Information-Based Optimal Subdata Selection

John Stufken

CIRM, Design of Experiments: New Challenges
April 30-May 4, 2018

ARIZONA STATE UNIVERSITY
SCHOOL OF **MATHEMATICAL AND STATISTICAL SCIENCES**

Main Reference

Based on "Information-Based Optimal Subdata Selection for Big Data Linear Regression", to appear in Journal of the American Statistical Association (JASA), with

Haiying Wang, U of Connecticut

Min Yang, U of Illinois Chicago

Outline

Motivation

Big Data Challenge

For "Big Data", how can we extract useful information under time and computational constraints?

The data size n and dimension p can both be very large

For us, $n \gg p$. For example, n may be on the order of a billion and p may be over a thousand (Raskutti and Mahoney, 2014).

Data reduction can be critical in such situations because:

- analyzing the full data may be computationally unfeasible
- a laptop or desktop may be all that is available
- storing all of the data may not be possible

Data reduction refers to using only some of the data points (subdata)

Data Reduction for Linear Regression

Goal: Select subdata consisting of k cases, $k <<< n$, and analyze the subdata

- What should the subdata size k be?
- How to select subdata of size k?

We focus primarily on the second question for given k

In the JASA paper, for the linear regression setting and small p, we propose a deterministic method for subdata selection, called Information-Based Optimal Subdata Selection (IBOSS)

Competing subsampling-based methods, such as uniform sampling (UNIF) and leveraged sampling (LEV), were developed earlier

Basic Setup

Linear regression model:

$$y_i = \beta_0 + \mathbf{z}_i^T \boldsymbol{\beta}_1 + \epsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\mathbf{z}_i$ is a $p \times 1$ covariate vector, and $\mathbf{x}_i = (1, \mathbf{z}_i^T)^T$.

Or

$$\mathbf{y} = \beta_0 \mathbf{1} + \mathbf{Z}\boldsymbol{\beta}_1 + \boldsymbol{\epsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Other assumptions: y_i's are uncorrelated given $\mathbf{Z}$; ϵ_i's have mean 0, variance σ^2

Subsampling-Based Methods

A subsampling method consists of

- selection probabilities π_i, $i = 1, \ldots, n$, $\sum_i \pi_i = 1$
- a weighted estimator $\tilde{\boldsymbol{\beta}} = \left(\sum_i \omega_i \eta_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \sum_i \omega_i \eta_i \mathbf{x}_i \mathbf{y}$, with weights ω_i (often $1/\pi_i$) and with η_i the number of times that the ith data point is selected

Uniform subsampling (UNI): $\pi_i = 1/n$, $\omega_i = 1$

Algorithmic leveraging (LEV): $\pi_i = h_{ii}/(p+1)$, $\omega_i = 1/\pi_i$, where $h_{ii} = \mathbf{x}_i^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{x}_i$; need to approximate the h_{ii}'s

Unweighted leveraging (LEVUNW): as LEV, but with $\omega_i = 1$

Shrinkage leveraging (SLEV): $\pi_i = \alpha h_{ii}/(p+1) + (1-\alpha)/n$ for some $\alpha \in [0, 1]$, $\omega_i = 1/\pi_i$ (Ma, Mahoney, Yu, 2015, JMLR)

IBOSS

IBOSS Approach

IBOSS: Select subdata judiciously to maximize the Fisher information matrix for the model parameters, in some sense

For linear regression, assuming normality and taking $\sigma^2 = 1$ for simplicity, the information matrix for β with subdata is

$$\mathbf{M}(\boldsymbol{\delta}) = \sum_{i=1}^{n} \delta_i x_i x_i^T = \mathbf{X}^T \boldsymbol{\Delta} \mathbf{X},$$

with δ_i an "inclusion" indicator, $\boldsymbol{\delta} = (\delta_1, ..., \delta_n)$ and $\boldsymbol{\Delta} = diag(\boldsymbol{\delta})$

Optimize this through a good choice for $\boldsymbol{\delta}$ subject to $\sum_i \delta_i = k$

Optimal Design of Experiments

As in optimal design of experiments, we aim to maximize a function of the information matrix.

D-optimality: Find $\boldsymbol{\delta}$, subject to $\sum_i \delta_i = k$, that maximizes $\det(\mathbf{M}(\boldsymbol{\delta}))$.

A difference with DOE is that we already have data, and are limited to a choice for the $\mathbf{z}_i$'s that appear in the data.

Another challenge is size: we need a computationally efficient algorithm to find, approximately, an optimal $\boldsymbol{\delta}$ (see next slide).

