The Canadian Journal of Statistics | La revue canadienne de statistique
Vol. 000, No. 00, 2026, e70067

Optimal subsampling for regression with mixed-type predictors

Jiaqing ZHU
Department of Statistics, School of Mathematics and Statistics, Shandong Normal University, Jinan, China

Lin WANG @
Department of Statistics, Purdue University, West Lafayette, Indiana, USA

Fasheng SUN
Department of Statistics, KLAS and School of Mathematics and Statistics, Northeast Normal University, Changchun, China

Abstract Subsampling has emerged as an appealing strategy to mitigate the computational and storage challenges imposed by large datasets. Recent subsampling techniques have shown notable computational gains for data dominated by numerical predictors. However, real-world datasets frequently contain both numerical and categorical predictors. Traditional treatments of categorical predictors commonly use dummy coding, which can lead to substantial dimensional expansion and added complexity during subsampling. To address these challenges, we propose a new mixed-type optimal subsampling technique (MOST) specifically designed for both numerical and categorical predictors. By enforcing combinatorial orthogonality, MOST effectively captures the informative structure of mixed-type predictors, enabling more efficient subdata estimators. Simulation studies and real-data applications demonstrate the superior accuracy and computational benefits of MOST compared to existing subsampling methods.

Résumé Le sous-échantillonnage s'est imposé comme une stratégie attrayante pour atténuer les difficultés de calcul et de stockage imposées par les grands jeux de données. Les techniques récentes de sous-échantillonnage ont montré des gains de calcul notables pour des données dominées par des prédicteurs numériques. Toutefois, les jeux de données réels contiennent fréquemment à la fois des prédicteurs numériques et catégoriels. Les traitements traditionnels des prédicteurs catégoriels recourent couramment au codage par variables muettes, ce qui peut entraîner une augmentation substantielle de la dimension et une complexité accrue lors du sous-échantillonnage. Pour relever ces défis, nous proposons une nouvelle technique de sous-échantillonnage optimal de type mixte (MOST), spécialement conçue pour les prédicteurs à la fois numériques et catégoriels. En imposant une orthogonalité combinatoire, MOST saisit efficacement la structure informative des prédicteurs de type mixte et produit des estimateurs sur sous-données plus performants. Des études de simulation et des applications sur données réelles démontrent la précision supérieure et les avantages de calcul de MOST par rapport aux méthodes de sous-échantillonnage existantes.

Keywords Data reduction; experimental design; orthogonal array; orthogonal subsampling.
MSC2020 Primary: 62K05; Secondary: 62K99.
@ Corresponding author linwang@purdue.edu

DOI: 10.1002/cjs.70067

1 Introduction

The exponential growth in dataset sizes has introduced unprecedented challenges for researchers. Traditional statistical and machine learning methods often become inefficient or even impractical for large datasets due to limitations in memory and computational resources (Fan et al., 2014; Meeker and Hong, 2014). Subsampling methods provide a practical solution for reducing storage and computational demands when analyzing large datasets. Beyond storage and computation, subsampling is crucial in measurement-constrained problems where labels or responses can be observed only for a subset of observations (Yu et al., 2025; Wang, 2026). Among existing subsampling methods, uniform subsampling selects subdata by assigning equal probability to each data point in the full dataset. Although uniform subsampling is simple to implement, its effectiveness is often limited, as it may fail to capture valuable information embedded in the full data. Ma and Sun (2015) proposed a leverage score–based subsampling method (LEV) and demonstrated its theoretical and numerical advantages for linear regression. Wang et al. (2019) proposed an information-based optimal subdata selection (IBOSS) method based on the D-optimality criterion, which outperforms LEV theoretically and numerically. Wang et al. (2021) proposed an orthogonal subsampling (OSS) approach and demonstrated its superiority for linear regression. Subsampling methods have also been extensively studied in a variety of other contexts. For instance, Ai et al. (2021) considered optimal sampling methods under the generalized linear model, Wang and Ma (2021) and Fan et al. (2021) investigated subsampling approaches for quantile regression, Ren et al. (2023) and Yang et al. (2024) presented subsampling methods for the multiplicative model, Zhu et al. (2024) focused on optimal subsampling for linear mixed models, and Chang (2023) developed an optimal subsampling method under a Gaussian process model. Ren et al. (2024) introduced a robust optimal subsampling framework based on double robustness. Subsampling for nonparametric regression has been addressed by Meng et al. (2020), Sun et al. (2021), and Zhang et al. (2024). For model-free scenarios, notable contributions include Mak and Joseph (2018), Song et al. (2022), Zhou et al. (2024), and Chang (2024). Furthermore, optimal subsampling methods in distributed environments have been studied by Zhang and Wang (2021), Yu et al. (2022), and Fan et al. (2024).

Most existing subsampling strategies focus on numerical predictors, with relatively little attention paid to categorical predictors. Handling categorical data often involves dummy coding, which creates a new variable for each level (category). Dummy coding can lead to substantial dimensional expansion and increased computational demands. Consequently, applying existing subsampling methods directly to dummy variables often intensifies the challenges associated with large datasets. To address this gap, Wang (2026) introduced balanced subsampling, a method specifically designed to select optimal subdata when working with categorical predictors. Despite this advancement, there remains a lack of optimal subsampling methods specifically designed for mixed-type big data containing both numerical and categorical predictors, which are commonly found in large datasets. For example, the airline dataset spanning from 2018 to April 2023 (https://www.kaggle.com/datasets/arvindnagaonkar/flight-delay/) contains over 30 million records. It has been extensively studied to reveal the dynamics of air travel, with representative analyses available in open-source projects such as the one available at https://github.com/alaahgag/BigData_Processing_Using_Spark/. In the modelling framework, predictors such as weekend/weekday status and flight distance can be regarded as categorical and numerical variables, respectively, which influence flight delays. Similarly, the diamond dataset (https://www.kaggle.com/datasets/shivam2503/diamonds/) includes about 54,000 records, with diamond colour as a categorical predictor and dimensions such as length and width as numerical predictors used to analyze factors affecting diamond price. The challenge of handling complex mixed-type data lies in addressing not only the unique issues associated with categorical and numerical predictors but also their interactions. Effectively resolving these challenges enables the efficient capture of valuable information within each type and across their interactions. This is closely related to the field of experimental design, where similar challenges are tackled using approaches such as mixed orthogonal arrays (MOA) (Hedayat et al., 1999), mixed-level supersaturated design (Sun et al., 2011), marginally coupled design (Deng et al., 2015; He et al., 2019), and optimal clustered-sliced Latin hypercube design (Huang et al., 2016). Building on the principles of these designs can help address these challenges and guide the selection of optimal subdata.

In this article, we propose our own mixed-type optimal subsampling technique (MOST), an optimal subsampling method designed for mixed-type data that includes both categorical and numerical predictors. MOST aims to select subdata that ensure combinatorial orthogonality within and across categorical and numerical predictors. This approach builds on the concept of approximating an MOA for optimal subdata selection, previously applied separately to numerical and categorical predictors in Wang et al. (2021) and Wang (2026).

