Statistica Sinica **34** (2024), 611-636
doi:https://doi.org/10.5705/ss.202022.0019

A SYSTEMATIC VIEW OF INFORMATION-BASED OPTIMAL SUBDATA SELECTION: ALGORITHM DEVELOPMENT, PERFORMANCE EVALUATION, AND APPLICATION IN FINANCIAL DATA

Li He, William Li*, Difan Song and Min Yang

Southwestern University of Finance and Economics,
Shanghai Jiao Tong University,
Georgia Institute of Technology and University of Illinois at Chicago

Abstract: The need to analyze large amounts of data without losing information is evidenced by the recent increase in attention for the information-based optimal subdata selection (IBOSS) approach. However, there are no systematic explorations of this framework, including characterizing the optimal subset when the model is more complex than first-order linear models. Motivated by a real finance case study on the effect of corporate attributes on firm value, we systematically explore the framework and steps required to use IBOSS for data reduction. In the context of second-order models, we develop a novel algorithm for selecting informative subdata. We also evaluate the performance of the proposed algorithm in terms of prediction and variable selection, the latter of which is important for complex models, but has not received sufficient attention in the IBOSS field. Empirical studies demonstrate that the proposed algorithm adequately addresses the trade-off between computation complexity and statistical efficiency, one of six core research directions for theoretical data science research proposed by the US National Science Foundation. The case study demonstrates the potential effect of the IBOSS strategy in scientific fields beyond statistics, particularly finance.

Key words and phrases: Algorithm, computation complexity, IBOSS, statistical efficiency.

1. Introduction

Massive data provide unprecedented opportunities for scientific discovery and advancement, but require new statistical methods and computational algorithms to "convert data into knowledge" (van Dyk et al. (2015)).

One method of dealing with massive data (full data) is to intelligently store/analyze a subset of the data (subdata), for example, using the sampling distribution-based optimal subsampling approach to select the subdata; see Drineas et al. (2012), Ma, Mahoney and Yu (2015), and Wang, Zhu and Ma (2018). However, this approach cannot take advantage of the rich information

*Corresponding author.

contained in big data sets. For linear models, Wang, Yang and Stufken (2019) developed the information-based optimal subdata selection (IBOSS) method, proving that if each independent variable has a distribution in the domain of attraction of the generalized extreme value distribution, the variance of the estimator of the slope parameters goes to zero, even though the size of the subdata is fixed. IBOSS-based approaches are receiving increasing attention; see Cheng, Wang and Yang (2020), who examine logistic regression models, Wang, Yang and Li (2021), who present a LASSO-IBOSS approach for models with a large number of variables, and Wang et al. (2021), who propose an orthogonal subsampling approach for selecting a subset under linear model setups.

The encouraging results of the aforementioned studies have built a strong theoretical foundation for IBOSS-based subdata selection. However, various challenges remain that are hindering the widespread use of IBOSS strategies in practice. First, most existing results are based on linear models that contain main effects only, which may not be adequate for complex big data problems. Thus, we require a general framework that provides guidance on how to develop an IBOSS-type algorithm for a given model. Second, existing studies assess the performance of IBOSS from the perspective of parameter estimation in model fitting. This may be sufficient when the model is relatively simple, but for a more complex model with many model terms, we may also wish to know the prediction capability of the model using IBOSS, and how IBOSS performs in variable selection procedures. Furthermore, we need a real case study that demonstrates how well IBOSS preserves the rich information in the full data and reduces the required computing time in practice.

This study makes three important contributions to the literature. First, we systematically explore the IBOSS framework, showing how to use IBOSS for data reduction. Owing to the relatively simple model format examined in most works in this field, few studies have explored how to characterize the informative points using optimal designs, arguably the most important and challenging step. The resulting characterization dictates the procedure used to obtain an appropriate algorithm. Here, we describe a general IBOSS framework of IBOSS that can be applied to any given model, consisting of three steps:

Step 1 : Derive the optimal (approximate) design in terms of an optimality criterion, say, the D-criterion;

Step 2 : Based on the characterization of the derived optimal design, develop a fast algorithm to efficiently select subdata of size $k < n$;

Step 3 : If possible, investigate the asymptotic properties of the resulting estimators.

Motivated by an important research question in finance, we apply the proposed framework in the context of optimal second-order designs. In Section 2, we show how to address important issues in the framework, and that the resulting IBOSS algorithm, which selects not only extreme end points, but also middle points, is a novel approach. Note that the same technique can be used to obtain optimal designs for other models, such as polynomial models and generalized linear models (GLMs). In the discussion section, we explain how to use the framework for a nonlinear model, and present some novel results.

Second, we provide a comprehensive evaluation of the IBOSS strategy from the standpoint of variable selection. In the context of second-order models, we assess the variable selection performance of our methods in terms of its sensitivity and specificity, and compare it with that of uniform sampling and leverage sampling. The results are encouraging for the proposed IBOSS strategy. For some model settings, it identifies nearly as many significant terms as when using the full data. Moreover, it exhibits higher specificity than when using the full data, implying that using the IBOSS subdata does not incorrectly identify nonsignificant model terms more often than when using the full data. Note that the time complexity of the proposed algorithm is $O(np + kp^4 + p^6)$, which is substantially better than that based on the full data set of $O(np^4 + p^6)$, because k is much smaller than n.

Third, we investigate applications of the IBOSS approach in scientific fields other than statistics. The motivating example for this work is a finance case study on the effects of corporate attributes on firm value. We investigate the relationship between firm value and other variables, such as a firm's assets, cash, capital expenditure, and leverage, because we suspect that the relationship between the response and many of these variables can be better represented by a second-order model. The results are very promising. Using 181,755 data points from all US nonfinancial public firms, we found several important second-order effects, both from the full data and from the IBOSS-selected subdata, that are largely neglected in the related finance literature. We show in Section 4 that the proposed IBOSS strategy preserves the rich information from the full data. Although using the IBOSS subdata results in a slightly higher prediction MSE, the computational savings are substantial (4.32 seconds vs. 79.59 seconds).

There are several important reasons why we chose to investigate using the IBOSS strategy in finance. The IBOSS strategy works particularly well when the distribution of the independent variables is heavy tailed, as is the case in many financial data. More importantly, speed is important in this field. For example, one estimate is that a 1 millisecond advantage can be worth USD 100 million to a major brokerage firm (Martin (2007)). Speed is affected by the proximity, hardware, and efficiency of the algorithms. The demand for faster trading speed has led to faster trading algorithms and a better trading infrastructure among high-frequency trading (HFT) firms. For instance,

Budish, Cramton and Shim (2015) found that the arbitrage opportunities between the S&P 500 index, essentially a benchmark for the US stock market, and S&P 500 futures declined substantially by over 92 percent, from 97 milliseconds in 2005 to 7 milliseconds in 2011, owing to HFT. At the same time, the efficiency of the algorithms is important because traders need to act on predictions promptly. Consequently, a good subset strategy, such as the proposed IBOSS algorithm, represents a promising opportunity in finance from the standpoints of both practice and research.

The remainder of the manuscript is organized as follows. In Section 2, we present a series of techniques to characterize a D-optimal design, and a computationally efficient algorithm based on the characterization. We examine the performance of the proposed algorithm using simulations in Section 3. In Section 4, we apply the proposed IBOSS strategy to finance data. Several important issues are discussed in Section 5. Most of technical proofs are presented in the online Supplementary Material.

2. Application of the Framework in Second-Order Models

In this section, we provide a step-by-step guideline on how to use the IBOSS framework for second-order models.

2.1. Model setup and information matrix

Let $(\boldsymbol{x}_i, y_i)$, for $i = 1, \ldots, n$, denote the full data, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ are the independent variables and y_i is the corresponding continuous response. The response y_i is modeled with interaction and quadratic terms, as follows:

$$y_i = \boldsymbol{\beta}^T \boldsymbol{f}(\boldsymbol{x}_i) + \varepsilon_i, \tag{2.1}$$

where $\boldsymbol{f}(\boldsymbol{x}_i)$ is a vector in the following order: $f_1(\boldsymbol{x}_i) = 1$; $f_{1+j}(\boldsymbol{x}_i) = x_{ij}^2$, for $1 \le j \le p$; and $f_{1+p+j}(\boldsymbol{x}_i) = x_{ij}$, for $1 \le j \le p$; for $1 \le l \le p(p-1)/2$, $f_{1+2p+l}(\boldsymbol{x}_i)$ consists of the terms $x_{ij}x_{ij'}$ and $1 \le j \le p-1$, for $j < j' \le p$; $\boldsymbol{\beta}$ is the corresponding vector of coefficients, with dimension $(p+1)(p+2)/2$; and ε_i is an error term satisfying $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{var}(\varepsilon_i) = \sigma^2$.

