Subdata selection for big data regression: an improved approach

Vasilis Chasiotis

Athens University of Economics and Business

School of Information Sciences and Technology

Department of Statistics

chasiotisv@aueb.gr

February 25, 2022

Table of contents

1. Introduction
   - Motivation example
2. Theoretical considerations
   - Generalized variance
3. Existing approaches
4. The new proposed algorithms
   - Algorithm 1
   - Variation of Algorithm 1
5. Simulation experiments
   - Evaluation of algorithms
   - Execution time of algorithms
   - About iterations of Alg1
6. Real data applications
   - Power consumption data
   - Chemical sensors data
7. Discussion
8. References

Introduction

- Big data analysis
- Computational power versus data volume
- Data reduction - Keep the most informative data points
- Random selection (Drineas *et al.*, 2011)
- Concept of optimal designs - IBOSS approach (Wang *et al.*, 2019)
- Orthogonal subsampling - OSS approach (Wang *et al.*, 2021)
- Orthogonal array (Ren and Zhao, 2021)

Which data points should one select?

- ◇ Covariates (p): 2
- ◇ Full data (n): 50
- ◇ Subdata size (k): 8

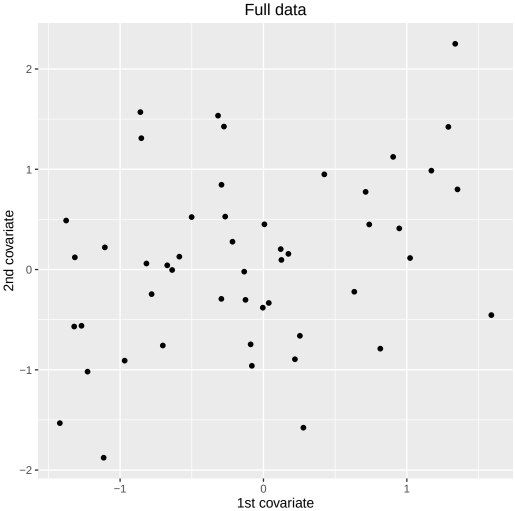


Figure 1: Full data between two covariates.

Motivation example

- ◇ Covariates (p): 2
- ◇ Full data (n): 50
- ◇ Subdata size (k): 8

Selection of data points with large convex hull

↓

Selected data points can have a large volume

↓

Maximize the determinant of the information matrix

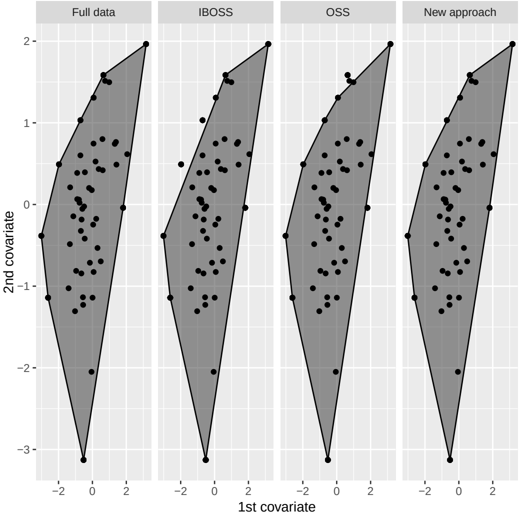


Figure 2: An example for the different approaches.

Theoretical considerations

◇ Linear regression model: $y_i = \beta_0 + \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}_1 + \epsilon_i, \quad i = 1, 2, \ldots, n$

◇ Covariate vectors: $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\mathsf{T}, \quad i = 1, 2, \ldots, n$

◇ Unknown parameters: $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^\mathsf{T})^\mathsf{T}, \boldsymbol{\beta}_1 = (\beta_1, \beta_2, \ldots, \beta_p)^\mathsf{T}$

- Under full data

$$\hat{\boldsymbol{\beta}}_{\text{Full}} = \left(\sum_{i=1}^{n} \mathbf{z}_i\mathbf{z}_i^\mathsf{T}\right)^{-1} \sum_{i=1}^{n} \mathbf{z}_i y_i, \quad \mathbf{z}_i = (1, \mathbf{x}_i^\mathsf{T})^\mathsf{T}$$

$$\mathbf{Q}_{\text{Full}} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{z}_i\mathbf{z}_i^\mathsf{T}$$

- Under subdata

$$\hat{\boldsymbol{\beta}}_{\text{Sub}} = \left(\sum_{i=1}^{n} \delta_i\mathbf{z}_i\mathbf{z}_i^\mathsf{T}\right)^{-1} \sum_{i=1}^{n} \delta_i\mathbf{z}_i y_i$$

$$\mathbf{Q}_{\text{Sub}} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \delta_i\mathbf{z}_i\mathbf{z}_i^\mathsf{T}$$

$$\delta_i = \begin{cases} 1, & \text{if } (\mathbf{x}_i, y_i) \text{ is included} \\ 0, & \text{if } (\mathbf{x}_i, y_i) \text{ is not included} \end{cases}, \ \sum_{i=1}^{n} \delta_i = k$$

