one eigenvalue of the information matrix goes to infinity when full data size increases even when the subdata size is fixed. The results demonstrate that the new approach effectively addresses the trade off between the computation complexity and statistical efficiency under logistic regression models.

Due to the intractable mathematical complication, the result is under the assumption that the covariates are from bivariate normal distribution. It indicates that such result likely holds for general cases. However, it would be rather challenging, if not impossible, to derive such asymptotic property due to the complexity.

The upper bound in Theorem 3.1 is based on the assumption that $c = X^T\boldsymbol{\beta}$ is unbounded. If we have more information about the range of c, the upper bound can be further improved. For example, if the dataset shows that $c = X^T\boldsymbol{\beta}$ can only take positive values, then the characterization of optimal subdata should be different. Consequently, the algorithm of selecting the subdata would also be different. An interesting question is to propose a general algorithm of selecting an informative subdata for arbitrary data set.

In the age of information, big data with complex structures are obtained via various sources. While they provide us more valuable information, the computational costs of analyzing them can be expensive and sometimes out of capacity. Efforts for developing subdata selection strategies have greatly improved the quality of the subdata which helps save tremendous computational costs. However, subdata selection strategies for

nonlinear models, like the logistic regreesion model considered in this paper, are still not well developed. We hope that this work can stimulate more ideas and attract more researches in this direction.

Acknowledgments

The authors are grateful for many insightful comments and suggestions from an anonymous referee, an associate editor, and editor, which helped to improve the article. Wang's research was supported by NSF grant DMS-1812013 and Yang's research was supported by NSF grant DMS-1811291.

References

[1] C. A. Mattmann, A. Hart, C. L., J. Lazio, S. Khudikyan, D. Jones, R. Preston, T. Bennett, B. Bulter, D. Harland, B. Glendenning, J. Kern, J. Robnett, Scalable data mining, archiving, and big data management for the next generation astronomical telescopes, in: W.-C. Hu, N. Kaabouch (Eds.), Big data management, technologies, and applications, IGI Global, 2014, pp. 196–221.

[2] N. Lin, R. Xi, Aggregated estimating equation estimation, Statistics and Its Interface 4 (2011) 73–83.

[3] X. Chen, M. ge Xie, A split-and-conquer approach for analysis of extraordinarily large data, Statistica Sinica 24 (2014) 1655–1684.

[4] E. D. Schifano, J. Wu, C. Wang, J. Yan, M.-H. Chen, Online updating of statistical inference in the big data setting, Technometrics 58 (3) (2016) 393–403.

[5] NSF, Transdisciplinary research in principles of data science ((tripods)), National Science Foundation.

[6] D. N. Politis, J. P. Romano, M. Wolf, Subsampling, Springer Science & Business Media, 1999.

[7] B. Efron, Bootstrap methods: Another look at the jackknife, The annals of Statistics 7(1) (1979) 1–26.

[8] P. J. Bickel, F. Götze, W. R. van Zwet, Resampling fewer than n observations: Gains, losses, and remedies for losses, Statistica Sinica 7 (1997) 1–31.

[9] A. Kleiner, A. Talwalkar, P. Sarkar, M. I. Jordan, A scalable bootstrap for massive data, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76(4) (2014) 795–816.

[10] F. Liang, Y. Cheng, Q. Song, J. Park, P. Yang, A resampling-based stochastic approximation method for analysis of large geostatistical data, Journal of the American Statistical Association 108(501) (2013) 325—-339.

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

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

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

[14] J. Kiefer, Optimum experimental designs, Journal of the Royal Statistical Society. Series B (Methodological) 21(2) (1959) 272—-319.

[15] A. A. Yildirim, C. Özdogan, D. Watson, Parallel data reduction techniques for big datasets, in: W.-C. Hu, N. Kaabouch (Eds.), Big data management, technologies, and applications, IGI Global, 2014, pp. 72–93.

[16] D. W. Hosmer, S. Lameshow, Applied logistic regression, Join Wiley and Sons, New York, 2000.

[17] C. Gourieroux, A. Monfort, Asymptotic properties of the maximum likelihood estimator in dichotomous models, Journal of Econometrics 17(1) (1981) 83–97.

[18] L. Nordberg, Asymptotic normality of maximum likelihood estimators based on independent, unequally distributed observations in exponential family models, Scand J Statist 24 (1980) 1655–1684.

Qianshun Cheng

Bayer(Monsanto) Company

E-mail: chengqianshun1@gmail.com

Haiying Wang

University of Conneticut

E-mail: haiying.wang@uconn.edu

Min Yang

University of Illinois at Chicago

E-mail: myang2@uic.edu

Supplemental Material

Qianshun Cheng, Haiying Wang and Min Yang

Monsanto(Bayer) Company, University of Connecticut and University of Illinois at Chicago

1. Theorem and Proof

1.1. Proof of Theorem 2.1

Note that $\mathbf{M_X} = \frac{1}{n}\sum_{i=1}^{n}\Psi(\hat{c}_i)X_iX_i^T$, $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}^0$ in probability, and $\Psi(\hat{c}_i)$ is continuous and bounded. Thus, from Lemma 1 of [1] and the law of large numbers, we know

$$\begin{aligned}\mathbf{M_X} = \frac{1}{n}\sum_{i=1}^{n}\Psi(c_i^0)X_iX_i^T + o_P(1) &= E\Big\{p_1(\boldsymbol{\beta}^0)(1-p_1(\boldsymbol{\beta}^0))X_1X_1^T\Big\} + o_P(1)\\ &= E\Big\{(y_1 - p_1(\boldsymbol{\beta}^0))^2X_1X_1^T\Big\} + o_P(1)\\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i(\boldsymbol{\beta}^0))^2X_iX_i^T + o_P(1)\\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i(\hat{\boldsymbol{\beta}}))^2X_iX_i^T + o_P(1),\end{aligned}$$

where $c^0 = X_i^T\boldsymbol{\beta}^0$ and $\boldsymbol{\beta}^0$ is the true parameter. Therefore, by the definitions of $\mathbf{M_X}$ and $\mathbf{V}_c$, the inverse of matrix $\mathbf{V}$ is

