(will be inserted by the editor)

Subdata Selection Algorithm for Linear Model Discrimination

Jun Yu · HaiYing Wang

Received: date / Accepted: date

Abstract A statistical method is likely to be sub-optimal if the assumed model does not reflect the structure of the data at hand. For this reason, it is important to perform model selection before statistical analysis. However, selecting an appropriate model from a large candidate pool is usually computationally infeasible when faced with a massive data set, and little work has been done to study data selection for model selection. In this work, we propose a subdata selection method based on leverage scores which enables us to conduct the selection task on a small subdata set. Compared with existing subsampling methods, our method not only improves the probability of selecting the best model but also enhances the estimation efficiency. We justify this both theoretically and numerically. Several examples are presented to illustrate the proposed method.

Keywords Bayesian Information Criterion · Big Data · Discrimination Design · D-optimal design · Entropy · Measurement constraints

Mathematics Subject Classification (2010) 62K05 · 62J05

1 Introduction

Understanding potential relationships among variables in massive data sets is an essential topic in the big data era. Despite the fact that more data seems to be better, diminishing marginal utility suggests that to achieve a preset statistical efficiency in estimation or hypothesis testing, we do not have to utilize all the observations. A smaller subdata set might be more cost effective while a researcher is exploring a large data set, because data visualizing and exploration are certainly easier and faster with fewer observations even when we have sufficient computer

Jun Yu
School of Mathematics and Statistics, Beijing Institute of Technology, Beijing, 100811, China
E-mail: yujunbeta@bit.edu.cn

HaiYing Wang
Department of Statistics, University of Connecticut, Storrs, CT 06269, USA
E-mail: haiying.wang@uconn.edu

resources (Boivin and Ng, 2006; Ng, 2017). Among various methods proposed to handle large data sets, subsampling has drawn the attentions of econometricians and statisticians in recent years (Lee and Ng, 2020).

Great efforts have been made to estimate unknown parameters for given models based on incredibly large data sets over the last decades. See Drineas et al. (2011); Mahoney (2012); Kleiner et al. (2015); Ma et al. (2015); Meng et al. (2017); Wang et al. (2018, 2019); Yao and Wang (2019); Ma et al. (2020); Meng et al. (2020a,b); Deldossi and Tommasi (2021); Li and Meng (2021); Yao and Wang (2021), among others. Most of the existing investigations focus on reducing the uncertainty brought by the sampling procedure, and they often ignore the uncertainty induced by the assumed models. However, model uncertainty can be quite influential for statistical inferences when the aforementioned methods are adopted. For example, as described in Wang (2019), the optimal subsampling method under the A-optimality criterion proposed in Wang et al. (2018) will be less efficient when the underlying model is misspecified. As another example, Fithian and Hastie (2014) pointed out that the case-control sampling method is inconsistent for the risk minimizer in the original population when the model is misspecified.

Model selection has been an important topic in regression problems. Two kinds of methods are routinely used by many data scientists to screen variables/models and fit regression models. One approach is the shrinkage based methods which are particularly suitable for the case that the true regression coefficient is sparse or nearly sparse, see Tibshirani (1996); Fan and Li (2001); Efron et al. (2004); Candes et al. (2007); Zhang (2010), among others. The other approach is the subset selection based methods. These methods allow the data-based selection of a single "best" model or a weighted average of multiple promising candidate models. Typical examples include F-tests for nested models, stepwise selection procedures, and model averaging. See Shao (1997); Miller (2002); Kadane and Lazar (2004); Yuan and Yang (2005); Claeskens and Hjort (2008); Zheng et al. (2019) and the references therein for examples. Besides the model discrimination investigations in observational studies, fruitful results have been achieved in design of experiments based on F-tests and Kullback-Leibler divergence. Important works include, but not limited to, Box and Hill (1967); Meyer et al. (1996); López-Fidalgo et al. (2007); Dette and Titoff (2009); Drovandi et al. (2014); Dette et al. (2015); Consonni and Deldossi (2016).

It is worth mentioning that all of the shrinkage based methods, and part of subset selection methods, such as stepwise selection, focus on picking a "good" (or useful) model. Despite the success of the aforementioned methods, the ability to discriminate between models in a candidate pool based on some criteria is also of interest in the big data era. An obvious benefit is that the discrimination results can help practitioners decide whether to use model averaging or model selection, and it is particularly useful when candidate models are not nested. More discussions can be found in Yuan and Yang (2005). These criteria are usually based on Kullback-Leibler divergence (Kullback and Leibler, 1951), such as Akaike information criterion (AIC) (Akaike, 1974), Bayesian information criterion (BIC) (Schwarz, 1978), and Kullback-Leibler information criterion (KLIC) (Sin and White, 1996). It is worth mentioning that when the data are governed by a parametric model in the candidate list, the BIC is consistent, i.e., it selects the best model with probability approaching one (Yang, 2005). To discriminate among

all possible models in a candidate pool, we focus on the subset selection method based on BIC in this work.

For massive data, computing resources become a bottleneck for statistical inference. Take a linear regression problem with n observations and p covariates as an example. If there are m candidate models to be compared, the computational cost is $O(np^2m)$. This may be infeasible for the case with a large n and moderately large m. To break the bottleneck of computation, one promising way is data reduction. However, when the inference is subject to model uncertainly, little work has been done to study how to perform data selection. Except uniform sampling, most existing strategies are designed for a prespecified model, and are not suitable to explore and prototype a variety of models. There are model-free subsample selection procedures such as support points (Mak and Joseph, 2018), Stein points (Chen et al., 2018), and the minimum energy design (Joseph et al., 2019). However, these methods rely on the empirical distribution of the full data and their computational cost is $O(n^2p)$ which is high for large n.

To address the issues mentioned above, this work studies a new subdata selection method for linear regression model selection based on BIC. Theoretically, we show that the proposed method is selection consistent under mild conditions. As illustrated in Section 5, compared with the uniform subsampling method, our method can detect a weaker signal which leads to a higher probability of selecting the best model. In addition, our method also improves the performance in estimating the true model. Extensive experiments on both simulated data sets and real-world examples show that the proposed method outperforms several state-of-the-art subsampling and subdata selection methods in terms of selection accuracy and mean square prediction error.

The rest of this article is organized as follows. In Section 2, we briefly review the model selection procedure via BIC. In Section 3, a subdata selection criterion based on maximum entropy sampling is proposed and a subdata selection algorithm based on leverage scores is designed. Theoretical results including the selection consistency and parameter estimation consistency based on the selected subdata are established in Section 4. Section 5 illustrates our methods via both simulated and real-world datasets. Section 6 concludes this article with some discussions. All technical proofs and additional simulation results are given in the Appendix.

2 Preliminaries

We assume that normally distributed independent observations $y_1, \ldots, y_n$ are generated from the process

$$y_i = \mu_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where ε_i is the normal distributed error term with mean zero and variance σ^2, and the true model for μ_i lies in a candidate set of models. Let z_i be the vector of observed values of the explanatory variables associated with each y_i, through the relationship $\mu_i = E(y_i|z_i) = f(z_i)^{\mathrm{T}}\beta$ where $f(z_i) = (f_1(z_i), \ldots, f_p(z_i))^{\mathrm{T}}$ and $f_j(z_i)$'s are p known and linearly independent functions of z_i. Here, the dimension of z_i does not have to be p. For ease of presentation, we define $\mu = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$, $x_i = f(z_i)$, $X = (x_1^{\mathrm{T}}, ..., x_n^{\mathrm{T}})^{\mathrm{T}}$, and $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$. In model selection problems, the intercept parameter, denoted as β_{int} is often not of interest, and it can

be eliminated in linear models by centralizing the full data (Y, X). Thus, the slope parameters β can be estimated by fitting the least squares without the intercept term on the centralized full data. When the intercept becomes relevant such as for prediction, it can be estimated with $\hat{\beta}_{\text{int}} = \bar{Y} - \bar{X}^{\text{T}}\hat{\beta}$, where $\bar{Y}$ is the full data mean of Y, and $\bar{X}$ is the column mean vector of X, and $\hat{\beta}$ is the least squares estimate of β from the centralized data. To ease the presentation, in the rest of the paper we assume that the full data are centralized, i.e., the column means of X and the mean of Y are zero.

Let (X^*, Y^*) be a subset of the centralized full data. The column means of X^* and the mean of Y^* may not be zero, but this does not affect the unbiasedness of the subdata estimator $\hat{\beta}_s = (X^{*\text{T}}X^*)^{-1}X^{*\text{T}}Y^*$. To see this, we first notice that if the subdata selection rule does not depend on Y, then the subdata set satisfies

$$\begin{aligned} Y^* &= \beta_{\text{int}}\mathbf{1} + X^*_{nc}\beta + \varepsilon^* - \bar{Y}_{nc}\mathbf{1} \\ &= \beta_{\text{int}}\mathbf{1} + X^*_{nc}\beta + \varepsilon^* - (\beta_{\text{int}} + \bar{X}_{nc}\beta + \bar{\varepsilon})\mathbf{1} \\ &= X^*\beta + \varepsilon^* - \bar{\varepsilon}\mathbf{1}, \end{aligned}$$

where the subscription $_{nc}$ indicates that the quantity is for the non-centralized data, ε^* is the error vector for the subdata, $\bar{\varepsilon}$ is the mean of ε_i, $i = 1, ..., n$, and $\mathbf{1}$ is a vector of ones. Thus, we know that $E(Y^*) = X^*\beta$ and therefore $E(\hat{\beta}_s) = \beta$.

In the following, we consider the model selection problem on X. Let S_k $(k = 1, \ldots, m)$ be a subset of the column indices of X, where m is the number of candidate models to consider. We use $\mathcal{M}$ to denote the set of all such candidate subsets. The candidate model corresponding to S_k has the form

$$\mu = X_{(k)}\beta_{(k)}, \tag{2}$$

where $X_{(k)}$ is an $n \times p_{(k)}$ submatrix of X and $\beta_{(k)}$ is the associated parameter vector. It is worth mentioning that for two different model S_{k_1} and S_{k_2}, the jth column of $X_{(k_1)}$ may be different from the jth column of $X_{(k_2)}$. We assume that both p and m are fixed and do not change with n. Typically, $m = 2^p$ if all possible models are considered and $m = p$ when only nested models are considered. Without loss of generality, we assume that the mth model is the widest model (also known as the encompassing model), i.e., $S_m = \{1, \ldots, p\}$, and we assume that it is always included in $\mathcal{M}$.