Generalized variance

◇ $\mathbf{x}_j^*$: jth covariate under the selected subdata

$$\det\left(\mathbf{Q}_{\text{Sub}}\right) = \frac{k^{p+1}}{\sigma^{2(p+1)}} \det\left(\begin{bmatrix} s^2_{\mathbf{x}_1^*} & s_{\mathbf{x}_1^*\mathbf{x}_2^*} & \cdots & s_{\mathbf{x}_1^*\mathbf{x}_p^*} \\ s_{\mathbf{x}_1^*\mathbf{x}_2^*} & s^2_{\mathbf{x}_2^*} & \cdots & s_{\mathbf{x}_2^*\mathbf{x}_p^*} \\ \vdots & \vdots & \ddots & \vdots \\ s_{\mathbf{x}_1^*\mathbf{x}_p^*} & s_{\mathbf{x}_2^*\mathbf{x}_p^*} & \cdots & s^2_{\mathbf{x}_p^*} \end{bmatrix}\right)$$

$$\bar{\mathbf{x}}_j^* = \frac{\sum_{i=1}^n \delta_i x_{ij}}{k},\ s^2_{\mathbf{x}_j^*} = \frac{\sum_{i=1}^n \delta_i (x_{ij} - \bar{\mathbf{x}}_j^*)^2}{k},\ s_{\mathbf{x}_j^*\mathbf{x}_t^*} = \frac{\sum_{i=1}^n \delta_i x_{ij} x_{it}}{k} - \bar{\mathbf{x}}_j^* \bar{\mathbf{x}}_t^*$$

↓

Cholesky decomposition

↓

Theorem 1. The generalized variance of covariates under the subdata is maximized by the selection of data points for which $s^2_{\mathbf{x}_j^*}$ is maximized for any $j = 1, 2, \ldots, p$, and $s_{\mathbf{x}_o^*\mathbf{x}_j^*} = 0$ for any $j > o = 1, 2, \ldots, j-1$, simultaneously.

Existing approaches

- The information-based optimal subdata selection (IBOSS) approach

  ◇ D-optimality

  ◇ Selection of data points with the smallest and largest values of all covariates sequentially

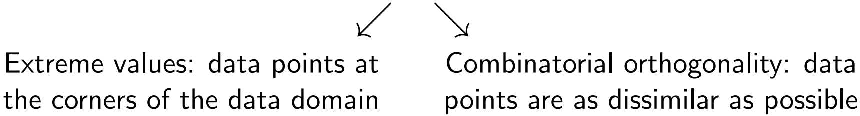
- The orthogonal subsampling (OSS) approach

  ◇ A two-level OA represents an optimal design for linear regression

  ◇ D- and A-optimality

  ◇ All covariates are scaled to $[-1, 1]$

  ◇ Elimination algortihm based on a discrepancy function



Extreme values: data points at the corners of the data domain

Combinatorial orthogonality: data points are as dissimilar as possible

Existing approaches

- The information-based optimal subdata selection (IBOSS) approach

  ◇ D-optimality

  ◇ Selection of data points with the smallest and largest values of all covariates sequentially

- The orthogonal subsampling (OSS) approach

  ◇ A two-level OA represents an optimal design for linear regression

  ◇ D- and A-optimality

  ◇ All covariates are scaled to $[-1, 1]$

  ◇ Elimination algortihm based on a discrepancy function

$\mathrm{OA}(4,3,2,2)$		
0	0	0
0	1	1
1	0	1
1	1	0

↙ ↘

Extreme values: data points at the corners of the data domain

Combinatorial orthogonality: data points are as dissimilar as possible

Alg1

Algorithm 1 Alg1

Input: subdata $\mathbf{S} = (\mathbf{s}_i), i = 1, 2, \ldots, k$ of the OSS approach, initial full data $\mathbf{D}_{\text{Full}}$, subdata size k, candidate data points K from each covariate