$$\begin{aligned}\mathbf{V}^{-1} = k&\left(\sum_{i=1}^{n}(y_i - p_i(\hat{\boldsymbol{\beta}}))^2X_iX_i^T\right)\left(\sum_{i=1}^{n}\frac{(y_i - p_i(\hat{\boldsymbol{\beta}}))^2X_iX_i^T}{\pi_i}\right)^{-1}\\ &\times\left(\sum_{i=1}^{n}(y_i - p_i(\hat{\boldsymbol{\beta}}))^2X_iX_i^T\right)\{1 + o_P(1)\}\end{aligned}$$

$$= k\left[(y_1 - p_1(\hat{\boldsymbol{\beta}}))X_1\sqrt{\pi_1}, \cdots, (y_n - p_n(\hat{\boldsymbol{\beta}}))X_n\sqrt{\pi_n}\right] \begin{bmatrix} \frac{(y_1-p_1(\hat{\boldsymbol{\beta}}))X_1^T}{\sqrt{\pi_1}} \\ \vdots \\ \frac{(y_n-p_n(\hat{\boldsymbol{\beta}}))X_n^T}{\sqrt{\pi_n}} \end{bmatrix}$$

$$\times \left(\left[\frac{(y_1 - p_1(\hat{\boldsymbol{\beta}}))X_1}{\sqrt{\pi_1}}, \cdots, \frac{(y_n - p_n(\hat{\boldsymbol{\beta}}))X_n}{\sqrt{\pi_n}} \right] \begin{bmatrix} \frac{(y_1-p_1(\hat{\boldsymbol{\beta}}))X_1^T}{\sqrt{\pi_1}} \\ \vdots \\ \frac{(y_n-p_n(\hat{\boldsymbol{\beta}}))X_n^T}{\sqrt{\pi_n}} \end{bmatrix} \right)^{-1}$$

$$\times \left[\frac{(y_1 - p_1(\hat{\boldsymbol{\beta}}))X_1}{\sqrt{\pi_1}}, \cdots, \frac{(y_n - p_n(\hat{\boldsymbol{\beta}}))X_n}{\sqrt{\pi_n}} \right]$$

$$\times \begin{bmatrix} (y_1 - p_1(\hat{\boldsymbol{\beta}}))X_1^T\sqrt{\pi_1} \\ \vdots \\ (y_n - p_n(\hat{\boldsymbol{\beta}}))X_n^T\sqrt{\pi_n} \end{bmatrix} \{1 + o_P(1)\}$$

Set $\mathbf{W} = \mathbf{diag}(\frac{(y_1-p_1(\hat{\boldsymbol{\beta}}))}{\sqrt{\pi_1}}, \cdots, \frac{(y_n-p_n(\hat{\boldsymbol{\beta}}))}{\sqrt{\pi_n}})$. Then $\mathbf{V}^{-1}$ can be re-written as

$$\mathbf{V}^{-1} = k\left[(y_1 - p_1(\hat{\boldsymbol{\beta}}))X_1\sqrt{\pi_1}, \cdots, (y_n - p_n(\hat{\boldsymbol{\beta}}))X_n\sqrt{\pi_n}\right] WX(X^TW^2X)^{-1}$$

$$\times X^TW \begin{bmatrix} (y_1 - p_1(\hat{\boldsymbol{\beta}}))X_1^T\sqrt{\pi_1} \\ \vdots \\ (y_n - p_n(\hat{\boldsymbol{\beta}}))X_n^T\sqrt{\pi_n} \end{bmatrix} \{1 + o_P(1)\}$$

$$= k\left[(y_1 - p_1(\hat{\boldsymbol{\beta}}))X_1\sqrt{\pi_1}, \cdots, (y_n - p_n(\hat{\boldsymbol{\beta}}))X_n\sqrt{\pi_n}\right]$$

$$\times \mathbf{Proj}_{WX} \begin{bmatrix} (y_1 - p_1(\hat{\boldsymbol{\beta}}))X_1^T\sqrt{\pi_1} \\ \vdots \\ (y_n - p_n(\hat{\boldsymbol{\beta}}))X_n^T\sqrt{\pi_n} \end{bmatrix} \{1 + o_P(1)\},$$

where $X = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix}$. Define $B_{WX} = \begin{bmatrix} w_1 X_1^T & \cdots \\ \vdots & \\ \cdots & w_n X_n^T \end{bmatrix}$, where w_i is the i-th diagonal element in W. Clearly the columns of WX are in the column space of B_{WX}. Thus, we have

$$\begin{aligned}
\mathbf{V}^{-1} &\leq k[(y_1 - p_1(\hat{\boldsymbol{\beta}}))X_1\sqrt{\pi_1}, \cdots, (y_n - p_n(\hat{\boldsymbol{\beta}}))X_n\sqrt{\pi_n}] \\
&\quad \times \mathbf{Proj}_{B_{WX}} \begin{bmatrix} (y_1 - p_1(\hat{\boldsymbol{\beta}}))X_1^T\sqrt{\pi_1} \\ \vdots \\ (y_n - p_n(\hat{\boldsymbol{\beta}}))X_n^T\sqrt{\pi_n} \end{bmatrix} \{1 + o_P(1)\} \\
&= k[(y_1 - p_1(\hat{\boldsymbol{\beta}}))\sqrt{\pi_1}X_1, \cdots, (y_n - p_n(\hat{\boldsymbol{\beta}}))\sqrt{\pi_n}X_n] \\
&\quad \times \begin{bmatrix} X_1^T(X_1X_1^T)^{-1}X_1 & \cdots \\ \vdots & \\ \cdots & X_n^T(X_nX_n^T)^{-1}X_n \end{bmatrix} \\
&\quad \times \begin{bmatrix} X_1^T(y_1 - p_1(\hat{\boldsymbol{\beta}}))\sqrt{\pi_1} \\ \vdots \\ X_n^T(y_n - p_n(\hat{\boldsymbol{\beta}}))\sqrt{\pi_n} \end{bmatrix} \{1 + o_P(1)\} \\
&= k\left(\sum_{i=1}^n (y_i - p_i(\hat{\boldsymbol{\beta}}))^2 \pi_i X_i X_i^T\right)\{1 + o_P(1)\} \\
&\leq \frac{k}{4}\left(\sum_{i=1}^n \pi_i X_i X_i^T\right)\{1 + o_P(1)\}
\end{aligned}$$

By taking inverse, we have

$$\mathbf{V} \geq \frac{4\{1 + o_P(1)\}}{k}\left(\sum_{i=1}^n \pi_i X_i X_i^T\right)^{-1}.$$

Here the inequalities are under the context of Lowering ordering.