The kth candidate model is said to be correct if $\beta_{(k)}$ satisfies that $\mu = X_{(k)}\beta_{(k)}$. Let $\mathcal{M}^C$ be the set of correct candidate models. If there are multiple models in $\mathcal{M}^C$, we wish to select a model with the fewest parameters, and we call the correct model with fewest parameters the true model. For convenience, we define $X_{(t)}$ to be the model matrix of the true model.

We provide a brief overview of the classical BIC method in the following. Let $Y = (y_1, \ldots, y_n)^{\text{T}}$ be the response vector, and $\pi(\beta)$ be the prior distribution of β. The likelihood under the kth candidate model in $\mathcal{M}$ is

$$p(Y|S_k) = \int f(Y|\beta, S_k)\pi(\beta)d\beta, \tag{3}$$

Let $P(S = S_k)$ be the prior probability that the kth model is the true model. The posterior probability that the kth model is the true model is

$$P(S = S_k|Y) = \frac{p(Y|S_k)P(S = S_k)}{\sum_{l=1}^{m} p(Y|S_l)P(S = S_l)}, \quad k = 1, \ldots, m. \tag{4}$$

The basic idea of BIC is to select the model with the largest posterior probability using the prior $P(S = S_k) = 1/m$ for all k. It is worth mentioning that when the number of candidate models is large, as suggested in Bingham and Chipman (2007), incorporating prior information in experimental design, such as effect forcing (Meyer et al., 1996) and effect heredity (Chipman and Hamada, 1996) will help narrow down the candidate set.

Under the BIC framework, this principle means that the model that maximizes the numerator $p(Y|S_k)$ should be selected. For the normal regression model (2), we write the selected model according to BIC based on the full data set as

$$\hat{S}_{\mathrm{BIC}} = \arg\min_{S_k} \left\{ n \log \left(n^{-1} \sum_{i=1}^{n} (y_i - \hat{\mu}_i^{(k)})^2 \right) + p_{(k)} \log n \right\}, \tag{5}$$

where $p_{(k)}$ is the cardinality of S_k, $\hat{\mu}_i^{(k)}$ is the ith element of $\hat{\mu}^{(k)} = X_{(k)}\hat{\beta}_{(k)}$, and $\hat{\beta}_{(k)}$ is the maximum likelihood estimator (MLE) under the kth candidate model.

3 Motivation and Methodology

When the full data is large and the computational resources are limited or when the response variable is expensive to measure (Zhang et al., 2020), a practical solution is to select a small subdata to save the computational cost or the data collection cost. To handle the case that only a small proportion of the responses can be measured, we are interested in both selecting a good model and achieving better estimation results using the selected model based on the subdata. For this purpose, we use entropy (Lindley, 1956) as a measure of the uncertainties coming from the model selection and estimation.

Note that for a random vector α with density function $f(\cdot)$, the Shannon entropy of α is defined as $\mathrm{Ent}(\alpha) = E_\alpha(-\log f(\alpha))$. To achieve the goals on both selecting a good model and achieving the smallest variance of the resultant estimator based on the selected subdata, we need to acquire the maximum amount of information about $\Theta = (S, \beta_1 \mathbb{1}(1 \in S), \ldots, \beta_p \mathbb{1}(p \in S))^{\mathrm{T}}$, where S is indices set of X, and $\mathbb{1}(j \in S) = 1$ if the jth covariate variable is in model S and $\mathbb{1}(j \in S) = 0$ otherwise. Let $\mathrm{Ent}(\Theta|X^*)$ be the entropy of Θ for the selected subdata of size r with the $r \times p$ model matrix X^*. The goal of our subdata selection problem is to choose subdata that minimize $\mathrm{Ent}(\Theta|X^*)$. However, the exact minimizer is not realistic due to the $\binom{n}{r}$ possible combinations of the possible subdata. The following theorem gives an equivalent representation of the subdata selection criterion, which will guide us to find a practical solution.

Theorem 1 *Suppose that* $\mathrm{Ent}(\Theta|X) < \infty$ *and the prior of the coefficient* β *under the widest model obeys* $N(\beta_{\mathrm{prior}}, \sigma_f^2 I_p)$*, where* I_p *stands for the* $p \times p$ *identity matrix and* β_{prior} *and* σ_f^2 *do not depend on* X*. Let* $X^*_{(t)}$ *be the* $r \times p_{(t)}$ *matrix of the selected subdata corresponding to* $X_{(t)}$*, where* $p_{(t)}$ *is the dimension of the true model. The optimal subdata selection problem of minimizing* $\mathrm{Ent}(\Theta|X^*)$ *is equivalent to finding* $X^*_{(t)}$ *that maximizes*

$$\det(X^{*\mathrm{T}}_{(t)} X^*_{(t)} + \sigma^2 \sigma_f^{-2} I_{p_{(t)}}), \tag{6}$$

where $\det(\cdot)$ *is the determinant of a square matrix.*

When we consider the estimation problem under the true model, the resultant criterion is known as Bayesian D-optimality criterion (Pukelsheim, 2006) in the language of experimental design. It is worth mentioning that Sebastiani and Wynn (2000) has shown that the Bayesian D-optimal design achieves the minimum entropy of $\beta_{(t)}$ under the true model. Here we further consider the problem of model selection and establish the equivalence between Bayesian D-optimality and the minimum entropy of Θ conditioning on the subdata. If $\sigma_f = \infty$, this criterion becomes the D-optimality criterion. For ease of presentation, we assume that $\sigma_f = \infty$ in the following discussion, which corresponds to a non-informative prior. The idea of using D-optimality criterion to select subsets of data under a given model is appealing and sensible, when the model is correctly specified in advance. For example, Dereziński and Warmuth (2018) studied the volume sampling procedure that samples the subdata proportional to the squared volume of the parallelepiped spanned by the rows of X^*. Thus, the probabilities of the subdata sets being selected are proportional to $\det(X^{*\mathrm{T}}X^*)$. Wang et al. (2019) also considered deterministically selecting subsamples based on D-optimality by selecting the "edge" points of the data, i.e., the extreme points in X. The D-optimality criterion also benefits model discrimination. On the one hand, the minimum volume of the confidence ellipsoid for the zero coefficients of the widest model can benefit the model selection. On the other hand, the minimum volume of the confidence ellipsoid for the non-zero coefficients of the widest model achieves the smallest variance of the true models' estimates. Atkinson and Fedorov (1975) found that D-optimal designs are useful for model discrimination in the presence of appreciable experimental error when the number of trials is limited.

Since $X_{(t)}$ is unknown, in working toward an approximate solution, we find a connection between the widest model matrix and the candidate model matrices via the leverage scores of the widest model. This will guide our later algorithm.

Theorem 2 *Assume that $X^{\mathrm{T}}X$ is a positive definite matrix. Let h_{ii} denote the leverage score for the ith data point, i.e., the (i,i)th entry of the hat matrix $X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. For the kth candidate model,*

$$X_{(k)}^{*\mathrm{T}}X_{(k)}^{*} \leq \min(\sum_{i=1}^{n}\delta_i h_{ii}, 1)X_{(k)}^{\mathrm{T}}X_{(k)}, \tag{7}$$

in the Loewner ordering, where $X_{(k)}^{*}$ is the design matrix under the kth candidate model for the selected subdata, and $\delta_i = 1$ if the ith data points is selected and $\delta_i = 0$ otherwise.

Theorem 2 reveals that $\det(X_{(t)}^{*\mathrm{T}}X_{(t)}^{*}) \leq (\sum_{i=1}^{n}\delta_i h_{ii})^{p_{(t)}}\det(X_{(t)}^{\mathrm{T}}X_{(t)})$. Therefore, data points with relatively small leverage scores of the widest model are much more likely form an inefficient design in the sense of D-optimality. This motivates us to select the data points corresponding to the r largest leverage scores of the widest model matrix. Moreover, as pointed out in the Kiefer-Wolfowitz equivalence theorem (Pukelsheim, 2006), a design ξ is D-optimal for the model $y = f(z)^{\mathrm{T}}\beta + \varepsilon$ if and only if $f(z)^{\mathrm{T}}M(\xi)^{-1}f(z) = p$ when z is a support point of ξ and $f(z)^{\mathrm{T}}M(\xi)^{-1}f(z) < p$ otherwise. Here, $M(\xi)$ is the information matrix under ξ. Therefore, finding large leverage scores tends to approximating the support points of the D-optimal design for the widest model.

This result also has an interpretation from the view of entropy sampling theory. As pointed out by Shewry and Wynn (1987), a larger entropy implies a higher prediction power. The prediction of y_i using the full data set based on the widest model can be expressed as $\hat{y}_i = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j$. Thus a data point with a higher h_{ii} is harder to predict by using other data points, and sampling high leverage data points makes the resultant subdata have higher prediction power.

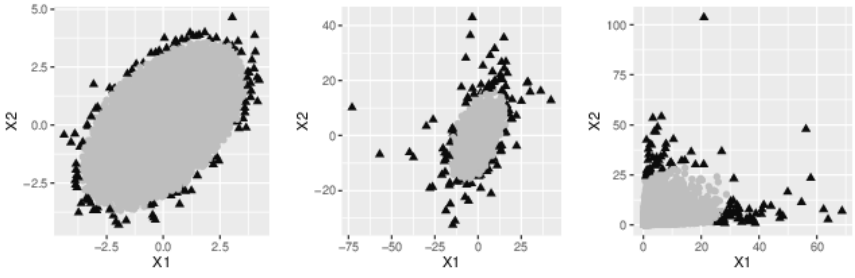
Let $\#(\Gamma)$ denote the cardinal number of a set Γ and $\kappa(A) := \lambda_{\max}(A)/\lambda_{\min}(A)$ denote the condition number of a matrix A, where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ are the maximum and minimum eigenvalues of a squared matrix, respectively. For simplicity, we define $\kappa(A) = \infty$ when A is a singular matrix. We summarize our idea in Algorithm 1.

