A Systematic View of Information-Based Optimal Subdata Selection: Algorithm Development, Performance Evaluation, and Application in Financial Data

Li He William Li Difan Song Min Yang *

July 26, 2022

Abstract

With the urgent need of analyzing extraordinary amount of data, the information-based optimal subdata selection (IBOSS) approach has gained considerable attention in the recent literature due to its ability to maintain rich information of the full data. On the other hand, there lacks a systematic exploration of the framework, especially the characterization of the optimal subset when the model is more complex than first-order linear models. Motivated by a real finance case study concerning the impact of corporate attributes on firm value, we systematically explore the framework consisting of the exact steps one can follow when employing the idea of IBOSS for data reduction. In the context of the second-order models, we develop a novel algorithm of selecting an

*Li He is Assistant Professor, Southwestern University of Finance and Economics, Chengdu, China (lhe@swufe.edu.cn). William Li is Professor, Shanghai Advanced Institute of Finance, Shanghai Jiao Tong University, Shanghai, China (wlli@saif.sjtu.edu.cn). Difan is Graduate Student, H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 (dfsong@gatech.edu). Min Yang is Professor, Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, IL 60607 (myang2@uic.edu). These authors contributed equally to this work and were listed alphabetically.

informative subdata. We also provide a thorough evaluation of the performance of the proposed algorithm from the standpoints of both predictions and variable selection, the latter of which is important for complex models but has not been given enough attention in the IBOSS field. Empirical studies including a real example demonstrate that the new algorithm adequately addresses the trade-off between the computation complexity and statistical efficiency, one of six core research directions for theoretical data science research proposed by the US National Science Foundation (NSF, 2016). The real case study demonstrates the potential impact of the IBOSS strategy in scientific fields beyond statistics. In particular, we note that finance field, where the speed is critically important, is a promising area for applications of IBOSS.

Keywords: Algorithm, Computation complexity, IBOSS, Statistical efficiency

1 Introduction

The extraordinary amount of data that are collected easily with the proliferation of electronic devices offers us unprecedented opportunities for scientific discovery and advancement. At the same time, using these massive data sets also presents unprecedented challenges due to not only the volume of data, but also the speed with which it must be analyzed. While computational resources have also been growing rapidly, under the traditional analysis approaches, the astonishing sizes of massive datasets dwarf the growth of computational resources. New statistical methods and computational algorithms are needed to "convert data into knowledge" (van Dyk et al., 2015).

One attractive idea that has received considerable attention for dealing with massive data (full data) is on intelligently storing/analyzing a subset of the data (subdata). The sampling distribution-based optimal subsampling approach is one way to select a subdata. Some exemplary works include 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 contained in big datasets. Under linear models, Wang, Yang, and Stufken (2019) developed a novel Information-Based Optimal Subdata Selection (IBOSS) method. They proved 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 have received increasing attention in recent years. Follow up works include Cheng, Wang, and Yang (2020) that dealt with logistic regression models, Wang, Yang, and Li (2021) that discussed a LASSO-IBOSS approach for models with a large number of variables, and Wang et al. (2021) that proposed an orthogonal subsampling approach to select a subset under linear model setups.

These encouraging results have built a strong theoretical foundation for the IBOSS-based subdata selection. However, there are still various challenges that need to be addressed before we observe more widespread use of IBOSS strategies in practice. First, most of existing

results were obtained based on the linear models that contains main effects only. Modern big data problems can be complex for which a main-effect model often is not adequate. It would be desirous to build a general framework that provides guidance on how to develop an IBOSS-type of algorithm for a given model. The second major limitation in existing studies assess the performance of IBOSS from the perspective of parameter estimation in model fitting. This was probably sufficient when the model is relatively simple. But for a more complex model consisting of a large number of model terms, we may also be interested in knowing the prediction capability of the model using IBOSS. It is also desirous to assess the performance of IBOSS from the standpoint of variable selection procedures. Another "missing piece" in the field of IBOSS was a real case study that showed not only IBOSS could preserve the rich information in the full data but also the savings in computing time from using IBOSS would be worth it in the practical situation.

This paper aims to make three important contributions from these aspects. First, we systematically explore the IBOSS framework consisting of the exact steps one can follow when employing the idea of IBOSS for data reduction. Due to the relatively simple model format used in most existing papers in this field, little attention has been paid to exploring how to characterize the informative points through optimal designs, arguably the most important and challenging step. The resulting characterization will dictate the procedure obtaining an appropriate algorithm. We describe a general framework of IBOSS that is applicable to any given model, consisting of three steps:

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

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

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

Motivated by an important research question in finance, we discuss the application of the

framework in the context of optimal second-order designs. Section 2 contains a step-by-step guidance on how to address important issues in the framework. We show that the resulting IBOSS algorithm, which selects not only extreme end points, but also middle points, is a novel approach that is different from all existing IBOSS strategies. We note that the same technique may be used for obtaining optimal designs for other models, such as polynomial models and generalized linear models. In the discussion section, we shall briefly discuss the use of the framework for a non-linear model and present some novel results.

The second contribution is to provide a comprehensive and thorough evaluation of the IBOSS strategy from the standpoint of variable selection. In the context of second-order models, we assessed the variable selection performance in terms of the sensitivity and specificity of our method and compared them with those of uniform sampling and leverage sampling. The results were encouraging for the proposed IBOSS strategy. For some model settings, the proposed IBOSS strategy can identify nearly as many significant terms as using the full data. At the same time, it can have higher specificity than using the full data, implying that using the IBOSS subdata does not incorrectly identify non-significant model terms more often than using the full data. Note that the time complexity of the new algorithm is $O(np + kp^4 + p^6)$, which represents substantial computational savings compared with the time complexity of analyzing the full dataset of $O(np^4 + p^6)$ as k is much smaller than n.

The third contribution we aim to make is to investigate the applications of IBOSS in other scientific fields beyond statistics. The motivating example of this project was a real finance case study concerning the impact of corporate attributes on firm value. We chose to investigate the relationship between firm value and other variables such as firm's asset, cash, capital expenditure, and leverage because it was suspected that the relationship between the response and many of those variables can be better represented by a second-order model. The results were very promising. Using 181,755 data points from all U.S. non-financial public firms, we found several important second-order effects both from the full data and from the IBOSS-selected subdata that had been largely neglected in the related finance literature. We

shall demonstrate in Section 4 that the proposed IBOSS strategy indeed can preserve the rich information from the full data. At the expense of slightly higher prediction MSE from the IBOSS subdata, the computational savings from IBOSS is substantial (4.32 seconds vs. 79.59 seconds).