1.2. Proof of Theorem 2.3

Applying theorem 2.1 and $\pi_i^{mVc} = \frac{|y_i - p_i(\hat{\boldsymbol{\beta}})|\|X_i\|}{\sum_{j=1}^n |y_j - p_j(\hat{\boldsymbol{\beta}})|\|X_j\|}$, we have

$$\begin{aligned} k^{-1}\mathbf{V}^{-1} &= \sum_{i=1}^n \frac{|y_i - p_i(\hat{\boldsymbol{\beta}})|\|X_i\|}{\sum_{j=1}^n |y_j - p_j(\hat{\boldsymbol{\beta}})|\|X_j\|} X_i X_i^T \\ &= \frac{1}{\sum_{j=1}^n |y_j - p_j(\hat{\boldsymbol{\beta}})|\|X_j\|} \sum_{i=1}^n \|X_i\||y_i - p_i(\hat{\boldsymbol{\beta}})| X_i X_i^T \qquad (1.1) \\ &\leq \frac{1}{\sum_{j=1}^n |y_j - p_j(\hat{\boldsymbol{\beta}})|\|X_j\|} \sum_{i=1}^n \|X_i\| X_i X_i^T \end{aligned}$$

Note that $n^{-1}\sum_{i=1}^n \|X_i\| X_i X_i^T \to E(\|X_1\| X_1 X_1^T)$ almost surely by the strong law of large numbers if $E(\|X_1\| X_1 X_1^T) < \infty$. From Lemma 1 of [1], $n^{-1}\sum_{j=1}^n |y_j - p_j(\hat{\boldsymbol{\beta}})|\|X_j\| = E(|y_1 - p_1(\boldsymbol{\beta}^0)|\|X_1\|) + o_P(1)$. Thus, (1.1) indicates that

$$k^{-1}\mathbf{V}^{-1} \leq \frac{1}{E(|y_1 - p_1(\boldsymbol{\beta}^0)|\|X_1\|)} E(\|X_1\| X_1 X_1^T) + o_P(1). \qquad (1.2)$$

The result in theorem follows by letting $a = E(|y_1 - p_1(\boldsymbol{\beta}^0)|\|X_1\|)$.

1.3. Proof of Theorem 3.1

With a little bit notation abuse, let (X_i, n_i), $i = 1, \ldots, s$ denote the distinct covariate vectors in the selected subset and the corresponding replications. Notice that $\sum_{i=1}^s n_i = k$. We consider a transformation on one of covariate. Without loss of generation, we transform the last covariate z_{im}. Let $C_i^T = (1, z_{i1}, \ldots, z_{i,m-1}, c_i)$, where $c_i = \beta_0 + \beta_1 z_{i1} + \ldots + \beta_m z_{im}$. Then we have $X_i = A(\boldsymbol{\beta}) C_i$, where

$A(\boldsymbol{\beta}) = \begin{pmatrix} I_m & 0 \\ A_1(\boldsymbol{\beta}) & 1/\beta_m \end{pmatrix}$ and $A_1(\boldsymbol{\beta}) = (-\beta_0/\beta_m, -\beta_1/\beta_m, \ldots, -\beta_{m-1}/\beta_m)$. By standard method, the information matrix for $\boldsymbol{\beta}$ under (2.1) can be written as

$$\begin{aligned} I_{\boldsymbol{\alpha}} &= k \sum_{i=1}^s \omega_i X_i \Psi(c_i) (X_i)^T \\ &= kA(\boldsymbol{\beta}) \left(\sum_{i=1}^s \omega_i C_i \Psi(c_i) (C_i)^T \right) A^T(\beta), \end{aligned} \qquad (1.3)$$

where $\Psi(c) = \frac{e^c}{(1+e^c)^2}$ and $\omega_i = n_i/k$. Notice we can do the similar transformations on the other covariates. Consequently, in the view of (1.3), it suffices to show that

$$|\sum_{i=1}^{s} \omega_i C_i \Psi(c_i)(C_i)^T| \leq \left\{ \frac{(c^*)^2 \Psi(c^*)^{m+1}}{4^{m-1}} \prod_{j=1}^{m-1} (Z_{(n)j} - Z_{(1)j})^2 \right\}. \tag{1.4}$$

For each covariates, we have $z_{ij} \in [Z_{(1),j}, Z_{(n),j}]$, $j = 1, \ldots, m$. Our conclusion follows if we can prove (1.4) when there is no constrain on z_{im}, i.e., $z_{im} \in (-\infty, +\infty)$. By Theorem 2 of [2], $|\sum_{i=1}^{s} \omega_i C_i \Psi(c_i)(C_i)^T|$ is maximized when $s = 2^m$, $X_i^* = (1, (Z_i^*)^T)^T$ and $\omega_i = 1/2^m$, $i = 1, \ldots, s$. Next we shall show that the maximum value is the right hand side of (1.4). Let

$$B = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 \\ -\frac{Z_{(n),1}+Z_{(1),1}}{Z_{(n),1}-Z_{(1),1}} & \frac{2}{Z_{(n),1}-Z_{(1),1}} & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{Z_{(n),m-1}+Z_{(1),m-1}}{Z_{(n),m-1}-Z_{(1),m-1}} & 0 & \ldots & \frac{2}{Z_{(n),m-1}-Z_{(1),m-1}} & 0 \\ 0 & 0 & \ldots & 0 & 1 \end{pmatrix}. \tag{1.5}$$

It can be verified that

$$B \left(\sum_{i=1}^{2^m} \frac{1}{2^m} C_i^* \Psi(c^*)(C_i^*)^T \right) B' = \Psi(c^*) \begin{pmatrix} I_m & 0 \\ 0 & (c^*)^2 \end{pmatrix}. \tag{1.6}$$

The desired conclusion follows by some routine algebra.

1.4. Proof of Theorem 3.2

Let $\mathcal{F}_{1j}$ and $\mathcal{F}_{2j}$ be the distribution function of z_{ij} and $-z_{ij}$ conditional on $c_i = X_i^T \boldsymbol{\beta} \in \mathbf{C}$ respectively, $j = 1, 2$. We first derive $f_{1j}(z)$ and $f_{2j}(z)$, the pdf of $\mathcal{F}_{1j}$ and $\mathcal{F}_{2j}$.

