Orthogonal Subsampling for Big Data Linear Regression

Hongquan Xu

UCLA Department of Statistics

Joint work with

Lin Wang, Jake Elmstedt, and Weng Kee Wong

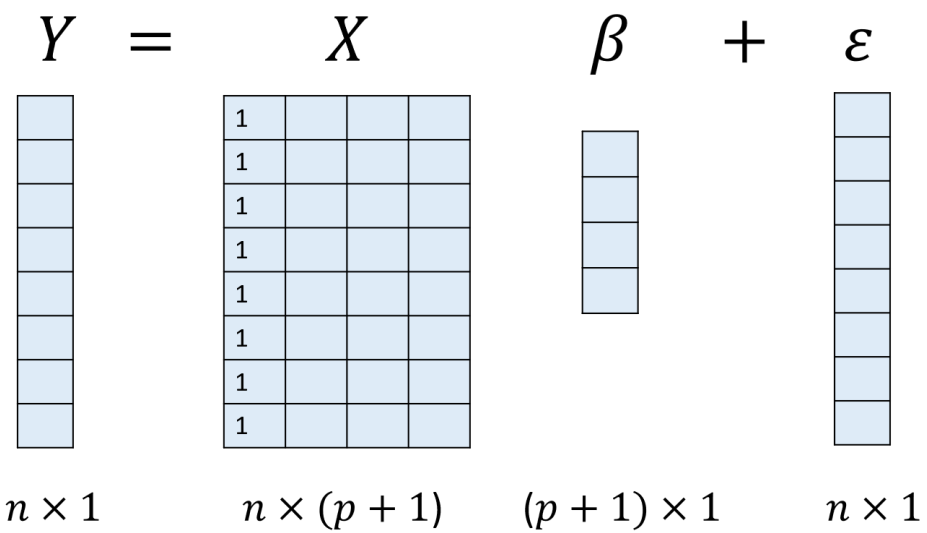
Motivation

- Data are big (sometimes redundant)
- Analyzing the full data may be computationally expensive
- Storing all of the data may not be possible
- There are a lot of circumstances under which X is big while the label (or response) Y is expensive to obtain

Big X	**Small Y**
Patients' records	Performance of a new medicine
Images as visual stimuli	Brain response

Linear Regression Setup

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$



Question:

1. If the budget only allows k responses (labels), $k \ll n$, choose which k to label? (Measurement-Constrained)
2. When all responses are available, to accelerate the computation, how to choose a subsample of size $k \ll n$?

Leverage Subsampling

A subsampling method consists of

- sampling probabilities π_i, $i = 1, \ldots, n$, $\sum_i^n \pi_i = 1$
- a weighted estimator $\hat{\beta}_s = (X_s^T W X_s)^{-1} X_s^T W Y_s$, where X_s is the subsample taken from X and $W = diag(w_1, \ldots, w_k)$ is a weight matrix

1. Uniform sampling: $\pi_i = 1/n$, $w_i = 1$
2. Leveraging: $\pi_i = h_{ii}/(p+1)$, $w_i = 1/\pi_i$, where

$$h_{ii} = (X(X^T X)^{-1} X^T)_{ii}$$

(Drineas et al., 2006; Ma et al., 2015)

Information-based Subsampling

The OLS for a subsample X_s is $\hat{\beta}_s = (X_s^T X_s)^{-1} X_s^T Y_s$

$$\mathrm{E}(\hat{\beta}_s) = \beta, \quad \mathrm{Var}(\hat{\beta}_s) = \sigma^2 (X_s^T X_s)^{-1}$$

- The Fisher information matrix for β with subdata is

$$M_s = X_s^T X_s$$

- D-optimality: to find X_s with k points that maximizes $\det(M_s)$ (that is, minimizes $\det(M_s^{-1})$)

Available approach: IBOSS (Wang H., Yang, and Stufken, 2018 JASA)
Include data points with extreme (largest and smallest) covariate values

Theoretical Results for Optimality

- D-optimality: Minimize the generalized variance of the estimates.
- A-optimality: Minimize the average variance of the estimates.

Theorem

Suppose each covariate is scaled to $[-1, 1]$. For a subsample X_s of size k,

$$\det(M_s) = \prod_{j=0}^{p} \lambda_j(M_s) \leq k^{p+1}, \tag{1}$$

$$trace(M_s^{-1}) = \sum_{j=0}^{p} \frac{1}{\lambda_j(M_s)} \geq \frac{p+1}{k}, \tag{2}$$

where $\lambda_j(M_s)$ are the eigenvalues of M_s. The equalities hold if and only if X_s forms a two-level OA with levels from $\{-1, 1\}$.

Orthogonal Array

- A matrix with entries from a fixed set of levels, say $\{-1, 1\}$
- All t-tuples of the levels appear the same number of times
- The number t is called the strength of the OA

A two-level OA, with strength $t = 2$

$$\begin{pmatrix} -1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \\ 1 & 1 & 1 \end{pmatrix}$$

- The optimality of orthogonal arrays inspires us to find a subsample that best approximates an orthogonal array

How to find an OA subsample?

- Exhaustive search requires to check $\binom{n}{k}$ subsamples and is infeasible
- An intuitive way is to select data points that match a prespecified OA

  Issue: the available data points typically do not allocate well in the subspace of a prespecified OA. For example, for a given OA with 10 columns, permuting the columns will generate $10! > 10^7$ OAs in up to 10^7 different subspaces. It is unlikely that the observed data allocate in any randomly specified subspace of them.
- We search for an OA that matches the data

Orthogonal Subsampling: A Discrepancy

Data points in an orthogonal array have two features:

(i). Extreme values: points are located at the corners of the data domain ($[-1,1]^p$) and have large distances from the center

(ii). Combinatorial orthogonality: points (or precisely, their signs) are as dissimilar as possible

Denote x_i^*, $i = 1, \ldots, k$, as the data points selected in X_s.

$$L(X_s) = \sum_{1 \le i < j \le k} \left[p - \|x_i^*\|^2/2 - \|x_j^*\|^2/2 + \delta(s(x_i^*), s(x_j^*)) \right]^2.$$

$$\delta(s(x_i^*), s(x_j^*)) = \sum_{l=1}^{p} \delta_1(s(x_{il}^*), s(x_{jl}^*))$$

δ_1 is the indicator function and $s(x)$ is the sign of x

Optimality

Theorem

For any k-point subsample X_s over $[-1,1]^p$,

$$L(X_s) \geq [k^2 p(p+1) - 4kp^2]/8,$$

with equality if and only if X_s forms a two-level orthogonal array.