Output: new obtained subdata $\mathbf{S}$

```
Step 1: Preperation
$\mathbf{S} = \text{convert}(\mathbf{S})$    ▷ convert subdata $\mathbf{S}$ to their initial values
$V = \det(\mathbf{Q}_{\text{Sub}})$    ▷ generalized variance of $\mathbf{S}$
$\mathbf{D} = \mathbf{D}_{\text{Full}} - \mathbf{S} = (d_{rj})$    ▷ remaining data points $\mathbf{d}_{r\cdot} = (d_{r1}, \ldots, d_{rp}) \notin \mathbf{S}$
$N_{\text{F}} = \text{nrow}(\mathbf{D})$    ▷ number of data points $\mathbf{d}_{r\cdot} \in \mathbf{D}$
$\mathbf{F} = \emptyset$    ▷ initialize the index set of candidate data points
Step 2: Find candidate data points
for $j$ in $1, \ldots, p$ do
    $\mathbf{d}_{\cdot j} = \text{sort}(\mathbf{d}_{\cdot j})$    ▷ sort $\mathbf{d}_{\cdot j} = (d_{1j}, \ldots, d_{N_{\text{F}}j})$
    $\mathbf{D} = \text{sort}(\mathbf{D})$    ▷ sort $\mathbf{D}$ based on $\mathbf{d}_{\cdot j}$
    $\mathbf{F} = \mathbf{F} \cup \mathbf{d}_{1\cdot} \cup \cdots \cup \mathbf{d}_{K/2\cdot}$
    $\mathbf{F} = \mathbf{F} \cup \mathbf{d}_{N_{\text{F}}-K/2+1\cdot} \cup \cdots \cup \mathbf{d}_{N_{\text{F}}\cdot}$
end for
$\mathbf{F} = \text{unique}(\mathbf{F})$    ▷ keep unique data points of $\mathbf{F} = (\mathbf{f}_w)$
$N_{\text{F}} = \text{nrow}(\mathbf{F})$    ▷ number of data points $\mathbf{f}_w \in \mathbf{F}$
Step 3: Main algorithm
for $i$ in $1, \ldots, k$ do
    for $w$ in $1, \ldots, N_{\text{F}}$ do
        $\mathbf{s}_i \leftrightarrow \mathbf{f}_w$    ▷ interchange data points $\mathbf{s}_i$ and $\mathbf{f}_w$
        $V_{\text{new}} = \det(\mathbf{Q}_{\text{Sub}})$    ▷ generalized variance of new $\mathbf{S}$
        if $V_{\text{new}} > V$ then
            $V = V_{\text{new}}$
            break
        else
            $\mathbf{s}_i \leftrightarrow \mathbf{f}_w$
        end if
    end for
end for
return $\mathbf{S}$
```

VAlg1

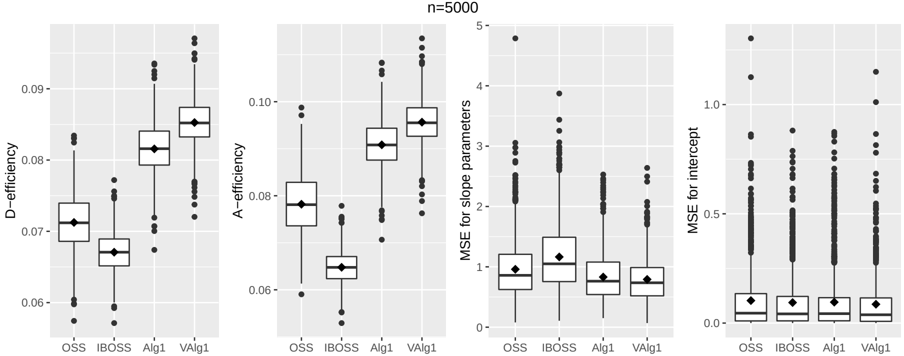
◇ Change as to which $\mathbf{f}_w$ is interchanged with $\mathbf{s}_i$

Algorithm 2 VAlg1

```
Steps 1 and 2:  Same as in Alg1
Step 3:  Main algorithm
for $i$ in $1, \dots, k$ do
    for $w$ in $1, \dots, N_{\mathrm{F}}$ do
        $\mathbf{s}_i \leftrightarrow \mathbf{f}_w$                ▷ interchange data points $\mathbf{s}_i$ and $\mathbf{f}_w$
        $V_{\mathrm{new}} = \det(\mathbf{Q}_{\mathrm{Sub}})$      ▷ generalized variance of new $\mathbf{S}$
        if $V_{\mathrm{new}} > V$ then
            $V = V_{\mathrm{new}}$
        else
            $\mathbf{s}_i \leftrightarrow \mathbf{f}_w$
        end if
    end for
end for
return $\mathbf{S}$
```

Evaluation of algorithms (1)

- ◇ $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{\Sigma})$
- ◇ Covariance matrix: $\mathbf{\Sigma} = (\Sigma_{ij}),\ i,j = 1, 2, \ldots, p$
- ◇ $\Sigma_{ij} = 1,\ i = j$ and $\Sigma_{ij} = 0.5,\ i \neq j$
- ◇ $k = 100,\ K = 25,\ p = 10,\ \boldsymbol{\beta} = (1, 1, \ldots, 1)^{\mathrm{T}},\ \sigma^2 = 3$
- ◇ 1000 simulations
- ◇ Alg1: 5 iterations - VAlg1: 1 iteration



Evaluation of algorithms (2)