Since $Z_i \sim \mathcal{N}(\mu, \boldsymbol{\Sigma})$, then $Z_i^t = (x_{i1}, X_i^T \boldsymbol{\beta}) \sim \mathcal{N}((u_1, u_c)^T, \boldsymbol{\Sigma}_t)$. According to the first stage procedure of new IBOSS algorithm,

$$C = \{c \mid |c - c^*| < \delta \text{ or } |c + c^*| < \delta\} = (a, b) \cup (-b, -a)$$

for some constants $a < b$.

All the proof works here is built with cases that $(a, b) \cap (-b, -a) = \emptyset$. For cases when $(a, b) \cap (-b, -a) \neq \emptyset$, one can rewrite $(a, b) \cap (-b, -a)$ as (a', b') for some constant a', b' and prove the same result using exactly the same framework.

Let σ_c^2 be $Var(c_i)$, σ_1^2 be $Var(z_{i1})$ with $\sigma_1, \sigma_c > 0$, and ρ_1 be the correlation coefficient between $c = X^T\boldsymbol{\beta}$ and z_1. By the assumption that Σ is nonsingular as well as $\beta_1 \neq 0$ and $\beta_2 \neq 0$, we have $|\rho_1| < 1$.

Utilizing the conditional distribution forms derived in [3], we can directly obtain that

$$f_{11}(z) = \frac{\frac{1}{\sigma_1} e^{-\frac{(z-u_1)^2}{2\sigma_1^2}} g_{11}(z)}{\Phi(\frac{b-u_c}{\sigma_c}) - \Phi(\frac{a-u_c}{\sigma_c}) + \Phi(\frac{-a-u_c}{\sigma_c}) - \Phi(\frac{-b-u_c}{\sigma_c})}$$

and

$$f_{21}(z) = \frac{\frac{1}{\sigma_1} e^{-\frac{(z+u_1)^2}{2\sigma_1^2}} g_{12}(z)}{\Phi(\frac{b+u_c}{\sigma_c}) - \Phi(\frac{a+u_c}{\sigma_c}) + \Phi(\frac{-a+u_c}{\sigma_c}) - \Phi(\frac{-b+u_c}{\sigma_c})},$$

where

$$g_{11}(z) = \Phi(\frac{\frac{b-u_c}{\sigma_c} - \rho_1\frac{z-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}}) - \Phi(\frac{\frac{a-u_c}{\sigma_c} - \rho_1\frac{z-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}}) + \Phi(\frac{\frac{-a-u_c}{\sigma_c} - \rho_1\frac{z-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}}) - \Phi(\frac{\frac{-b-u_c}{\sigma_c} - \rho_1\frac{z-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})$$

and

$$g_{12}(z) = \Phi(\frac{\frac{b+u_c}{\sigma_c} - \rho_1\frac{z+u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}}) - \Phi(\frac{\frac{a+u_c}{\sigma_c} - \rho_1\frac{z+u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}}) + \Phi(\frac{\frac{-a+u_c}{\sigma_c} - \rho_1\frac{z+u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}}) - \Phi(\frac{\frac{-b+u_c}{\sigma_c} - \rho_1\frac{z+u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}}).$$

Similarly, we can obtain

$$f_{12}(z) = \frac{\frac{1}{\sigma_2} e^{-\frac{(z-u_2)^2}{2\sigma_2^2}} g_{21}(z)}{\Phi(\frac{b-u_c}{\sigma_c}) - \Phi(\frac{a-u_c}{\sigma_c}) + \Phi(\frac{-a-u_c}{\sigma_c}) - \Phi(\frac{-b-u_c}{\sigma_c})}$$

and

$$f_{22}(z) = \frac{\frac{1}{\sigma_2} e^{-\frac{(z+u_2)^2}{2\sigma_2^2}} g_{22}(z)}{\Phi(\frac{b+u_c}{\sigma_c}) - \Phi(\frac{a+u_c}{\sigma_c}) + \Phi(\frac{-a+u_c}{\sigma_c}) - \Phi(\frac{-b+u_c}{\sigma_c})},$$

where

$$g_{21}(z) = \Phi(\frac{\frac{b-u_c}{\sigma_c} - \rho_2 \frac{z-u_2}{\sigma_1}}{\sqrt{1-\rho_2^2}}) - \Phi(\frac{\frac{a-u_c}{\sigma_c} - \rho_2 \frac{z-u_2}{\sigma_1}}{\sqrt{1-\rho_2^2}}) + \Phi(\frac{\frac{-a-u_c}{\sigma_c} - \rho_2 \frac{z-u_2}{\sigma_1}}{\sqrt{1-\rho_2^2}}) - \Phi(\frac{\frac{-b-u_c}{\sigma_c} - \rho_2 \frac{z-u_2}{\sigma_1}}{\sqrt{1-\rho_2^2}})$$

and

$$g_{22}(z) = \Phi(\frac{\frac{b+u_c}{\sigma_c} - \rho_2 \frac{z+u_2}{\sigma_1}}{\sqrt{1-\rho_2^2}}) - \Phi(\frac{\frac{a+u_c}{\sigma_c} - \rho_2 \frac{z+u_2}{\sigma_1}}{\sqrt{1-\rho_2^2}}) + \Phi(\frac{\frac{-a+u_c}{\sigma_c} - \rho_2 \frac{z+u_2}{\sigma_1}}{\sqrt{1-\rho_2^2}}) - \Phi(\frac{\frac{-b+u_c}{\sigma_c} - \rho_2 \frac{z+u_2}{\sigma_1}}{\sqrt{1-\rho_2^2}}).$$

Now we investigate the information matrix of subdata from the new IBOSS algorithm. By implementing the new algorithm, we will pick the subdata in which $c_i = X_i^T \boldsymbol{\beta} \in C$. And for the two dimension case discussed here, the second stage procedure is to pick the data rows with largest $\left\lceil \frac{k}{2*p} \right\rceil$ values and smallest $\left\lceil \frac{k}{2*p} \right\rceil$ values from jth covariate sequentially to build the final subdata with size around k, where $j = 1, 2$.