When using full data with n observations $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, the least squares estimator resulting from Model (2.1) is $\hat{\boldsymbol{\beta}}_{full} = (\sum_{i=1}^n \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i))^{-1} \sum_{i=1}^n \boldsymbol{f}(\boldsymbol{x}_i)y_i$. Its covariance matrix is $\sigma^2\mathbf{M}_{full}^{-1}$, where $\mathbf{M}_{full} = \sum_{i=1}^n \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i)$. Under the additional assumption that $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, this matrix is the Fisher information matrix. Although we do not impose the normality assumption, we still refer to $\mathbf{M}_{full}$ as the information matrix, for simplicity.

For very large n, we aim to use subdata with k observations for a regression. Let δ_i be an indicator variable, $\delta_i = 1$ if the ith data point is in the subdata, and $\delta_i = 0$ otherwise, and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$. Using the notation $\boldsymbol{\delta}$ to denote the

subdata, the resulting estimator can be expressed as

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\delta}) = \left(\sum_{i=1}^{n} \delta_i \boldsymbol{f}(\boldsymbol{x}_i) \boldsymbol{f}^T(\boldsymbol{x}_i) \right)^{-1} \sum_{i=1}^{n} \delta_i \boldsymbol{f}(\boldsymbol{x}_i) y_i, \tag{2.2}$$

with covariance matrix $\sigma^2 M(\boldsymbol{\delta})^{-1}$, where

$$\mathbf{M}(\boldsymbol{\delta}) = \sum_{i=1}^{n} \delta_i \boldsymbol{f}(\boldsymbol{x}_i) \boldsymbol{f}^T(\boldsymbol{x}_i). \tag{2.3}$$

2.2. Step 1: characterization

The first step in the IBOSS framework is to derive the associated optimal design. In general, optimal designs are available only for specific cases (e.g., the first-order linear model and the logistic model with one independent variable). For most general models, obtaining an optimal design may be difficult. Here, we use a new approach called "complete classes" of designs (Yang and Stufken (2009); Yang (2010); Dette and Melas (2011); Yang and Stufken (2012); Dette and Schorning (2013)). The new tools greatly simplify the process of deriving optimal designs, and can be used to derive most of the available optimality results for GLMs and nonlinear models as special cases.

Without loss of generality, we assume that the design region is the cubic region $\bigcap_{j=1}^{p}[-1,1]$. The objective is to construct a D-optimal approximate design, denoted by $\xi = \{(\boldsymbol{z_i}, w_i), i = 1, \ldots, q\}$, where w_i is the weight on the design point $\boldsymbol{z}_i \in \bigcap_{j=1}^{p}[-1,1]$, q is the number of support points, and $\sum_{i=1}^{q} w_i = 1$. Under Model (2.1), the corresponding information matrix can be written as

$$I(\xi) = \sum_{i=1}^{q} w_i \boldsymbol{f}(\boldsymbol{z}_i) \boldsymbol{f}^T(\boldsymbol{z}_i). \tag{2.4}$$

Maximizing $I(\xi)$ directly for a general model is complicated. In the appendix, we present a series of results for the characterization of a D-optimal design for a second-order model. The same techniques can be applied to more general models, which we discuss further in the last section.

The three lemmas given in the appendix involve simplifying the optimization of $I(\xi)$ by setting a large number of nonzero off-diagonal elements in the information matrix to zero, as well as reducing the number of *distinct* nonzero elements in the matrix. This can be achieved by exploring the symmetry of the variables for the model considered. Using a two-dimensional example for illustration, this consists of three steps: (i) split a point (x_{i1}, x_{i2}, w_i) into four points $(\pm x_{i1}, \pm x_{i2}, w_i/4)$; (ii) move points to extreme points, such as $\{-1, 0, 1\}$ for the quadratic model; and (iii) explore the symmetry between the variables of x_1 and x_2 to reduce the number of distinct parameters in the information matrix

that need to be optimized.

Employing the procedure described above, we can show that the optimal design for the quadratic model (2.1) has all support points in $\bigcap_{j=1}^{p}\{-1,0,1\}$. Let Θ_l denote the set of design points with l elements equal to ± 1, and the remaining $p-l$ elements equal to zero. Then, the optimal design for model (2.1) can be found from those with the form

$$\bar{\xi} = \begin{Bmatrix} \boldsymbol{z}_{i,0} \in \Theta_0 & \boldsymbol{z}_{i,1} \in \Theta_1 & \cdots & \boldsymbol{z}_{i,p} \in \Theta_p \\ \pi_0 & \pi_1 & \cdots & \pi_p \end{Bmatrix}, \tag{2.5}$$

where $\pi_l \geq 0$ is the equal weight assigned to each design point in Θ_l, for $l = 0, 1, \ldots, p$. Furthermore, it can be shown that the information matrix of $\bar{\xi}$ is given by

$$I(\bar{\xi}) = \begin{bmatrix} A & O \\ O & B \end{bmatrix}, \tag{2.6}$$

where

$$A = \begin{bmatrix} 1 & a & a & \cdots a \\ a & a & b & \cdots b \\ a & b & a & \cdots b \\ & & \cdots & \\ a & b & b & \cdots a \end{bmatrix}_{(p+1)\times(p+1)}, \tag{2.7}$$

and $B = \operatorname{diag}(\underbrace{a, \ldots, a}_{p}, \underbrace{b, \ldots, b}_{p(p-1)/2})$. In (2.7), $a = \sum_{i=1}^{q} w_i z_{ij}^2$ and $b = \sum_{i=1}^{q} w_i z_{ij}^2 z_{ij'}^2$, for any $j \neq j'$. Maximizing the determinant of the information matrix $I(\bar{\xi})$ results in optimal values of a^* and b^*:

$$a^* = \frac{p+3}{4(p+1)(p+2)^2}\left[(2p^2+3p+7) + (p-1)\sqrt{4p^2+12p+17}\right], \tag{2.8}$$

$$b^* = \frac{p+3}{8(p+1)(p+2)^3}\left[(4p^3+8p^2+11p-5) + (2p^2+p+3)\sqrt{4p^2+12p+17}\right]. \tag{2.9}$$

Finally, in order to find the optimal design ξ^* from the class of designs satisfying (2.5), we need to find optimal weights π_i^* $(i = 0, 1, \ldots, p)$. Because the support points are in $\bigcap_{j=1}^{p}\{-1,0,1\}$, the optimal π_i^* can be obtained by solving three equations:

$$\sum_{l=0}^{p}\binom{p}{l}2^l\pi_l = 1, \sum_{l=1}^{p}\binom{p-1}{l-1}2^l\pi_l = a^*, \sum_{l=2}^{p}\binom{p-2}{l-2}2^l\pi_l = b^*. \tag{2.10}$$

Theorem 1 (D-optimality). *Let $\xi^* = \{((z_{i1}^*, \ldots, z_{ip}^*)^T, \pi_i^*), i = 1, \ldots, q\}$, where z_{ij}^* takes the value $-1, 0,$ or 1, and π_i^* satisfies* (2.10)*, for which a^* and b^* are*

Table 1. Examples of the designs of Kiefer (1961) and Kôno (1962).

	$p = 3$		$p = 4$	
	Kiefer's	Kôno's	Kiefer's	Kôno's
Corner (π_p)	0.0720	0.0638	0.0370	0.0282
Midpoint of edge (π_{p-1})	0.0190	0.0353	0.0038	0.0157
Center of face (π_{p-2})	0.0328	0	0.0118	0
Origin (π_0)	0	0.0656	0	0.0474

determined by (2.8) *and* (2.9)*, respectively. Then,* ξ^* *is a D-optimal design for* $\boldsymbol{\beta}$ *under Model* (2.1)*.*

Theorem 1 shows the optimal design for model (2.1). Although several previous works have studied optimal designs for second-order models (Kiefer (1961); Kôno (1962); Farrell, Kiefer and Walbran (1967)), they begin with a *guessed* optimal design, and then verify the optimality using the equivalence theorem. In comparison, we derive the optimality from Lemmas 1–3. There are three equations in (2.10), indicating that solutions are not unique when $p \geq 3$. One way of solving this problem is to allow only some weights to be nonzero, which is the approach taken by Kiefer (1961) and Kôno (1962). More specifically, the former approach considers designs in which support points are restricted to corners, midpoints of edges, and centers of two-dimensional faces, such that $\pi_p, \pi_{p-1}, \pi_{p-2} > 0$. The latter approach provides solutions for designs with support on corners, midpoints of edges, and the origin ($\pi_p, \pi_{p-1}, \pi_0 > 0$). Table 1 shows the numerical results for $p = 3$ and $p = 4$, as an example. From Theorem 1, there are more solutions than those given in Table 1. For example, any linear combination of the two designs given in the table is also an optimal design.

More general results are given in Farrell, Kiefer and Walbran (1967); most of our results in this subsection are similar to theirs. Again, the difference is that they use geometry arguments and the general equivalence theorem. As stated in Farrell, Kiefer and Walbran (1967, p. 113), *"Our main way of finding D- and G-optimum designs and of verifying their optimality is thus to guess a* ξ^* *(perhaps by minimizing det* $M(\xi)$ *over some subset of designs depending on only a few parameters) and then to verify (1.4)."* Our use of the complete class approach is based on a series of lemmas that can be adapted easily to obtain optimal designs for more general models. One such example for nonlinear models is given in Section 5.1.