There are several important reasons why we chose to investigate the use of the IBOSS strategy in finance. It has been noted that the IBOSS strategy works particularly well when the distribution of the independent variables is heavy tailed, which is exactly what happens in many financial data. More importantly, speed is critically important in this field. For example, it has been stated that a 1 millisecond advantage can be worth $100 million to a major brokerage firm (Martin, 2007). Three factors that impact the speed are proximity, hardware, and highly efficient algorithms. The demand for faster trading speed induces an arms race for faster trading algorithms and better trading infrastructure among high-frequency trading (HFT) firms. For instance, Budish, Cramton, and Shim (2015) found that the arbitrage opportunities between S&P 500 index, which is essentially the benchmark of U.S. stock market, and S&P 500 futures, declined substantially by over 92 percent, from 97 milliseconds in 2005 to 7 milliseconds in 2011, due to HFT. At the same time, the efficiency of algorithms has been widely considered as critically important, as predictability does not mean anything if the trader cannot act on those predictions promptly. Consequently, a good subset strategy such as the proposed IBOSS algorithm represents a very 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 the D-optimal design and a computationally efficient algorithm based on the characterization. The performance of the proposed algorithm is examined through extensive simulations in Section 3. Section 4 is devoted to the application of the proposed IBOSS strategy in the finance data. Several important issues are discussed in Section 5. Most technical proofs are presented in the supplementary material.

2 Application of the framework in second-order models

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

2.1 Model setup and information matrix

Let $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$, 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:

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

where $\boldsymbol{f}(\boldsymbol{x}_i)$ is a vector in the following order: $f_1(\boldsymbol{x}_i) = 1$; $f_{1+j}(\boldsymbol{x}_i) = x_{ij}^2$, $1 \le j \le p$; $f_{1+p+j}(\boldsymbol{x}_i) = x_{ij}$, $1 \le j \le p$; 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 $1 \le j \le p-1, j < j' \le p$; $\boldsymbol{\beta}$ 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 $\hat{\boldsymbol{\beta}}_{full} = \left(\sum_{i=1}^n \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i)\right)^{-1} \sum_{i=1}^n \boldsymbol{f}(\boldsymbol{x}_i) y_i$. 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 $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, this matrix is the Fisher information matrix. While we do not impose the normality assumption, we still call $\mathbf{M}_{full}$ the information matrix for simplicity.

For extraordinary size of n, we aim to use a subdata with k observations for regression. Let δ_i be an indicator variable, $\delta_i = 1$ if the ith data point is in the subdata, $\delta_i = 0$ otherwise, and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$. Using the notation $\boldsymbol{\delta}$ to denote the subdata, the resulting estimator can be written 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 sporadic cases (e.g., first-order linear model, logistic model with one independent variable). For most general models, obtaining optimal designs may be very difficult. Here, we utilize a new approach called "complete classes" of designs, developed by a series of papers (Dette & Melas, 2011; Dette & Schorning, 2013; Yang, 2010; Yang & Stufken, 2009, 2012). The new tools greatly simplify the process of deriving optimal designs, and most of the available optimality results for GLMs and nonlinear models can be readily derived as special cases with the new tools.

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 the D-optimal approximate design, denoted by $\xi = \{(\boldsymbol{z_i}, w_i), i = 1,\ldots,q\}$, in which 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

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

It is rather complicated to maximize $I(\xi)$ directly for a general model. In the appendix, we present a series of results for the characterization of the D-optimal design for the second-order model. The same techniques are also applicable to more general models, which will be discussed in more details in the last section.

The key ideas of the three lemmas given in the appendix involve simplifying the optimization of $I(\xi)$ by making a large number of non-zero off-diagonal elements in the information matrix to zero, as well as reducing the number of *distinct* non-zero elements in the matrix.

This can be achieved by exploring the symmetry of variables for the model considered. Using a two-dimensional example for illustration, it 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, like {-1, 0, 1} for the quadratic model; (iii) explore symmetry between 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 0. 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 $\Theta_l, l = 0, 1, \ldots, p$. Further, 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

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

$B = \text{diag}(\underbrace{a, \ldots, a}_{p}, \underbrace{b, \ldots, b}_{p(p-1)/2})$. 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 $\bar{\xi}$, we need to find optimal weights π_i^* $(i = 0, 1, \ldots, p$. As the support points are in $\bigcap_{j=1}^{p}\{-1, 0, 1\}$, it can be shown that optimal π_i^* can be obtained by solving the 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* $\xi^* = \left\{\left((z_{i1}^*, \ldots, z_{ip}^*)^T, \pi_i^*\right), i = 1, \ldots, q\right\}$*, where* z_{ij}^* *takes values -1, 0, and 1, and* π_i^* *satisfies* (2.10)*, for which* a^* *and* b^* *are 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). Notice several papers have studied optimal designs for second-order models in earlier years in the literature (Kiefer 1961; Kôno 1962; Farrell, Kiefer, and Walbran 1967). Their main ideas were to start from a *guessed* optimal design and then verify the optimality using the equivalence theorem. In comparison, we derived them from Lemmas 1–3. There are three equations in (2.10), which shows that solutions are not unique when $p \geq 3$. One way of solving this problem is to allow only some of the weights to be non-zero, which was the approach taken by Kiefer (1961) and Kôno (1962). More specifically, the former approach considered designs where support points are restricted to corners, midpoints of edges, and centers of two-dimensional faces, so that $\pi_p, \pi_{p-1}, \pi_{p-2} > 0$. And the latter approach provided solutions for designs with support on corners, midpoints of edges, and the origin $(\pi_p, \pi_{p-1}, \pi_0 > 0)$. We showed their numerical results for $p = 3$ and $p = 4$ in Table 1 for illustration. As seen in 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.

The more general results are given in Farrell, Kiefer, and Walbran (1967), and most of our results in this subsection are similar to theirs. Again, the difference is that they utilized some geometry arguments and the general equivalence theorem. As stated in Farrell, Kiefer, and Walbran (1967, page 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 be more easily adapted to obtaining optimal designs for more general models. One such example for non-linear models will be given in Section 5.1.

2.3 Step 2: Algorithm and its properties

There are two challenges in selecting optimal subdata under the model (2.1). When $p \geq 3$, there is an infinite number of optimal designs to choose from. Furthermore, after an optimal design is chosen, it usually requires a substantial number of points where multiple variables take the extreme values, which may not exist in the full data.

Fortunately, all optimal designs ξ^* having the form of (2.5), independent of π_i, satisfy a common property when the design space is projected onto a 1-dimensional space of each independent variable. As can be seen in (2.7) and the definitions of a and b given below (2.7), 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 the support points of the design. Along with the condition

that $z_{ij} = -1, 0, 1$, the support points always have

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

as one-dimensional projections. We now propose the main IBOSS algorithm for the 2nd-order model.

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

1. *Determine a^* according to Equation (2.8). Then calculate $r_1 = \lceil r \cdot a^* \rceil, r_2 = r - r_1$.*

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 $\frac{x_{(1)1}+x_{(n)1}}{2}$.*

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 $\frac{x_{(1)j}+x_{(n)j}}{2}$.*

4. *Let $\delta_i, i = 1, \ldots, n$ be an indicator variable, $\delta_i = 1$ if $(\boldsymbol{x}_i, y_i)$ is selected in the previous steps, and $\delta_i = 0$ otherwise.*

Compared to existing IBOSS algorithms proposed previously in the literature, there are two key differences. First, for each independent variable, it chooses *three* types of values: the largest, the smallest, and the middle values. In comparison, almost all existing IBOSS algorithms select only the largest and the smallest values. Second, the weights assigned to each of the three types of values depends on the number of factors p. Results in Table 2 are interesting and, to certain extent, surprising. One might expect the weights given to three values of $-1, 0$, and $+1$ to be the same. Instead, it shows that the weight for the middle number should be smaller than the weights for extreme values. In addition, the weight allocation is a function of p.