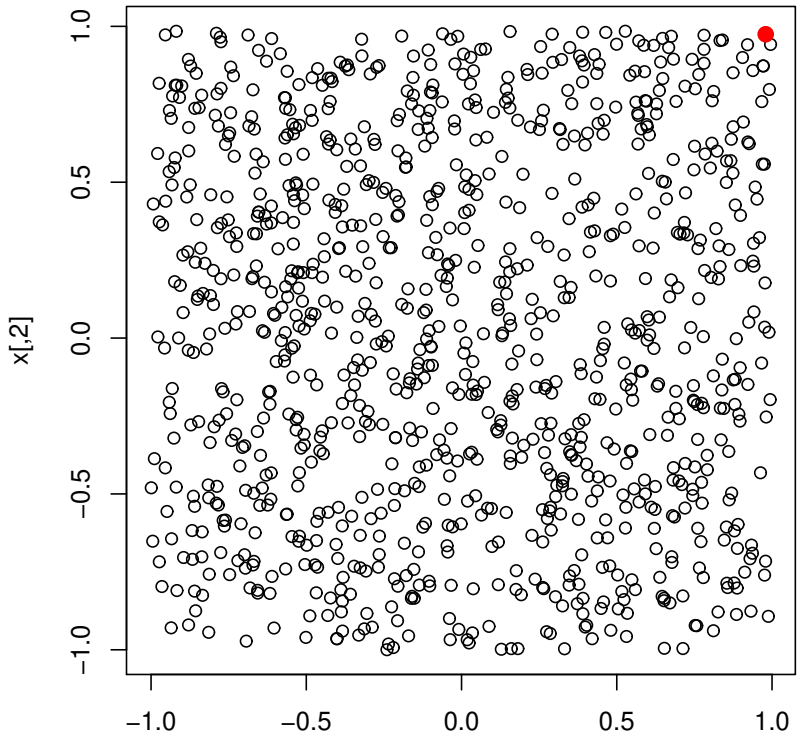
The subsampling problem can be presented as the following optimization problem:

$$\begin{aligned} X_s^* &= \arg\min_{X_s} L(X_s), \\ & s.t. \quad X_s \text{ contains } k \text{ points.} \end{aligned}$$

This enables sequential selection of the subsample points.

Algorithm: Select and eliminate points simultaneously

Step 1. Let $i = 1$. Find the point in X with the largest Euclidean norm, denoted as x_1^*. Include x_1^* in X_s and remove it from X. Let $\mathscr{D} = (0, \ldots, 0)$ be an $(n-1)$-vector with each component corresponding to each data point in X.

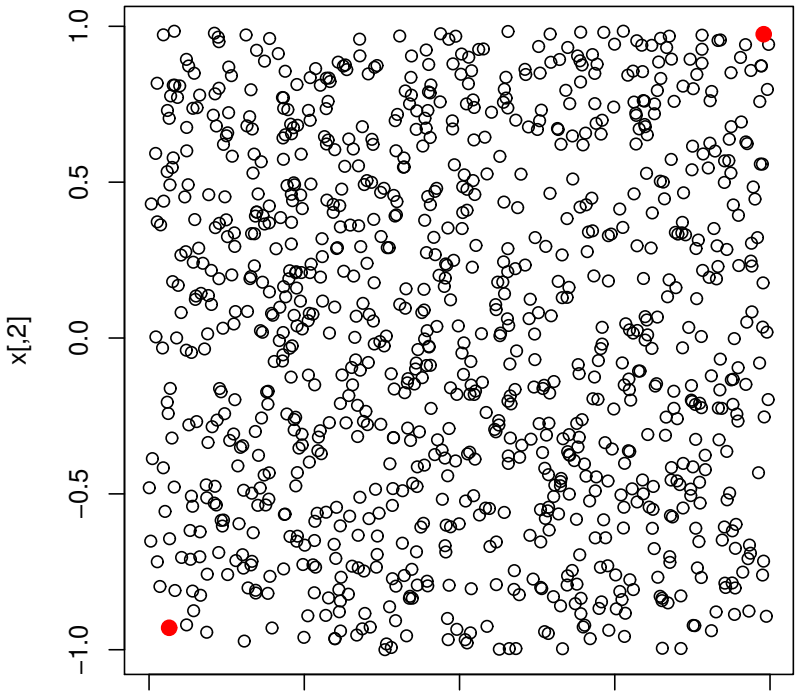


Algorithm

Step 2. Increase i by 1. For each $x \in X$, add the score

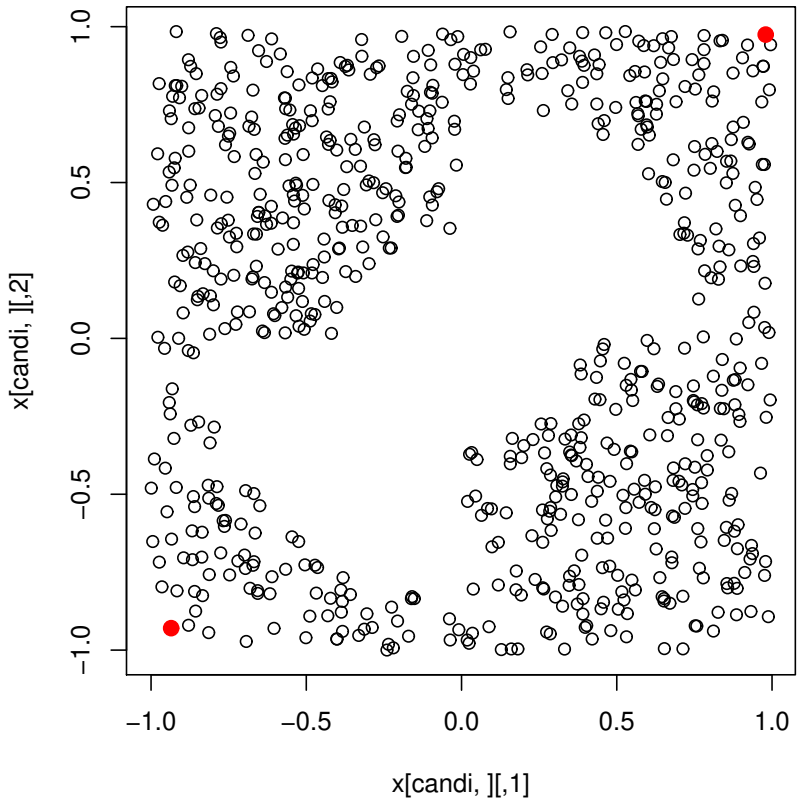
$$L(x, x_{i-1}^*) = \left[p - \|x\|^2/2 - \|x_{i-1}^*\|^2/2 + \delta(s(x), s(x_{i-1}^*))\right]^2$$