2.3. Step 2: Algorithm and its properties

There are two difficulties when selecting optimal subdata under model (2.1). First, when $p \geq 3$, there is an infinite number of optimal designs to choose from. Second, after an optimal design is chosen, it usually requires a substantial number

Algorithm 1

Suppose $r = k/(2p)$ is an integer. Denote $x_{(1)j}$ and $x_{(n)j}$ as the minimum and maximum, respectively, of x_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Perform the following steps:

1. Determine a^* according to Equation (2.8). Then, calculate $r_1 = \lceil r \cdot a^* \rceil$ and $r_2 = r - r_1$.
2. For $x_{i1}, i = 1, \ldots, n$, select r_1 data points with the smallest x_{i1} values, r_1 data points with the largest x_{i1} values, and $2r_2$ data points closest to $(x_{(1)1} + x_{(n)1})/2$.
3. For $j = 2, \ldots, p$, exclude previously selected data points. From the remainder, select r_1 data points with the smallest x_{ij} values, r_1 data points with the largest x_{ij} values, and $2r_2$ data points closest to $(x_{(1)j} + x_{(n)j})/2$.
4. Let $\delta_i, i = 1, \ldots, n$ be an indicator variable, where $\delta_i = 1$ if $(\boldsymbol{x}_i, y_i)$ is selected in the previous steps, and $\delta_i = 0$ otherwise.

of points in which multiple variables take extreme values, which may not exist in the full data.

Fortunately, all optimal designs ξ^* of the form of (2.5), independent of π_i, satisfy a common property when the design space is projected onto a one-dimensional space of each independent variable. As shown in (2.7) and the definitions of a and b, the optimal designs satisfy

$$\sum_{i=1}^{q} w_i z_{ij}^2 = a^*, j = 1, \ldots, p, \tag{2.11}$$

where the sum is taken over all support points of the design. Along with the condition that $z_{ij} = -1, 0, 1$, the support points always have

$$\begin{Bmatrix} -1 & 0 & 1 \\ \dfrac{a^*}{2} & 1 - a^* & \dfrac{a^*}{2} \end{Bmatrix} \tag{2.12}$$

as one-dimensional projections. We now present the main IBOSS algorithm for second-order models.

This algorithm differs from existing IBOSS algorithms in two ways. First, for each independent variable, it chooses *three* types of values: the largest, smallest, and middle values. In comparison, almost all existing IBOSS algorithms select only the largest and smallest values. Second, the weight assigned to each type of value depends on the number of factors p. The results reported in Table 2 are interesting and, to a certain extent, surprising. One might expect the weights given to the three values $-1, 0$, and 1 to be the same. Instead, the weight for the middle number needs to be smaller than the weights for the extreme values. In addition, the weight allocation is a function of p.

Table 2. Relationship between p, a^*, and the weights on 0 and ± 1 in the D-optimal design. On the original scale, ± 1 correspond to the extreme points, and 0 corresponds to the median.

p	a^*	0	± 1
2	0.7435	0.2565	0.3717
3	0.7930	0.2070	0.3965
4	0.8271	0.1729	0.4136
5	0.8518	0.1482	0.4259
6	0.8705	0.1295	0.4352
7	0.8850	0.1150	0.4425
8	0.8967	0.1033	0.4484
9	0.9062	0.0938	0.4531
10	0.9142	0.0858	0.4571
		...	
20	0.9537	0.0463	0.4768
		...	
40	0.9759	0.0241	0.4880

2.4. Step 3: Asymptotic properties of the algorithm

The proposed algorithm is a partition-based selection algorithm that needs to identify three groups of values for each independent variable: the largest, the smallest, and the middle values. As with any newly proposed algorithm, we wish to measure the statistical efficiency of the selected subdata. An ideal solution is to measure how the variance of each element of $\hat{\boldsymbol{\beta}}$ changes asymptotically as a function of n, that is, the asymptotic properties of the inverse of the information matrix. Unfortunately, the resulting information matrix is much more complicated than the main-effects only model because of the additional quadratic and interaction effects. Consequently, some well-known criteria, such as D-, A-, or E-criteria, are intractable under the resulting information matrix.

One alternative choice is the T-criterion, defined as the trace of the information matrix (Pukelsheim (2006)). This approach is feasible because the criterion is tractable in general, which is crucial for a complicated information matrix. The T-criterion also has an attractive property. If the trace of a resulting information matrix goes to infinity as a function of n, then this implies that the sum of all eigenvalues of the matrix goes to infinity. Consequently, at least one of the eigenvalues of the corresponding covariance matrix goes to zero as a function of n. In other words, there exists at least one linear combination of the elements of $\hat{\boldsymbol{\beta}}$, such that its variance goes to zero when n goes to infinity, even when k is finite.

Under certain distribution assumptions of $\boldsymbol{x}$, we have the following theorem.

Theorem 2. *Let* $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ *and* $\boldsymbol{\Sigma} = \boldsymbol{\sigma}\boldsymbol{\rho}\boldsymbol{\sigma}$, *where* $\boldsymbol{\sigma} = diag(\sigma_1, \ldots, \sigma_p)$ *is a diagonal matrix of standard deviations, and* $\boldsymbol{\rho}$ *is a correlation matrix. Assume*

that $\boldsymbol{x}_i$, for $i = 1, \ldots, n$, are independent and identically distributed (i.i.d.) with a distribution specified below. The following results hold for $\mathbf{M}(\boldsymbol{\delta})_{jj}$, the jth diagonal element of the information matrix for $\hat{\boldsymbol{\beta}}(\boldsymbol{\delta})$, which is the estimator from the proposed algorithm.

(i) For multivariate normal independent variables, that is, $\boldsymbol{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

$$\mathbf{M}(\boldsymbol{\delta})_{jj} = \begin{cases} O_p((\log n)^2) & \text{for } 2 \le j \le p+1, \\ O_p(\log n) & \text{for } p+2 \le j \le 2p+1, \\ O_p((\log n)^2) & \text{for } 2p+2 \le j \le \dfrac{(p+1)(p+2)}{2}. \end{cases} \tag{2.13}$$

(ii) For multivariate log-normal independent variables, that is, $\boldsymbol{x}_i \sim LN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

$$\mathbf{M}(\boldsymbol{\delta})_{jj} = \begin{cases} O_p(4\sigma_j\sqrt{2\log n}) & \text{for } 2 \le j \le p+1, \\ O_p(2\sigma_j\sqrt{2\log n}) & \text{for } p+2 \le j \le 2p+1, \\ O_p\Big(\max\limits_{1\le l\le p}\{|2\rho_{lm}\sigma_m + 2\rho_{lm'}\sigma_{m'}|\}\sqrt{2\log n}\Big), & 1 \le m \le p-1, m < m' \le p, \\ & \text{for } 2p+2 \le j \le \dfrac{(p+1)(p+2)}{2}. \end{cases} \tag{2.14}$$

Theorem 2 shows that, under the T-criterion, the resulting information matrix increases as a function of n, even when k is fixed. That is, $\mathrm{Var}(L'\hat{\boldsymbol{\beta}}) \to 0$ when $n \to \infty$, for some nonzero vector L. Theoretically, it is not as strong as the A-criterion, which minimizes the sum of the variance of each element of $\hat{\boldsymbol{\beta}}$ (except the intercept). However, our extensive simulation studies, discussed in the next section, show that the proposed algorithm actually demonstrates the desired asymptotic properties of $\hat{\boldsymbol{\beta}}$, and is sufficient in practice.

3. Simulation Studies

3.1. Estimation and prediction MSE

The first part of the simulations focuses on the mean squared error (MSE) criterion. We generate independent samples $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ under three covariance structures: a multivariate normal $N(\mathbf{0}, \Sigma)$, multivariate log-normal $LN(\mathbf{0}, \Sigma)$, and multivariate t-distribution with two degrees of freedom $t_2(\mathbf{0}, \Sigma)$. In our study, $p = 10$ and $\Sigma_{jj'} = 0.5^{\mathbb{I}(j \ne j')}$, where $\mathbb{I}(\cdot)$ is the indicator function. The training set of size n and the test set with fixed size $n_{test} = 10{,}000$ are generated independently. Responses are generated according to Model (2.1), with $\boldsymbol{\beta}$ being a vector of unity; the noise ε_i is i.i.d. $N(0, 3^2)$.

The data-generating process is repeated $S = 100$ times. Let $\boldsymbol{\beta}_{-0}$ denote the

vector of all parameters except the intercept term, and $\boldsymbol{\beta}_{-0}^{(s)}$ and $\boldsymbol{\beta}^{(s)}$ denote the estimators of $\boldsymbol{\beta}_{-0}$ and $\boldsymbol{\beta}$, respectively, at the sth repetition. For each method, calculate the estimation MSE as $\text{MSE}_e = (1/S)\sum_s \|\hat{\boldsymbol{\beta}}_{-0}^{(s)} - \boldsymbol{\beta}_{-0}\|^2$ and the prediction MSE as $\text{MSE}_p = (1/n_{test}) \cdot S \sum_{s,i}(\boldsymbol{\beta}^T \boldsymbol{f}(\boldsymbol{x}_i) - (\hat{\boldsymbol{\beta}}^{(s)})^T \boldsymbol{f}(\boldsymbol{x}_i))^2$.