Algorithm 1: Deterministic Leverage Score Selection Algorithm

Input: The widest model matrix $X \in \mathbb{R}^{n\times p}$, target sample size $r(> p)$, threshold $T(\geq 1)$.
Output: the selected index set Γ and the subsample design matrix X^*.
Initialization: $\Gamma = \emptyset; U_\Gamma = \emptyset; \kappa(U_\Gamma^{\mathrm{T}} U_\Gamma) = \infty$.
Perform a singular value decomposition of X as $X = U\Sigma V^{\mathrm{T}}$, calculate the leverage scores $h_{ii} := \|U_{i\cdot}\|^2$, where $U_{i\cdot}$ denotes the ith row of U, and sort h_{ii}'s to have $h_{(11)} \geq \ldots \geq h_{(nn)}$.
for $i = 1, ..., n$ **do**
  if $\#(\Gamma) \leq r$ *or* $\kappa(U_\Gamma^{\mathrm{T}} U_\Gamma) \geq T$ **then**
    Add the index of the data point corresponding to $h_{(ii)}$ to set Γ.
    Update the U_Γ as the selected rows of U in Γ.
  else
    Break.

Clearly, the proposed subdata selection algorithm only relies on the covariates, i.e., X's information. Thus Algorithm 1 is also applicable for the scenario of measurement constraints. This is a situation that all the covariates are available and the responses are expensive or time-consuming to collect. A typical example is semi-supervised learning problems in linear models (see Chakrabortty and Cai, 2018, for more details).

Note that a large value of the condition number may lead to a ill-conditioned matrix and thus cause multicollinearity. One purpose to set the stopping criterion for Algorithm 1 on the condition number is to ensure that the subsample design matrix is not ill-conditioned. As a result, a small condition number leads to a stable estimator. From a geometrical perspective, the condition number measures the level of "space spanning" for the selected data. A small condition number implies that the column space of the subdata matrix is well spanned, and a threshold on the condition number prevents the case that the subdata lies in a low rank subspace. Some two-dimensional synthetic samples are demonstrated in Figure 1 where Z is generated from a two-dimensional normal, t_3 or log-normal distribution, respectively, and X is column-centralized version of Z. The selected subdata are marked by triangles and the rest are marked by circles. Clearly, the selected subdata well span the two dimensional space. We do not recommend setting the stopping criterion on the condition number of X^*, because the condition number of X^* is unbounded and the threshold will be harder to determine.



(a) Normal (b) t_3 (c) Log normal

Fig. 1 Illustration of the selected subdata points through Algorithm 1 with $r = 100, T = 4$ (triangle) among 100,000 points (circle) that are randomly generated from two dimensional normal, t_3 and lognormal distributions. Here X is the centralized model matrix with two columns.

Remark 1 If the covariates are from the family of elliptically contoured distributions (Fang et al., 1990), the covariate space of the selected data expands to the covariate space of the full data quickly. For this scenario, the criterion on the condition number is not critical because the condition number for the subdata drops to be close to the condition number for the full data very quickly.

In general cases, if Algorithm 1 produces a subsample with more than r data points and we want to strictly restrict the subsample size at r, the following procedure can be adopted. Let $\tilde{\Gamma}$ be the index set with $\tilde{r} > r$ elements from the original Algorithm 1. Select r elements from $\tilde{\Gamma}$ via simple random sampling as Γ. It can be shown that $\kappa(U_\Gamma^{\mathrm{T}} U_\Gamma) - \kappa(U_{\tilde{\Gamma}}^{\mathrm{T}} U_{\tilde{\Gamma}})$ converges to zero in probability as $r \to \infty, n \to \infty$. Thus the resultant subdata set satisfies the constraint on the condition number asymptotically. On the other hand, comparing with simple random sampling directly from the full data, the $\sum_{i\in\Gamma} h_{ii}$ is larger, which implies that the selected subdata set still enjoys the benefit brought by the leverage scores.

4 Selection Consistency and Parameter Estimation

Analogous to (5), for the selected subdata through Algorithm 1, the selected model according to BIC, denoted as $\hat{S}^*_{\mathrm{BIC}}$, satisfies that

$$\hat{S}^*_{\mathrm{BIC}} = \arg\min_{S_k} \left\{ r \log \left(r^{-1} \sum_{i=1}^{r} (y_i^* - \hat{\mu}_i^{*(k)})^2 \right) + p_{(k)} \log r \right\}, \tag{8}$$

where $\hat{\mu}_i^{*(k)}$ is the ith element of $\hat{\mu}^{*(k)} = X^*_{(k)} \hat{\beta}^*_{(k)}$, and $\hat{\beta}^*_{(k)}$ is the MLE of the kth candidate model based on the selected subdata. In the following, we will show the model selection consistency and estimation consistency based on the selected subdata via Algorithm 1.

Recall that the true model refers to the model in the set of correct candidate models $\mathcal{M}^C$ with the fewest parameters. For convenience, we define $\beta_{(t)}$ to be the parameter vector of the true model. For selection consistency, we mean that a selection criterion selects the true model with probability approaching one. Let $\hat{\sigma}^{*2}_{(k)}$ be the MLE of σ^2 based on the selected subdata. Recall that $h_{(ii)}$ defined

in Algorithm 1 is the order statistic of full data leverage scores such that $h_{(11)} \geq h_{(22)} \geq \ldots \geq h_{(nn)}$. The following theorem states the selection consistency of $\hat{S}^*_{\text{BIC}}$ defined in (8) based on the subdata selected via Algorithm 1.

Theorem 3 *Assume that $n^{-1}X^{\mathrm{T}}X$ goes to a positive definite matrix, and $\hat{\sigma}^{*2}_{(k)} \not\to 0$, $\hat{\sigma}^{*2}_{(k)} \not\to \infty$ and $\liminf_{n,r\to\infty} \min_j (n\sum_{i=1}^{r} h_{(ii)})\|\beta_{(t)j}\|^2/\log r \to \infty$, for $k = 1,\ldots,m$, where $\beta_{(t)j}$ is the jth component of $\beta_{(t)}$. As $r \to \infty, n \to \infty$, with probability approaching one, the model selection based on the selected subdata according to Algorithm 1 is consistent. That is, the probability that $\hat{S}^*_{BIC}$ is correct and has the smallest dimension, goes to one.*

Theorem 3 reveals that the signal that can be detected based on X^* mainly relies on the distribution of the leverage scores. For the uniform random subsampling method, the detection condition is $\liminf_{r\to\infty} \min_j r\|\beta_{(t)j}\|^2/\log r \to \infty$. Note that $r^{-1}\sum_{i=1}^{r} h_{(ii)} \geq n^{-1}\sum_{i=1}^{n} h_{(ii)}$ because $h_{(11)} \geq \ldots \geq h_{(nn)}$. This implies that $pr \leq n\sum_{i=1}^{r} h_{(ii)}$. Recall that the true model can be selected via the subdata obtained by Algorithm 1 when $\liminf_{r\to\infty} \min_j (n\sum_{i=1}^{r} h_{(ii)})\|\beta_{(t)j}\|^2/\log r$ goes to infinity. Thus the worst performance of our method is as good as the uniform subsampling in the sense of the rate for $\|\beta_{(t)j}\|$ to go to zero. For very nonuniform leverage scores with power law decay $h_{(ii)} = i^{-a}h_{(11)}$ with some $a > 0$ (Papailiopoulos et al., 2014), it is possible that $n\sum_{i=1}^{r} h_{(ii)} = O(n)$. Thus our method has some "super" efficiency. That is to say, our method may select the true model under the condition $n\|\beta_{(t)j}\|^2/\log(n) \to \infty$, when only $r(r \ll n)$ subsample are used, which is the weakest signal that the full data BIC selector can detect.

Now we turn our attention to the estimation efficiency. Note that the selection rule in Algorithm 1 is auxiliary, i.e., the selection procedure does not depend on Y, so the MLE of β based on the subdata in the true model is unbiased and efficient. Since Algorithm 1 has selection consistency, the following theorem focuses on the variance of the estimator for the selected model, which is the true model with probability approaching one.

Theorem 4 *Assume that $n^{-1}X^{\mathrm{T}}X$ goes to a positive definite matrix, $\hat{\sigma}^{*2} \not\to 0$, and $\hat{\sigma}^{*2} \not\to \infty$. Let $\hat{\beta}^*_s$ be the estimator under the selected model obtained from the data selected by Algorithm 1. As $r \to \infty, n \to \infty$, the following result holds*

$$\text{var}(\hat{\beta}^*_s) = O\left(\frac{1}{n\sum_{i=1}^{r} h_{(ii)}}\right). \tag{9}$$

Theorem 4 reveals that the quality of the subdata MLE based on the selected model also mainly relies on the distribution of the leverage scores. It gives a guarantee that our method outperforms the uniform random subsampling method since $pr \leq n\sum_{i=1}^{r} h_{(ii)} \leq np$.

5 Numerical Studies

5.1 Simulation Results

In this section, we use simulation to evaluate the finite sample performance of Algorithm 1. We assume that there are seven potential covariates to be considered,

and an intercept term is included in the model with the true value fixed at $\beta_{\text{int}} = 0.25$. To be precised, we generate $\mu_i = \beta_{\text{int}} + \sum_{j=1}^{p} z_{ij}\beta_j$ for $i = 1, \ldots, n$. The responses y_i's are generated from $y_i = \mu_i + \varepsilon_i$ with ε_i's being i.i.d. $N(0, 1)$. Here we opt to simulate the case with an intercept since it is more common in practice. For the model slope parameters in each repetition of the simulation, we generate the values of β_1 and β_2 independently from distribution $\text{Unif}(0.5, 1)$, generate the values of β_3 and β_4 independently from distribution $\text{Unif}(0.05, 0.1)$, and set $\beta_j = 0$ for $j = 5, 6$ and 7. We assume that any non-empty subset of these variables can be a candidate set of active variables. Therefore there are $2^7 - 1 = 127$ candidate models. Here we use all-subset regression to illustrate our method. The results of model selection via the forward regression are quite similar and thus we relegate them to Appendix B. We consider the following six scenarios to generate the covariates $z_i = (z_{i1}, \ldots, z_{i7})^{\text{T}}$ for full (training) data sets with $n = 500,000$.