Algorithm for D-optimality

To maximize $\det(\mathbf{M}(\delta))$, include data points with large and small covariate values, equally distributed over the extremes

For a fixed subdata size k, using a partition-based selection algorithm, for $j = 1, ..., p$, select the $k/(2p)$ largest and smallest values for the jth regression variable, and include these data points in the subdata

Estimate β by $\hat{\beta}^D = (\mathbf{X}^T \mathbf{\Delta} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{\Delta} \mathbf{y}$

Computational complexity for selection of subdata is $O(np)$; overall $O(kp^2 + np)$, or $O(np)$ if $n > kp$. Better than LEV

Can select subdata one regression variable at a time (no duplication) or in parallel (possibly less than k data points due to duplication)

Theory

Theoretical Results

D-optimal IBOSS can be used no matter what the distribution of the covariates is ...

... but its performance is affected by it

Let $\mathbf{z}_1, ..., \mathbf{z}_n$ be iid, and consider 3 scenarios:

1. **Normal**, $\mathbf{z}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
2. **Lognormal**, $\mathbf{z}_i \sim LN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
3. **Multivariate t with ν df**, $\mathbf{z}_i \sim t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

For all scenarios, $Var(\hat{\beta}_0^D|\mathbf{Z})$ is proportional to $1/k$ when $n \to \infty$

But the story is different for $Var(\hat{\beta}_1^D|\mathbf{Z})$ ...

Theoretical Results

Elements of $Var(\hat{\beta}_1^D|\mathbf{Z})$ converge to 0 when $n \to \infty$ in all cases (even though the subdata size k is fixed)

For scenario 1 (normal), elements converge to 0 as $1/\log(n)$

For scenario 2 (lognormal), the element in position (j_1, j_2) converges to 0 as $\exp(-(\sigma_{j_1} + \sigma_{j_2})\sqrt{2\log(n)})$

For scenario 3 (t-distribution), elements converge to 0 as $n^{-2/\nu}$

Similar results typically do not hold for subsampling methods UNI, LEV

Simulations

Simulation setup

$p = 50$, $\beta = \mathbf{1}_{51\times 1}$, $\epsilon_i \sim N(0, \sigma^2)$ with $\sigma^2 = 9$, $\Sigma = (.5^{I(i\neq j)})$.

$\mathbf{z}_i$'s are generated from the following distributions.