to the corresponding component in $\mathscr{D}$. Find x_i^* with the smallest component in $\mathscr{D}$ and add it to X_s.



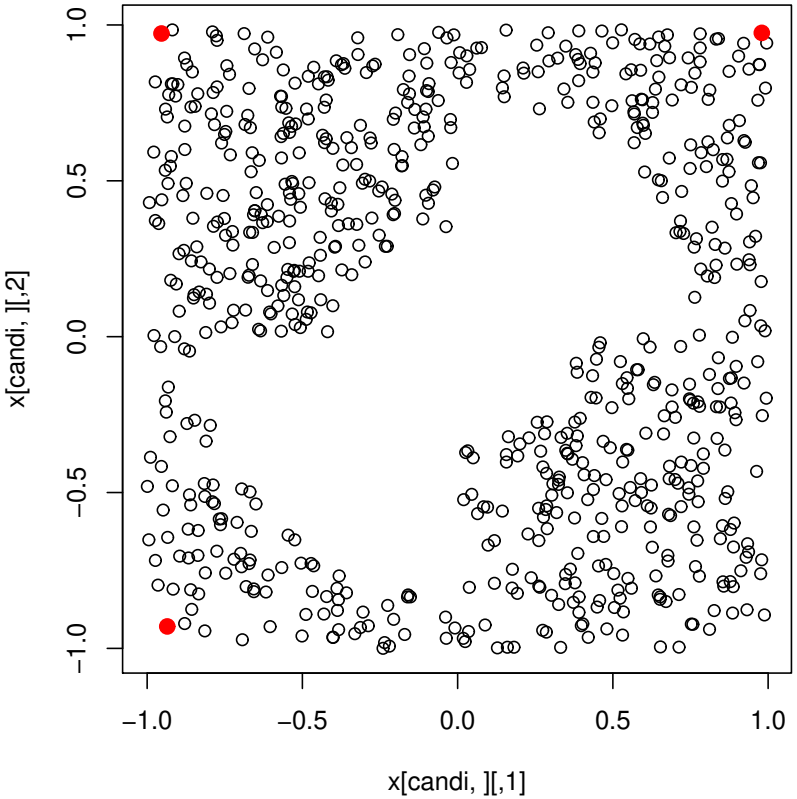
Algorithm

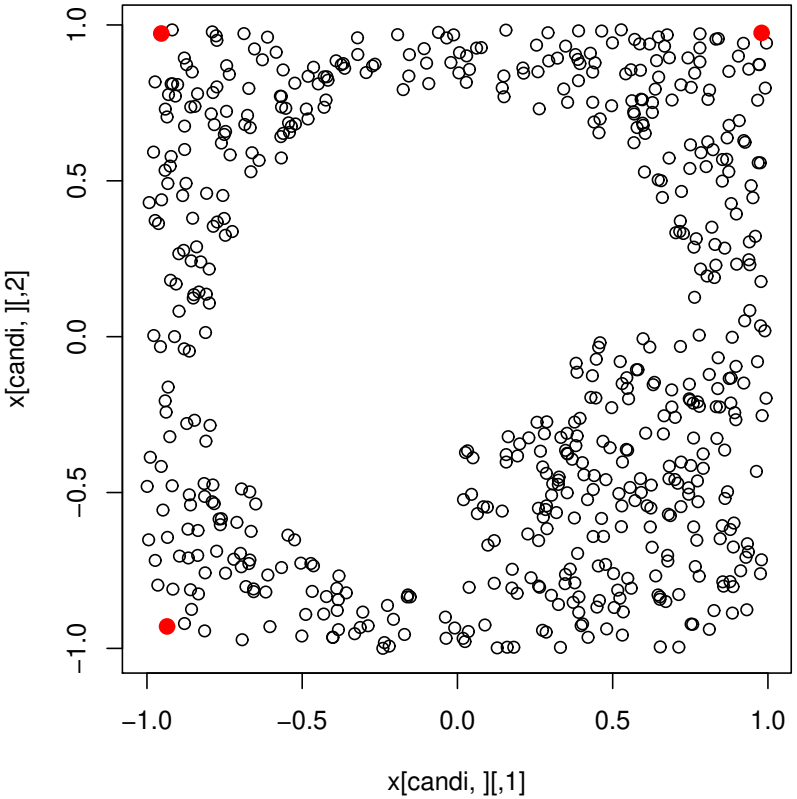
Step 3. Keep t_i points in X with t_i smallest components in $\mathscr{D}$. Remove x_i^* and other points from X as well as their corresponding components from $\mathscr{D}$.