DOI: 10.1002/cjs.70067

Our key contribution is to extend the optimality of MOAs to linear regression with mixed-type predictors, thereby ensuring accurate parameter estimation and reliable predictions based on the selected subdata. Extensive numerical results show that MOST outperforms existing subsampling methods in minimizing the mean squared error (MSE) for both parameter estimation and prediction.

The remainder of the article is organized as follows. Section 2 introduces the notation and background. Section 3 investigates the optimal properties of MOAs, based on which we present the proposed method, MOST, and describe its implementation algorithm. Sections 4 and 5 evaluate the method through simulation and real-world studies. Section 6 concludes the article with some discussion. All technical proofs and additional simulation results are provided in Supplementary Material.

2 Notation and background

Denote the full data as $F = \{(x_i, u_i, y_i)\}_{i=1}^N$, where $x_i = (x_{i1}, \dots, x_{ip_1})$ consists of p_1 numerical predictors, $u_i = (u_{i1}, \dots, u_{ip_2})$ consists of p_2 categorical predictors, and y_i is the corresponding response. Each categorical predictor $u_{\cdot j}$, having q_j levels, is coded into $q_j - 1$ dummy variables. Let $z_{i,jk}$ be the value of the kth dummy variable for u_{ij}, $j \in \{1, \dots, p_2\}$ and $k \in \{1, \dots, q_j - 1\}$. We consider the following linear model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p_1} x_{ip_1} + \sum_{k=1}^{q_1-1} \beta_{1k} z_{i,1k} + \dots + \sum_{k=1}^{q_{p_2}-1} \beta_{p_2 k} z_{i,p_2 k} + \varepsilon_i, \quad i = 1, \dots, N, \tag{1}$$

where ε_i is the independent normal random error with mean 0 and variance σ^2. Let $z_i = (1, x_{i1}, \dots, x_{ip_1}, z_{i,11}, \dots, z_{i,p_2(q_{p_2}-1)})^\top$, $Z = (z_1, \dots, z_N)^\top$, $\beta = (\beta_0, \beta_1, \dots, \beta_{p_1}, \beta_{11}, \dots, \beta_{p_2(q_{p_2}-1)})^\top$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_N)^\top$, and $y = (y_1, \dots, y_N)^\top$. The matrix form of the model in (1) is given by

$$y = Z\beta + \varepsilon. \tag{2}$$

We are commonly interested in the estimator of β, whose ordinary least squares (OLS) estimator based on the full data is given by

$$\hat{\beta} = (Z^\top Z)^{-1} Z^\top y.$$

The dimensionality of Z is $p = 1 + p_1 + \sum_{i=1}^{p_2}(q_j - 1)$, so the estimator $\hat{\beta}$ requires a time complexity of $O(Np^2 + p^3)$. When N is large, the term $O(Np^2)$ dominates the computational cost. The magnitude of q_j is squared in p^2 and then scaled by N. Therefore, even moderate values of q_j can significantly increase the computational burden, particularly for large N. As a result, the high computational demands of calculating $\hat{\beta}$ make it challenging to compute in practice.

Now consider taking a subset of size n from the full data. Denote the subdata as $S = \{(x_i^*, u_i^*, y_i^*)\}_{i=1}^n$. Let Z_S be the submatrix of Z corresponding to the subdata S and y_S be the corresponding response. The OLS estimator based on the subdata is given by

$$\hat{\beta}^* = (Z_S^\top Z_S)^{-1} Z_S^\top y_S.$$

Taking expectation and variance for $\hat{\beta}^*$, respectively, we have

$$\mathrm{E}(\hat{\beta}^*) = \beta, \text{ and } \mathrm{Var}(\hat{\beta}^*) = (Z_S^\top Z_S)^{-1} = M_S^{-1},$$

where

$$M_S = Z_S^\top Z_S$$

is the information matrix of the subdata. The optimal subdata should minimize $\mathrm{Var}(\hat{\beta}^*)$ in some manner, which can be obtained by minimizing an optimality function of M_S^{-1}. Denote ψ as the optimality function. Finding the optimal subdata is to solve the following optimization problem:

$$S_{opt} = \arg\min_S \psi(M_S^{-1}) \quad s.t. \ \ S \text{ contains } n \text{ points.} \tag{3}$$

The approach of selecting an optimal set of subdata through (3) aligns with the idea of optimal experimental design proposed by Kiefer (1959). Common optimality criteria include D-optimality and A-optimality. D-optimality seeks to minimize the uncertainty in estimating the parameter β by compressing the volume of the confidence ellipsoid, which is achieved by setting the optimization function ψ as the determinant function. A-optimality aims to reduce the average variance of parameter estimates by defining the optimization function ψ as the trace function. Wang et al. (2021) and Wang (2026) used the same optimization problem to identify optimal subdata for datasets containing only numerical or only categorical predictors, which are special cases considered in this article. For datasets with only numerical predictors (i.e., $p_2 = 0$), assume each predictor is scaled to the range $[-1, 1]$. Wang et al. (2021) defined a discrepancy function

$$\Delta^{(1)}(S) = \sum_{1\le i<j\le n}\left[p_1 - \left\|x_i^*\right\|^2/2 - \left\|x_j^*\right\|^2/2 + \sum_{k=1}^{p_1} I\left(\text{sign}(x_{ik}^*), \text{sign}(x_{jk}^*)\right)\right]^2, \tag{4}$$

where $I(\cdot)$ is the indicator function and $\text{sign}(x)$ is the sign of x. The OSS algorithm proposed by Wang et al. (2021) selects subdata by minimizing the discrepancy function defined in (4). For datasets with only categorical predictors (i.e., $p_1 = 0$), Wang (2026) defined a discrepancy function

$$\Delta^{(2)}(S) = \sum_{1\le i<j\le n}\left[\delta\left(u_i^*, u_j^*\right)\right]^2, \tag{5}$$

where

$$\delta\left(u_i^*, u_j^*\right) = \sum_{k=1}^{p_2} q_k I\left(u_{ik}^*, u_{jk}^*\right), \tag{6}$$

$I\left(u_{ik}^*, u_{jk}^*\right)$ is the indicator function that equals 1 if $u_{ik}^* = u_{jk}^*$ and 0 otherwise, and q_k is the number of levels of the kth categorical predictor. Balanced subsampling proposed by Wang (2026) selects subdata by minimizing the discrepancy function defined in (5). Both discrepancies in (4) and (5) measure the distortion of the combinatorial orthogonality of the subdata. Minimizing these discrepancies ensures the selection of subdata with the highest combinatorial orthogonality. For datasets containing both numerical and categorical predictors, an intuitive approach is to select a subset of the data that minimizes a combined discrepancy function by summing (4) and (5). However, this method ensures optimality only for the numerical and categorical predictors separately, without accounting for interactions between them. To address this, we propose an optimal subsampling algorithm that simultaneously accounts for combinatorial orthogonality within each predictor type and across their interactions, enabling the selection of truly optimal subdata.