Case 1 Covariates are generated from the multivariate normal distribution $N(0, \Sigma_1)$ with the (i, j)th entry of Σ_1 being $0.5^{|i-j|}$.

Case 2 Covariates are generated from the multivariate normal distribution $N(0, \Sigma_2)$ with the (i, j)th entry of Σ_2 being $0.5^{\mathbb{1}(i \neq j)}$ where $\mathbb{1}(\cdot)$ is the indicator function.

Case 3 Covariates are generated from a mixture multivariate normal distribution $0.5N(0, \Sigma_1) + 0.5N(0, \Sigma_2)$, where Σ_1 and Σ_2 are defined in Cases 1 and 2, respectively.

Case 4 Covariates are generated from the multivariate t-distribution with three degrees of freedom. The mean parameter is 0, and the scale matrix parameter Σ_1 is defined in Case 1.

Case 5 Covariates are generated from a multivariate t-distribution with three degrees of freedom. The mean parameter is 0, and the scale matrix parameter Σ_2 is defined in Case 2.

Case 6 Covariates are generated from the log-normal distribution with parameters 0 and Σ_2, which is defined in Case 2.

Data generated in Cases 1–3 have different correlation structures and leverage scores are more uniform. In Cases 4 and 5, leverage scores are less-uniform as data are generated from heavy tailed distributions. The distribution of the covariates in Case 6 is asymmetric and skewed to the right.

For all the above six cases, we centralize $Z = (z_1^{\text{T}}, \ldots, z_n^{\text{T}})^{\text{T}}$ and Y, so the widest model matrix X input in Algorithm 1 is centralized Z. For each candidate model S_k, we first estimate the slope parameters $\hat{\beta}^*_{(k)}$ using the selected subdata from the centralized full data set, and then the intercept term $\beta_{\text{int(k)}}$ is estimated by $\bar{Y} - \bar{Z}^{\text{T}}_{(k)}\hat{\beta}^*_{(k)}$, where $\bar{Z}_{(k)}$ is the column mean vector of $Z_{(k)}$.

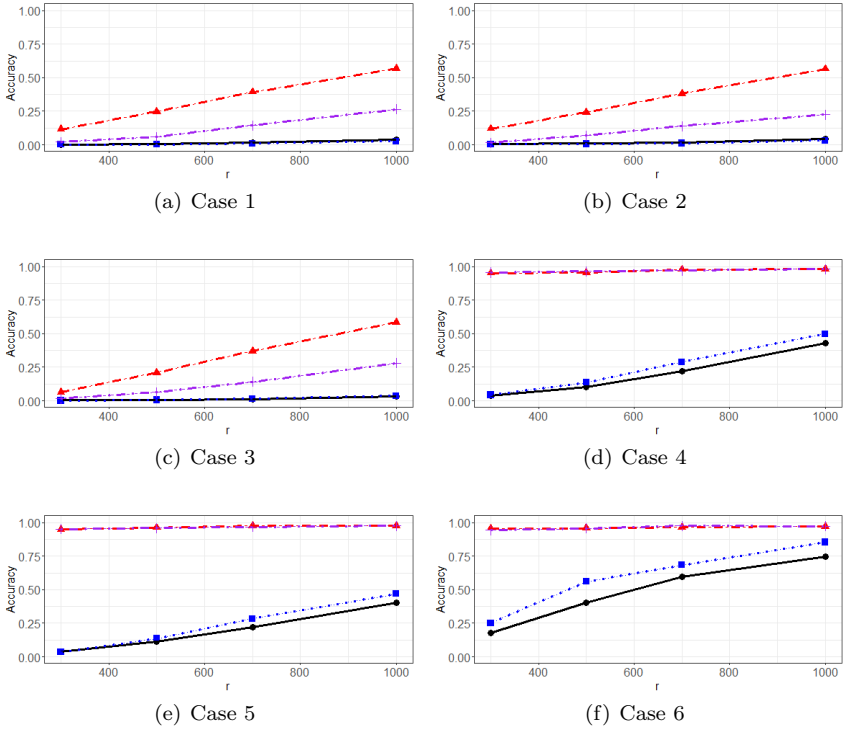
The performance of a subdata selection/sampling strategy is evaluated by the following two criteria:

(i) Accuracy: The selection probability of the true model.

(ii) MSPE: The mean squared prediction error for the observations in the test sample. To be precise, MSPE $= n_{\text{test}}^{-1} \sum_{i=1}^{n_{\text{test}}} \|\mu_{i,\text{test}} - \hat{\beta}^*_{\text{int(k)}} - z_{i(k),\text{test}}\hat{\beta}^*_{(k)}\|^2$ where n_{test} is the size of test data, $\mu_{i,\text{test}}$ is the conditional mean of the test data, k stands for the selected model S_k, and $z_{i(k),\text{test}}$ are the covariates of the kth model in the test data.

For comparison, we consider Algorithm 1 (LEVSS) with $T = 10$; the uniform subsampling (UNIF) in which the sampling probabilities are $1/n$; leverage score

subsampling (LEV) in which the sampling probabilities are $h_{ii}/\sum_{i=1} h_{ii}$ (Ma et al., 2015); and the IBOSS method proposed in Wang et al. (2019). For the leverage score subsampling, the BIC is calculated through the sampling version of BIC which is proposed in Xu et al. (2013). We repeat 1,000 times for each method under each setting throughout this section. Computations are performed using `R`.



(a) Case 1 (b) Case 2

(c) Case 3 (d) Case 4

(e) Case 5 (f) Case 6

Fig. 2 Selection accuracies for different r based on UNIF (solid line with circles), LEVSS (dashed line with triangles), LEV (dotted line with squares), and IBOSS (dot-dash line with plus signs) methods for the six cases listed in Section 5.1. The models are selected based on all-subset procedures via BIC.

Figure 2 reports the model selection results on the accuracy. We see that the selection accuracies of the four methods increase as r increases in general, which echoes the result in Theorem 3. It is clear to see that our method is uniformly better than the uniform subsampling method. Our method and the IBOSS method have a similar performance in selecting the true model, because both of them are influenced by the D-optimality criterion. They have a higher probabilities in selecting the true model than the UNIF and LEVSS methods.

To see the prediction results based on the selected model, we report the corresponding log MSPE in Figure 3 with $n_{\text{test}} = 500$. We clearly see that the IBOSS and our method (LEVSS) are uniformly better than the uniform subsampling method. It is worth mentioning that our method performs the best. The leverage score subsampling has a similar performance to the uniform subsampling method

in Cases 1–3. For Cases 4–6, the leverage score subsampling performs slightly better than uniform subsampling and worse than the IBOSS and our method (LEVSS). This phenomenon arises due to the following two reasons. The first is that IBOSS and our method have the advantage in selecting the true model compared with the uniform subsampling and leverage score sampling methods. These echo the model selection results in Figure 2. Second, for Cases 4–6, both uniform subsampling and leverage score subsampling methods result in a root r consistent estimator while our method has a higher efficiency with more nonuniform leverage scores as discussed in Theorem 4.

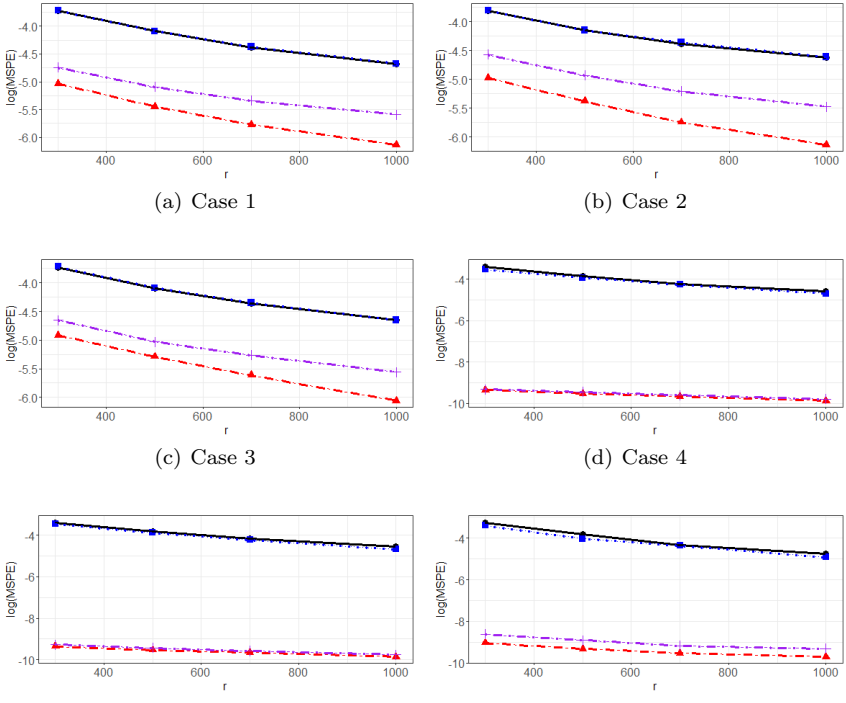


Fig. 3 Log MSPEs for different r based on UNIF (solid line with circles), LEVSS (dashed line with triangles), LEV (dotted line with squares), and IBOSS (dotdash line with plus signs) methods for the six cases listed in Section 5.1. The models are selected based on the all-subset regression via BIC.