Following the above procedure, we conduct two simulations. In the first simulation, we generate training data of sizes $n = 5000$, 10^4, 10^5, and 10^6, while fixing the subdata size at $k = 1{,}000$. Figure 1 compares the approaches in terms of MSE_e and MSE_p (in logarithm scale) as n increases. For the estimation MSE, Panel (a) of Figure 1 shows that the IBOSS approach outperforms uniform sampling and leverage sampling for all three distributions considered. The MSE_e values of the IBOSS approach decrease with an increase in n, and the convergence rate is faster when the independent variables are more heavy tailed. In particular, when $\boldsymbol{x} \sim t_2(\mathbf{0}, \Sigma)$, the IBOSS approach yields a comparable MSE_e to that based on the full data. For the prediction MSE, Panel (b) of Figure 1 shows very similar patterns.

In the second simulation, we fix the full data size at $n = 10^6$, and select subdata of sizes $k = 500$, 1000, 2000, 3000, and 5000. The plots of MSE_e and MSE_p (in logarithm scale) are given in Panels (a) and (b), respectively, of Figure 2. As expected, the performance of all subsampling approaches improves as k increases, and the IBOSS approach outperforms the uniform and leverage sampling methods consistently. The advantage of the IBOSS approach becomes more significant when the distribution of $\boldsymbol{x}$ has heavier tails. For MSE_p, the prediction MSE value for a given k becomes larger for uniform sampling when the distribution changes from normal to the t-distribution, as observed from the red curve (corresponding to uniform sampling) moving upward from the top to the bottom in the three figures in Panel (b). In comparison, the three green curves corresponding to leverage sampling barely move, and the three blue curves corresponding to the IBOSS approach actually move downward, indicating smaller prediction errors for heavy-tailed distributions.

Table 3 compares the computing times of the three sampling methods ($k = 1{,}000$) with that using the full data. For the full-data approach, each number represents the time taken to fit the second-order regression model. For the IBOSS and other subsampling approaches, each number is the total CPU time required to select the subdata and then fit the model. For the IBOSS approach, we use a C++ function to determine the desired quantiles of a given vector. All other approaches are implemented in R on a laptop with an Intel® Core™ i7-10710U processor and 16 GB memory. Not surprisingly, uniform sampling and the IBOSS approach have a short computation time. Leverage sampling shows no reduction in computation time, because calculating the leverage values has the same complexity as fitting the full model. Note that when n or k increases, the IBOSS and full-data approaches both require longer computation times. However, the computation time of the IBOSS method increases at a much slower

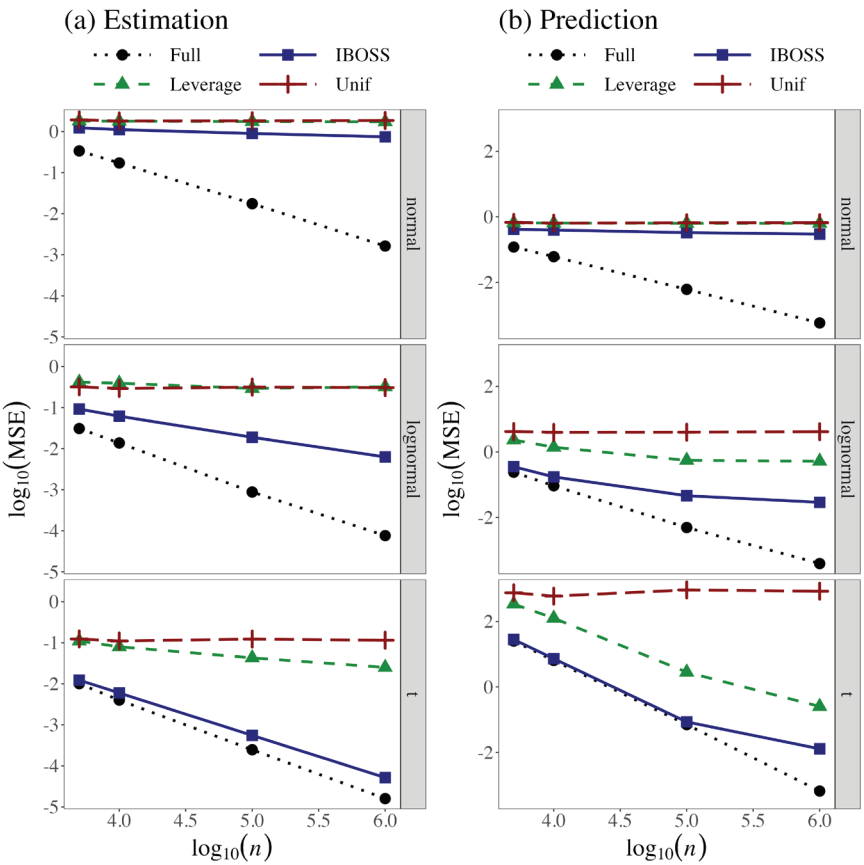

Figure 1. MSEs when estimating the slope parameter (Panel a) and out-of-sample prediction errors (Panel b) for three distributions of the independent variables. The subdata size is fixed at $k = 1{,}000$ and the full data size n changes.

rate than fitting with the full data. Overall, together with the MSE_e and MSE_p results, the IBOSS approach appears to achieve satisfactory statistical efficiency at a small computational cost.

3.2. Sensitivity and specificity

Most prior works examine the performance of the IBOSS strategy by focusing on the estimation error, with few examining its variable selection capability. At the time, most models were relatively easy, and it was often assumed that all effects in the model were significant. However, for a more complicated model that has many model terms, we also need to assess the model using the IBOSS subdata to effectively identify the significant effects.

Variable selection has been studied extensively; see, for example, Tibshirani (1996), Fan and Li (2001), and Choi, Li and Zhu (2010). More recently, Wang, Yang and Li (2021) proposed a LASSO approach for an IBOSS subdata

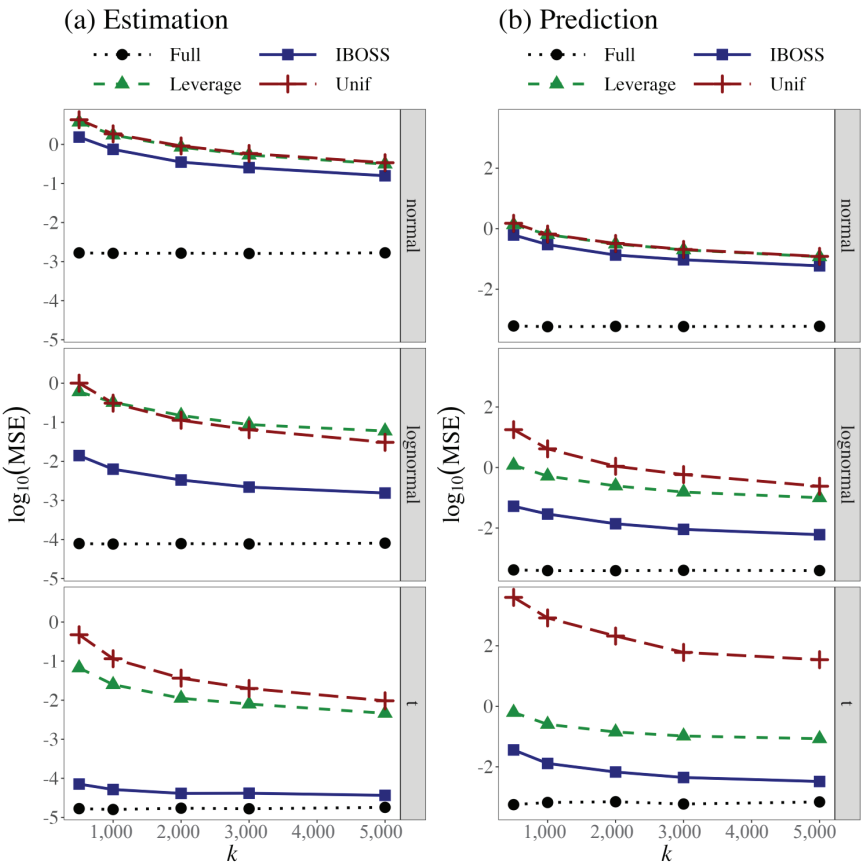

Figure 2. MSEs when estimating the slope parameter (Panel a) and out-of-sample prediction errors (Panel b) for three distributions of the independent variables. The full data size is fixed at $n = 10^6$ and the subdata size k changes.

Table 3. CPU times for different approaches; subdata size fixed at $k = 1{,}000$.