3 Main method

We first establish the theoretical optimization of MOAs under model (2), which serves as the basis for developing the optimal subsampling method.

An $MOA(n, q_1^{d_1} q_2^{d_2} \cdots q_v^{d_v}, t)$ is an array of size $n \times d$, where $d = d_1 + d_2 + \cdots + d_v$ is the total number of predictors, in which the first d_1 predictors (columns) have q_1 levels, the next d_2 predictors have q_2 levels, and so on, with the property that every possible t-tuple appears equally often in any $n \times t$ subarray. Here, t is called the strength of an MOA, and the levels can be represented with any symbols. In the special case where all q_i's are equal, the MOA is called a fixed-level orthogonal array (OA). The reader is referred to Hedayat et al. (1999) for a comprehensive introduction to MOAs and fixed-level OAs. Here is an example of an $MOA(8, 2^2 4, 2)$:

$$\begin{pmatrix} -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 2 & 2 & 3 & 3 & 4 & 4 \end{pmatrix}^{\top}.$$

The first two predictors have two levels, represented by -1 and 1, and the last predictor has four levels, represented by 1,2,3, and 4. The symbols used to represent factor levels are purely notational. For each predictor, any one-to-one relabelling of its levels preserves the MOA property. In the MOA literature, it is common to use

the same set of symbols across factors to simplify notation and presentation. In this article, we use $\{-1, 1\}$ for two-level predictors for reasons discussed below. This choice does not affect the orthogonality or combinatorial balance of the MOA.

Now consider selecting optimal subdata from the full data. Recall that we denote the full data as $F = \{(x_i, u_i, y_i)\}_{i=1}^{N}$, where $x_i = (x_{i1}, \dots, x_{ip_1})$ consists of p_1 numerical predictors and $u_i = (u_{i1}, \dots, u_{ip_2})$ consists of p_2 categorical predictors. Without loss of generality, we assume that each numerical predictor is scaled to $[-1, 1]$. The following theorem, which guides our main method, establishes the optimality of MOAs for model (2).

Theorem 1. *For full data with p_1 numerical predictors and p_2 categorical predictors, a subdata set $S = \{(x_i^*, u_i^*, y_i^*)\}_{i=1}^{n}$ is D- and A-optimal if its design matrix $\{(x_i^*, u_i^*)\}_{i=1}^{n}$ forms an $MOA(n, 2^{p_1} q_1 \cdots q_{p_2}, 2)$, where the two levels corresponding to the p_1 numerical predictors are $\{-1, 1\}$, and $q_1, \cdots, q_{p_2}$ are the number of levels of the p_2 categorical predictors.*

By Theorem 1, the D- and A-optimal subdata should exhibit combined orthogonality, suggesting the selection of subdata that approximate an MOA of strength 2. First, the numerical predictors should form a two-level $OA(n, 2^{p_1}, 2)$, where the two levels represent the extreme values of the numerical predictors. Since the numerical predictors are scaled to $[-1, 1]$, these two levels are represented using $\{-1, 1\}$. This approach aligns with the findings in Wang et al. (2021). Second, the categorical predictors form an $MOA(n, q_1 \cdots q_{p_2}, 2)$, where the levels can be represented using any symbols. Additionally, these two parts should jointly form an $MOA(n, 2^{p_1} q_1 \cdots q_{p_2}, 2)$, ensuring combinatorial orthogonality between the numerical and categorical predictors. At this point, the information matrix M_S achieves its maximum under both D- and A-optimality criteria, with the estimator $\hat{\beta}^*$ minimizing the corresponding covariance matrix.

Practical questions in the study of MOAs concern whether arrays exist for specified design sizes and level structures, and how to construct them. Substantial progress has been made over the past few decades. For strength 2, Wu (1991) and Hedayat et al. (1992) established existence conditions and proposed constructive algorithms. Recursive techniques that scale to high dimensions were introduced by Suen et al. (2001). Comprehensive summaries of construction strategies, especially for multifactor and mixed-level settings, are provided in Hedayat et al. (1999). For higher strengths, Pang et al. (2021) developed efficient algorithms. A recent overview that consolidates core concepts, applications, and links to related areas appears in Lin and Stufken (2025). Despite these advances, existence and construction remain challenging for many parameter combinations and warrant further study. Additionally, these theoretical challenges are compounded in real applications because exact MOA structures are rarely present in full datasets. Our aim, therefore, is not to recover an MOA but to construct a subdata set that closely approximates the optimal structure and attains asymptotic optimality. This approach preserves the key theoretical guarantees while remaining broadly applicable in real-world settings.

Inspired by Theorem 1 and the above discussion, we propose a mixed-type optimal subsampling technique (MOST) to select subdata. One advantage of our method is that it does not rely on the existence of an MOA. Instead, we use a discrepancy measure $L(S)$ to quantify the distinction between the subdata and an MOA. Define the discrepancy function for subdata containing mixed-type predictors as

$$L(S) = \sum_{1 \le i < j \le n} \left[2p_1 - ||x_i^*||^2 - ||x_j^*||^2 + \delta\Big(\text{sign}(x_i^*), \text{sign}(x_j^*)\Big) + \delta\Big(u_i^*, u_j^*\Big) \right]^2, \tag{7}$$

where δ is defined in (6). Note that

$$\delta\Big(\text{sign}(x_i^*), \text{sign}(x_j^*)\Big) = 2 \sum_{k=1}^{p_1} I\Big(\text{sign}(x_{ik}^*), \text{sign}(x_{jk}^*)\Big)$$

because we take the signs (or extreme values) of a numerical predictor as two levels. The proposed MOST approach seeks the subdata S that minimizes $L(S)$. That is, MOST solves the following optimization problem:

$$S_{opt} = \arg\min_{S} L(S),$$
$$s.t.\ S \text{ contains } n \text{ points.} \tag{8}$$

The discrepancy function $L(S)$ in (7) can be decomposed as the sum of three terms:

$$L(S) = \sum_{1\le i<j\le n}\Big[2p_1 - ||x_i^*||^2 - ||x_j^*||^2 + \delta\Big(\text{sign}(x_i^*), \text{sign}(x_j^*)\Big)\Big]^2 + \sum_{1\le i<j\le n}\Big[\delta\Big(u_i^*, u_j^*\Big)\Big]^2$$
$$+ \sum_{1\le i<j\le n}\Big[2p_1 - ||x_i^*||^2 - ||x_j^*||^2 + \delta\Big(\text{sign}(x_i^*), \text{sign}(x_j^*)\Big)\Big]\Big[\delta\Big(u_i^*, u_j^*\Big)\Big]. \tag{9}$$

The first term in (9) is actually $\Delta^{(1)}(S)$, where minimization ensures combinatorial orthogonality of the numerical predictors. The second term corresponds exactly to $\Delta^{(2)}(S)$, and its minimization ensures combinatorial orthogonality of the categorical predictors. The third term, representing the cross product, ensures combinatorial orthogonality between the numerical and categorical predictors. When setting $p_1 = 0$ and $p_2 = 0$, respectively, $L(S)$ reduces to the discrepancy functions of balanced subsampling and OSS, which means that these two subsampling methods are special cases of our MOST. Minimizing $L(S)$ reduces all three terms, thus producing the optimal subdata. This discussion is supported by the following theorem.