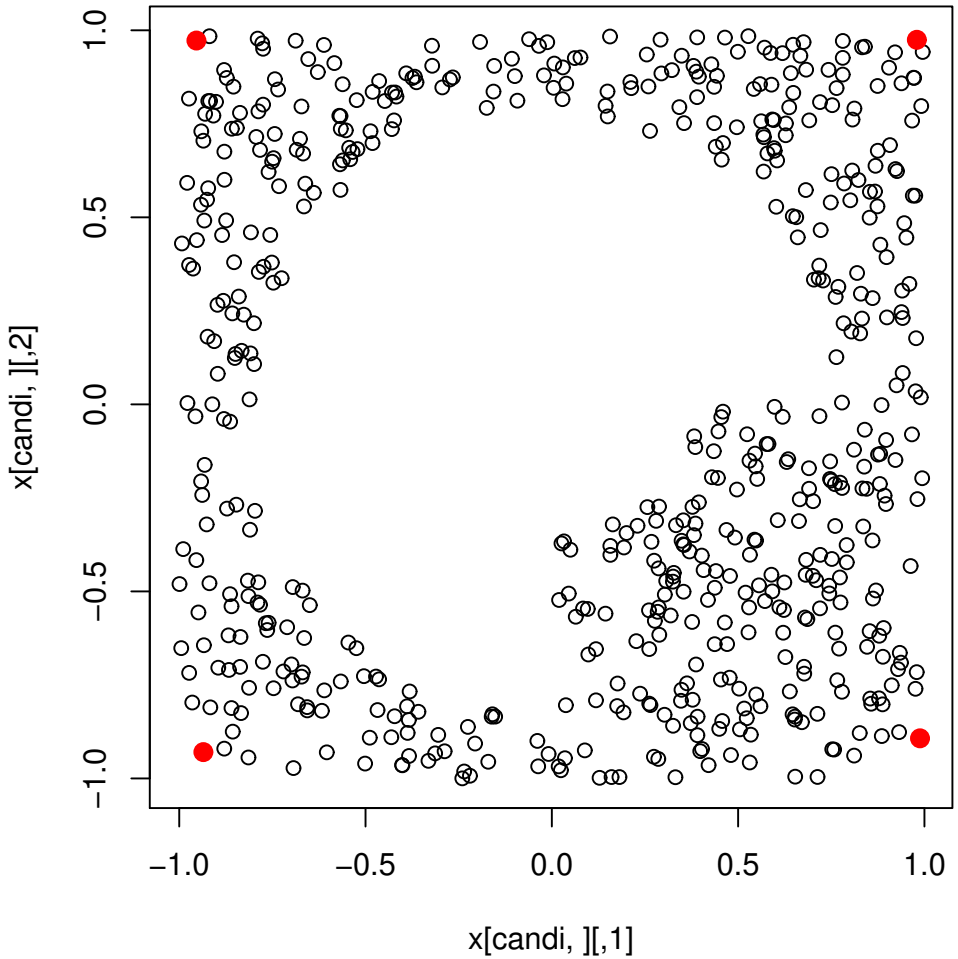
Algorithm

Step 4. Iterate Steps 2 and 3 until X_s contains k points.





Algorithm



Algorithm

- For t_i,
  - if $n \gg k$, say $n > k^2$, then we have far more candidate points than needed, eliminate a large proportion
  - otherwise, eliminate a small portion

$$r = \log(n/k), \quad t_i = \begin{cases} n/i, & \text{if } n \geq k^2; \\ n/i^{r-1}, & \text{if } n < k^2. \end{cases}$$

- When n is large, the complexity of the algorithm is $O(np\log(k))$

Parallel Computing

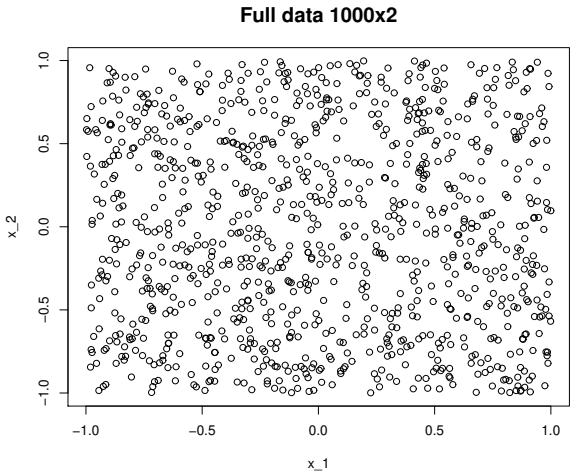
The combination of orthogonal arrays would still be an orthogonal array.

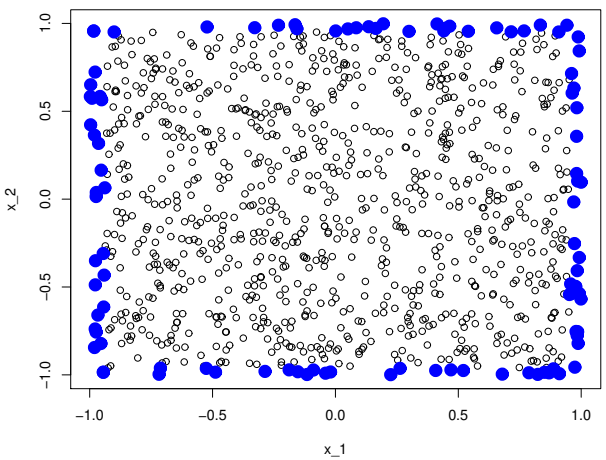
- The proposed algorithm is well suited for distributed and parallel computing.
- Separate the full dataset into g groups and apply the proposed algorithm simultaneously on the groups, from each of which k/g data points are selected.
- Combine all selected points would lead to a subsample of size k.

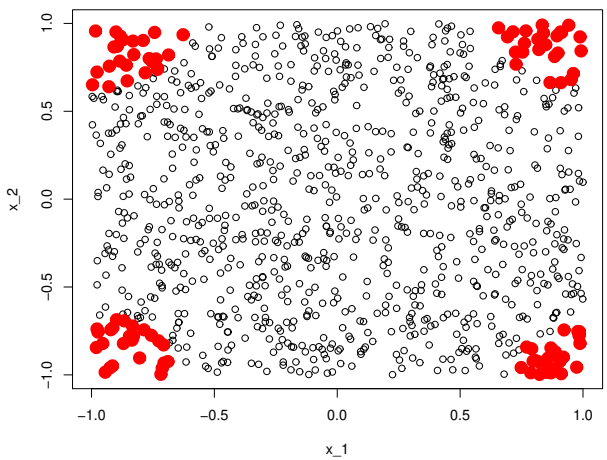
A Toy Example

Setup: $x_1, x_2 \overset{\text{iid}}{\sim} \text{Unif}[-1, 1]$, $n = 1000$, $p = 2$, $k = 100$, $\varepsilon \sim N(0, 1)$