(a) CPU times for different n with $p = 10$				
n	Unif	Leverage	IBOSS	Full
5×10^4	0.01	1.08	0.05	0.39
10^5	0.01	1.73	0.05	0.84
10^6	0.02	16.74	0.42	6.87
(b) CPU times for different p with $n = 10^6$				
p	Unif	Leverage	IBOSS	Full
5	0.01	4.12	0.25	1.50
10	0.02	16.74	0.42	6.87
15	0.04	65.21	0.78	26.30

strategy under the first-order linear model. We now evaluate the performance of the IBOSS subdata under the second-order model (2.1). Two traditionally

Table 4. Setup for simulations comparing sensitivity and specificity.

Settings	1	2	3	4	5
# of variables	10	10	10	10	10
# of non-zero main effects	5	5	0	7	5
# of non-zero interaction effects	10	10	10	21	10
# of non-zero quadratic effects	5	5	5	7	5
Coef of non-zero main effects	1	5 or 1*	-	1	1
Coef of non-zero interaction effects	0.5	2.5 or 0.5*	2.5 or 0.5*	0.5	0.5
Coef of non-zero quadratic effects	0.5	2.5 or 0.5*	2.5 or 0.5*	0.5	0.5
Full data size	10,000	10,000	10,000	10,000	50,000
Subdata size	1,000	1,000	1,000	1,000	1,000

* In Settings 2 and 3, the coefficients are not equal. In Setting 2, the coefficients of X_1, X_2 are 5, while the coefficients of $X_3 - X_5$ are 1; the coefficients of X_1^2, X_2^2 are 2.5, while the coefficients of $X_3^2 - X_5^2$ are 0.5; and the nonzero second-order effects involving X_1, X_2 are 2.5, and the others are 0.5. Setting 3 is the same as Setting 2, except that all main effects are zero.

important variable selection criteria are *sensitivity* and *specificity*, defined respectively as follows:

$$\text{sensitivity} = \frac{\text{number of selected significant effects}}{\text{total number of true significant effects}},$$
$$\text{specificity} = \frac{\text{number of unselected nonsignificant effects}}{\text{total number of true nonsignificant effects}}.$$

We adopt a similar simulation setting to those of Choi, Li and Zhu (2010) and Chen, Li and Wang (2020). Five settings are considered in Table 4. There are $p = 10$ log-normally distributed variables in the model. However, in each of the five settings, we assume that only $p_1 < p$ main effects are significant. We further assume that the corresponding p_1 quadratic effects and $\binom{p_1}{2}$ two-factor interactions between them are significant. Among the five settings, Setting 1 can be considered a "base" setting; Setting 2 represents a model with larger coefficients for some of the significant terms; Setting 3 is similar to Setting 2, but only second-order effects are assumed to be significant; Setting 4 increases p_1 from five to seven; and Setting 5 increases the full data size from 10,000 to 50,000. In all settings, the error terms are assumed to follow a normal distribution with $\sigma = 100$. We use the stepwise regression approach with the AIC criterion for model fitting.

Figure 3 compares box plots of the distributions of sensitivity and specificity over 100 tries for four methods: full data, the subdata using IBOSS, uniform sampling, and leverage sampling ($k = 1{,}000$). For all settings, the sensitivity of IBOSS is comparable with that of the full data, both of which are significantly better than those of uniform sampling and leverage sampling. Compared with the base setting of Setting 1, when the coefficients of some significant terms increase

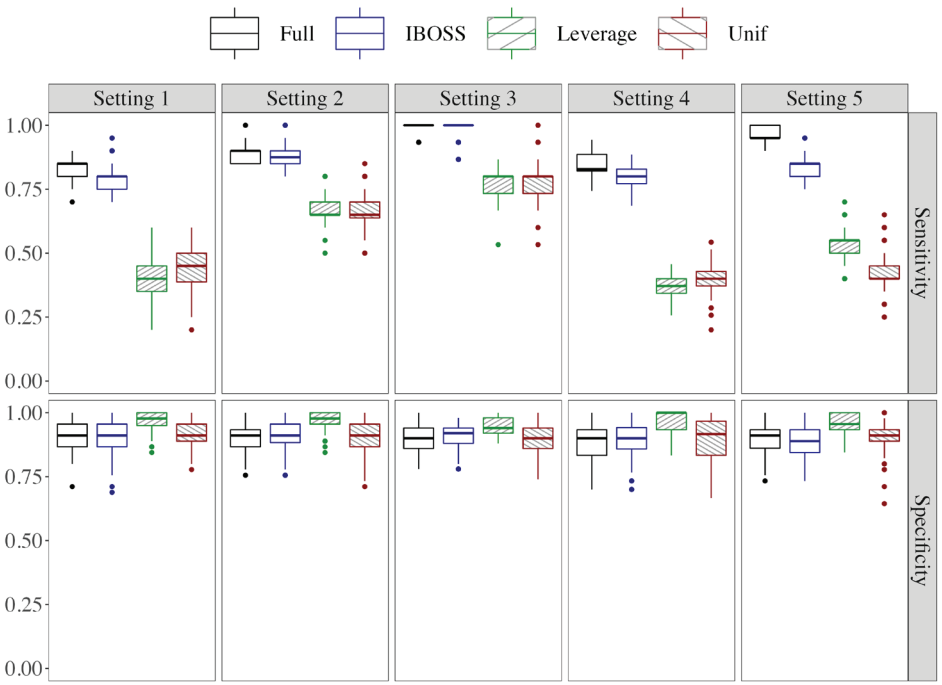

Figure 3. Sensitivity and specificity box plots for Settings 1-5.

in Setting 2, the sensitivities of all four methods increase, but the overall trend remains the same as that in Setting 1. In Setting 3, we require that all main effects are zero, but some second-order terms are significant. In this case, the sensitivities of the full data and IBOSS are near 100% for most tries, indicating that both have a strong ability to identify significant second-order effects. In Setting 4, the number of significant variables is increased from $p_1 = 5$ to $p_1 = 7$. The sensitivities of the methods are all similar to those of Setting 1, as expected. The results in Setting 5 are interesting. In this setting, we increase the overall data size n from 10,000 to 50,000. Consequently, the sensitivity of the full data increases, as expected, but the sensitivity of IBOSS also improves, even if the size of the subdata is unchanged at $k = 1{,}000$. This shows that IBOSS can take advantage of a larger data set. With more candidate points available, the points selected by IBOSS become more informative, making this approach advantageous for big data regression.

The specificity values of the four methods are similar in all settings. Note that in all settings except Setting 5, the specificity of IBOSS is slightly better than that of the full data. In particular, the specificity of IBOSS is noticeably better than that of the full data in Setting 3, suggesting that IBOSS does not incorrectly choose nonsignificant second-order effects as often as the full data approach does. The specificity of uniform sampling and leverage sampling are

better than that of the IBOSS approach. However, given their poor performance in terms of sensitivity, which is arguably more important, these methods are clearly inferior.

To assess the overall performance of both sensitivity and specificity, we also plot sensitivity and "1− specificity" in Figure 4. In each figure, a dot in the upper-left corner indicates that the corresponding method exhibits good performance overall in terms of both sensitivity and specificity. In all five settings, the overall performance of IBOSS is very close to that of the full data, and significantly better than that of uniform sampling and leverage sampling.

The settings in Table 4 assume that all variables follow a log-normal distribution. Unreported simulation results based on other distribution assumptions show similar patterns, although IBOSS performs better for distributions that are more heavy tailed, which is consistent with the findings of most existing IBOSS studies. Interestingly, many financial data are indeed heavy tailed. For example, in the finance case we study here, the main variable of interest, financial leverage, is heavy tailed with a kurtosis of 27.42. In fact, it has been argued that many results in finance based on normal distribution assumptions may not be valid. For example, Deakin (1976) showed that many important financial ratios are non-normally distributed, urging researchers to be cautious when using these financial data in empirical studies.

Our simulation results have shown that the proposed IBOSS subdata strategy clearly outperforms the alternative uniform sampling and leverage sampling methods. Table 2 shows that when p increases, the weight assigned to the center points is closer to zero. Figure 5 provides further information about a^* and p. In the simulations, the proposed Algorithm 1 may select similar points to those selected by the IBOSS algorithm of Yang and Stufken (2009), and the similarity is stronger for larger p. This phenomenon stems from the smaller weights assigned to the center points, and both IBOSS algorithms using a one-variable-at-a-time approach. The final selected subdata are only a proxy for the theoretically optimal solution. Nonetheless, we show in the next two sections that the proposed algorithm enjoys clear advantages in real cases when the true model is more complicated than that used in the simulation. Furthermore, our algorithm has better robustness properties against possible missing terms.