Now we evaluate the computational efficiency of the subsampling strategies. We implemented all methods using the R programming language and recorded the computing times of the four subsampling strategies using the `Sys.time()` function. Computations were carried out on a laptop computer with macOS and an 8-Core Intel Core i9 processor. For brevity, in Table 1 we only report the computing times for Case 1. For reference, we also implemented the Lasso method based on the full data set, and recorded its computing time as a benchmark. Note that the Lasso method is much faster than the all-subset regression on the full data set.

Table 1 Computational time (in seconds) with different r for the four methods in Case 1. The average time spent on the model selection step based on the BIC valus are reported in parentheses.

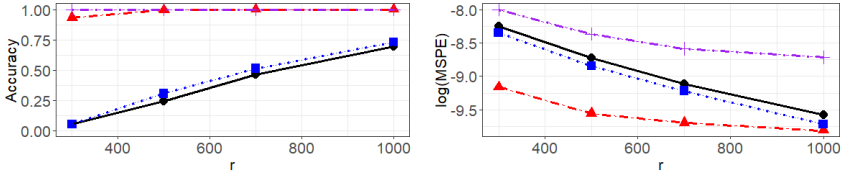
r	UNIF	LEVSS	LEV	IBOSS
300	0.062 (0.059)	0.208 (0.065)	0.179 (0.065)	0.165 (0.062)
500	0.188 (0.160)	0.348 (0.149)	0.278 (0.155)	0.260 (0.138)
700	0.334 (0.309)	0.458 (0.287)	0.465 (0.331)	0.424 (0.281)
1000	0.711 (0.661)	0.773 (0.600)	0.780 (0.646)	0.704 (0.596)
Full		3.444		

Calculating the estimator using a subsample requires $O(rp^2)$ time for each candidate model, which is a significant reduction in comparison with the computational cost of the full data estimator which is $O(np^2)$. As expected, it is seen that subsample based methods are faster than the full data selection method. It is worth mentioning that our method has comparable performance with the uniform subsampling method when r is not that small. We also record the time spent on calculating the BIC for all candidate models. More than 80% of the time was spent on calculating BIC values with the uniform subsampling method. The time spent on calculating the BIC values increases rapidly as r increases. When $r = 1000$, more than 80% of the time was spent on calculating BIC values for the candidate models with all the four methods. Thus when r is not very small, our method does not take much more time than the uniform subsampling method. Here the leverage scores are calculated through the singular value decomposition of the widest model matrix of the full data. The proposed method can be further accelerated if some approximating or parallel computing methods such as those in Drineas et al. (2006); Meng et al. (2014) are adopted. This will make our algorithm scalable to large datasets.

5.2 Real Data Study

Understanding factors that affect the price of diamond is important because each diamond is unique and it is hard to find a reference price of a new diamond. The price of a new diamond has to be determined by its attributes; this is different from the determination of price for most manufactured products. The `diamonds` data set in the `ggplot2` package contains the prices and the specifications for more than 50,000 diamonds. The problem of interest is to identify important factors that affect the diamond pricing. Here are the seven factors in this data set. The first is the carat (z_1) which is the weight of the diamond, ranges from 0.2 to 5.01, and a cube-root transform is made on this factor; the second factor indicates the quality of the diamond cut (z_2) and it is coded as one if the quality is better than " Good" and zero otherwise; the third factor is the level of diamond color (z_3) with z_3 being one if the quality is better than "level F" and zero otherwise; the fourth factor is a measurement on how clear the diamond is (z_4) with z_4 being one if the quality is better than "VS2" and zero otherwise; the fifth and sixth factors are the total depth percentage (z_5) and the width at the widest point (z_6) i.e., the "depth" and "table" columns in the data set, respectively; the last factor is the volume of the diamond (z_7) which is roughly calculated as the product of the diamond's length, width, and depth. The response variable y is log10 of the price. Note that some

special attribute affects the price significantly and needs to be considered case by case. For example, the gemological institute of America pointed out that prices of diamonds are affected by the "magic sizes" (Mamonov and Triantoro, 2018). Some outliers such as the NO.24068 are excluded since the corresponding diamond has an unusually large width that makes the price too high. As in Section 5.1, we assume that any non-empty subset of these variables can be a candidate set of active variables so there are $2^7 - 1 = 127$ candidate models. Again, we regard the model with the smallest BIC value based on the full data set as the true model, which is $\hat{y} = 1.087 + 3.760z_1 + 0.041z_2 - 0.081z_3 + 0.126z_4 - 0.007z_5 - 0.006z_6 - 0.003z_7$. The model selection and prediction results with 500 replications are reported in Figure 4.



(a) Selection Accuracies (b) Log MSPE

Fig. 4 Empirical model selection accuracies and Log MSPE for the diamond data set with different r for the UNIF (solid line with circles), LEVSS (dashed line with triangles), LEV (dotted line with squares), and IBOSS (dot-dash line with plus signs) methods through BIC.

6 Conclusion and remarks

In this work, guided by the experimental design theory, we have developed a deterministic data selection method using leverage scores for linear model selection and estimation problems. Compared with random sampling, the proposed method better identifies the correct model with weaker signals. When the leverage scores are highly non-uniform, an analysis based on the selected subdata can be arbitrarily close to the full data analysis. Furthermore, Theorem 4 shows that the variance of the subdata estimator can converge to zero under a mild assumption even with a fixed subsample size r. Extensive numerical results have demonstrated the superiority of the proposed method to other state-of-the-art algorithms.

Investigating deterministic leveraging for linear model selection is the first step to understand this approach. Although our analysis has focused on the linear regression, we believe that the implications of leveraging apply to more complex models and estimation problems. For example, if the covariates are the spline bases in B-splines, then the idea of deterministic leveraging can be extended to the generalized additive models (Hastie and Tibshirani, 1990; Truong et al., 2005). Moreover, the proposed method can also be extended to include the weighted least squares method and its variants, which are commonly used for generalized linear models (McCullagh and Nelder, 1989) and varying coefficient models (Hastie and Tibshirani, 1993).

In this work, we have carried out model selection via the BIC and measure the uncertainty with entropy under the assumption that the true model is included in the candidate set. As a referee pointed out, lurking variables may exist in practice and as a result the true model is not in the candidate set. For this scenario, the AIC is more appropriate and the theoretical and practical performance warrant further investigations. In addition, high leverage observations may contain outliers and thus do not give a representative picture of the full data. A pre-processing step that tries to eliminate outliers will be helpful in this situation. We address the model discrimination framework with a fixed dimension in this paper, which occurs in many practical problems and real datasets. The case that both p and n are growing rapidly is also common in practice. How to conduct subdata selection, especially in determining the subdata size, is an interesting and important problem that warrants future investigations.

Acknowledgements The authors sincerely thank the editor, associate editor, and referees for their valuable comments and insightful suggestions, which led to further improvement of this article. This work is supported by NSFC grants 12001042 and Beijing Institute of Technology Research Fund Program for Young Scholars.

Appendix

A Technical details

Proof of Theorem 1. For any given subdata X^*, by applying the entropy decomposition in information theory (Sebastiani and Wynn, 2000, Equation (2)), the joint entropy of Y^* and the parameter Θ can be decomposed as

$$\begin{aligned} \mathrm{Ent}(Y^*, \Theta|X^*) &= \mathrm{Ent}(\Theta|X^*) + E_\Theta\{\mathrm{Ent}(Y^*|\Theta, X^*)\} \\ &= \mathrm{Ent}(\Theta) + \mathrm{Ent}(\boldsymbol{\varepsilon}^*), \end{aligned} \tag{A.1}$$

where $\boldsymbol{\varepsilon}^*$ stands for the corresponding error term of (Y^*, X^*) in model (1). The second equality holds by the model assumption since all the randomness of Y^* comes from the error term conditional on Θ, X^* and Θ is functionally independent of X^*. This implies that $\mathrm{Ent}(Y^*, \Theta|X^*)$ is a constant up to the subdata size r.

Also note that $\mathrm{Ent}(Y^*, \Theta|X^*)$ can also be decomposed as

$$\mathrm{Ent}(Y^*, \Theta|X^*) = \mathrm{Ent}(Y^*|X^*) + E_{Y^*}\{\mathrm{Ent}(\Theta|Y^*, X^*)\}. \tag{A.2}$$

That is to say maximizing $\mathrm{Ent}(Y^*|X^*)$ indicates minimizing the overall expected deviance loss $E_{Y^*}\{\mathrm{Ent}(\Theta|Y^*, X^*)\}$.

Now we turns to calculate $\mathrm{Ent}(Y^*|X^*)$. Without loss of generality, we assume that the first $p_{(t)}$ columns of X be the model matrix of the true model. Thus the prior of $\beta_{(t)}$ comes from $N(\beta_{\mathrm{prior},(\mathrm{t})}, \sigma_f^2 I_{p_{(t)}})$, where $\beta_{\mathrm{prior},(\mathrm{t})}$ corresponds to the first $p_{(t)}$ entries of β_{prior}. Note that $Y^* = X^*_{(t)}\beta_{(t)} + \boldsymbol{\varepsilon}$ and the prior of $\beta_{(t)}$ obeys $N(\beta_{\mathrm{prior},(\mathrm{t})}, \sigma_f^2 I_{p_{(t)}})$. Thus the marginal distribution of Y^* is normal with mean $X^*_{(t)}\beta_{\mathrm{prior},(\mathrm{t})}$ and variance $\sigma^2 I_t + \sigma_f^2 X^{*\mathrm{T}}_{(t)} X^*_{(t)}$ under model (1). The desired results come from the facts

$$\begin{aligned} \mathrm{Ent}(Y^*|X^*) &= \log\det(\sigma^2 I_r + \sigma_f^2 X^*_{(t)} X^{*\mathrm{T}}_{(t)}) + c_1 \\ &= \log\det(\sigma^{-2}\sigma_f^2 X^{*\mathrm{T}}_{(t)} X^*_{(t)} + I_t) + c_2, \\ &= \log\det(X^{*\mathrm{T}}_{(t)} X^*_{(t)} + \sigma^2\sigma_f^{-2} I_t) + c_3, \end{aligned} \tag{A.3}$$

where c_1, c_2, c_3 are some constant up to the subdata size r. The second equality comes from the matrix determinant lemma, i.e., $\det(A + BC) = \det(A)\det(I + CA^{-1}B)$ for some matrices A, B, C with $A > 0$. □