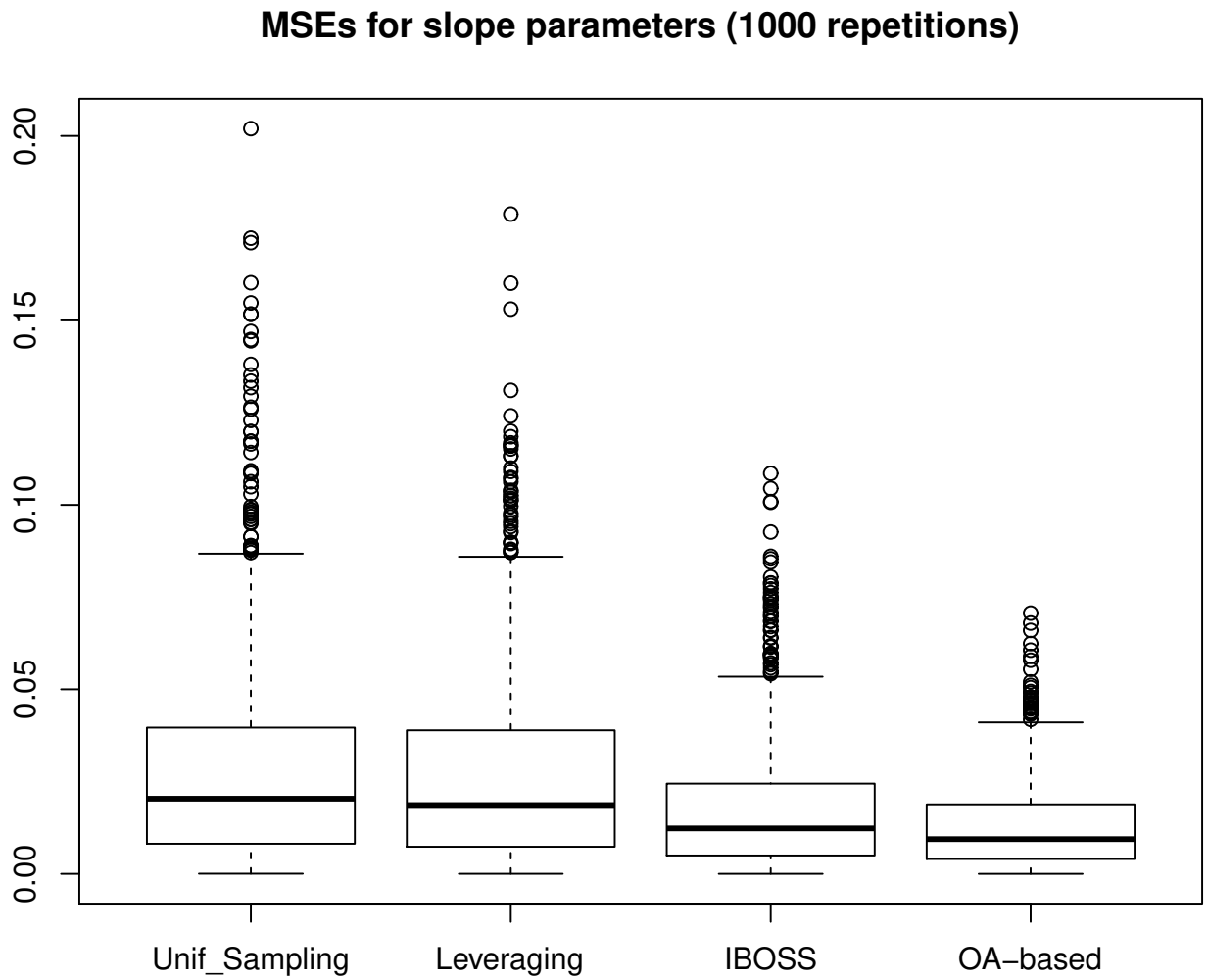
$$y = 1 + 2x_1 + 2x_2 + \varepsilon$$







Comparison


MSEs for slope parameters (1000 repetitions)
0.00
0.05
0.10
0.15
0.20
Unif_Sampling
Leveraging
IBOSS
OA−based

Possible Outliers

- May select outliers
- The proposed algorithm can be combined with outlier diagnostic methods for better performance
- The selected subsample follows the same underlying regression model as the full data, so outliers are very likely to be identified in the subsample if they exist in the full data (as long as the subsample size is not too small).

Simulation

Covariates are generated according to the following scenarios.

Case 1. x_i's are independent and have a multivariate uniform distribution with all covariates independent.

Case 2. x_i's have a multivariate normal distribution: $x_i \sim N(0, \Sigma)$, with

$$\Sigma = \left(0.5^{1-\delta(i,j)}\right).$$

- The y is generated from the linear model $y = X\beta + \epsilon$.
- True value of β is a 51 dimensional vector of unity. An intercept is included so $p = 50$.
- $\sigma^2 = 9$
- The simulation is repeated $T = 1000$ times

Simulation

Compare three subsampling approaches: uniform subsampling (Unif), IBOSS, and orthogonal subsampling (OSS).

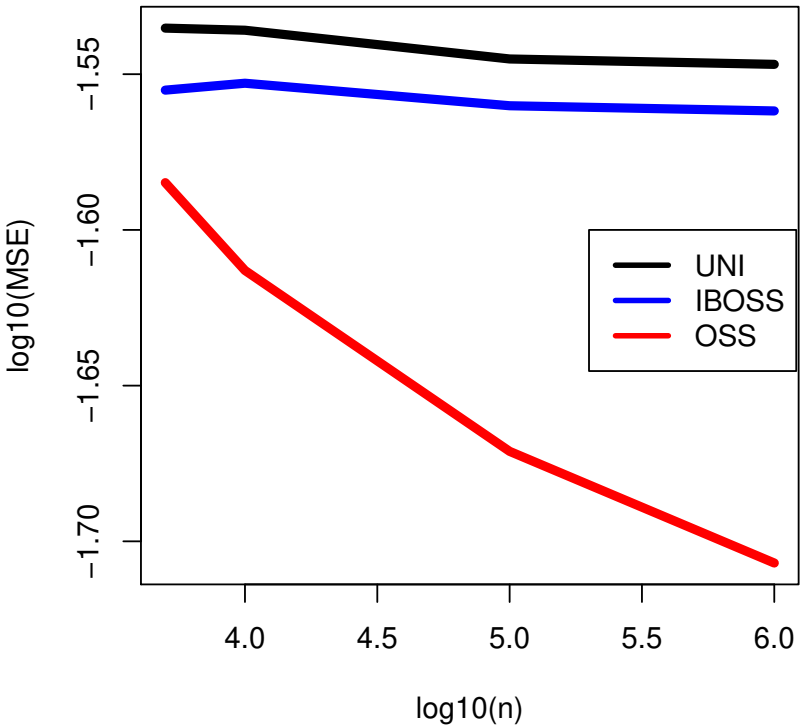
$$\mathrm{MSE}_{\beta_{-0}} = T^{-1} \sum_{t=1}^{T} \|\hat{\beta}_{-0}^{(t)} - \beta_{-0}\|^2,$$