Theorem 2. *For an $n \times (p_1 + p_2)$ subdata set S, $L(S)$ attains its minimum if S forms an MOA as defined in Theorem 1.*

Next, we propose an algorithm to approximate the optimal subdata by sequentially solving (8). We start with the point (x_1^*, u_1^*) with the maximum Euclidean norm in the numerical part and select the next $n-1$ points step by step. Suppose we have already selected m points for subdata S_m, then the $(m+1)$th point is selected by

$$(x_{m+1}^*, u_{m+1}^*) = \arg\min_{(x,u)\in F} L((x,u)|S_m),$$

where

$$L((x,u)|S_m) = \sum_{(x_i^*,u_i^*)\in S_m}\Big[2p_1 - ||x||^2 - ||x_i^*||^2 + \delta\big(\text{sign}(x), \text{sign}(x_i^*)\big) + \delta\big(u, u_i^*\big)\Big]^2$$

is the discrepancy introduced by adding (x,u) to S_m, and the minimization is over $(x,u) \in F \setminus S_m$. Because $L((x,u)|S_m) = L((x,u)|S_{m-1}) + \ell((x,u)|(x_m^*, u_m^*))$, and we have calculated $L((x,u)|S_{m-1})$ when searching for (x_m^*, u_m^*) at the mth iteration, we only need to compute

$$\ell((x,u)|(x_m^*, u_m^*)) = \big[2p_1 - ||x||^2 - ||x_m^*||^2 + \delta(\text{sign}(x), \text{sign}(x_m^*)) + \delta(u, u_m^*)\big]^2 \tag{10}$$

at the $(m+1)$th iteration. The computational complexity of each iteration is $O(N(p_1+p_2))$. To reduce the computation, we retain only κ_m points in $F \setminus S_m$ with the smallest values of $L((x,u)|S_m)$, discarding all other data points so they will not be considered in the $(m+2)$th iteration.

The parameter κ_m controls the number of points retained in $F \setminus S_m$. A larger value of κ_m generally improves the quality of the subdata but increases the computational cost, whereas a smaller κ_m reduces the computation time at the risk of suboptimal performance. Therefore, κ_m serves as a tuning parameter that balances statistical accuracy against computational efficiency. When $N \gg n$, many candidate points can be efficiently discarded, and we set $\kappa_m = N/m$, which is much smaller than N. Otherwise, only a limited number of points can be excluded. In this case, we set $\kappa_m = N/m^{r-1}$, where $r = \log(N)/\log(n)$. As demonstrated in our simulation and real-data studies, this choice yields favourable estimation and prediction accuracy while keeping the computational cost reasonable. Furthermore, this selection strategy is consistent with the practical guidance suggested in the OSS algorithm (Wang, 2026). The complete procedure of the MOST algorithm is summarized in Algorithm 1.

Remark 1. When $N \geqslant n^2$, $\kappa_m = N/m$ data points are retained in the mth iteration, requiring $O(N(p_1+p_2)/m)$ time to compute (x_{m+1}^*, u_{m+1}^*), so the complexity of the subsampling process in Algorithm 1 is $O(N(p_1+p_2)/1) + \cdots + O(N(p_1+p_2)/n) = O(N(p_1+p_2)\ln n)$. When $N \leqslant n^2$, $\kappa_m = N/m^{r-1}$ points are retained in the mth iteration, and the complexity becomes $O(N(p_1+p_2)/1) + \cdots + O(N(p_1+p_2)/n^{r-1}) = O(N(p_1+p_2)(n^{2-r}-1)/(2-r))$. Notably, the subsampling complexity of MOST does not depend on the levels of the categorical predictors q_j, which can be a considerable advantage in scenarios in which some q_j are large. In comparison, the subsampling process of IBOSS carries a complexity of $O(Np)$, with

Algorithm 1. MOST

Input: Full data $F = \{(x_i, u_i, y_i)\}_{i=1}^{N}$,subdata size n

Output: The subdata-based OLS estimator of $\hat{\beta}^*$

Scale values of each numerical predictor to $[-1, 1]$
Select $S_1 \leftarrow (x_1^*, u_1^*)$ which has the largest Euclidean norm in F
Calculate $\ell((x,u)|S_1)$ by (10), for all $(x,u) \in F \setminus S_1$
for $m = 1$ to $n-1$ **do**
  $(x_{m+1}^*, u_{m+1}^*) \leftarrow \arg\min_{(x,u)\in F\setminus S_m} L((x,u)|S_m)$
  $\{S_{m+1}, y_{m+1}^*\} \leftarrow \{S_m, y_m^*\} \bigcup \{(x_{m+1}^*, u_{m+1}^*, y_{m+1}^*)\}$
  if $N \geqslant n^2$ **then**
    $\kappa_m = N/m$
  else
    $\kappa_m = N/m^{r-1}$, where $r = \ln(N)/\ln(n)$
  end if
  Keep κ_{m+1} points in $F \setminus S_{m+1}$ with κ_{m+1} smallest components in $L((x,u)|S_m)$
  Compute $L((x,u)|S_{m+1}) \leftarrow L((x,u)|S_m) + \ell\left((x,u)|(x_{m+1}^*, u_{m+1}^*)\right)$
end for
Let Z_S be the variable obtained by a dummy coding of the categorical part of S_n and $y_S = y_n^*$. Estimate the coefficient β using

$$\hat{\beta}^* = (Z_S^\top Z_S)^{-1} Z_S^\top y_S. \tag{11}$$

$p = 1 + p_1 + \sum_{j=1}^{p_2}(q_j - 1)$, while LEV's subsampling step requires $O(Np^2)$ when using singular value decomposition (SVD) to compute leverage scores.