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, i.e., the asymptotic properties of 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 as the criterion is generally tractable, which is crucial for a complicated information matrix. The T-criterion also has an attractive property. If the trace of resulting information matrix goes to infinity as a function of n, it implies that the sum of all eigenvalues of the matrix goes to infinity. Consequently, at least one of 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 $\boldsymbol{\mu} = (\mu_1, ..., \mu_p)$ and $\boldsymbol{\Sigma} = \boldsymbol{\sigma}\boldsymbol{\rho}\boldsymbol{\sigma}$, where $\boldsymbol{\sigma} = diag(\sigma_1, ..., \sigma_p)$ is a diagonal matrix of standard deviations and $\boldsymbol{\rho}$ is a correlation matrix. Assume that $\boldsymbol{x}_i$'s, $i = 1, ..., n$, are i.i.d. with a distribution specified below. The following results hold for $\mathbf{M}(\boldsymbol{\delta})_{jj}$, the j-th diagonal element of the information matrix for $\hat{\boldsymbol{\beta}}(\boldsymbol{\delta})$, the estimator from the proposed algorithm.*

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

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

(ii) For multivariate lognormal independent variables, i.e., $\boldsymbol{x}_i \sim LN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, then*

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

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 non-zero 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, the extensive simulation studies in the next section show that the proposed algorithm actually demonstrates the desirous asymptotic properties of $\hat{\boldsymbol{\beta}}$ and is sufficient from the practical standpoint.

3 Simulation Studies

3.1 Estimation and Prediction MSE

The first part of simulations is focused on the Mean Squared Error (MSE) criteria. We generate independent samples $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ under three different covariance structures: multivariate normal $N(\mathbf{0}, \Sigma)$, multivariate lognormal $LN(\mathbf{0}, \Sigma)$, and multivariate t-distribution

with 2 degrees of freedom $t_2(\mathbf{0}, \Sigma)$. In our study, $p = 10$, $\Sigma_{jj'} = 0.5^{\mathbb{I}(j \neq j')}$, where $\mathbb{I}(\cdot)$ is the indicator function. The training set with 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's are i.i.d. $N(0, 3^2)$.

The data generating process is repeated for $S = 100$ times. Let $\boldsymbol{\beta}_{-0}$ denote the vector of all parameters except the intercept term, $\boldsymbol{\beta}_{-0}^{(s)}$ and $\boldsymbol{\beta}^{(s)}$ denote the estimators of $\boldsymbol{\beta}_{-0}$ and $\boldsymbol{\beta}$ at the sth repetition, respectively. For each method, calculate the estimation MSE as $\text{MSE}_e = \frac{1}{S}\sum_s \left\| \hat{\boldsymbol{\beta}}_{-0}^{(s)} - \boldsymbol{\beta}_{-0} \right\|^2$ and prediction MSE as $\text{MSE}_p = \frac{1}{n_{test} \cdot S}\sum_{s,i} \left(\boldsymbol{\beta}^T \boldsymbol{f}(\boldsymbol{x}_i) - (\hat{\boldsymbol{\beta}}^{(s)})^T \boldsymbol{f}(\boldsymbol{x}_i) \right)^2$.

Following the above procedure, we conduct two simulations. In the first simulation, we generate training data of sizes $n = 5,000$, 10^4, 10^5 and 10^6 while fixing the subdata size at $k = 1,000$. Figure 1 compares all approaches in terms of MSE_e and MSE_p (in logarithm scale) when n increases. For estimation MSE, Panel (a) of Figure 1 shows that the IBOSS approach dominates uniform sampling and leverage sampling for all three distributions considered. The MSE_e values of the IBOSS approach decrease with the 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 comparable MSE_e as using the full data. For 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) of Figure 2, respectively. 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 IBOSS 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 yellow curve (corresponding to uniform sampling) moving upwards from the top to the bottom in the three figures in Panel (b). In comparison, the three green curves corresponding to leverage sampling hardly move, and

the three blue curves corresponding to IBOSS actually move downwards, indicating smaller prediction errors for heavy tailed distributions.

Table 3 compares computing times of the three sampling methods ($k = 1000$) with using the full data. For the full data approach, each number amounts to the time used for fitting the second-order regression model. For the IBOSS and other subsampling approaches, each number is the total CPU time for selecting the subdata and then fitting the model. For the IBOSS approach, a C++ function is implemented to find out the desired quantiles of a given vector. All other codes are implemented in R and carried out on a laptop with Intel® Core™ i7-10710U processor and 16GB memory. Not surprisingly, uniform sampling and the IBOSS approach take a fractional of the computation time using the full data. Leverage sampling shows no reduction in computational time since calculating the leverage values has the same complexity as fitting the full model. Note that when n or k increases, both the IBOSS and full data approaches requires larger computation time. But the computation time of the IBOSS method increases at a much slower rate than fitting with the full data. Overall, combined with the results in terms of MSE_e and MSE_p, the IBOSS approach appears to achieve satisfactory statistical efficiency at a small computational cost.

3.2 Sensitivity and specificity

Most of existing papers examined the performance of the IBOSS strategy focusing on estimation error, and not much attention has been given to its variable selection capability. This was not an issue as most models in existing studies 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, it is also important to assess the capability of the model using the IBOSS subdata to effectively identify the significant effects.

Variable selection has been studied extensively in the literature. See, for example, Tibshirani (1996), Fan and Li (2001), and Choi, Li, and Zhu (2010). More recently, (Wang, Yang, & Li, 2021) proposed a LASSO approach for an IBOSS subdata strategy under the

first-order linear model. We now evaluate the performance of IBOSS subdata under the second-order model (2.1). Traditionally, two important variable selection criteria are *sensitivity* and *specificity*, which are defined respectively as follows:

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

$$\text{specificity} = \frac{\text{number of unselected insignificant effects}}{\text{total number of true insignificant effects}}.$$

We adopt a similar simulation setting as the ones in Choi, Li, and Zhu (2010) and Chen, Li, and Wang (2020). Five settings are considered in Table 4. There are $p = 10$ lognormally 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 $\binom{p_1}{2}$ 2-factor interactions between them are significant. Among the five settings, Setting 1 can be considered as a "base" setting; Setting 2 represents a model with larger coefficients for a part of the five 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 5 to 7; 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 $\sigma = 100$. The stepwise regression approach is used with the AIC criterion in model fitting.