$$\hat{\beta}_0 = \bar{y} - \bar{x}^T \hat{\beta}_{-0},$$

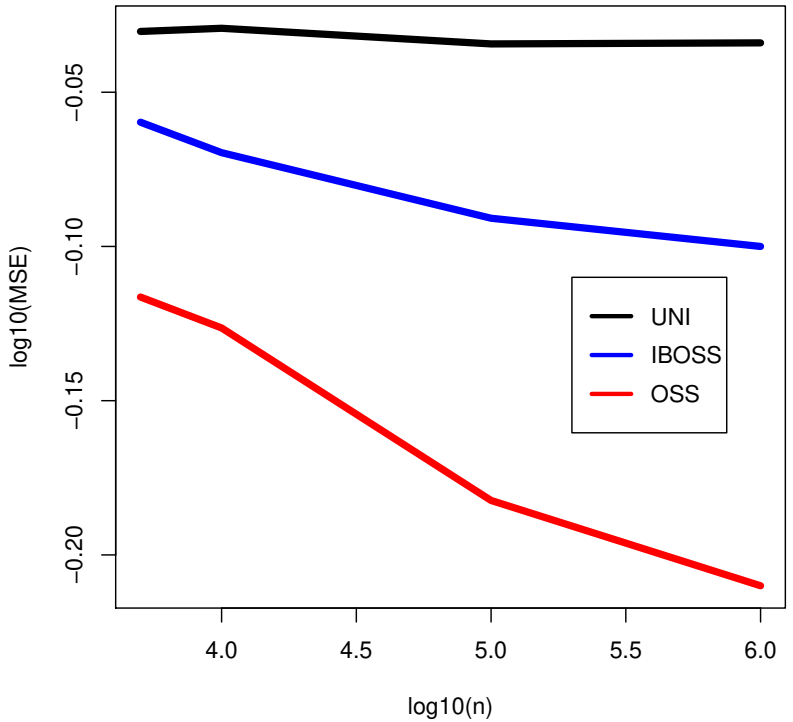
$$\mathrm{MSE}_{\beta_0} = T^{-1} \sum_{t=1}^{T} \|\hat{\beta}_{0}^{(t)} - \beta_{0}\|^2,$$