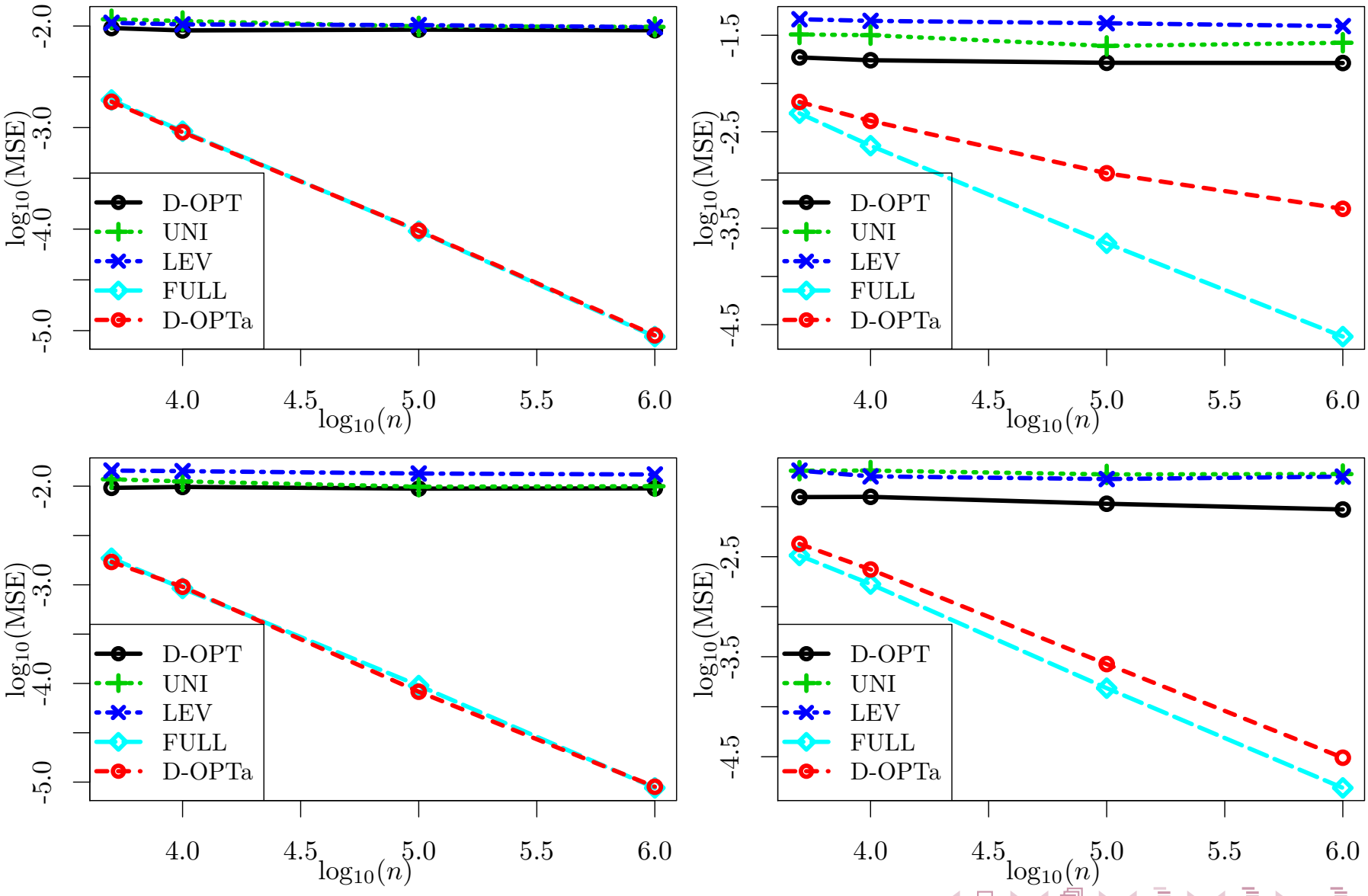
1. **Normal**, $\mathbf{z}_i \sim N(\mathbf{0}, \Sigma)$;
2. **Lognormal**, $\mathbf{z}_i \sim LN(\mathbf{0}, \Sigma)$;
3. **Multivariate t with 2 df**, $\mathbf{z}_i \sim t_2(\mathbf{0}, \Sigma)$;
4. **Mixture**, $\mathbf{z}_i$'s have a mixture distribution of $N(\mathbf{1}, \Sigma)$, $t_2(\mathbf{1}, \Sigma)$, $t_3(\mathbf{1}, \Sigma)$, Unif$[\mathbf{0}, \mathbf{2}]$ and $LN(\mathbf{0}, \Sigma)$ with equal proportions.

Each simulation was repeated $S = 1000$ times

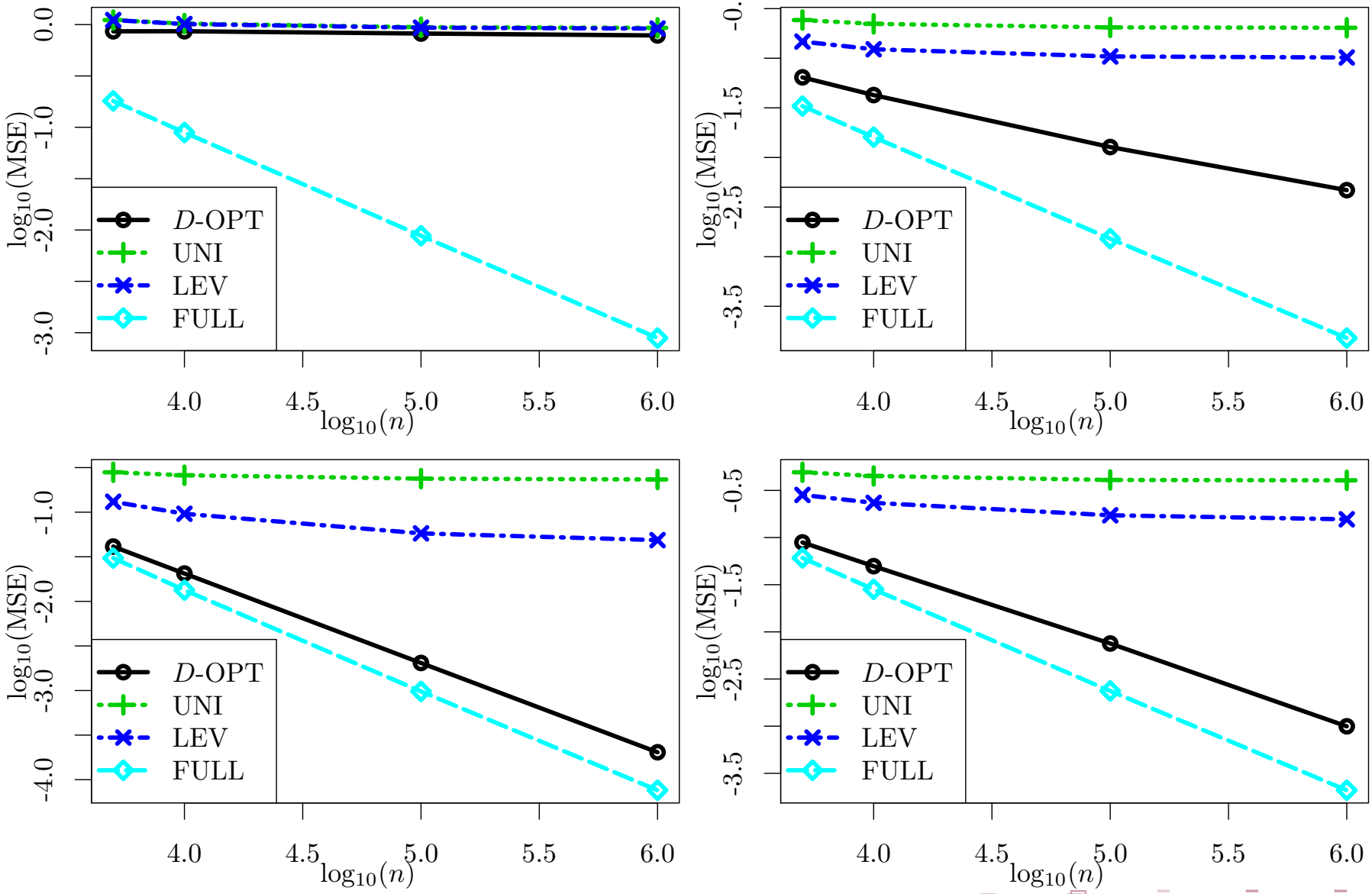
Empirical mean squared errors (MSE) are compared