Remark 2. In the special case where many of the categorical predictors have many levels, so that the total number of dummy variables $\sum_{j=1}^{p_2}(q_j - 1)$ is large relative to $p_1 + p_2$, the expanded matrix Z becomes high dimensional and sparse. Each observation contains at most one nonzero indicator per categorical predictor, so the number of nonzero entries in Z, denoted by $\text{nnz}(Z)$, is approximately $N(p_1 + p_2 + 1)$. In this case, the OLS problem can be accelerated using iterative Krylov methods such as least squares QR decomposition (LSQR; Paige and Saunders, 1982) and least squares minimal residual decomposition (LSMR; Fong and Saunders, 2011). Each iteration requires one multiplication by Z and one by $Z^\top$, yielding $O(\text{nnz}(Z))$ cost per iteration and $O(\tau \cdot \text{nnz}(Z))$ total time to reach a target tolerance, where τ is the number of iterations required for convergence. If the only objective is to speed up computation when all responses of the full data are observed, these iterative solvers can be a competitive alternative to subsampling. However, if the primary objective is to select subdata for labelling (measurement-constrained problem), subsampling remains necessary. In this case, the computational complexity of MOST and IBOSS is unchanged. For LEV, sparsity can reduce the cost of score computation when leverage scores are approximated using sparse operations, but exact leverage scores require a sparse QR or singular value decomposition whose fill-in can make the runtime superlinear in $\text{nnz}(Z)$ and, in the worst case, approach dense $O(Np^2)$ time (Holodnak et al., 2015; Davis et al., 2016).

4 Simulation studies

In this section, we conduct simulation studies to evaluate the performance of the MOST algorithm. We consider several different combinations of (p_1, p_2), where p_1 denotes the number of numerical predictors and p_2 denotes the number of categorical predictors, which have $q_j = j + 1$ levels for $j = 1, \ldots, p_2$. Specifically, we examine the settings $(p_1, p_2) = (50, 5), (10, 1), (5, 10)$, and $(100, 5)$ to show the robustness under different scenarios. To avoid overloading the article, we focus on the setting $(50, 5)$ here and defer the additional results to the Supplementary Material. The following cases are examined to generate the design matrix for the full dataset.

> Case 1. The numerical predictors are independent and have a multivariate uniform distribution: $x_{ik} \sim U[-1, 1]$, $k \in \{1, \ldots, p_1\}$. The categorical predictors are independent, and for each predictor, the q_j levels have probabilities $1/q_j$.

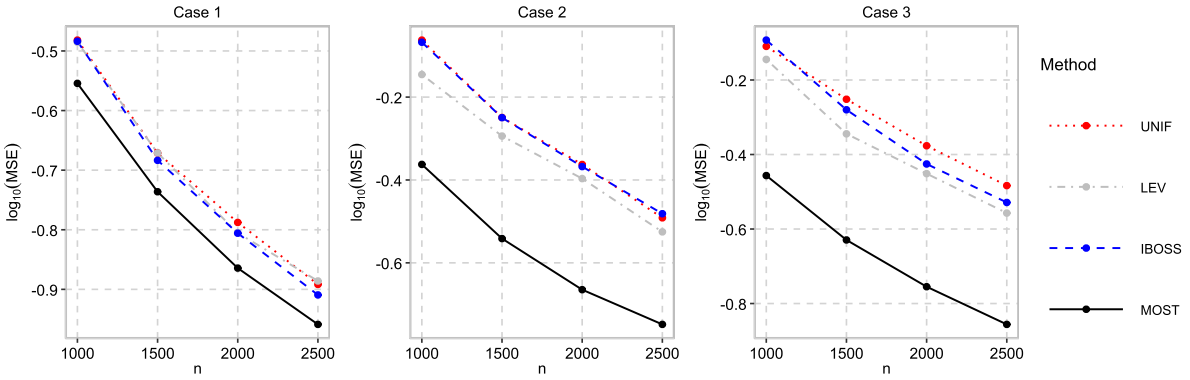

Figure 1: The $\log_{10}$(MSE) of the estimated parameters for different subdata sets of size n. The full data size is $N = 5 \times 10^4$.

Case 2. The numerical predictors are generated in the same way as in **Case 1**. The categorical predictors follow the multivariate normal distribution: $u_i \sim N(0, \Sigma)$, with

$$\Sigma = \left(0.5^{I(k \neq k')}\right), \tag{12}$$

where $I(k \neq k')$ is equal to 0 if $k = k'$ and 1 otherwise. Discretize $[-3, 3]$ to q_j intervals of equal length, and let $u_{ij} = t$ if u_{ij} falls into the tth interval. Let $u_{ij} = 1$ if $u_{ij} < -3$ and $u_{ij} = q_j$ if $u_{ij} > 3$.

Case 3. The numerical predictors have a multivariate normal distribution: $x_i \sim N(0, \Sigma)$, where Σ is defined in (12). The categorical predictors are generated in the same way as in **Case 2**.

The response data are generated from the linear model (1) with the true value of β being a vector of unity, which includes an intercept, 50 slope parameters for numerical predictors, and $\sum_{j=1}^{5}(q_j - 1) = 15$ slope parameters for the categorical predictors. The error term is $\varepsilon_i \sim N(0, 1)$. The simulation is repeated $B = 200$ times. We compare MOST with three different subsampling methods: UNIF (uniform subsampling), LEV, and IBOSS. Since LEV and IBOSS cannot be directly applied to mixed-type predictors, they are instead applied to the matrix combining the numerical predictors and the dummy-coded variables of the categorical predictors. In the bth repetition, $b = 1, \ldots, 200$, we select subdata using each subsampling method, train model (1), and compute the OLS estimator $\hat{\beta}^{*(b)}$. The empirical MSE of $\hat{\beta}^*$ is then calculated as

$$\text{MSE} = B^{-1} \sum_{b=1}^{B} ||\hat{\beta}^{*(b)} - \beta||^2. \tag{13}$$

Figure 1 plots the estimation performance of the different subsampling methods for various subdata sets of sizes $n = 10^3, 1.5 \times 10^3, 2 \times 10^3$, and 2.5×10^3, with the full dataset size being size $N = 5 \times 10^4$. In all three cases, MOST consistently exhibits smaller estimation errors than the other subsampling methods, although the MSEs for all four methods decrease at the same rate as n increases. Notably, the superiority of MOST is more pronounced in Cases 2 and 3 than in Case 1. In Case 2, where the categorical predictors are derived from a correlated normal distribution, the combinatorial orthogonality achieved by MOST in selecting subdata enhances the estimation performance. In Case 3, where both the categorical and numerical predictors exhibit correlations, the advantage of MOST becomes even more evident.

Figure 2 evaluates the performance of the subsampling methods as the full data size increases, with $N = 5 \times 10^3, 10^4, 10^5$, and 5×10^5, and a fixed subdata size of $n = 4 \times 10^3$. MOST consistently exhibits smaller MSEs than the subdata sets derived from the other methods. Notably, only MOST shows a decrease in MSE as the full dataset size N increases. This trend highlights that MOST effectively extracts more information from larger full data, leading to improved estimation accuracy. In contrast, while IBOSS also shows this trend for datasets containing only numerical predictors, its efficiency is significantly reduced in mixed-type data containing both categorical and numerical predictors, with performance inferior to LEV. This is because the extreme-value rule of IBOSS can cluster similar dummy patterns and affect the conditioning of $Z_S^\top Z_S$, whereas leverage-based sampling uses joint geometry and selects boundary points in the dummy space.