Proof of Theorem 2. It is sufficient to show that $X^{*\mathrm{T}}X^* \leq (\sum_{i=1}^n \delta_i h_{ii}) X^{\mathrm{T}}X$ in the sense of Loewner ordering. Let x_i be the ith row of X. For any $a \in \mathbb{R}^p$, noting that $X^{\mathrm{T}}X$ is a full rank matrix, a can be represent as $a = (X^{\mathrm{T}}X)^{-1/2}b$ for some $b \in \mathbb{R}^p$. Then,

$$\begin{aligned} a^{\mathrm{T}}x_i x_i^{\mathrm{T}}a &= b^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1/2}x_i^{\mathrm{T}}x_i(X^{\mathrm{T}}X)^{-1/2}b \\ &\leq \mathrm{tr}\{(X^{\mathrm{T}}X)^{-1/2}x_i^{\mathrm{T}}x_i(X^{\mathrm{T}}X)^{-1/2}\}\|b\|_2^2 \\ &= h_{ii}\{b^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1/2}(X^{\mathrm{T}}X)(X^{\mathrm{T}}X)^{-1/2}b\} \qquad \text{(A.4)} \\ &= h_{ii}a^{\mathrm{T}}(X^{\mathrm{T}}X)a, \qquad \text{(A.5)} \end{aligned}$$

where $\mathrm{tr}(\cdot)$ is the trace operator and $(X^{\mathrm{T}}X)^{-1/2}(X^{\mathrm{T}}X)^{-1/2} = (X^{\mathrm{T}}X)^{-1}$. Therefore, $x_i x_i^{\mathrm{T}} \leq h_{ii}X^{\mathrm{T}}X$ and the desired result comes from summing over the both side of the inequality. □

For clarity, we begin with the proof of the following lemma since some results in the following lemma will be used in the proof of Theorem 3.

Lemma 1 *Assume that $n^{-1}X^TX$ goes to a positive definite matrix. Let $\hat{\beta}_k^*$ be the MLE based on selected subdata set according to Algorithm 1 for the kth candidate model. As $r \to \infty, n \to \infty$, the following result holds:*

$$\mathrm{Var}(\hat{\beta}_k^*) = O\left(\frac{1}{n\sum_{i=1}^r h_{(ii)}}\right). \qquad \text{(A.6)}$$

Proof of Lemma 1 According to Algorithm 1, $X^* = U_\Gamma \Sigma V^{\mathrm{T}}$. Then it is sufficient to show that

$$c\sum_{i=1}^r h_{(ii)} \leq \lambda_{\min}(U_\Gamma^T U_\Gamma) \leq \lambda_{\max}(U_\Gamma^T U_\Gamma) \leq \sum_{i=1}^r h_{(ii)}, \qquad \text{(A.7)}$$

for some constant c, where $\lambda_{\max}(A)$, $\lambda_{\min}(A)$ stand for the maximum and minimum eigenvalue of A, respectively. Since $U_\Gamma^{\mathrm{T}}U_\Gamma$ is positive definite through Algorithm 1, therefore $\lambda_{\max}(U_\Gamma^{\mathrm{T}}U_\Gamma) \leq \mathrm{tr}(U_\Gamma^{\mathrm{T}}U_\Gamma) = \sum_{i=1}^r h_{(ii)}$. By the definition of the condition number, it holds that

$$\lambda_{\min}(U_\Gamma^{\mathrm{T}}U_\Gamma) = \lambda_{\max}(U_\Gamma^{\mathrm{T}}U_\Gamma)/\kappa(U_\Gamma^{\mathrm{T}}U_\Gamma) \geq \mathrm{tr}(U_\Gamma^{\mathrm{T}}U_\Gamma)/(pT),$$

where the last inequality comes from the fact that $\mathrm{tr}(U_\Gamma^{\mathrm{T}}U_\Gamma) \leq p\lambda_{\max}(U_\Gamma^{\mathrm{T}}U_\Gamma)$.

From (A.7), it follows that

$$c\sum_{i=1}^r h_{(ii)}V\Sigma^2V^T \leq X^{*T}X^* \leq \sum_{i=1}^r h_{(ii)}V\Sigma^2V^T, \qquad \text{(A.8)}$$

and the desired results follows by noting $X^TX = V\Sigma^2V^T$ and $\mathrm{Var}(\hat{\beta}_k) = (P^TX^{*T}X^*P)^{-1}$ for the projection matrix P such that $X_{(k)} = XP$ where $X_{(k)}$ is the design matrix for model S_k. □

Now, let us turn to proof Theorem 3.

Proof of Theorem 3. Denote $\mathcal{M}^C$ be the set of correct candidate models, and $\mathcal{M}^I = \mathcal{M} - \mathcal{M}^C$ be the set of incorrect candidate models. We first show that

$$\Delta(k) = \liminf_r \min_{S_k \in \mathcal{M}^I} \|\mu^* - H_{(k)}^*\mu^*\|^2/\log r \to \infty, \qquad \text{(A.9)}$$

where μ^* stands for the mean of the selected data, $H_{(k)}^* = X_{(k)}^*(X_{(k)}^{*\mathrm{T}}X_{(k)}^*)^{-1}X_{(k)}^{*\mathrm{T}}$.

For any candidate model in $\mathcal{M}^I$, say S_k as an example, let the model matrix for the closest correct model be $\tilde{X}_{(k)}^* := (X_{(\check{c})}^*, X_{(k)}^*)$. Here $X_{(\check{c})}^*$ stands for the "complementary" design, which consists of the columns of $X_{(t)}^*$ that are not included in $X_{(k)}^*$. Denote the regression coefficient vector corresponding to $X_{(\check{c})}^*$ as $\beta_{(\check{c})}$, which is a subvector of $\beta_{(t)}$. Direct calculation yields

$$\|\mu^* - H_{(k)}^*\mu^*\|^2 = \inf_\alpha \|X_{(\check{c})}^*\beta_{(\check{c})} - X_{(k)}^*\alpha\|^2 \qquad \text{(A.10)}$$

$$= \inf_{\alpha}\{(\beta_{(\check{c})}^{\mathrm{T}}, \alpha^{\mathrm{T}})(\tilde{X}_{(k)}^{*\mathrm{T}}\tilde{X}_{(k)}^{*})(\beta_{(\check{c})}^{\mathrm{T}}, \alpha^{\mathrm{T}})^{\mathrm{T}}\}. \tag{A.11}$$

Utilizing the results in (A.8), we have

$$\left(c_2 n \sum_{i=1}^{r} h_{(ii)}\right)\left(\frac{1}{n}X^{\mathrm{T}}X\right) \leq X^{*\mathrm{T}}X^{*} \leq \left(n\sum_{i=1}^{r} h_{(ii)}\right)\left(\frac{1}{n}X^{\mathrm{T}}X\right), \tag{A.12}$$

for some constant c_2. Note that the $\tilde{X}_{(k)}$ is a submatrix of X up to a column permutation. Thus $\lambda_{\min}(\tilde{X}_{(k)}^{*\mathrm{T}}\tilde{X}_{(k)}^{*}) \geq \lambda_{\min}(\tilde{X}^{*\mathrm{T}}\tilde{X}^{*}) = O\left(n\sum_{i=1}^{r} h_{(ii)}\right)$, where $\lambda_{\min}(\cdot)$ stands for the smallest eigenvalue of a squared matrix. From (A.10), we have

$$\liminf_{n,r\to\infty} \min_{S_k \in \mathcal{M}^I} \|\mu^* - H_{(k)}^{*}\mu^*\|^2/\log r \geq \liminf_{n,r\to\infty} \min_{j}(n\sum_{i=1}^{r} h_{(ii)})\|\beta_{(t)j}\|^2/\log r \to \infty,$$

which implies (A.9) holds.

For convenience, let $\mathrm{BIC}^*(S_k) = r\log\left(r^{-1}\sum_{i=1}^{r}(y_i^* - \hat{\mu}_i^{*(k)})^2\right) + (p_{(k)}+1)\log r$. From (3.7) in Shao (1997), for any model S_k in $\mathcal{M}^I$, we have

$$\sum_{i=1}^{r}(y_i^* - \hat{\mu}_i^{*(k)})^2 - \sum_{i=1}^{r}(y_i^* - \hat{\mu}_i^{*(t)})^2 \geq \Delta(k) \geq p\log r > 0, \tag{A.13}$$

which implies $\log(\sum_{i=1}^{r}(y_i^* - \hat{\mu}_i^{*(k)})^2) - \log(\sum_{i=1}^{r}(y_i^* - \hat{\mu}_i^{*(t)})^2) > r\log(1 + p\log r/r)$ under the assumption $\hat{\sigma}^* \not\to 0$. Therefore,

$$\mathrm{BIC}^*(S_k) - \mathrm{BIC}^*(S_t) \geq r\log(1 + p\log r/r) - (p - p_{(k)})\log r \to \infty. \tag{A.14}$$

Similarly, for any model $S_{k'}$ in $\mathcal{M}^C$ with $p_{(k')} > p_{(t)}$, where $p_{(t)}$ is the column dimension of $X_{(t)}$, it is straightforward to see that

$$r\log\left(r^{-1}\sum_{i=1}^{r}(y_i^* - \hat{\mu}_i^{*(k')})^2\right) - r\log\left(r^{-1}\sum_{i=1}^{r}(y_i^* - \hat{\mu}_i^{*(t)})^2\right) \to \chi^2_{p_{(k')}-p_{(t)}}, \tag{A.15}$$

according to the log likelihood ratio test (see van der Vaart, 1998, Chapter 16). Therefore, it holds that

$$\mathrm{BIC}^*(S_{k'}) - \mathrm{BIC}^*(S_t) = O(\log r) \to \infty. \tag{A.16}$$