Figure 3 compares box plots of distributions of sensitivity and specificity over 100 tries for four methods: full data, the subdata using IBOSS, uniform sampling and leverage sampling ($k = 1000$). For all settings, the sensitivity of IBOSS is comparable to the full data, both of which are significantly better than the uniform sampling and leverage sampling. Compared to the base setting of Setting 1, when the coefficients of some significant terms increase in Setting 2, the sensitivities of all four methods increase, but the overall trend remains the same as 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 both the full data and IBOSS was near 100% for most tries, indicating that both appeared to have strong capability of

identifying 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 all the methods are similar to Setting 1, as expected. The results in Setting 5 are very 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 dataset. With more candidate points available, the points selected by IBOSS also becomes more informative. This feature makes the IBOSS approach advantageous for big data regression.

The specificity values of the four methods are similar in all settings. It is interesting to notice 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 the full data in Setting 3. This suggests that IBOSS does not incorrectly choose insignificant second-order effects as often as the full data approach. The specificity of uniform sampling and leverage sampling are better than the IBOSS approach. However, given their poor performance on sensitivity, which is arguably more important than specificity by many scholars, these methods are clearly inferior.

To assess the overall performance of both sensitivity and specificity, we also plotted sensitivity and "1− specificity" in Figure 4. In each figure, a dot at the upper left corner would mean that the corresponding method has an overall good performance in both sensitivity and specificity. It is clear that in all five settings, the overall performance of IBOSS is very close to the full data, and significantly better than uniform sampling and leverage sampling.

The settings considered in Table 4 assume that all variables follow a lognormal distribution. Unreported simulation results based on other distribution assumptions shows similar patterns, although IBOSS performs better for more heavy tailed distributions, consistent with findings in most existing IBOSS studies. Interestingly, many financial data are indeed heavy tailed. For example, in the finance case we studied in this paper, 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 that were based on normal distribution assumptions may not be valid. For example, Deakin (1976) showed that many important financial ratios were proved to be non-normally distributed. The authors urged researchers to be cautious when using these financial data in empirical studies.

In sum, the simulation results showed that the proposed IBOSS subdata strategy clearly outperformed alternative uniform sampling and leverage sampling methods. As noted by an anonymous reviewer, Table 2 shows that when p increases, the weight assigned to center points is closer to zero. Figure 5 provides further information between a^* and p. In the simulations, we noted that the proposed Algorithm 1 may select similar points as those selected using the IBOSS algorithm of (Yang & Stufken, 2009), and the similarity is more notable for larger p. This phenomenon stems from not only the smaller weight assigned to center points, but also that both IBOSS algorithms used a one-variable-at-a-time approach, and the final selected subdata are only a proxy for the theoretically optimal solution. Nonetheless, we will show in the two next sections that the proposed algorithm enjoys clearly advantages for real cases when the true model is more complicated than used in the simulation. Further, it has better robustness properties against possible missing terms.

4 A finance case study

One of the fundamental research questions in finance literature is the impact of corporate attributes on firm value. We study the relationship between the firm value and several important variables by using a second-order model. Typical approaches in finance usually involved identifying one main independent variable and several control variables and then running a linear regression models. More often than not, first-order models were used, and less attention was paid to quadratic and interaction effects. However, second-order effects may also be significant in corporate finance studies. For instance, financial leverage has been

considered as one of the key variables that are related to firm value. But its relationship with firm value may not be linear. In fact, there have been contradicting results reported in the literature. The Modigliani-Miller Theorem (Modigliani & Miller, 1958) stated that the value of the firm is *independent* of the firm's capital structure. Then Baxter (1967) showed the *negative* effect of leverage, which increases the financial distress costs before reaching the optimal debt-equity ratio; and Jensen (1986) reported the *positive* effect of leverage on firm value due to agency costs.

This motivated us to consider a second-order model consisting of leverage and several other critical variables, whose impact on firm value have been previously studied. The variables included in the model are: 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 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 shareholder's 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 was how to measure the firm value. Tobin's Q, which is 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 behind the notion of Tobin's Q is that it captures profitable investment opportunities. Higher Tobin's Q values suggest that the firm uses the economics resources more effectively because the market value created by firm's assets is higher than the cost of reproducing the firm's underlying assets.

We follow the conventional way of many finance researchers who have extensively used the Compustat Fundamentals database to examine the effect of various firm-specific attributes on Tobin's Q. The Compustat database collects the financial statement and financial market

data of all the U.S. publicly traded companies and is published by Standard & Poor's Global. We select the data of all U.S. non-financial public firms for 1980 through 2020 and delete the observations with missing values. The final data consists of 181,755 firm-year observations, representing 20,117 distinct firms. We calculate the dependent variable Tobin's Q as the market value of a company (common shares outstanding multiple with 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 amounts to approximately 5% of the full data, in Figure 6. In the calculation of 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 criterion. For $k = 10,000$, the MSE for using the IBOSS data is 2.456, which was only slightly higher than the MSE value of 2.434 for using the full data. However, the CPU time for the IBOSS approach was only 4.32 seconds, which is substantially smaller than 79.59 seconds for using the full data. In an industry where the improvement is often measured by milliseconds and one millisecond advantage could be worth $100 million (Martin, 2007), the CPU time savings from the IBOSS strategy have enormous financial implications.

Figure 6 also compares the proposed IBOSS strategy (labeled as "IBOSS (quadratic)") with leverage sampling, the existing IBOSS of Wang, Yang, and Stufken (2019) that was developed for the linear model (labeled as "IBOSS (linear)"), and the linear model using the full data. MSE values are computed for various k values from 2,000 to 10,000. We first note that the proposed IBOSS strategy is superior to leverage sampling for all subdata sizes. Not only does the former have smaller MSE than the latter, leverage sampling sometimes becomes unstable, as evidenced by the surprisingly increased MSE when k increases from 8,000 to 10,000. Another interesting observation is that the IBOSS (linear) approach performs nicely even if the true model is quadratic, indicating that the IBOSS algorithm of Wang, Yang, and Stufken (2019) can have some nice robust properties against model misspecifications. We shall provide with more detailed discussions on robustness in the next section. Finally,

Figure 6 also demonstrates that it is important not to ignore second-order terms in a model. As mentioned previously, in the research of the finance field, scholars often focus on the first-order linear model. This would result in significantly higher MSE values even if the full data was used. Finally, Table 5 assesses the performance of the proposed IBOSS strategy from the standpoint of variable selection. The results in the table correspond to $k = 10,000$. It can be seen that the IBOSS subdata allows us to identify most of the terms that are shown to be significant using the full model. IBOSS identifies all the main effects and most of second-order terms that are identified to be significant from using the full data.