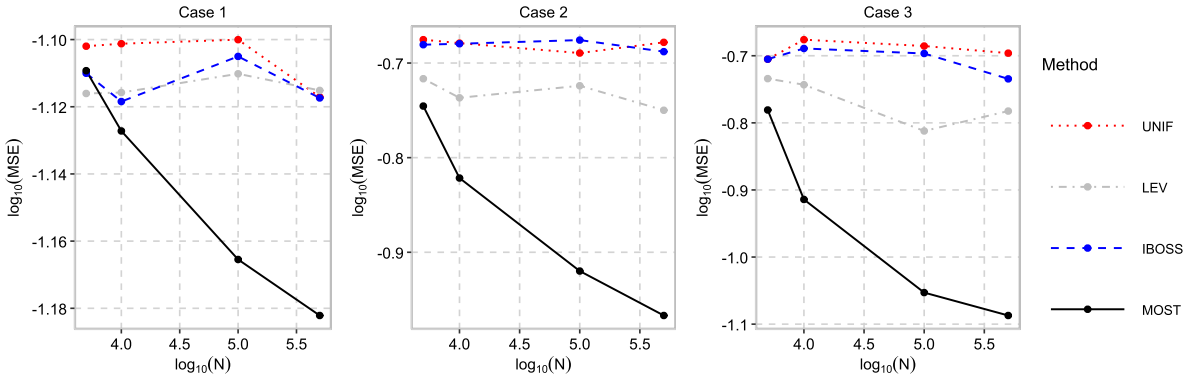

Figure 2: The $\log_{10}$(MSE) of the estimated parameters for different full data of size N. The subdata size is $n = 4 \times 10^3$.

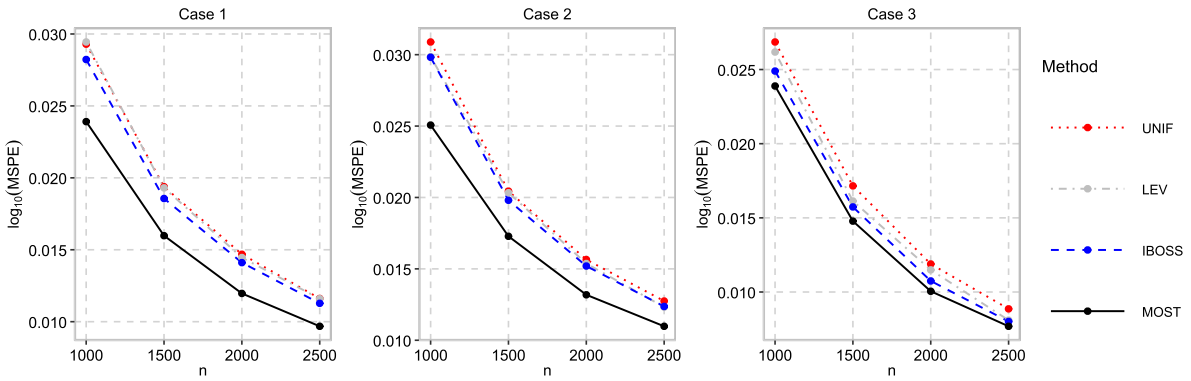

Figure 3: The $\log_{10}$(MSPE) of test data response for different subdata sets of size n. The full data size is $N = 5 \times 10^4$.

To demonstrate the predictive performance with the MOST subdata, we generate independent test data of size $N_{test} = 10^5$ for each case and obtain the predicted response from the model trained on each subdata set. We calculate the mean squared predictive errors (MSPE) of the prediction from

$$\text{MSPE} = \frac{1}{BN_{test}} \sum_{b=1}^{B} ||\hat{Y}^{(b)} - Y_{test}||^2, \tag{14}$$

where Y_{test} is the vector of observed responses for the test data, and $\hat{Y}^{(b)}$ is the vector of predicted responses in the bth repetition. Figure 3 displays the predictive performance of the different subsampling methods, measured by MSPE, as the subdata set sizes increase across three cases. Although the MSPEs of all the methods generally decrease with larger subdata set sizes, MOST consistently achieves the lowest MSPE. This indicates that MOST selects subdata with superior predictive performance.

We further examine the predictive performance of MOST as the full data size increases, as shown in Figure 4. MOST consistently exhibits smaller MSPEs than the other subdata. Although IBOSS exhibits a decreasing trend in Case 3, likely due to an increase in extreme numerical values, the balance between the categorical and numerical predictors in the MOST subdata sets yields more accurate predictions. As a result, MOST maintains its superiority with a rapidly declining trend of MSPE.

Tables 1 and 2 compare the computation times (including subsampling and computing estimators). The predictors are generated as in Case 1. Each reported time is the average CPU time of 200 repetitions. All computations are carried out on a laptop running Windows 11 Pro with an Intel Core i7-1280P processor and 32GB memory. As indicated in Table 1, when the q_j values are relatively small, as in the above simulation studies, MOST runs faster than LEV but slower than IBOSS. Meanwhile, Table 2 shows that as q_j increases, the

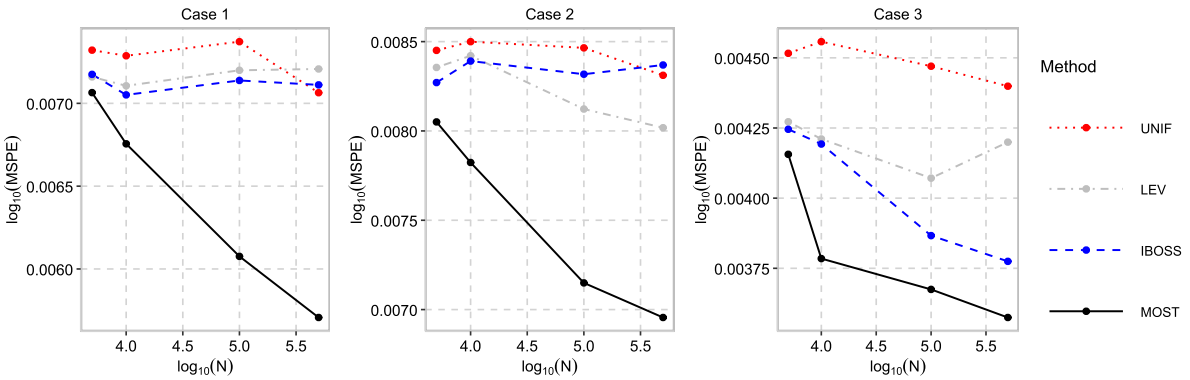

Figure 4: The $\log_{10}$(MSPE) of test data response for different full data of size N. The subdata size is $n = 4 \times 10^3$.

Table 1: CPU times (in seconds) of subsampling methods with $n = 1000$ and $q_j = j + 1$.

Method	UNIF	LEV	IBOSS	MOST
$N = 10^4$	0.048	0.120	0.078	1.048
$N = 10^5$	0.480	1.412	0.836	2.880
$N = 5 \times 10^5$	1.381	9.946	2.747	7.701

Table 2: CPU times (in seconds) of subsampling methods with $n = 1000$ and $N = 10^6$.