With the remaining datalines $(X'_1, \cdots, X'_{n_1})$ after first stage procedure, denote the subdata we select from the j-th covariate as $(X'^{1j}, X'^{2j}, \cdots, X'^{\lceil \frac{k}{2*p} \rceil j})$ and $(X'^{1(j)}, \cdots, X'^{\lceil \frac{k}{2*p} \rceil (j)})$. $(X'^{1j}, \cdots, X'^{\lceil \frac{k}{2*p} \rceil j})$ represents the selected rows with the largest value and $(X'^{1(j)}, \cdots, X'^{\lceil \frac{k}{2*p} \rceil (j)})$ represents the selected rows with the smallest value on jth covariate in remaining data. Then

$$\begin{aligned} \mathbf{I}^{IBOSS} &= \sum_{i=1}^{n} \alpha_i \Psi(c_i) X_i X_i^T \\ &= \sum_{j=1}^{2} \sum_{i=1}^{\lceil \frac{k}{2*p} \rceil} \Psi(c'_{ij}) X'^{ij} (X'^{ij})^T + \sum_{j=1}^{2} \sum_{i=1}^{\lceil \frac{k}{2*p} \rceil} \Psi(c'_{i(j)}) X'^{i(j)} (X'^{i(j)})^T, \end{aligned} \tag{1.7}$$

where $c'_{ij} = (X'^{ij})^T\boldsymbol{\beta}$ and $c'_{i(j)} = (X'^{i(j)})^T\boldsymbol{\beta}$.

Next we focus on the explicit form for $X'^{ij} = (1, {z'_1}^{ij}, {z'_2}^{ij})$. Suppose ${z'_1}^{11}$ is the largest value among $(z'_{11}, \cdots, z'_{n_1 1})$ and z'_{i1} are independently generate from F_{11}. If we can prove that $\mathcal{F}_{11}$ belongs to the Gumbel type and $F_{11}^{-1}(1-\frac{1}{n_1}) \to \infty$ as $n_1 \to \infty$, then the distribution of ${z'_1}^{11}$ will satisfy

$$F_{{z'_1}^{11}}(a_{n_1} z + b_{n_1}) = e^{-e^{-z}} \tag{1.8}$$

when $n_1 \to \infty$, where $a_{n_1} = \frac{\sigma_1^2(1-\rho_1^2)}{F_{11}^{-1}(1-\frac{1}{n_1})}$ and $b_{n_1} = F_{11}^{-1}(1-\frac{1}{n_1})$.

By plugging in $z = \sqrt{F_{11}^{-1}(1-\frac{1}{n})}$ and $z = -\sqrt{F_{11}^{-1}(1-\frac{1}{n})}$ to (1.8), one can obtain $z'^{11}_1 = F_{11}^{-1}(1-\frac{1}{n_1}) + o(1)$. Then by theorem 2.8.1 and theorem 2.8.2 [4], we can derive

$${z'_1}^{i1} = F_{11}^{-1}(1-\frac{1}{n_1}) + o(1) \text{ for } i = 1, \cdots, \left\lceil \frac{k}{2*p} \right\rceil. \tag{1.9}$$

Similarly, one can derive

$${z'_2}^{i2} = F_{12}^{-1}(1-\frac{1}{n_1}) + o(1) \text{ for } i = 1, \cdots, \left\lceil \frac{k}{2*p} \right\rceil. \tag{1.10}$$

For $X'^{i(j)} = (1, {z'_1}^{i(j)}, {z'_2}^{i(j)})$, consider random variables $\mathbf{V}^T = (-1, v_{1j} = -z'_{1j}, \cdots, v_{n_1 j} = -z'_{n_1 j})$ and $c_v = \mathbf{V}^T\boldsymbol{\beta} = -c$. Thus v_{ij} follows distribution F_{2j} for $i = 1, \cdots, n_1$. Reorder $(v_{1j}, \cdots, v_{n_1 j})$ as $(v_j^1, \cdots, v_j^{n_1})$ in descending order. Similarly, by assuming that F_{2j} belongs to the Gumbel type and $F_{2j}^{-1}(1-\frac{1}{n_1}) \to \infty$, we can get explicit forms for v_i^1

$$v_j^1 = F_{2j}^{-1}(1-\frac{1}{n_1}) + o(1).$$

Again by theorem 2.8.1 and theorem 2.8.2 from [4],

$$-{z'_j}^{i(j)} = F_{2j}^{-1}(1-\frac{1}{n_1}) + o(1) \to \infty \text{ for } i = 1, \cdots, \left\lceil \frac{k}{2*p} \right\rceil,$$

which is equivalent to

$${z'_j}^{i(j)} = -F_{2j}^{-1}(1-\frac{1}{n_1}) + o(1) \text{ for } i = 1, \cdots, \left\lceil \frac{k}{2*p} \right\rceil. \tag{1.11}$$

By applying (1.9), (1.10), (1.11) and the fact that $c \in C$ is bounded, we have

$$z'_{j_2}{}^{ij_1} = -\frac{\beta_{j_1} F_{1j_1}^{-1}(1-\frac{1}{n_1})}{\beta_{j_2}} + O(1) \quad \text{for } i = 1, \cdots, \left\lceil \frac{k}{2*p} \right\rceil \quad \text{and } j_1 \neq j_2 \tag{1.12}$$

and

$$z'_{j_2}{}^{i(j_1)} = \frac{\beta_{j_1} F_{2j_1}^{-1}(1-\frac{1}{n_1})}{\beta_{j_2}} + O(1) \quad \text{for } i = 1, \cdots, \left\lceil \frac{k}{2*p} \right\rceil \quad \text{and } j_1 \neq j_2. \tag{1.13}$$

Let e denote the minimum value for $\Psi(\boldsymbol{\beta}) = p(\boldsymbol{\beta})(1-p(\boldsymbol{\beta})) = \frac{e^{X^T\boldsymbol{\beta}}}{(1+e^{X^T\boldsymbol{\beta}})^2}$ in range C. Thus apply (1.9), (1.10), (1.11), (1.12), (1.13) to (1.7), we can show that

$$\mathbf{I}^{\text{IBOSS}} \geq e \begin{pmatrix} k & \mathbf{I}_{12} \\ \mathbf{I}_{12}^T & \mathbf{I}_{22} \end{pmatrix}, \tag{1.14}$$

where the diagonal elements of $\mathbf{I}_{22}$ are

$$\left\lceil \frac{k}{2*p} \right\rceil \left((F_{11}^{-1}(1-\frac{1}{n_1}))^2 + (F_{21}^{-1}(1-\frac{1}{n_1}))^2 \right.$$
$$\left. + \left(\frac{\beta_2}{\beta_1}\right)^2 \left((F_{12}^{-1}(1-\frac{1}{n_1}))^2 + (F_{22}^{-1}(1-\frac{1}{n_1}))^2 \right) + o(F) \right)$$

and the off diagonal elements of $\mathbf{I}_{12}$ are

$$\left\lceil \frac{k}{2*p} \right\rceil \left((F_{12}^{-1}(1-\frac{1}{n_1}))^2 + (F_{22}^{-1}(1-\frac{1}{n_1}))^2 \right.$$
$$\left. + \left(\frac{\beta_1}{\beta_2}\right)^2 \left((F_{11}^{-1}(1-\frac{1}{n_1}))^2 + (F_{21}^{-1}(1-\frac{1}{n_1}))^2 \right) + o(F) \right).$$