5 Discussion

5.1 Characterization for a non-linear model

We have been focusing on developing an IBOSS strategy for the second-order model. However, the proposed framework is applicable to a more general model. As an example, we now briefly discuss how to develop an IBOSS strategy for a non-linear 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 and were mainly focused on the model containing main effects only (Sitter & Torsney, 1995; Yang, Zhang, & Huang, 2011). As far as we know, there is no optimality result available when interaction effects are present in a multivariate logistic regression model. In this section, utilizing the aforementioned strategy, we 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 covariates level $\boldsymbol{x}_i$, $p \geq 3$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p, \beta_{12}, \ldots, \beta_{p-2,p-1})$ are unknown parameters. We assume the first $p-1$ covariates are bounded,

i.e., $x_{ij} \in [L_j, U_j]$, $j = 1, \ldots, p-1$ and there is no constraint on the last covariate, i.e., $x_{ip} \in (-\infty, \infty)$. Such an assumption is typical for deriving optimal designs under multivariate logistic regression models (Sitter & Torsney, 1995; Yang, Zhang, & 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 $\xi = \{(\boldsymbol{c}_i, \omega_i), i = 1, ..., k\}$. Let

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

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. One can 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 other existing IBOSS approaches in the literature, the characterization of optimal designs in the general framework depend on model assumptions. An important issue which, as far as we know, has not been adequately addressed previously, is how robust they are against model misspecifications. For example, is the IBOSS algorithm proposed by Wang, Yang, and Stufken (2019), which was based on the linear model, also effective for the second-order model (2.1)? This important question was raised by an anonymous reviewer. For the

finance example, we compared the performance of two IBOSS approaches. As shown in Figure 6, the IBOSS (linear) approach performed surprisingly well even if the true model has significant second-order terms. Across all selected subsample sizes, it has slightly higher prediction MSE values than the proposed algorithm, labeled as IBOSS (quadratic), but outperforms both random sampling and leverage sampling. This demonstrates that the IBOSS algorithm proposed by Wang, Yang, and Stufken (2019) is robust against possible important second-order terms in the true model. One possible explanation is that their IBOSS 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, and some middle points would be selected inevitably.

By the same token, the true distribution of weights in our proposed algorithm, which also takes up the one-variable-at-a-time approach, would be different from those theoretic results shown in Table 2. Thus, we suspected that the inclusion of middle points in the proposed algorithm based on the second-order model would make the algorithm even more robust than the one of Wang, Yang, and Stufken (2019). We investigated this by 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 would attempt to fit a quadratic regression model.

Figure 7 presents boxplots for out-of-sample prediction errors over 100 repetitions. For both linear IBOSS and quadratic IBOSS, we select a subdata of size of 1,000. The left and right panel has full data size of 10,000 and 100,000, respectively. We can see that the prediction MSE values resulting from using the quadratic IBOSS algorithm are much

smaller than those using the linear IBOSS algorithm. Furthermore, when the full data size increases, there are more extreme values that lead to a further deterioration of linear IBOSS. In contrast, the prediction MSE values resulting from quadratic IBOSS remain stable. This example demonstrates that the additional points selected in the middle by the quadratic IBOSS models may provide a certain level of robustness against model misspecification.

References

Agresti, A. (2002). *Categorical data analysis* (2nd). John Wiley; Sons, New Work.

Baxter, N. D. (1967). Leverage, risk of ruin and the cost of capital. *the Journal of Finance*, *22*(3), 395–403.

Brainard, W. C., & Tobin, J. (1968). Pitfalls in financial model building. *The American Economic Review*, *58*(2), 99–122.

Budish, E., Cramton, P., & Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics*, *130*(4), 1547–1621.

Chen, K., Li, W., & 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., & 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., & Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, *105*(489), 354–364.

Deakin, E. B. (1976). Distributions of financial accounting ratios: Some empirical evidence. *The Accounting Review*, *51*(1), 90–96.

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

Dette, H., & 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., & Woodruff, D. (2012). Faster approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, *13*, 3475–3506.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348–1360.

Farrell, R., Kiefer, J., & Walbran, A. (1967). Optimum multivariate designs. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, *1*, 113–138.

Jensen, M. C. (1986). Agency costs of free cash flow, corporate finance, and takeovers. *The American economic review*, *76*(2), 323–329.

Kiefer, J. (1961). Optimum designs in regression problems, ii. *The Annals of Mathematical Statistics*, *32*(1), 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*(2), 114–122.

Ma, P., Mahoney, M., & 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*(21), 07.

Modigliani, F., & Miller, M. H. (1958). The cost of capital, corporation finance and the theory of investment. *The American economic review*, *48*(3), 261–297.

NSF. (2016). Transdisciplinary research in principles of data science phase i (tripods). https://www.nsf.gov/pubs/2016/nsf16615/nsf16615.htm

Pukelsheim, F. (2006). *Optimal design of experiments*. Society for Industrial; Applied Mathematics (SIAM), Philadelphia, PA.

Sitter, R. R., & 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*(1), 267–288.

Tobin, J. (1969). A general equilibrium approach to monetary theory. *Journal of money, credit and banking*, *1*(1), 15–29.

van Dyk, D., Fuentes, M., Jordan, M., Newton, M., Ray, B., Lang, D., & Wickham, H. (2015). ASA statement on the role of statistics in data science. *Amstat News*.

Wang, H., Yang, M., & 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., & 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., & Xu, H. (2021). Orthogonal subsampling for big data linear regression. *arXiv preprint arXiv:2105.14647*.

Wang, X., Yang, M., & Li, W. (2021). Efficient data reduction strategies for big data and high-dimensional lasso regressions. *under review*.

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

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

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

Tables

Table 1: Examples of Kiefer (1961)'s and Kôno (1962)'s designs.

	$p = 3$		$p = 4$	
	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

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

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

Table 4: Setup details 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, 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 non-zero 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 0.

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

	IBOSS	Full
Main Effects	1, 2, 3, 4, 5, 6, 7, 8 12, 16, 17, 23, 26, 28	1, 2, 3, 4, 5, 6, 7, 8 13, 14, 15, 16, 23, 24, 25
Interaction Effects	35, 37, 45, 46, 47, 48 56, 67, 78	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