Results

MSEs of the slope parameters with $p = 50, k = 1000$



case 1: uniform

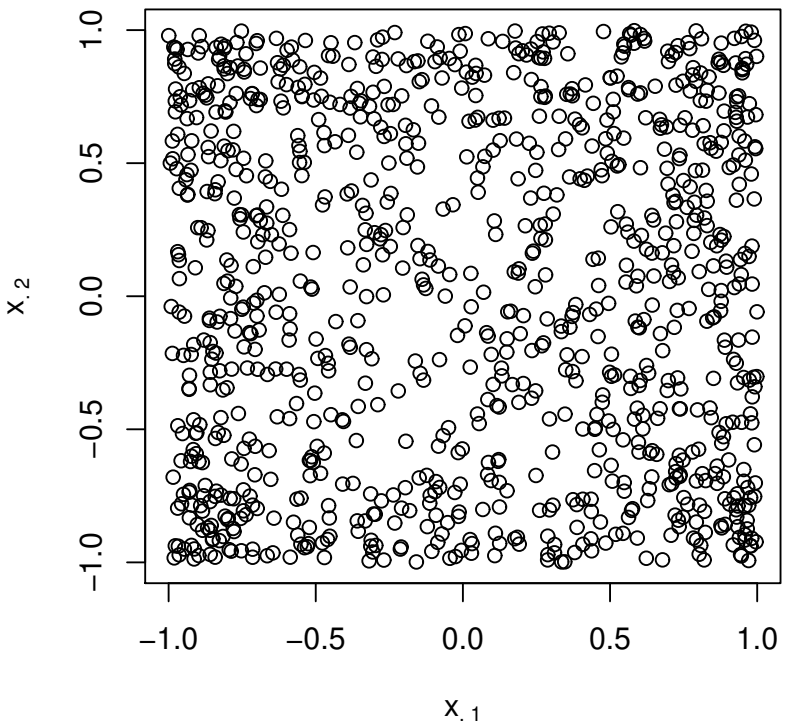


case 2: normal

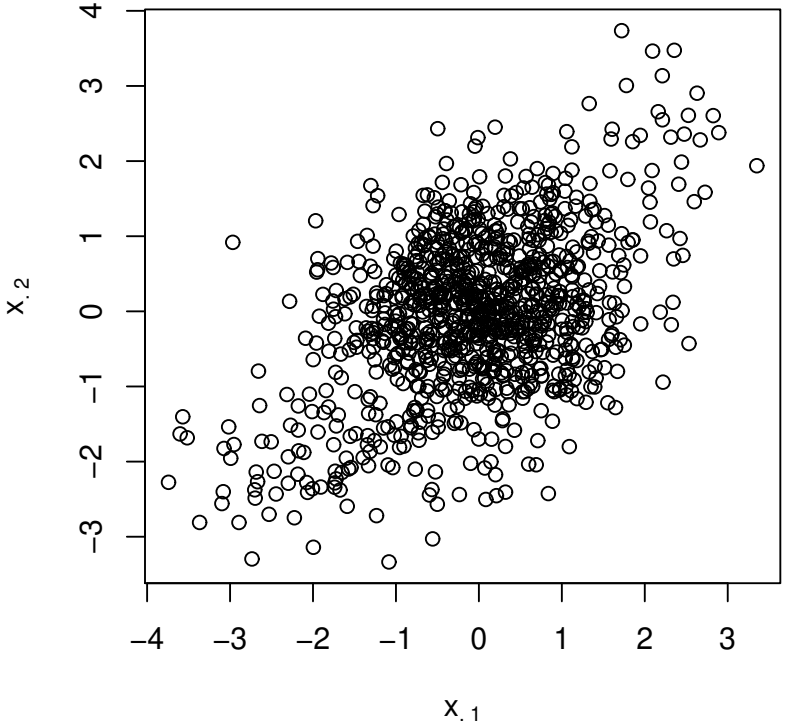
Results

Two dimensional projection plots of the subsamples

$(p = 50, k = 1000)$



case 1: uniform



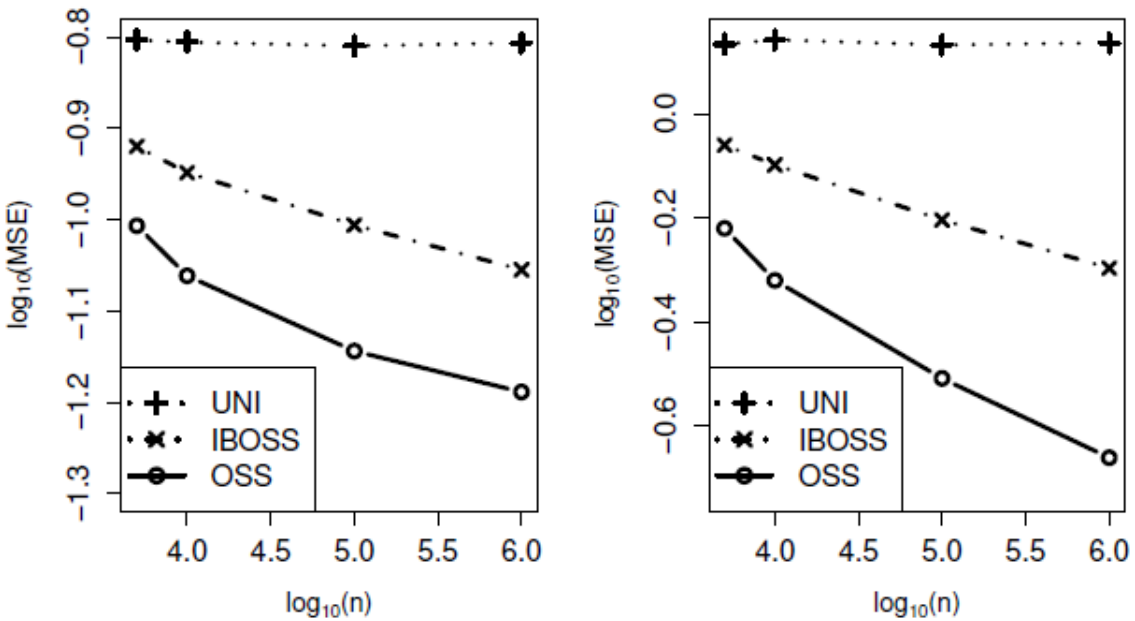
case 2: normal

Model misspecification: Robust to Interactions

The true model is

$$y = \beta_0 + X\tilde{\beta}_1 + X_I\tilde{\beta}_2 + \varepsilon,$$

where $p = 10, X : n \times p, X_I : n \times \binom{p}{2}$, but we still use a first-order model $y = \beta_0 + X\beta_1 + \varepsilon$ to select a subsample with $k = 1000$.



$\text{MSE}(\tilde{\beta}_1)$ $\text{MSE}(\tilde{\beta}_2)$

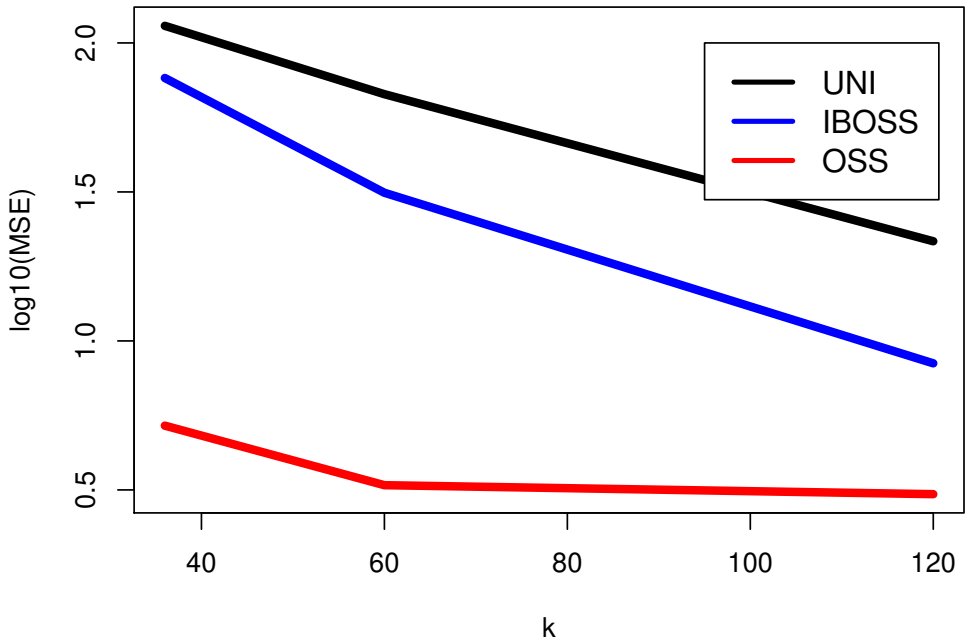
Running times (in seconds)

(a) $n = 10^6$, $k = 4p$, modeling without interactions				
p	UNI	OSS	IBOSS	FULL
50	0.004	1.243	2.022	4.704
100	0.009	2.086	4.270	14.310
500	0.359	13.536	24.220	—
(b) $p = 50$, $k = 1000$, modeling without interactions				
n	UNI	OSS	IBOSS	FULL
10^5	0.007	0.223	0.151	0.392
10^6	0.013	1.264	2.014	4.695
10^7	2.213	17.641	23.660	64.468
(c) $n = 10^6$, $k = 1000$, modeling with interactions				
p	UNI	OSS	IBOSS	FULL
10	0.008	0.448	2.263	4.871
20	0.043	0.694	9.426	52.421
30	0.136	1.030	21.476	—