Here $F = max(\{F_{ij}^{-1}(1-\frac{1}{n})\}) \to \infty$. Since $F_{lj}^{-1}(1-\frac{1}{n_1}) \to \infty$, for $l = 1, 2$ and $j = 1, 2$, then the diagonal elements of $\mathbf{I}_{22}$ goes to ∞.

Let $\lambda_1, \ldots, \lambda_3$ be the three eigenvalues of $\begin{pmatrix} k & \mathbf{I}_{12} \\ \mathbf{I}_{12}^T & \mathbf{I}_{22} \end{pmatrix}$. Then $\sum_{i=1}^{3} \lambda_i$ goes to ∞. This implies at least one of the three eigenvalues goes to ∞.

Now the remaining part is to prove, for $l = 1, 2$ and $j = 1, 2$,

- $\mathcal{F}_{lj}$ belongs to Gumbel type.

- $F_{lj}^{-1}(1-\frac{1}{n_1}) \to \infty$.

Here we will just show the framework to prove these assumptions with $\mathcal{F}_{11}$ and one can similarly prove them using similar framework for other $\mathcal{F}_{lj}$'s.

By [5], the necessary and sufficient condition for distribution $\mathcal{F}_{11}$ to be the Gumbel type is that

$$\lim_{t\to\infty} \frac{1-F_{11}(t+xr(t))}{1-F_{11}(t)} = e^{-x} \text{ for } x \in \Re, \tag{1.15}$$

where $r(t)$ is a positive function when t is big enough. Thus as long as (1.15) holds for F_{11}, the first assumption will holds for distribution $\mathcal{F}_{11}$.

Set $r(t) = \sigma_1^2(1-\rho_1^2)/(t-u_1)$. Then $r(t) > 0$ for t big enough and $\lim_{t\to\infty} r(t) = \lim_{t\to\infty} r'(t) = 0$. Thus

$$\begin{aligned}
\lim_{t\to\infty} \frac{1-F_{11}(t+xr(t))}{1-F_{11}(t)} &= \lim_{t\to\infty} \frac{f_{11}(t+xr(t))(1+xr'(t))}{f_{11}(t)} \\
&= \lim_{t\to\infty} \frac{e^{-\frac{(t+xr(t)-u_1)^2}{2\sigma_1^2}} g_{11}(t+xr(t))(1+xr'(t))}{e^{-\frac{(t-u_1)^2}{2\sigma_1^2}} g_{11}(t)} \\
&= \lim_{t\to\infty} \frac{e^{-\frac{(t+xr(t)-u_1)^2}{2\sigma_1^2}}}{e^{-\frac{(t-u_1)^2}{2\sigma_1^2}}} \lim_{t\to\infty} \frac{g_{11}(t+xr(t))}{g_{11}(t)}.
\end{aligned} \tag{1.16}$$

Consider $\lim_{t\to\infty} \frac{e^{-\frac{(t+xr(t)-u_1)^2}{2\sigma_1^2}}}{e^{-\frac{(t-u_1)^2}{2\sigma_1^2}}}$ first, one can directly derive that

$$\begin{aligned}
\lim_{t\to\infty} \frac{e^{-\frac{(t+xr(t)-u_1)^2}{2\sigma_1^2}}}{e^{-\frac{(t-u_1)^2}{2\sigma_1^2}}} &= \lim_{t\to\infty} e^{-\frac{(xr(t))^2}{2\sigma_1^2}} e^{-\frac{(xr(t)(t-u_1))}{\sigma_1^2}} \\
&= \lim_{t\to\infty} e^{-\frac{(xr(t)(t-u_1))}{\sigma_1^2}}.
\end{aligned} \tag{1.17}$$

Thus by plugging in $r(t) = \sigma_1^2(1-\rho_1^2)/(t-u_1)$, we can obtain

$$\lim_{t\to\infty} \frac{e^{-\frac{(t+xr(t)-u_1)^2}{2\sigma_1^2}}}{e^{-\frac{(t-u_1)^2}{2\sigma_1^2}}} = e^{-(1-\rho_1^2)x}. \tag{1.18}$$

Then to calculate $\lim_{t\to\infty}\frac{g_{11}(t+xr(t))}{g_{11}(t)}$, first consider

$$\lim_{t\to\infty}\frac{\Phi(\frac{\frac{b-u_c}{\sigma_c}-\rho_1\frac{t+xr(t)-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})-\Phi(\frac{\frac{a-u_c}{\sigma_c}-\rho_1\frac{a+xr(t)-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})}{\Phi(\frac{\frac{b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})-\Phi(\frac{\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})}.$$