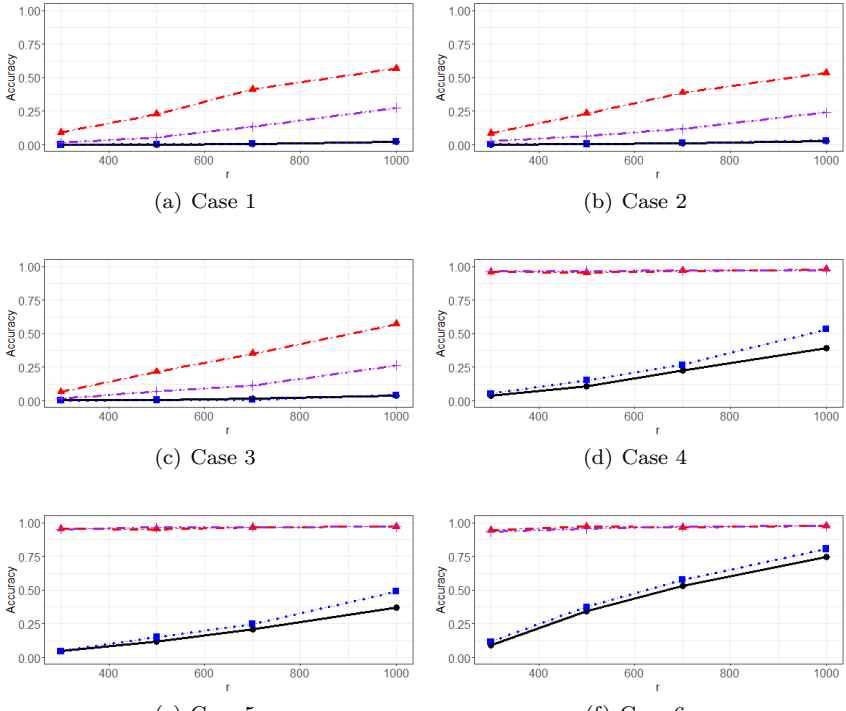
Combining (A.14) and (A.16), we can get the desired result. □

Proof of Theorem 4. This is the direct result from Lemma 1. □

B Additional Simulation Results on forward regression

Since the number of possible models increases exponentially with p, all-subset regression is only feasible for the cases that p is relatively small. Alternatively, a forward selection approach is usually adopted. More precisely, the forward regression starts from the null model, and iteratively adds one variable to the currently "best" model which yields the lowest value for the BIC at a time. This process is repeated until no more variables should be added into the currently "best" model. In this part, we adapt the forward regression to illustrate the proposed method. Of course, a backward elimination procedure and a step-wise regression procedure can also be adopted. Since the three methods have similar performance, we only report the results on forward regression.

In accordance with Section 5.1, we also demonstrate our method as well as the uniform subsampling, leveraging score subsampling and IBOSS through the six cases. The selection accuracies for the six cases are presented in Figure A.1. The log MSPEs are also provided in Figure A.2 to evaluate the performance of prediction.



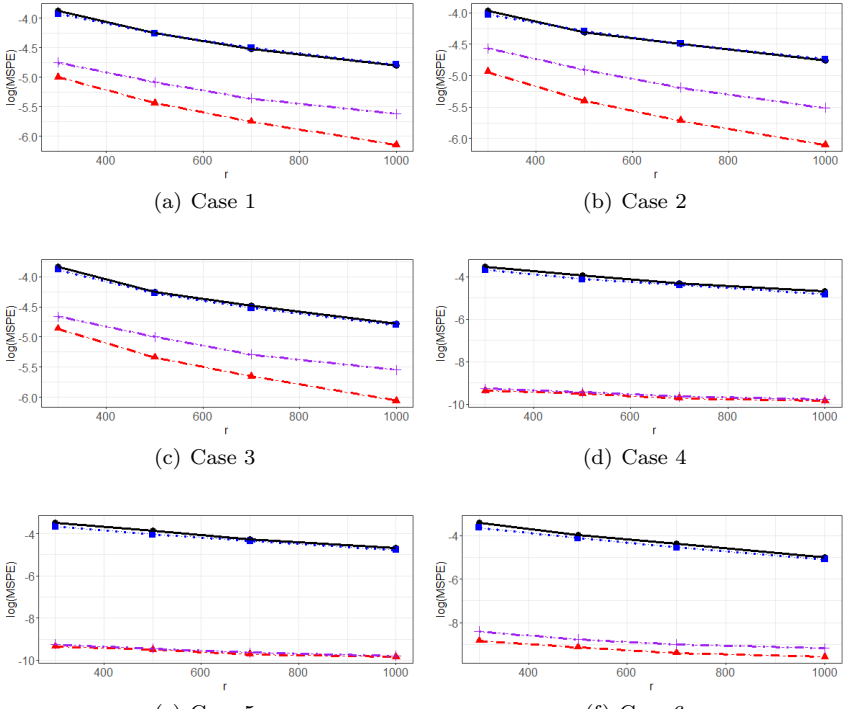
(a) Case 1 (b) Case 2

(c) Case 3 (d) Case 4

(e) Case 5 (f) Case 6

Fig. A.1 Selection accuracies for different r based on UNIF (solid line with circles), LEVSS (dashed line with triangles), LEV (dotted line with squares) and IBOSS (dotdash line with plus signs) method for the six cases listed in Section 5.1. The models are selected based on the forward regression procedure via BIC.

From Figures A.1, and A.2, one can see that the forward regression results based on BIC are very similar to the results on the all-subset regressions on BIC.



(a) Case 1 (b) Case 2

(c) Case 3 (d) Case 4

(e) Case 5 (f) Case 6

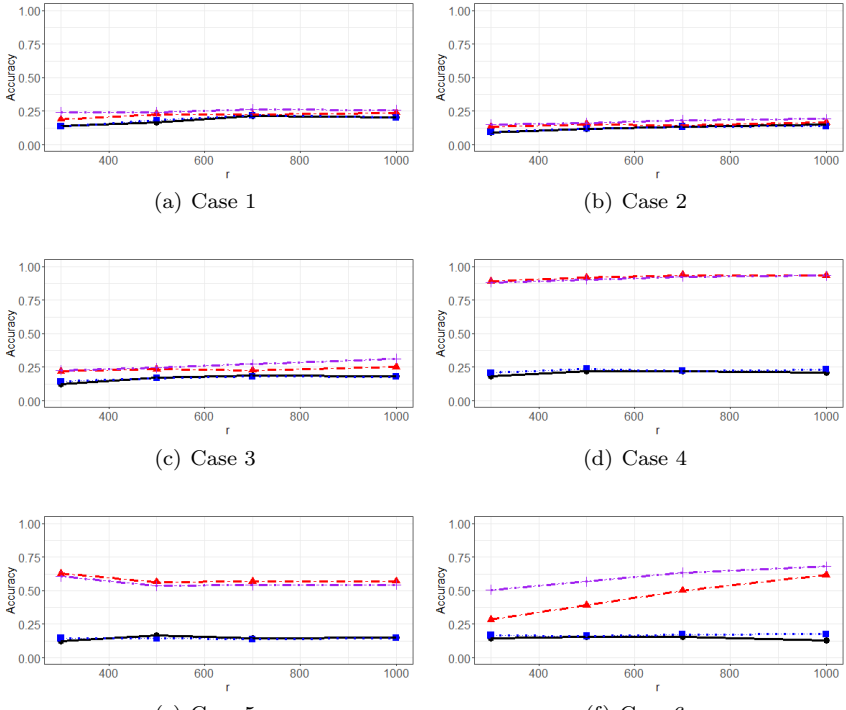
Fig. A.2 Log MSPEs for different r based on UNIF (solid line with circles), LEVSS (dashed line with triangles), LEV (dotted line with squares) and IBOSS (dotdash line with plus signs) methods for the six cases listed in Section 5.1. The models are selected based on the forward regression via BIC.

C Additional Simulation Results on Lasso

Now, we will exam our method's performance on model selection according to Lasso (Tibshirani, 1996).

To be aligned with the settings described in Section 5.1, we also demonstrate our methods as well as the other three methods (i.e., uniform subsampling, leverage score subsampling, and the IBOSS) on the six cases listed at the beginning of Section 5.1 and evaluate the selection performance through the selection accuracies and MSPEs. The Lasso method is conducted through the `glmnet` package (Simon et al., 2011) and the tuning parameters are selected through 10-fold cross-validation according to the `cv.glmnet()` function. As for the leverage score subsampling, the Lasso is conduct as in Leng and Leung (2011).

Results on the selection accuracies are presented in Figure A.3. It can be seen that the selection results based on Lasso are very similar to the results on the subset selection based on BIC.



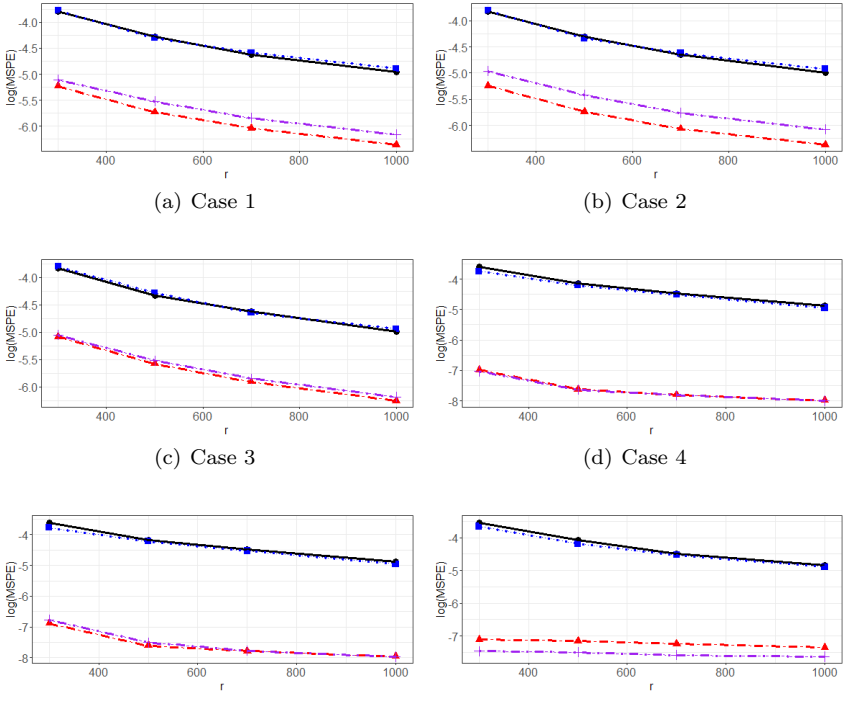
(a) Case 1 (b) Case 2

(c) Case 3 (d) Case 4

(e) Case 5 (f) Case 6

Fig. A.3 Selection accuracies for different r based on UNIF (solid line with circles), LEVSS (dashed line with triangles), LEV (dotted line with squares) and IBOSS (dotdash line with plus signs) method for the six cases listed in Section 5.1. The models are selected via Lasso.