Figures

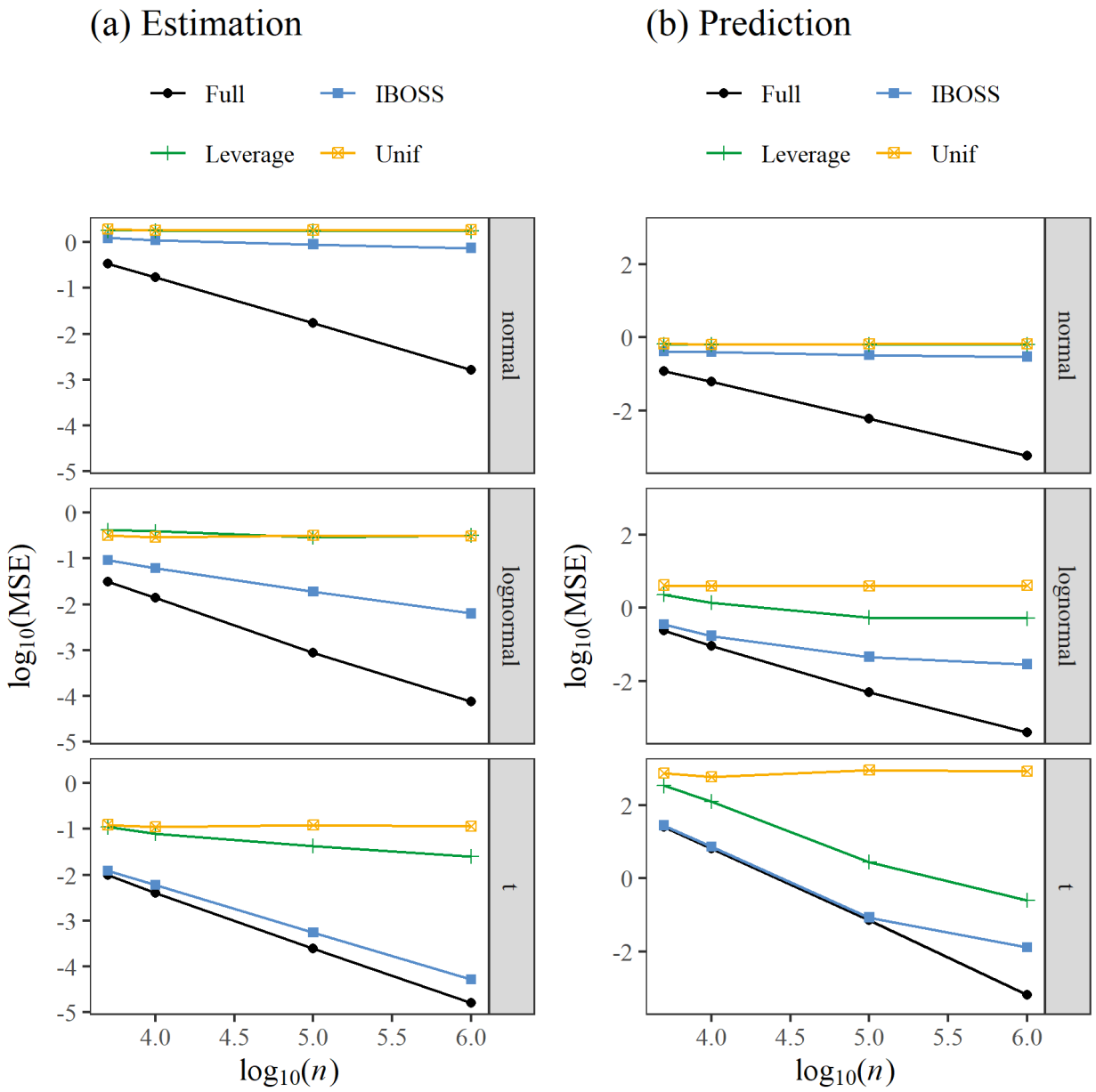

Figure 1: MSEs for estimating the slope parameter (Panel a) Out-of-sample prediction errors (Panel b) for three different distributions of the independent variables. The subdata size is fixed at $k = 1,000$ and the full data size n changes.

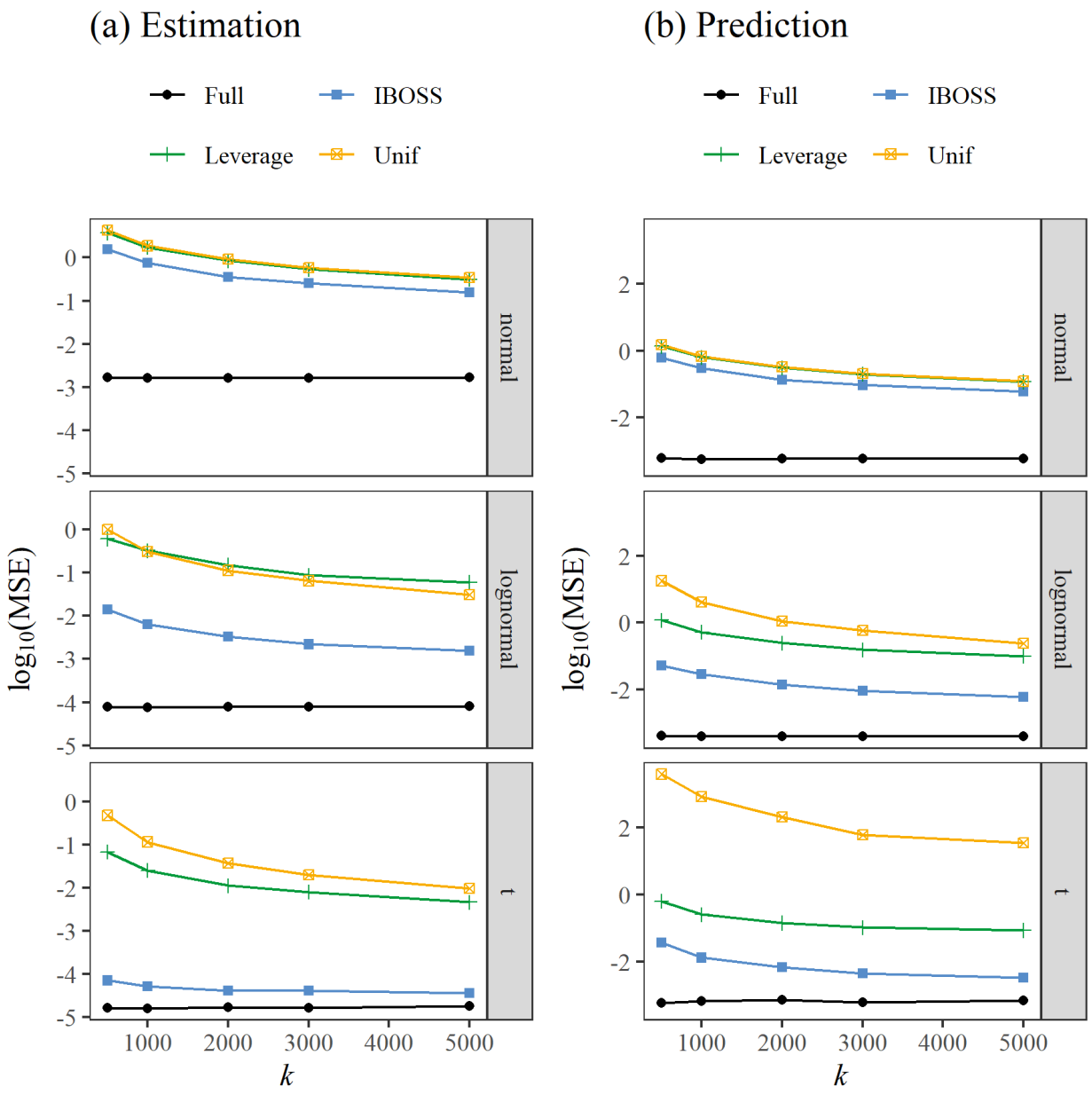

Figure 2: MSEs for estimating the slope parameter (Panel a) Out-of-sample prediction errors (Panel b) for three different distributions of the independent variables. The full data size is fixed at $n = 10^6$ and the subdata size k changes.