One can derive that

$$\begin{aligned}
&\lim_{t\to\infty}\frac{\Phi(\frac{\frac{b-u_c}{\sigma_c}-\rho_1\frac{t+xr(t)-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})-\Phi(\frac{\frac{a-u_c}{\sigma_c}-\rho_1\frac{a+xr(t)-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})}{\Phi(\frac{\frac{b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})-\Phi(\frac{\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})}\\
&=\lim_{t\to\infty}\frac{\Phi'(\frac{\frac{b-u_c}{\sigma_c}-\rho_1\frac{t+xr(t)-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})-\Phi'(\frac{\frac{a-u_c}{\sigma_c}-\rho_1\frac{t+xr(t)-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})}{\Phi'(\frac{\frac{b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})-\Phi'(\frac{\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})}\\
&=\lim_{t\to\infty}\frac{e^{-\frac{(\frac{b-u_c}{\sigma_c}-\rho_1\frac{t+xr(t)-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}\frac{-\rho_1(1+xr'(t))}{\sqrt{1-\rho_1^2}\sigma_1}-e^{-\frac{(\frac{a-u_c}{\sigma_c}-\rho_1\frac{t+xr(t)-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}\frac{-\rho_1(1+xr'(t))}{\sqrt{1-\rho_1^2}\sigma_1}}{e^{-\frac{(\frac{b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}\frac{-\rho_1}{\sqrt{1-\rho_1^2}\sigma_1}-e^{-\frac{(\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}\frac{-\rho_1}{\sqrt{1-\rho_1^2}\sigma_1}}\\
&=\lim_{t\to\infty}\frac{e^{-\frac{(\frac{b-u_c}{\sigma_c}-\rho_1\frac{t+xr(t)-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}-e^{-\frac{(\frac{a-u_c}{\sigma_c}-\rho_1\frac{t+xr(t)-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}}{e^{-\frac{(\frac{b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}-e^{-\frac{(\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}}(1+xr'(t))\\
&=\lim_{t\to\infty}\frac{e^{-\frac{(\frac{b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}e^{-\frac{(\rho_1\frac{xr(t)}{\sigma_1})^2}{2(1-\rho_1^2)}}e^{\frac{(\frac{b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})(\rho_1\frac{xr(t)}{\sigma_1})}{(1-\rho_1^2)}}-e^{-\frac{(\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}e^{-\frac{(\rho_1\frac{xr(t)}{\sigma_1})^2}{2(1-\rho_1^2)}}e^{\frac{(\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})(\rho_1\frac{xr(t)}{\sigma_1})}{(1-\rho_1^2)}}}{e^{-\frac{(\frac{b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}-e^{-\frac{(\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}}(1+xr'(t)).
\end{aligned}$$

If $\rho_1>0$, one can show that

$$\begin{aligned}
\lim_{t\to\infty}\frac{e^{-\frac{(\frac{b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}}{e^{-\frac{(\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}}&=\lim_{t\to\infty}\frac{e^{-\frac{(\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}e^{-\frac{(\frac{b-a}{\sigma_c})^2}{2(1-\rho_1^2)}}e^{-\frac{(\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})(\frac{b-a}{\sigma_c})}{(1-\rho_1^2)}}}{e^{\frac{(\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}} \qquad (1.19)\\
&=\infty.
\end{aligned}$$

Thus

$$\lim_{t\to\infty}\frac{\Phi(\frac{\frac{b-u_c}{\sigma_c}-\rho_1\frac{t+xr(t)-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})-\Phi(\frac{\frac{a-u_c}{\sigma_c}-\rho_1\frac{a+xr(t)-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})}{\Phi(\frac{\frac{b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})-\Phi(\frac{\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})}$$

$$=\lim_{t\to\infty}(\frac{e^{-\frac{(\frac{b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}e^{-\frac{(\rho_1\frac{xr(t)}{\sigma_1})^2}{2(1-\rho_1^2)}}e^{\frac{(\frac{b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})(\rho_1\frac{xr(t)}{\sigma_1})}{(1-\rho_1^2)}}}{e^{-\frac{(\frac{b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}}$$

$$-\frac{e^{-\frac{(\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}e^{-\frac{(\rho_1\frac{xr(t)}{\sigma_1})^2}{2(1-\rho_1^2)}}e^{\frac{(\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})(\rho_1\frac{xr(t)}{\sigma_1})}{(1-\rho_1^2)}}}{e^{-\frac{(\frac{b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}})(1+xr'(t)) \tag{1.20}$$

$$=\lim_{t\to\infty}(e^{-\frac{(\rho_1\frac{xr(t)}{\sigma_1})^2}{2(1-\rho_1^2)}}e^{\frac{(\frac{b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})(\rho_1\frac{xr(t)}{\sigma_1})}{(1-\rho_1^2)}}$$

$$-\frac{e^{-\frac{(\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}e^{-\frac{(\rho_1\frac{xr(t)}{\sigma_1})^2}{2(1-\rho_1^2)}}e^{\frac{(\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})(\rho_1\frac{xr(t)}{\sigma_1})}{(1-\rho_1^2)}}}{e^{-\frac{(\frac{b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}})(1+xr'(t)).$$

Since $\lim_{t\to\infty} r(t)=0$, we have

$$\lim_{t\to\infty}e^{-\frac{(\rho_1\frac{xr(t)}{\sigma_1})^2}{2(1-\rho_1^2)}}e^{\frac{(\frac{b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})(\rho_1\frac{xr(t)}{\sigma_1})}{(1-\rho_1^2)}}=\lim_{t\to\infty}e^{-\frac{(-\rho_1\frac{t-u_1}{\sigma_1})(\rho_1\frac{xr(t)}{\sigma_1})}{(1-\rho_1^2)}}$$
$$=e^{-\rho_1^2x}. \tag{1.21}$$

and

$$\lim_{t\to\infty}e^{-\frac{(\rho_1\frac{xr(t)}{\sigma_1})^2}{2(1-\rho_1^2)}}e^{\frac{(\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})(\rho_1\frac{xr(t)}{\sigma_1})}{(1-\rho_1^2)}}=\lim_{t\to\infty}e^{-\frac{(-\rho_1\frac{t-u_1}{\sigma_1})(\rho_1\frac{xr(t)}{\sigma_1})}{(1-\rho_1^2)}}$$
$$=e^{-\rho_1^2x}. \tag{1.22}$$

Apply (1.19),(1.21), (1.22) to (1.20),

$$\lim_{t\to\infty}\frac{\Phi(\frac{\frac{b-u_c}{\sigma_c}-\rho_1\frac{t+xr(t)-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})-\Phi(\frac{\frac{a-u_c}{\sigma_c}-\rho_1\frac{a+xr(t)-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})}{\Phi(\frac{\frac{b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})-\Phi(\frac{\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})}$$

$$=\lim_{t\to\infty}(1+xr'(t))(e^{-\rho_1^2x}-e^{-\rho_1^2x}\lim_{t\to\infty}\frac{e^{-\frac{(\frac{a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}}{e^{-\frac{(\frac{b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1})^2}{2(1-\rho_1^2)}}}). \tag{1.23}$$