Light blue = full data; black = IBOSS with D-optimality; green = uniform sampling; blue = leveraged sampling

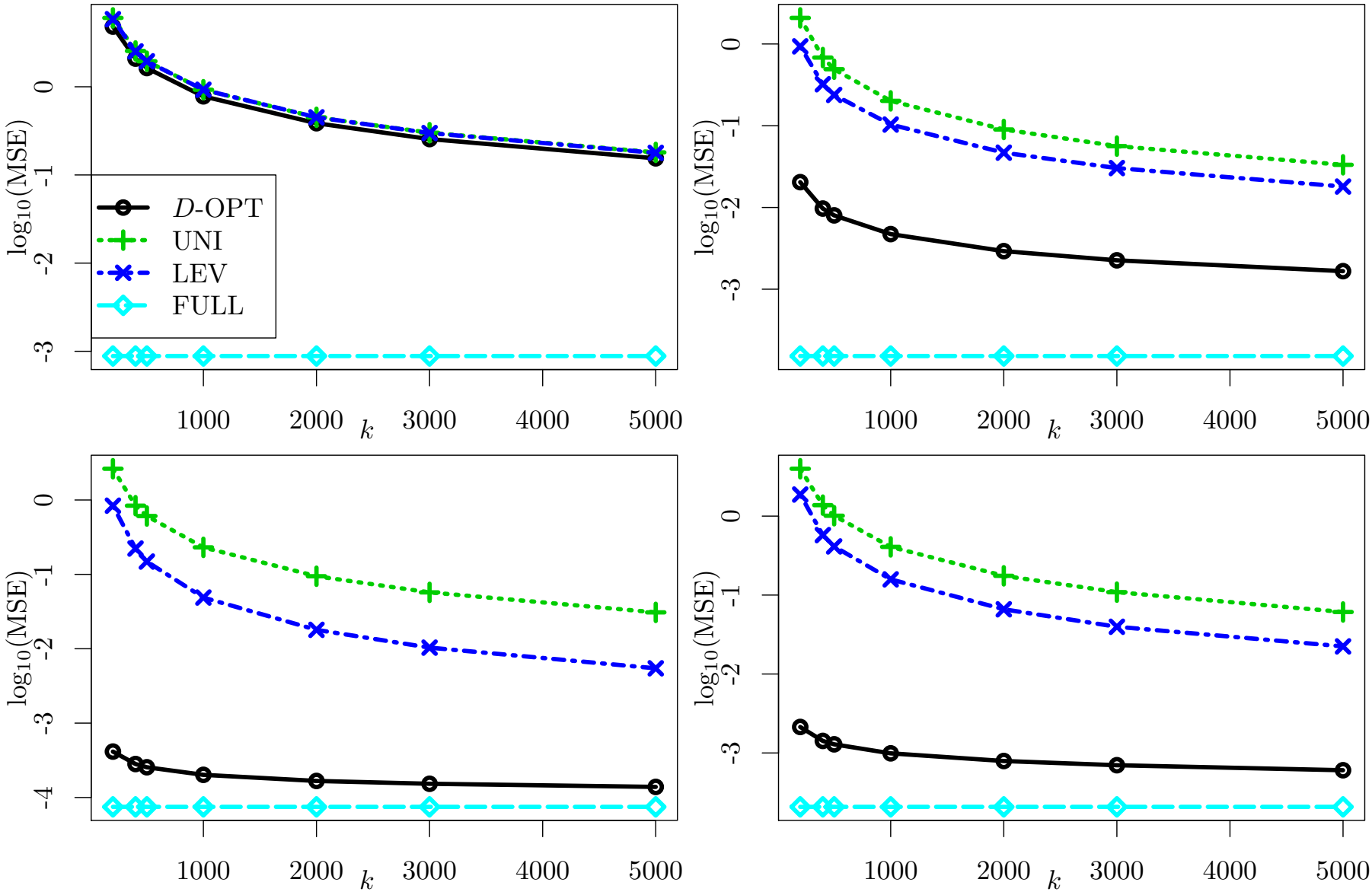
MSE of the intercept estimator with $k = 1000$