To see the benefits of the model selection, we also report the log MSPEs in Figure A.4 with $n_{\text{test}} = 500$. From Figure A.4, we can clearly see that the IBOSS and our method (LEVSS) are uniformly performance better than the uniform subsampling method.



(a) Case 1 (b) Case 2

(c) Case 3 (d) Case 4

(e) Case 5 (f) Case 6

Fig. A.4 Log MSPEs for different r based on UNIF (solid line with circles), LEVSS (dashed line with triangles), LEV (dotted line with squares) and IBOSS (dotdash line with plus signs) methods for the six cases listed in Section 5.1. The models are selected via Lasso.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19:716–723

Atkinson AC, Fedorov VV (1975) The design of experiments for discriminating between two rival models. Biometrika 62:57–70

Bingham DR, Chipman HA (2007) Incorporating prior information in optimal design for model selection. Technometrics 49:155–163

Boivin J, Ng S (2006) Are more data always better for factor analysis? Journal of Econometrics 132:169 – 194

Box GEP, Hill WJ (1967) Discrimination among mechanistic models. Technometrics 9:57–71

Candes E, Tao T, et al. (2007) The Dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics 35:2313–2351

Chakrabortty A, Cai T (2018) Efficient and adaptive linear regression in semi-supervised settings. The Annals of Statistics 46:1541 – 1572

Chen WY, Mackey L, Gorham J, Briol FX, Oates C (2018) Stein points. In: Dy J, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, vol 80, pp 844–853

Chipman HA, Hamada MS (1996) Discussion: Factor-based or effect-based modeling? implications for design. Technometrics 38:317–320

Claeskens G, Hjort NL (2008) Model Selection and Model Averaging. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press

Consonni G, Deldossi L (2016) Objective bayesian model discrimination in follow-up experimental designs. TEST 25:397–412

Deldossi L, Tommasi C (2021) Optimal design subsampling from big datasets. Journal of Quality Technology in press
Dereziński M, Warmuth MK (2018) Reverse iterative volume sampling for linear regression. Journal of Machine Learning Research 19:1–39
Dette H, Titoff S (2009) Optimal discrimination designs. The Annals of Statistics 37:2056–2082
Dette H, Melas VB, Guchenko R (2015) Bayesian T-optimal discriminating designs. The Annals of Statistics 43:1959–1985
Drineas P, Kannan R, Mahoney MW (2006) Fast monte carlo algorithms for matrices I: Approximating matrix multiplication. SIAM Journal on Computing 36:132–157
Drineas P, Mahoney MW, Muthukrishnan S, Sarlós T (2011) Faster least squares approximation. Numerische Mathematik 117:219–249
Drovandi CC, McGree JM, Pettitt AN (2014) A sequential Monte Carlo algorithm to incorporate model uncertainty in Bayesian sequential design. Journal of Computational and Graphical Statistics 23:3–24
Efron B, Hastie T, Johnstone I, Tibshirani R, et al. (2004) Least angle regression. The Annals of statistics 32:407–499
Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association 96:1348–1360
Fang KT, Kotz S, Ng KW (1990) Symmetric Multivariate and Related Distributions. Monographs on Statistics and Applied Probability, Springer
Fithian W, Hastie T (2014) Local case-control sampling: Efficient subsampling in imbalanced data sets. The Annals of Statistics 42:1693–1724
Hastie T, Tibshirani R (1993) Varying-coefficient models. Journal of the Royal Statistical Society: Series B 55:757–779
Hastie TJ, Tibshirani RJ (1990) Generalized additive models, vol 43. CRC press
Joseph VR, Wang D, Gu L, Lyu S, Tuo R (2019) Deterministic sampling of expensive posteriors using minimum energy designs. Technometrics 61:297–308
Kadane JB, Lazar NA (2004) Methods and criteria for model selection. Journal of the American Statistical Association 99:279–290
Kleiner A, Talwalkar A, Sarkar P, Jordan MI (2015) A scalable bootstrap for massive data. Journal of the Royal Statistical Society: Series B 76:795–816
Kullback S, Leibler RA (1951) On information and sufficiency. The Annals of Mathematical Statistics 22:79–86
Lee S, Ng S (2020) An econometric perspective on algorithmic subsampling. Annual Review of Economics 12:45–80
Leng C, Leung DHY (2011) Model selection in validation sampling: An asymptotic likelihood-based lasso approach. Statistica Sinica 21:659–678
Li T, Meng C (2021) Modern subsampling methods for large-scale least squares regression. arXiv preprint arXiv:210501552
Lindley DV (1956) On a measure of the information provided by an experiment. The Annals of Mathematical Statistics 27:986–1005
López-Fidalgo J, Tommasi C, Trandafir PC (2007) An optimal experimental design criterion for discriminating between non-normal models. Journal of the Royal Statistical Society: Series B 69:231–242
Ma P, Mahoney MW, Yu B (2015) A statistical perspective on algorithmic leveraging. Journal of Machine Learning Research 16:861–919
Ma P, Zhang X, Xing X, Ma J, Mahoney MW (2020) Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. arXiv preprint arXiv:200210526
Mahoney MW (2012) Randomized algorithms for matrices and data. Foundations & Trends in Machine Learning 3:647–672
Mak S, Joseph VR (2018) Support points. The Annals of Statistics 46:2562–2592
Mamonov S, Triantoro T (2018) Subjectivity of diamond prices in online retail: insights from a data mining study. Journal of theoretical and applied electronic commerce research 13:15–28
McCullagh P, Nelder JA (1989) Generalized Linear Models. Monographs on Statistics and Applied Probability 37. Chapman & Hall
Meng C, Wang Y, Zhang X, Mandal A, Ma P, Zhong W (2017) Effective statistical methods for big data analytics. Handbook of Research on Applied Cybernetics and Systems Science pp 280–299
Meng C, Xie R, Mandal A, Zhang X, Zhong W, Ma P (2020a) Lowcon: A design-based subsampling approach in a misspecified linear model. Journal of Computational and Graphical

Statistics in press
Meng C, Zhang X, Zhang J, Zhong W, Ma P (2020b) More efficient approximation of smoothing splines via space-filling basis selection. Biometrika 107:723–735
Meng X, Saunders MA, Mahoney MW (2014) LSRN: A parallel iterative solver for strongly over-or underdetermined systems. SIAM Journal on Scientific Computing 36:C95–C118
Meyer RD, Steinberg DM, Box G (1996) Follow-up designs to resolve confounding in multifactor experiments. Technometrics 38:303–313
Miller A (2002) Subset selection in regression. CRC Press
Ng S (2017) Opportunities and challenges: Lessons from analyzing terabytes of scanner data. Tech. rep., National Bureau of Economic Research
Papailiopoulos D, Kyrillidis A, Boutsidis C (2014) Provable deterministic leverage score sampling. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 997–1006
Pukelsheim F (2006) Optimal design of experiments. Society for Industrial and Applied Mathematics
Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6:461–464
Sebastiani P, Wynn HP (2000) Maximum entropy sampling and optimal Bayesian experimental design. Journal of the Royal Statistical Society: Series B 62:145–157
Shao J (1997) An asymptotic theory for linear model selection. Statistica Sinica 7:221–264
Shewry MC, Wynn HP (1987) Maximum entropy sampling. Journal of Applied Statistics 14:165–170
Simon N, Friedman J, Hastie T, Tibshirani R (2011) Regularization paths for cox's proportional hazards model via coordinate descent. Journal of Statistical Software 39:1–13
Sin CY, White H (1996) Information criteria for selecting possibly misspecified parametric models. Journal of Econometrics 71:207 – 225
Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B 58:267–288
Truong Y, Kooperberg C, Stone C, Hansen M (2005) Statistical Modeling with Spline Functions: Methodology and Theory. Springer Series in Statistics, Springer Verlag New York
van der Vaart A (1998) Asymptotic statistics. Cambridge University Press
Wang H (2019) More efficient estimation for logistic regression with optimal subsamples. Journal of Machine Learning Research 20:1–59
Wang H, Zhu R, Ma P (2018) Optimal subsampling for large sample logistic regression. Journal of the American Statistical Association 113:829–844
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
Xu C, Chen J, Mantel H (2013) Pseudo-likelihood-based Bayesian information criterion for variable selection in survey data. Survey Methodology 39:303–321
Yang Y (2005) Can the strengths of AIC and BIC be shared? a conflict between model indentification and regression estimation. Biometrika 92:937–950
Yao Y, Wang H (2019) Optimal subsampling for softmax regression. Statistical Papers 60:585–599
Yao Y, Wang H (2021) A selective review on statistical techniques for big data. In: Modern Statistical Methods for Health Research, Springer in press
Yuan Z, Yang Y (2005) Combining linear regression models: When and how? Journal of the American Statistical Association 100:1202–1214
Zhang C (2010) Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics 38:894–942
Zhang T, Ning Y, Ruppert D (2020) Optimal sampling for generalized linear models under measurement constraints. Journal of Computational and Graphical Statistics in press
Zheng C, Ferrari D, Yang Y (2019) Model selection confidence sets by likelihood ratio testing. Statistica Sinica 29:827–851