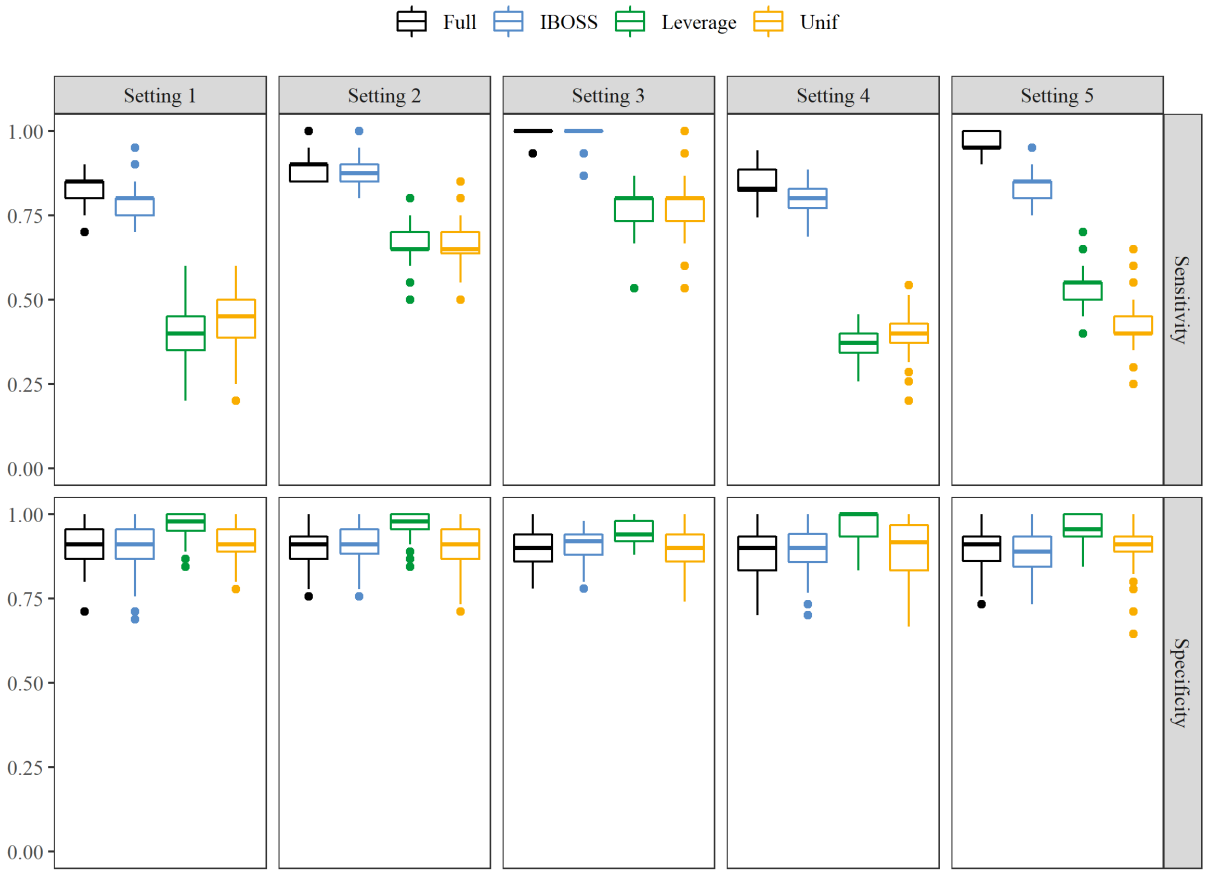

Figure 3: Sensitivity and specificity boxplots of Settings 1 - 5.

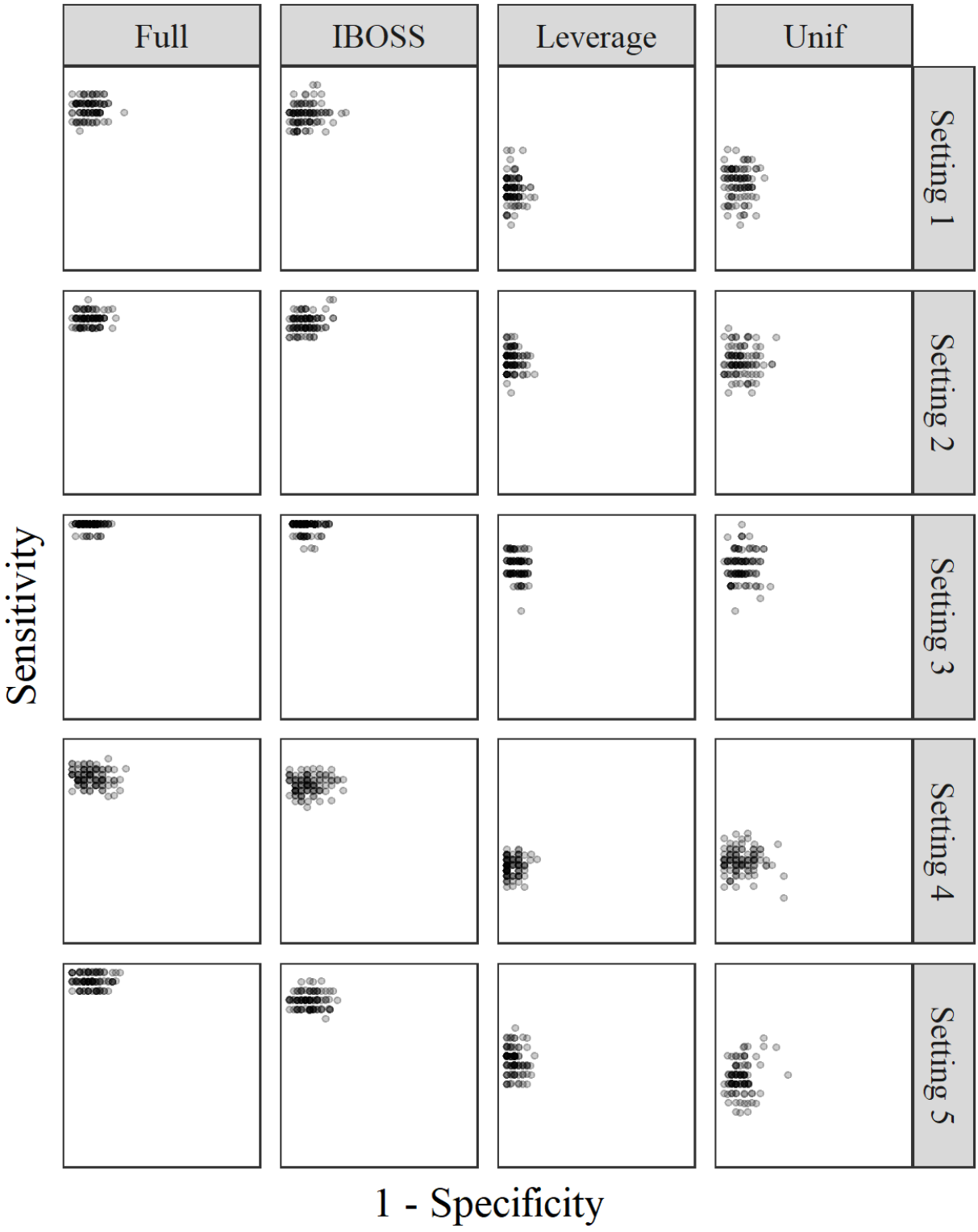

Figure 4: Sensitivity plotted against 1− specificity for Settings 1 - 5, 100 runs each. Points are jittered. Upper left points have best performances.