Method	UNIF	LEV	IBOSS	MOST
$q_j = j + 1$	2.545	12.341	4.133	13.533
$q_j = 5 \times (j + 1)$	3.312	48.980	7.967	13.208
$q_j = 10 \times (j + 1)$	4.466	137.345	13.412	13.508
$q_j = 20 \times (j + 1)$	7.682	431.101	25.532	13.594

computation time for MOST remains unchanged while the times for IBOSS and LEV increase and eventually become much higher than that of MOST.

5 Real data analysis—Diamonds data

We evaluate the performance of MOST by analyzing the diamond dataset, which records 53,940 features of diamonds, facilitating the analysis of various factors that affect diamond prices. More information on the dataset can be found at https://www.kaggle.com/datasets/shivam2503/diamonds/. Our focus is on examining the relationship between the diamond price (y) and six continuous numerical features: weight (x_1), total depth percentage (x_2), width of the top of the diamond relative to its widest point (x_3), length (x_4), width (x_5), and depth (x_6), as well as three categorical features: cut quality (u_1), colour (u_2), and clarity (u_3), which have $q_1 = 5$, $q_2 = 7$, and $q_3 = 8$ levels, respectively. Given the skewed distribution of the weight and price of diamonds, we perform a logarithmic transformation on x_1 and y. Following the application of dummy coding to the three categorical predictors, we consider fitting the data using the model below:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \sum_{k=1}^{4} \beta_{1k} u_{i,1k} + \sum_{k=1}^{6} \beta_{2k} u_{i,2k} + \sum_{k=1}^{7} \beta_{3k} u_{i,3k} + \varepsilon_i, \ i \in \{1, \ldots, 53940\}.$$

We consider subdata sizes $n = 280, 350, 420$, and 490 and evaluate subsampling methods by computing the difference between the estimator obtained from the subdata and that from the full data. Let

$$\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_{11}, \ldots, \beta_{14}, \beta_{21}, \ldots, \beta_{26}, \beta_{31}, \ldots, \beta_{37})^\top.$$

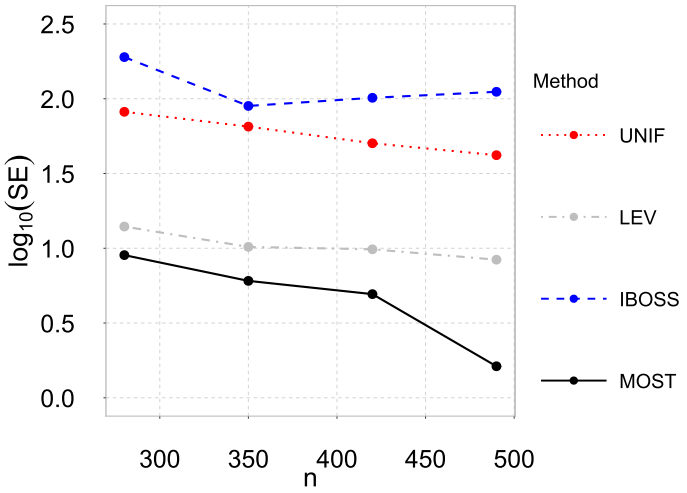

Figure 5: Squared error (SE) of estimated parameters for different subsampling methods.

Table 3: CPU times (in seconds) of subsampling methods with $n = 490$ for the real data study.

Method	UNIF	LEV	IBOSS	MOST
Time	0.049	0.145	0.096	0.106

We consider the squared error (SE)

$$SE = \|\hat{\beta}^* - \hat{\beta}\|^2,$$

where $\hat{\beta}$ is the OLS estimator of the parameters based on the full data, and $\hat{\beta}^*$ is the estimator based on the subdata. For the UNIF and LEV methods, being random subsampling techniques, we repeat the process 200 times and compute the average SE. MOST and IBOSS are deterministic methods and are executed only once. Figure 5 plots the SE for the different subsampling methods, which demonstrates the superior performance of MOST in minimizing SE across all subdata sizes.

Table 3 shows the CPU times of the different subsampling methods for $n = 490$. Because the subdata size n is relatively small, MOST exhibits a computation time comparable to IBOSS, and both are faster than LEV.

6 Conclusion

In this article, we introduced a novel subsampling algorithm, MOST, designed to handle both numerical and categorical predictors. By leveraging principles from experimental design to maintain combinatorial orthogonality, MOST effectively ensures the approximately D- and A-optimality of the selected subdata. Extensive simulation studies and a real-world application confirmed that MOST consistently outperforms existing methods in minimizing the MSE of parameter estimation. Moreover, beyond its robustness in parameter estimation, MOST also exhibits strong predictive accuracy, making it a comprehensive tool for regression analysis in large-scale mixed-type data settings.

Beyond existing studies, our work made three distinctive contributions. First, we established the optimality of $MOA(n, 2^{p_1}q_1 \cdots q_{p_2}, 2)$ under mixed-type data and demonstrated its effectiveness in minimizing the deviation function, thereby providing a solid theoretical foundation for algorithm design. Second, we proposed a new discrepancy function that accommodates both numerical and categorical predictors, ensuring that the selected subdata better approximates an MOA. Finally, numerical results consistently showed that MOST outperforms algorithms developed for single-type data in both estimation and prediction.

Although in this article we assumed binary dummy coding for categorical predictors, all theoretical results remain valid under any nonsingular coding scheme, such as orthogonal coding frameworks. This flexibility ensures that the superiority of a balanced subsample does not depend on the specific coding system. The

core advantage of MOST lies in its capacity to preserve combinatorial orthogonality, which is invariant under nonsingular transformations of the predictor space. This generalizability further reinforces the broad applicability and robustness of the proposed method.

We focus on MOAs of strength 2. Understanding how other design properties, especially higher strengths, affect theoretical guarantees and empirical performance is an important direction. Higher strengths provide better coverage of interactions and can reduce bias in models that include higher order terms, but they typically require larger arrays. Because our goal is to estimate main effects under subsampling, strength 2 is sufficient to balance the main effects in this study.

Several aspects of this study warrant further investigation. First, we focused on continuous response variables for simplicity, yet categorical response variables also arise frequently in practice. Developing design-based subsampling methods that address the unique challenges posed by categorical predictors and categorical responses is therefore an important direction for future work. Second, real-world datasets often contain missing values, resulting in sparse and incomplete data. Designing subsampling methods that effectively accommodate sparsity and missingness in the presence of categorical predictors remains a compelling direction for further research.

Data sharing

The data used in this article were obtained from public domain resources via available in Kaggle https://www.kaggle.com/datasets/shivam2503/diamonds/. R source code to implement the proposed method is provided at https://github.com/zhu1137/MOST/.