As $\lim_{t\to\infty} r'(t)=0$,

$$=e^{-\rho_1^2x}.$$

Similarly, we can prove that

$$\lim_{t\to\infty} \frac{\Phi(\frac{\frac{-a-u_c}{\sigma_c}-\rho_1\frac{t+xr(t)-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}}) - \Phi(\frac{\frac{-b-u_c}{\sigma_c}-\rho_1\frac{a+xr(t)-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})}{\Phi(\frac{\frac{-a-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}}) - \Phi(\frac{\frac{-b-u_c}{\sigma_c}-\rho_1\frac{t-u_1}{\sigma_1}}{\sqrt{1-\rho_1^2}})} \tag{1.24}$$

$$= e^{-\rho_1^2 x}.$$

Apply (1.23), (1.24) to $\lim_{t\to\infty}\frac{g_{11}(t+xr(t))}{g_{11}(t)}$, we can obtain

$$\lim_{t\to\infty}\frac{g_{11}(t+xr(t))}{g_{11}(t)} = e^{-\rho_1^2 x}$$

for the $\rho_1 > 0$ case.

For the $\rho_1 < 0$ case, one can follow similar frame work and get $\lim_{t\to\infty}\frac{g_{11}(t+xr(t))}{g_{11}(t)} = e^{-\rho_1^2 x}$.

For the $\rho_1 = 0$ case, one can easily find $\lim_{t\to\infty}\frac{g_{11}(t+xr(t))}{g_{11}(t)} = 1 = e^{-\rho_1^2 x}$.

Thus we have

$$\lim_{t\to\infty}\frac{g_{11}(t+xr(t))}{g_{11}(t)} = e^{-\rho_1^2 x} \tag{1.25}$$

as long as $\rho_1 \neq \pm 1$.

With (1.25) and (1.17), (1.15) can be written as

$$\lim_{t\to\infty}\frac{1-F_{11}(t+xr(t))}{1-F_{11}(t)} = e^{-x} \text{ for } x \in \Re.$$

So the necessary and sufficient condition (1.15) holds and therefore the first assumption holds for distribution $\mathcal{F}_{11}$.

Now we prove the second assumption that $F_{11}^{-1}(1-\frac{1}{n_1}) \to \infty$ as $n_1 \to \infty$.

Suppose $F_{11}^{-1}(1-\frac{1}{n_1}) \to h < \infty$. Then there exists N, for all $n_1 > N$, we have $|F_{11}^{-1}(1-\frac{1}{n_1}) - h| < \epsilon_0$, where ϵ_0 is a fixed positive constant. Then consider

$$\begin{aligned}\int_{h+\epsilon_0}^{\infty} f_{11}(z)\,dx_1 &\geq \int_{h+\epsilon_0}^{h+2\epsilon_0} f_{11}(z)\,dz \\ &= \int_{h+\epsilon_0}^{h+2\epsilon_0} \frac{\frac{1}{\sigma_1}e^{-\frac{(z-u_1)^2}{2\sigma_1}} g_{11}(z)}{\Phi(\frac{b-u_c}{\sigma_c}) - \Phi(\frac{a-u_c}{\sigma_c}) + \Phi(\frac{-a-u_c}{\sigma_c}) - \Phi(\frac{-b-u_c}{\sigma_c})}\end{aligned}$$

$$\times \ dz$$

$$= \frac{1}{\Phi(\frac{b-u_c}{\sigma_c}) - \Phi(\frac{a-u_c}{\sigma_c}) + \Phi(\frac{-a-u_c}{\sigma_c}) - \Phi(\frac{-b-u_c}{\sigma_c})} \times \int_{h+\epsilon_0}^{h+2\epsilon_0} \frac{1}{\sigma_1} e^{-\frac{(z-u_1)^2}{2\sigma_1^2}} g_{11}(z) \ dz.$$

Since $g_{11}(z)$ is a positive continuous function of z and z is bounded, thus the minimum value of $g(z)$ is $g_0 > 0$. Then

$$\begin{aligned}
\int_{h+\epsilon_0}^{\infty} f_{11}(z) \ dz &\geq \frac{g_0}{\Phi(\frac{b-u_c}{\sigma_c}) - \Phi(\frac{a-u_c}{\sigma_c}) + \Phi(\frac{-a-u_c}{\sigma_c}) - \Phi(\frac{-b-u_c}{\sigma_c})} \times \int_{h+\epsilon_0}^{h+2\epsilon_0} \frac{1}{\sigma_1} e^{-\frac{(z-u_1)^2}{2\sigma_1^2}} \ dx_1 \\
&= \frac{g_0}{\Phi(\frac{b-u_c}{\sigma_c}) - \Phi(\frac{a-u_c}{\sigma_c}) + \Phi(\frac{-a-u_c}{\sigma_c}) - \Phi(\frac{-b-u_c}{\sigma_c})} \times \sqrt{2\pi}(\Phi(\frac{h+2\epsilon_0-u_1}{\sigma_1}) - \Phi(\frac{h+\epsilon_0-u_1}{\sigma_1})) \\
&\geq \epsilon_1 > 0.
\end{aligned}$$

Thus as long as we take any n_1 satisfy $n_1 \geq \frac{1}{\epsilon_1}$, we can obtain $F_{11}^{-1}(1-\frac{1}{n_1}) > h+\epsilon_0$, which is conflict with our assumption. Thus $F_{11}^{-1}(1-\frac{1}{n_1}) \to \infty$ if $\rho_1 \neq \pm 1$. The second assumption also holds for distribution $\mathcal{F}_{11}$.

References

References

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

[2] M. Yang, B. Zhang, S. Huang, Optimal designs for binary response experiments with multiple variables, JASA 21 (2011) 1415–1430.

[3] B. C. Arnold, R. J. Beaver, R. A. Groeneveld, W. Q. Meeker, The nontruncated marginal of a truncated bivariate normal distribution, Psychometrika 58 (1993) 471–488.

[4] J. Galambos, The asymptotic theory of extreme order statistics., Florida: Robert E. Krieger, 1987.

[5] M. R. Leadbetter, G. Lindgren, H. Rootz´en, Extremes and related properties of random sequences and processes, Springer, New York, 1983.

Qianshun Cheng

Bayer(Monsanto) Company

E-mail: chengqianshun1@gmail.com

Haiying Wang

University of Conneticut

E-mail: haiying.wang@uconn.edu

Min Yang

University of Illinois at Chicago

E-mail: myang2@uic.edu