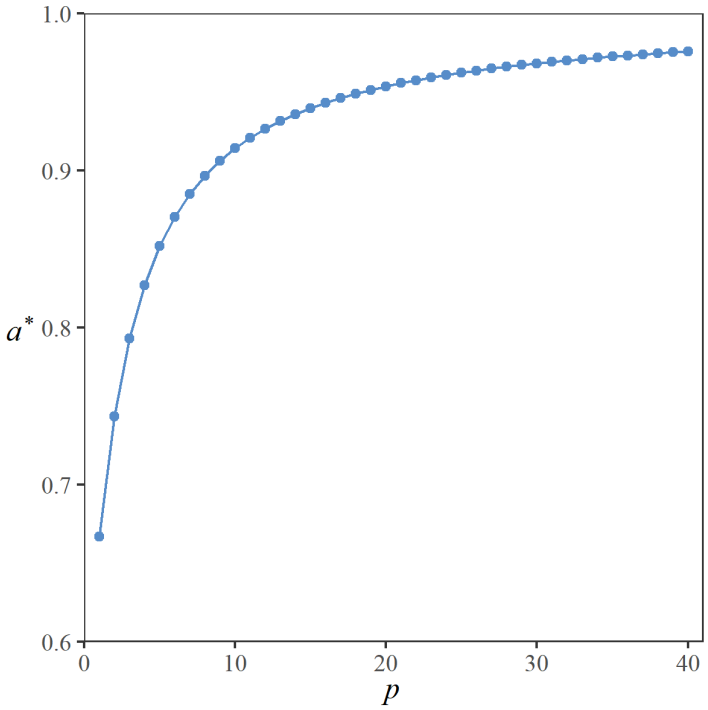

Figure 5: Relationship between p and a^*.

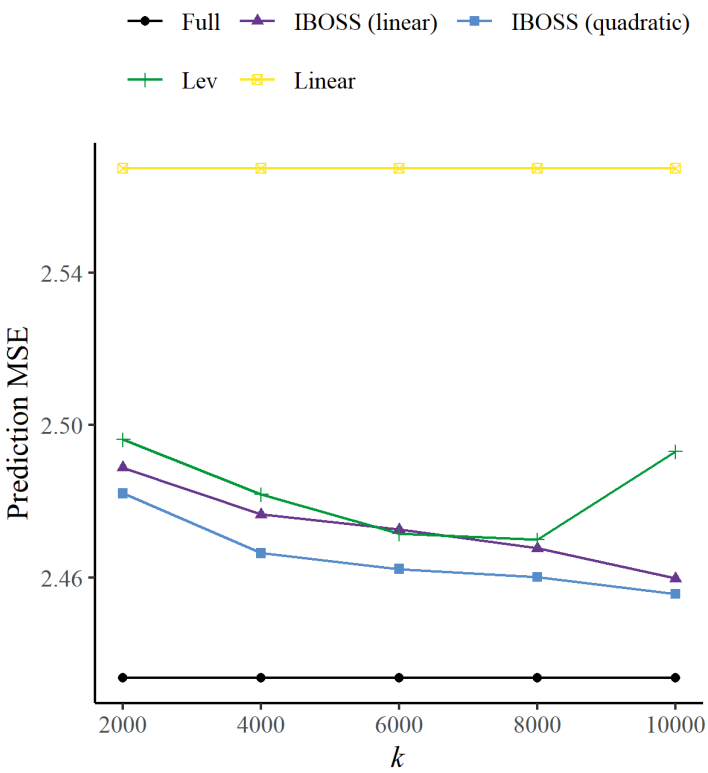

Figure 6: Out-of-sample prediction errors of different methods in real data.

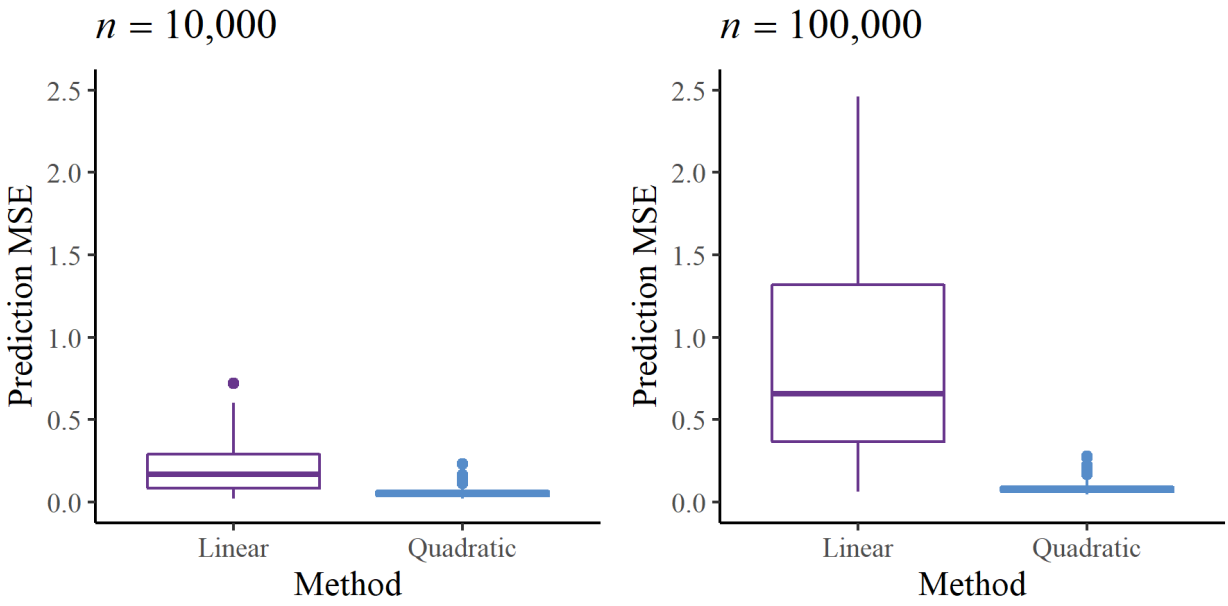

Figure 7: Out-of-sample prediction errors when the model is misspecified.

Appendix

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

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

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 $i = 1, ..., n$, $j = 1, ..., p$, 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 $x_j^{(i)l} = x_{sj}$, $i = 1, ..., n$.

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

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

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

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

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