Individual Household Electric Power Consumption Data

To model the metering: $p = 6, n = 2,075,259$

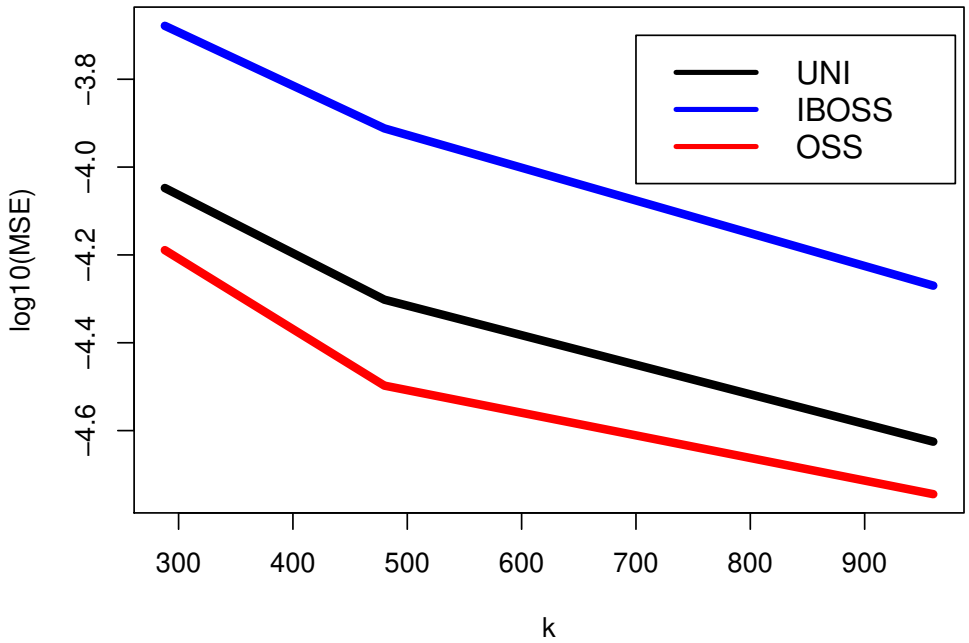
Bootstrap MSE for $k = 4p, 6p, 10p, 20p$ and 100 bootstrap samples



Wave Energy Converters Data

$p = 48, n = 288,000$, y is the total power output of the farm

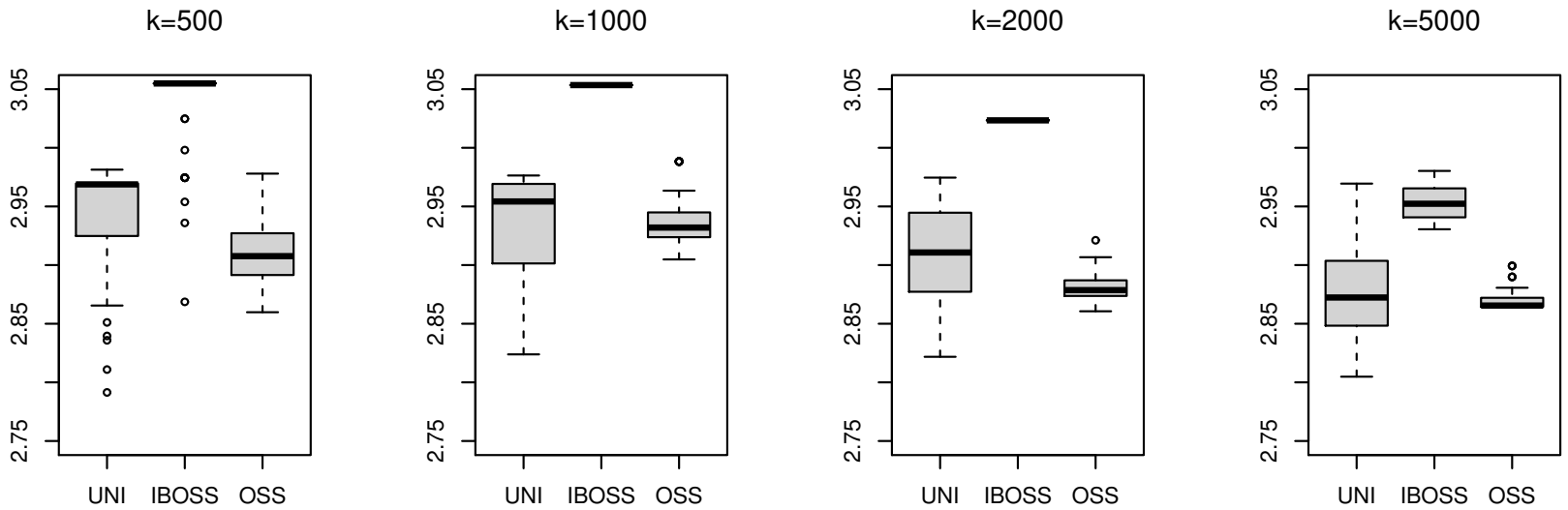
Bootstrap MSE for $k = 4p, 6p, 10p, 20p$ and 100 bootstrap samples



Blog Feedback Data

$p = 280, n = 52,397$, y is the number of comments for a blog post in the upcoming 24 hours

A test dataset with 7,624 data points is provided. We consider the prediction for the test data with Lasso regression.



Summary

- An orthogonal subsampling framework which is optimal in minimizing the average variance of estimators and the variance of predications
- Save running time with the proposed subsampling method
- Tackle the issue of storage capacity of individual's computing resources
- Suitable for the measurement-constrained regression
- Given some existing work, subsampling for big data is still in its infancy stage, and starting with linear models is a necessary first step. Subsampling for other goals and models is interesting and are extensively studied.

Reference: Wang, L., Elmstedt, J., Wong, W. K., and Xu, H. (2021). Orthogonal Subsampling for Big Data Linear Regression. *Annals of Applied Statistics*, 15(3): 1273-1290.