MSE of the slope estimators with $k = 1000$



MSE of the slope estimators with $n = 10^6$



CPU times for different n, p and $k = 1000$

Table: CPU times (seconds) for different n with $p = 500$

n	D-opt	UNI	LEV	FULL
5×10^3	1.19	0.33	0.88	1.44
5×10^4	1.36	0.29	2.20	13.39
5×10^5	8.89	0.31	21.23	132.04

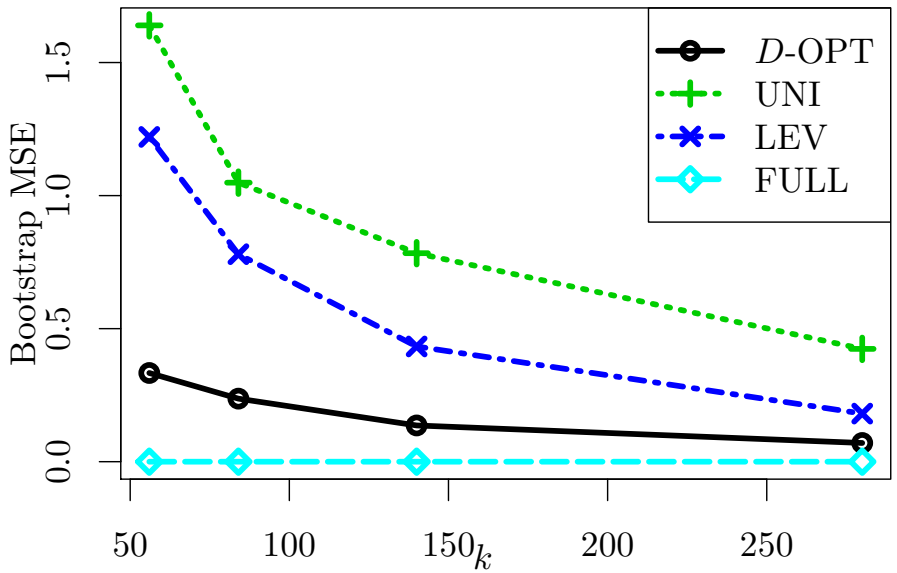
Table: CPU times (seconds) for different p with $n = 5 \times 10^5$

p	D-opt	UNI	LEV	FULL
10	0.19	0.00	1.94	0.21
100	1.74	0.02	4.66	6.55
500	9.30	0.31	21.94	132.47

Chemical Sensors Data

Chemical sensors data (Fonollosa et al., 2015), with $n = 4,188,261$ and $p = 14$

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



Discussion

Discussion

IBOSS works great for linear models with a modest number of covariates p. But ...

- ... develop a better algorithm for the D-optimal IBOSS approach
- ... consider other optimization goals (prediction; other criteria) and corresponding algorithms
- ... consideration of independent categorical variables
- ... combine IBOSS with variable selection methods if p is large
- ... consideration of outliers
- ... model inadequacy or other models (interaction terms, pure quadratic terms, heteroscedastic errors, nonlinear model, dependencies)
- ... nonparametric approach

THANK YOU