4. A Finance Case Study

A fundamental research question in the finance literature is the effect of corporate attributes on firm value. Here, we examine the relationship between firm value and several important variables using a second-order model. Typical approaches in finance usually involve identifying one main independent variable and several control variables, and then running linear regression models. More often than not, first-order models are used, with less attention given to quadratic

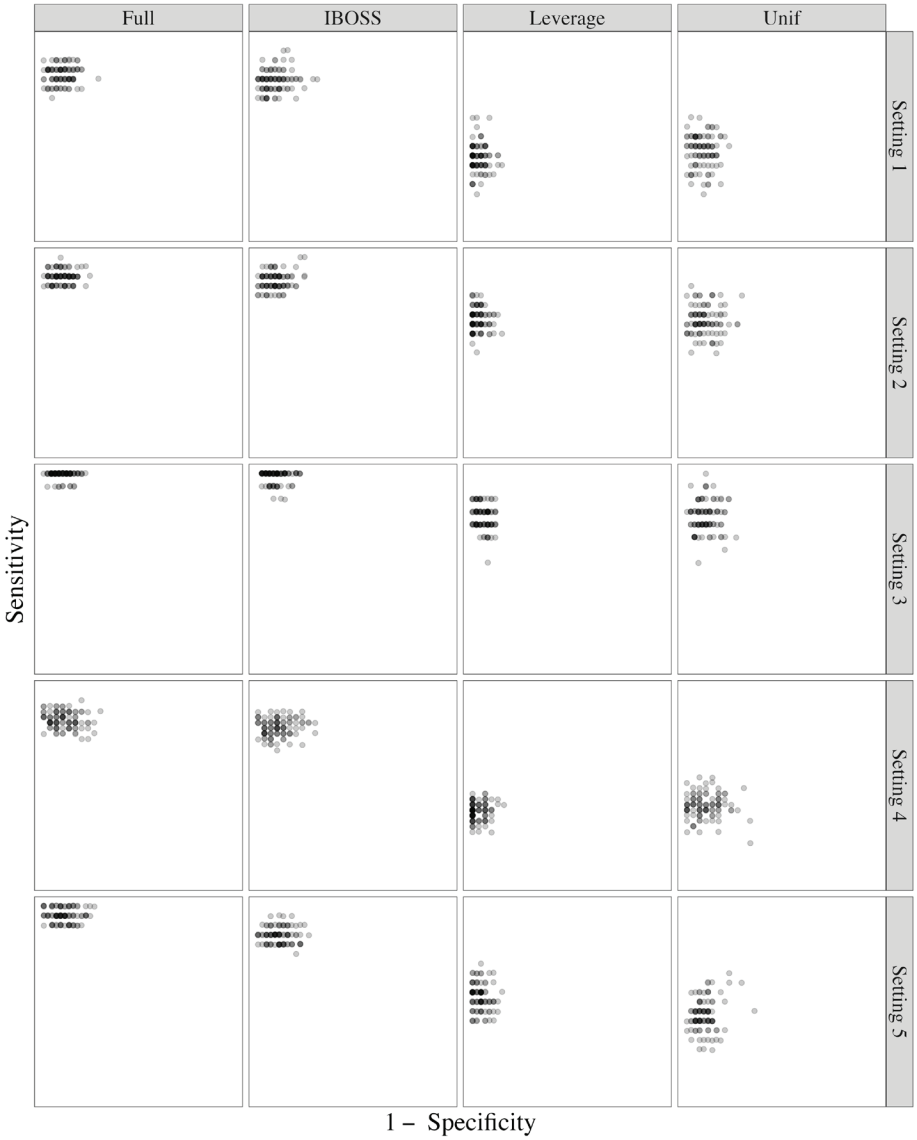

Figure 4. Sensitivity plotted against 1− specificity for Settings 1-5, with 100 runs each. Points are jittered. Upper left points exhibit the best performance.

and interaction effects. However, second-order effects may also be significant in corporate finance studies. For instance, financial leverage is considered to be a key variable related to firm value. However, its relationship with firm value may not be linear, and there have been contradicting results reported in the literature. The Modigliani–Miller theorem (Modigliani and Miller (1958)) states that the value of a firm is *independent* of the firm's capital structure. However, Baxter (1967) finds a *negative* effect of leverage that increases financial distress costs before reaching the optimal debt-equity ratio, and Jensen (1986) finds a

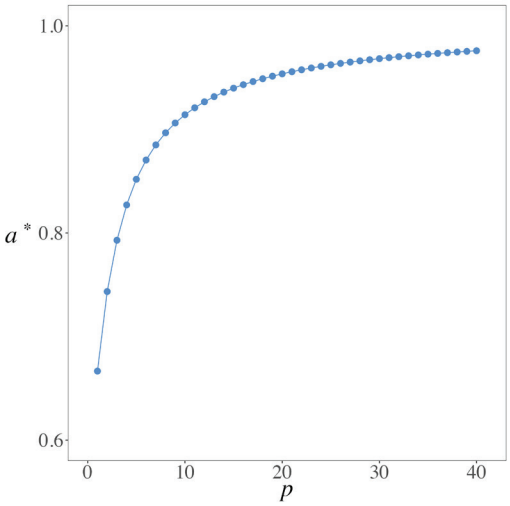

Figure 5. Relationship between p and a^*.

positive effect of leverage on firm value, owing to agency costs.

This motivated us to consider a second-order model consisting of leverage and several other critical variables that affect firm value: LEVERAGE (X_1), measured as total liabilities divided by total assets; SIZE (X_2), measured as the logarithm of total assets (in millions); CASH (X_3), measured as total cash and cash equivalent holding, scaled by total assets; PPE (X_4), measured as the net value of property, plant, and equipment, scaled by total assets; CAPEX (X_5), measured as capital expenditure, scaled by total assets; ROE (X_6), measured as net income divided by shareholders' equity; RD (X_7), measured as research and development costs, scaled by total assets; and AGE (X_8), which is the firm age.

In empirical financial studies, a key issue is how to measure firm value. Here, Tobin's Q, the ratio of the market value of the financial claims on the firm to the replacement cost of the firm's assets, has been widely accepted as a fundamental performance metric since its introduction by Brainard and Tobin (1968) and Tobin (1969). Of particular importance is that Tobin's Q captures profitable investment opportunities. A higher Tobin's Q value suggests that the firm uses its economics resources effectively, because the market value created by the firm's assets is higher than the cost of reproducing its underlying assets.

We follow convention among finance researchers, and use the Compustat Fundamentals database to examine the effect of various firm-specific attributes on Tobin's Q. The Compustat database collects financial statement and financial market data on all US publicly traded companies, and is published by Standard & Poor's Global. We select the data on all US nonfinancial public firms for the period 1980 through 2020, and delete observations with missing values. The final data comprises 181,755 firm-year observations, representing 20,117 distinct

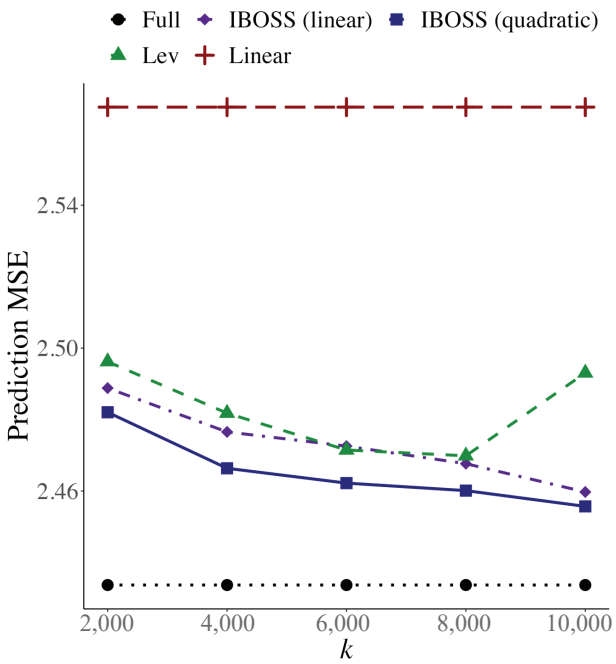

Figure 6. Out-of-sample prediction errors of various methods based on real data.

firms. We calculate the dependent variable Tobin's Q as the market value of a company (common shares outstanding multiplied by the fiscal-year end share price), divided by the book value of the net assets.

We first examine the prediction capability of the IBOSS subdata for $k =$ 10,000, which is approximately 5% of the full data; see Figure 6. In the calculation of the MSE, we randomly select 20% of the full data as the test set, and the remaining 80% constitute the training set. We then employ the stepwise model selection procedure using the AIC. For $k = 10{,}000$, the MSE for the IBOSS data is 2.456, which is only slightly higher than the MSE value of 2.434 for the full data. However, the CPU time for the IBOSS approach is only 4.32 seconds, which is substantially smaller than the 79.59 seconds for the full data.

Figure 6 also compares the proposed IBOSS strategy (labeled "IBOSS (quadratic)") with leverage sampling, the existing IBOSS of Wang, Yang and Stufken (2019) developed for a linear model (labeled "IBOSS (linear)"), and the linear model using the full data. The MSE values are computed for various k from 2,000 to 10,000. First, note that the proposed IBOSS strategy is superior to leverage sampling for all subdata sizes. Not only does the former have a smaller MSE than the latter does, leverage sampling sometimes becomes unstable, as evidenced by the surprisingly higher MSE when k increases from 8,000 to 10,000. The IBOSS (linear) approach performs nicely even if the true model is quadratic, indicating that it is robust against model misspecifications. We discuss

Table 5. Terms included in the final model of forward selection.