Funding

Zhu is supported by the National Natural Science Foundation of China (No. 12501343). Wang is supported by the US National Science Foundation (DMS-2413741) and the Central Indiana Corporate Partnership AnalytiXIN Initiative. Sun is supported by the National Natural Science Foundation of China (No. 12371259).

ORCID

Jiaqing Zhu: https://orcid.org/0000-0001-9008-9093
Lin Wang: https://orcid.org/0000-0003-0888-6232
Fasheng Sun: https://orcid.org/0000-0003-2410-4018

Supplementary Material

The proofs of Theorems 1 and 2, along with additional simulation results, are available as Supplementary Material.

References

Ai, M., Yu, J., Zhang, H., and Wang, H. (2021). Optimal subsampling algorithms for big data regressions. *Statistica Sinica*, 31(2):749–772.

Chang, M.-C. (2023). Predictive subdata selection for computer models. *Journal of Computational and Graphical Statistics*, 32(2):613–630.

Chang, M.-C. (2024). Supervised stratified subsampling for predictive analytics. *Journal of Computational and Graphical Statistics*, 33(3):1017–1036.

Davis, T. A., Rajamanickam, S., and Sid-Lakhdar, W. M. (2016). A survey of direct methods for sparse linear systems. *Acta Numerica*, 25:383–566.

Deng, X., Hung, Y., and Lin, C. D. (2015). Design for computer experiments with qualitative and quantitative factors. *Statistica Sinica*, 25:1567–1581.

Fan, J., Han, F., and Liu, H. (2014). Challenges of big data analysis. *National Science Review*, 1(2):293–314.

Fan, Y., Lin, N., and Yu, L. (2024). Distributed quantile regression for longitudinal big data. *Computational Statistics*, 39(2):751–779.

Fan, Y., Liu, Y., and Zhu, L. (2021). Optimal subsampling for linear quantile regression models. *The Canadian Journal of Statistics*, 49(4):1039–1057.

Fong, D. C.-L. and Saunders, M. (2011). LSMR: An iterative algorithm for sparse least-squares problems. *SIAM Journal on Scientific Computing*, 33(5):2950–2971.

He, Y., Lin, C. D., and Sun, F. (2019). Construction of marginally coupled designs by subspace theory. *Bernoulli*, 25(3):2163–2182.

Hedayat, A., Pu, K., and Stufken, J. (1992). On the construction of asymmetrical orthogonal arrays. *The Annals of Statistics*, 20:2142–2152.

Hedayat, A., Sloane, N., and Stufken, J. (1999). *Orthogonal Arrays: Theory and Applications*, Springer Science & Business Media, New York, USA.

Holodnak, J. T., Ipsen, I. C., and Wentworth, T. (2015). Conditioning of leverage scores and computation by QR decomposition. *SIAM Journal on Matrix Analysis and Applications*, 36(3):1143–1163.

Huang, H., Lin, D. K., Liu, M., and Yang, J. (2016). Computer experiments with both qualitative and quantitative variables. *Technometrics*, 58(4):495–507.

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society, Series B*, 21:272–319.

Lin, C. and Stufken, J. (2025). Orthogonal arrays: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 17(2):e70029.

Ma, P. and Sun, X. (2015). Leveraging for big data regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(1):70–76.

Mak, S. and Joseph, V. R. (2018). Support points. *The Annals of Statistics*, 46(6A):2562–2592.

Meeker, W. Q. and Hong, Y. (2014). Reliability meets big data: Opportunities and challenges. *Quality Engineering*, 26(1):102–116.

Meng, C., Zhang, X., Zhang, J., Zhong, W., and Ma, P. (2020). More efficient approximation of smoothing splines via space-filling basis selection. *Biometrika*, 107(3):723–735.

Paige, C. C. and Saunders, M. A. (1982). LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software (TOMS)*, 8(1):43–71.

Pang, S., Wang, J., Lin, D. K., and Liu, M. (2021). Construction of mixed orthogonal arrays with high strength. *The Annals of Statistics*, 49(5):2870–2884.

Ren, M., Zhao, S., and Wang, M. (2023). Optimal subsampling for least absolute relative error estimators with massive data. *Journal of Complexity*, 74:101694.

Ren, M., Zhao, S., Wang, M., and Zhu, X. (2024). Robust optimal subsampling based on weighted asymmetric least squares. *Statistical Papers*, 65(4):2221–2251.

Song, D., Xi, N. M., Li, J. J., and Wang, L. (2022). scSampler: Fast diversity-preserving subsampling of large-scale single-cell transcriptomic data. *Bioinformatics*, 38(11):3126–3127.

Suen, C., Das, A., and Dey, A. (2001). On the construction of asymmetric orthogonal arrays. *Statistica Sinica* 11:241–260.

Sun, F., Lin, D. K. J., and Liu, M. (2011). On construction of optimal mixed-level supersaturated designs. *The Annals of Statistics*, 39(2):1310–1333.

Sun, X., Zhong, W., and Ma, P. (2021). An asymptotic and empirical smoothing parameters selection method for smoothing spline ANOVA models in large samples. *Biometrika*, 108(1):149–166.

Wang, H. and Ma, Y. (2021). Optimal subsampling for quantile regression in big data. *Biometrika*, 108(1): 99–112.

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

Wang, L. (2026). Balanced subsampling for big data with categorical covariates. *Statistica Sinica,* 36:1389–1406. doi: 10.5705/ss.202023.0434.

Wang, L., Elmstedt, J., Wong, W., and Xu, H. (2021). Orthogonal subsampling for big data linear regression. *Annals of Applied Statistics*, 15(3):1273–1290.

Wu, C. (1991). Balanced repeated replications based on mixed orthogonal arrays. *Biometrika*, 78(1):181–188.

Yang, Z., Wang, H., and Yan, J. (2024). Subsampling approach for least squares fitting of semi-parametric accelerated failure time models to massive survival data. *Statistics and Computing*, 34(2):77.

Yu, J., Wang, H., Ai, M., and Zhang, H. (2022). Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association*, 117(537):265–276.

Yu, J., Ye, Z., Ai, M., and Ma, P. (2025). Optimal subsampling for data streams with measurement constrained categorical responses. *Journal of Computational and Graphical Statistics*, 34(3):994–1004.

Zhang, H. and Wang, H. (2021). Distributed subdata selection for big data via sampling-based approach. *Computational Statistics & Data Analysis*, 153:107072.

Zhang, Y., Wang, L., Zhang, X., and Wang, H. (2024). Independence-encouraging subsampling for nonparametric additive models. *Journal of Computational and Graphical Statistics*, 33(4):1424–1433.

Zhou, Z., Yang, Z., Zhang, A., and Zhou, Y. (2024). Efficient model-free subsampling method for massive data. *Technometrics*, 66(2):240–252.

Zhu, J., Wang, L., and Sun, F. (2024). Group-orthogonal subsampling for hierarchical data based on linear mixed models. *Journal of Computational and Graphical Statistics*, 33(3):1037–1046.

Received 29 March 2025 — Accepted 11 February 2026