	IBOSS	Full
Main Effects	1, 2, 3, 4, 5, 6, 7, 8	1, 2, 3, 4, 5, 6, 7, 8
Interaction Effects	12, 16, 17, 23, 26, 28 35, 37, 45, 46, 47, 48 56, 67, 78	13, 14, 15, 16, 23, 24, 25 26, 27, 28, 34, 35, 37, 38, 45, 48, 58, 67, 68, 78
Quadratic Effects	1, 2, 4, 5, 7, 8	1, 2, 3, 4, 5, 7, 8
Adjusted R^2 with linear	20.76%	16.71%
Adjusted R^2 (linear+lev^2)	20.76%	16.79%
Adjusted R^2 with second-order	26.94%	21.65%
Time (seconds)	4.32	79.59

X_1 : LEVERAGE X_2 : SIZE X_3 : CASH X_4 : PPE
X_5 : CAPEX X_6 : ROE X_7 : RD X_8 : AGE

robustness further in the next section. Figure 6 shows that it is important not to ignore second-order terms in a model. Focusing on a first-order linear model results in significantly higher MSE values, even if the full data are used. Finally, Table 5 assesses the performance of the proposed IBOSS strategy in terms of variable selection. The results correspond to $k = 10{,}000$, and show that the IBOSS subdata identify most of the terms shown to be significant using the full model. IBOSS identifies all of the main effects and most of the second-order terms identified as significant based on the full data.

5. Discussion

5.1. Characterization for a nonlinear model

Thus far, we have focused on developing an IBOSS strategy for a second-order model. However, the proposed framework can be applied to more general models. As an example, we develop an IBOSS strategy for a nonlinear model. Consider a multivariate logistic regression model with binary responses, where a subject is administered p covariates at level $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})'$ (Agresti (2002)). The associated theoretical optimality results for such a model are relatively scarce, tending to focus on the model containing main effects only (Sitter and Torsney (1995); Yang, Zhang and Huang (2011)). To the best of our knowledge, no optimality results are available when interaction effects are present in a multivariate logistic regression model. In this section, we use the aforementioned strategy to provide an optimality result for the following model:

$$logit(y_i = 1) = \beta_0 + \sum_{k=1}^{p-1} \beta_k x_{ik} + \sum_{k=1}^{p-2} \sum_{l=k+1}^{p-1} \beta_{kl} x_{ik} x_{il} + \beta_p x_{ip}. \tag{5.1}$$

Here, y_i is the response of subject i with covariate $\boldsymbol{x}_i$, for $p \geq 3$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p, \beta_{12}, \ldots, \beta_{p-2,p-1})$ are unknown parameters. We assume the first

$p-1$ covariates are bounded, that is, $x_{ij} \in [L_j, U_j]$, for $j = 1, \ldots, p-1$, and there is no constraint on the last covariate, that is, $x_{ip} \in (-\infty, \infty)$. Such an assumption is common for optimal designs under multivariate logistic regression models (Sitter and Torsney (1995); Yang, Zhang and Huang (2011)).

In the locally optimal design context, there is a one-to-one mapping between $\boldsymbol{x}_i$ and $\boldsymbol{c}_i$, where $\boldsymbol{c}_i = (1, x_{i1}, \ldots, x_{i,p-1}, c_i)'$. Here, $c_i = \beta_0 + \sum_{k=1}^{p-1} \beta_k x_{ik} + \sum_{k=1}^{p-2} \sum_{l=k+1}^{p-1} \beta_{kl} x_{ik} x_{il} + \beta_p x_{ip}$. It is convenient to denote the design ξ as $\xi = \{(\boldsymbol{c}_i, \omega_i), i = 1, \ldots, k\}$. Let

$$a_{l,j} = \begin{cases} L_j & \left\lceil \dfrac{l}{2^{p-1-j}} \right\rceil \text{ is odd,} \\ U_j & \left\lceil \dfrac{l}{2^{p-1-j}} \right\rceil \text{ is even,} \end{cases} \qquad l = 1, \ldots, 2^{p-1};\ j = 1, \ldots, p-1, \tag{5.2}$$

where $\lceil a \rceil$ is the smallest integer greater than or equal to a.

Theorem 3. *Under Model* (5.1), ξ^* *is a D-optimal design of parameter vector* $\boldsymbol{\beta}$ *if* $\xi^* = \{(\boldsymbol{c}_{l1}^*, 1/2^p) \& (\boldsymbol{c}_{l2}^*, 1/2^p), l = 1, \ldots, 2^{p-1}\}$, *where* $(\boldsymbol{c}_{l1}^*)^T = (1, a_{l,1}, \ldots, a_{l,p-1}, c^*)$ *and* $(\boldsymbol{c}_{l2}^*)^T = (1, a_{l,1}, \ldots, a_{l,p-1}, -c^*)$, $a_{l,j}$ *is defined in* (5.2), c^* *minimizes* $c^{-2}(\Psi(c))^{-m}$, *and* $m = (p^2 - p + 4)/2$.

The characterization in Theorem 3 lays the theoretical foundation for developing subdata selection algorithms. Follow Step 2 outlined in Section 2.3 and Step 3 outlined in Section 2.4 to develop an efficient subdata selection algorithm, as well as some theoretical properties.

5.2. Robustness

Like all existing IBOSS approaches, the characterization of optimal designs in the general framework depends on the model assumptions. Here, we examine how robust they are against model misspecifications. For example, is the IBOSS algorithm proposed by Wang, Yang and Stufken (2019), which is based on a linear model, also effective for the second-order model (2.1)? Here, we compare the performance of two IBOSS approaches for the finance example. As shown in Figure 6, the IBOSS (linear) approach performs surprisingly well, even when the true model has significant second-order terms. Across all selected subsample sizes, the IBOSS (linear) approach has slightly higher prediction MSE values than those of the proposed algorithm, IBOSS (quadratic), but outperforms both random sampling and leverage sampling. This shows that the IBOSS algorithm proposed by Wang, Yang and Stufken (2019) is robust against possible important second-order terms in the true model. A possible explanation for this is that their algorithm selects points one variable at a time. Thus, even if the characterization of the D-optimal design for the linear model requires that only end points be chosen, in reality, the true weight distributions may resemble those in Table 2,

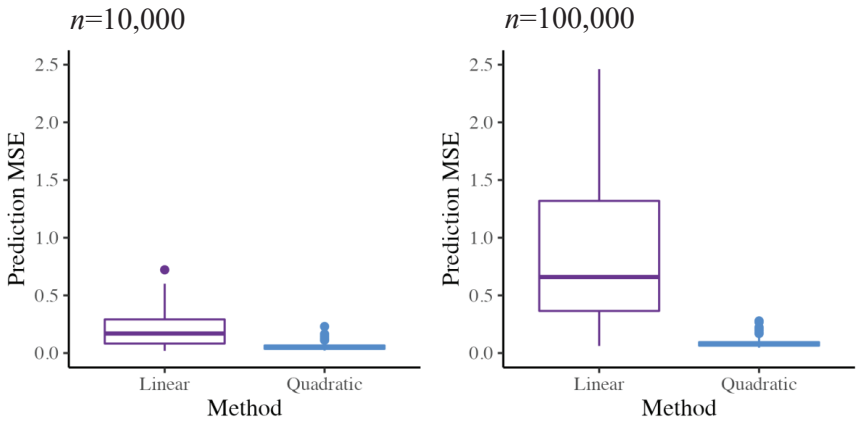

Figure 7. Out-of-sample prediction errors when the model is misspecified.

and some middle points are inevitably selected.

By the same token, the true distribution of the weights in our proposed algorithm, which also selects points one variable at a time, differs from the theoretic results shown in Table 2. Thus, we suspect that including middle points in the proposed algorithm based on the second-order model will make the algorithm even more robust than that of Wang, Yang and Stufken (2019). We investigate this using a simulated example, in which we generate independent samples $\boldsymbol{x}_i$ from a bivariate normal distribution:

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right). \tag{5.3}$$

The responses are generated from the following model: $y = x_1 + x_2 + \cos(x_1) + \varepsilon$, where $\varepsilon \sim N(0, 3^2)$. Suppose that, without knowledge of the true data-generating process, we attempt to fit a quadratic regression model.

Figure 7 presents box plots for the out-of-sample prediction errors over 100 repetitions. For both the linear IBOSS and the quadratic IBOSS, we select subdata of size 1,000. The left and right panels show the full data size of 10,000 and 100,000, respectively. The prediction MSE values from the quadratic IBOSS algorithm are much smaller than those from the linear IBOSS algorithm. Furthermore, when the full data size increases, there are more extreme values, which leads to a further deterioration of the linear IBOSS. In contrast, the prediction MSE values from the quadratic IBOSS remain stable. This example demonstrates that the additional points selected in the middle by the quadratic IBOSS model may provide a certain level of robustness against model misspecification.

Supplementary Material

The online Supplementary Material provides proofs for Lemmas 1–3 and Theorem 3.

Acknowldgments

We thank the editor, an associate editor, and the referees for the helpful comments that have led to substantive improvement in the article. Yang's research was supported by NSF grant DMS-1811291 and DMS-2210546. He's research was supported by National Natural Science Foundation of China (NSFC-72002178). The authors contributed equally to this work and are listed alphabetically.

Appendix

Lemma 1. *For any given design* $\xi = \{((z_{i1}, \ldots, z_{ip})^T, w_i), i = 1, \ldots, q\}$, *there exists a design* $\tilde{\xi}$ *such that* $|I(\xi)| \leq |I(\tilde{\xi})|$. *Here*

$$\tilde{\xi} = \left\{ \left((\pm z_{i1}, \ldots, \pm z_{ip})^T, \frac{w_i}{2^p} \right), i = 1, \ldots, q \right\}. \tag{A.1}$$

Lemma 1 states that, in an optimal design, each independent variable should be symmetric in its possible range of values. The updated design $\tilde{\xi}$ requires splitting the weight of one design point to 2^p design points, resulting in a much larger number of support points.

Lemma 2. *For a design* $\tilde{\xi}$ *described in Lemma* 1, *there exists a design* $\bar{\xi}$ *such that* $|I(\tilde{\xi})| \leq |I(\bar{\xi})|$, *and all variables of the design points in* $\bar{\xi}$ *take the values* $-1, 0, 1$.

This lemma states that, for a given design in the form of $\tilde{\xi}$, we can always find a better design $\bar{\xi}$ with at most 3^p support points. Thus, a design that is optimal in the design space $\bigcap_{j=1}^{p} \{-1, 0, 1\}$ is also optimal among all designs in $\bigcap_{j=1}^{p} [-1, 1]$.

Lemma 3. *For* $l = 0, 1, \ldots, p$, *let* Θ_l *denote the set of design points with* l *elements equal to* ± 1, *and the remaining* $p - l$ *elements equal to* 0. *An optimal design assigns equal weight to all points that belong to the same set* Θ_l.

Lemma 3 states that all points in the same Θ_l should receive the same weight in an optimal design. Therefore, even though there are 3^p design points, we only have p weights to determine (one for each Θ_l, and the last weight can be decided by the constraint that all weights sum to 1).

Proof of Theorem 2. For $i = 1, \ldots, n$, $j = 1, \ldots, p$, let $x_{(i)j}$ be the ith order statistic for $x_{1j}, \ldots, x_{nj}$. For $l \neq j$, let $x_j^{(i)l}$ be the concomitant of $x_{(i)l}$ for x_j, i.e., if $x_{(i)l} = x_{sl}$ then $x_j^{(i)l} = x_{sj}$, $i = 1, \ldots, n$.

When $\boldsymbol{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, by the similar argument as that of Proof Theorem 6 in Wang, Yang and Stufken (2019), we obtain

$$\begin{aligned} x_{(i)j} &= \mu_j - \sigma_j\sqrt{2\log n} + o_P(1), \quad i = 1, \ldots, r, \\ x_{(i)j} &= \mu_j + \sigma_j\sqrt{2\log n} + o_P(1), \quad i = n-r+1, \ldots, n, \\ x_j^{(i)l} &= \mu_j - \rho_{lj}\sigma_j\sqrt{2\log n} + O_P(1), \quad i = 1, \ldots, r, \\ x_j^{(i)l} &= \mu_j + \rho_{lj}\sigma_j\sqrt{2\log n} + O_P(1), \quad i = n-r+1, \ldots, n. \end{aligned} \tag{A.2}$$

By Equation (A.2), and the definitions of $M(\boldsymbol{\delta})$ and $\boldsymbol{f}(\boldsymbol{x}_i)$ (Equations (2.3) and (2.1), respectively), we can directly verify Equation (2.13) holds.

When $\boldsymbol{x}_i \sim \mathrm{LN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, also by the similar argument as that of Proof Theorem 6 in Wang, Yang and Stufken (2019), we obtain

$$\begin{aligned} x_{(i)j} &= \exp(-\sigma_j\sqrt{2\log n})O_P(1), \quad i = 1, \ldots, r, \\ x_{(i)j} &= \exp(\sigma_j\sqrt{2\log n})O_P(1), \quad i = n-r+1, \ldots, n. \\ x_j^{(i)l} &= \exp(-\rho_{lj}\sigma_j\sqrt{2\log n})O_P(1), \quad i = 1, \ldots, r, \\ x_j^{(i)l} &= \exp(\rho_{lj}\sigma_j\sqrt{2\log n})O_P(1), \quad i = n-r+1, \ldots, n. \end{aligned} \tag{A.3}$$

We can directly verify Equation (2.14) using the same strategy as that of (2.13).

References

Agresti, A. (2002). *Categorical Data Analysis*. 2nd Edition. John Wiley and Sons, New York.

Baxter, N. D. (1967). Leverage, risk of ruin and the cost of capital. *The Journal of Finance* **22**, 395–403.

Brainard, W. C. and Tobin, J. (1968). Pitfalls in financial model building. *The American Economic Review* **58**, 99–122.

Budish, E., Cramton, P. and Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics* **130**, 1547–1621.

Chen, K., Li, W. and Wang, S. (2020). An easy-to-implement hierarchical standardization for variable selection under strong heredity constraint. *Journal of Statistical Theory and Practice* **14**, 1–32.

Cheng, Q., Wang, H. and Yang, M. (2020). Information-based optimal subdata selection for big data logistic regression. *Journal of Statistical Planning and Inference* **209**, 112–122.

Choi, N. H., Li, W. and Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association* **105**, 354–364.

Deakin, E. B. (1976). Distributions of financial accounting ratios: Some empirical evidence. *The Accounting Review* **51**, 90–96.

Dette, H. and Melas, V. (2011). A note on the de la Garza phenomenon for locally optimal designs. *The Annals of Statistics* **39**, 1266–1281.

Dette, H. and Schorning, K. (2013). Complete classes of designs for nonlinear regression models and principal representations of moment spaces. *The Annals of Statistics* **41**, 1260–1267.

Drineas, P., Magdon-Ismail, M., Mahoney, M. and Woodruff, D. (2012). Faster approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research* **13**, 3475–3506.

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

Farrell, R., Kiefer, J. and Walbran, A. (1967). Optimum multivariate designs. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* **1**, 113–138. University of California Press, Berkeley.

Jensen, M. C. (1986). Agency costs of free cash flow, corporate finance, and takeovers. *The American Economic Review* **76**, 323–329.

Kiefer, J. (1961). Optimum designs in regression problems, II. *The Annals of Mathematical Statistics* **32**, 298–325.

Kôno, K. (1962). Optimum design for quadratic regression on k-cube. *Memoirs of the Faculty of Science, Kyushu University. Series A, Mathematics* **16**, 114–122.

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

Martin, R. (2007). Wall street's quest to process data at the speed of light. *Information Week* **4**, 07.

Modigliani, F. and Miller, M. H. (1958). The cost of capital, corporation finance and the theory of investment. *The American Economic Review* **48**, 261–297.

Pukelsheim, F. (2006). *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia.

Sitter, R. R. and Torsney, B. (1995). Optimal designs for binary response experiments with two design variables. *Statistica Sinica* **5**, 405–419.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267–288.

Tobin, J. (1969). A general equilibrium approach to monetary theory. *Journal of Money, Credit and Banking* **1**, 15–29.

van Dyk, D., Fuentes, M., Jordan, M., Newton, M., Ray, B., Lang, D. et al. (2015). ASA statement on the role of statistics in data science. *Amstat News*. Web: https:// magazine. amstat.org/blog/2015/10/01/asa-statement-on-the-role-of-statistics-in-data -science/.

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

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

Wang, L., Elmstedt, J., Wong, W. K. and Xu, H. (2021). Orthogonal subsampling for big data linear regression. *arXiv preprint arXiv:2105.14647*.

Wang, X., Yang, M. and Li, W. (2021). Efficient data reduction strategies for big data and high-dimensional Lasso regressions. *arXiv: 2401.11070*.

Yang, M. (2010). On the de la Garza phenomenon. *The Annals of Statistics* **38**, 2499–2524.

Yang, M. and Stufken, J. (2009). Support points of locally optimal designs for nonlinear models with two parameters. *The Annals of Statistics* **37**, 518–541.

Yang, M. and Stufken, J. (2012). Identifying locally optimal designs for nonlinear models: A simple extension with profound consequences. *The Annals of Statistics* **40**, 1665–1681.

Yang, M., Zhang, B. and Huang, S. (2011). Optimal designs for binary response experiments with multiple variables. *Statistica Sinica* **21**, 1415–1430.

Li He

Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China.

E-mail: lhe@swufe.edu.cn

Difan Song

Georgia Institute of Technology, Atlanta, GA 30332, USA.

E-mail: dfsong@gatech.edu

William Li

Shanghai Advanced Institute of Finance, Shanghai Jiao Tong University, Xuhui District, Shanghai 200030, China.

E-mail: wlli@saif.sjtu.edu.cn

Min Yang

Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA.

E-mail: myang2@uic.edu

(Received January 2022; accepted July 2022)