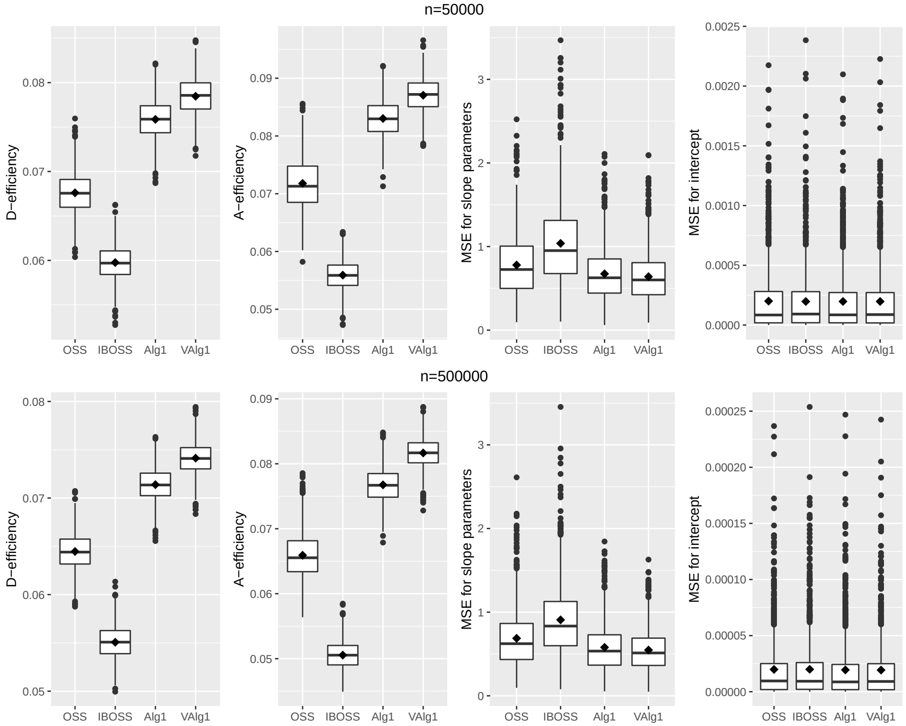


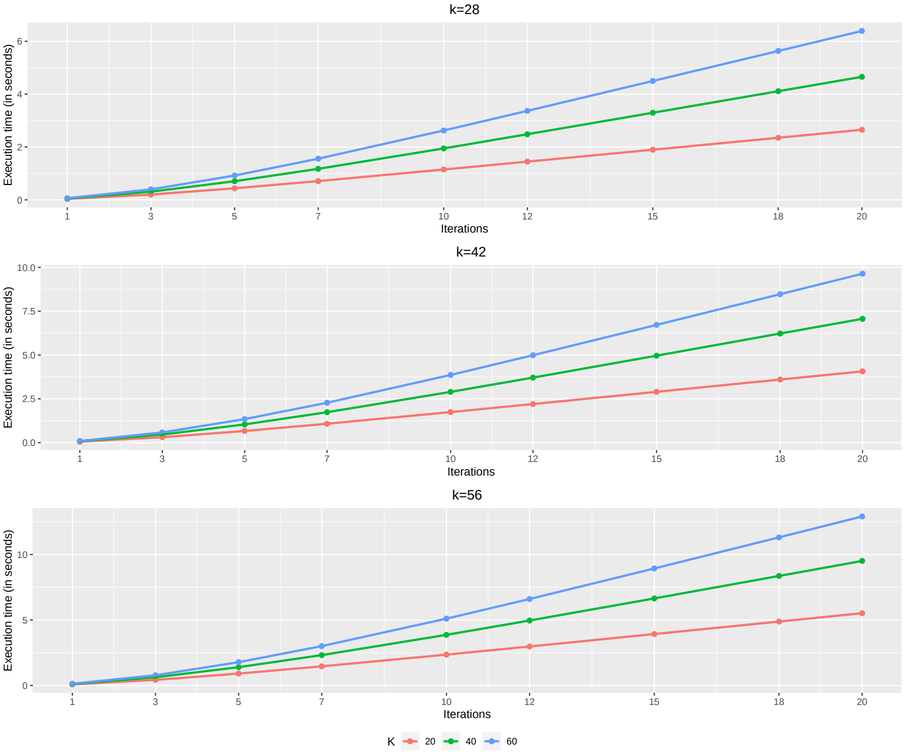
Figure 3: The MSEs, D- and A-efficiencies for the subdata selected by different approaches.

Execution time of Alg1

◇ $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{\Sigma})$

◇ $\mathbf{\Sigma} = (\Sigma_{ij}),\ i, j = 1, 2, \ldots, p,\ \Sigma_{ij} = 1,\ i = j$ and $\Sigma_{ij} = 0.5,\ i \neq j$

◇ $n = 1000,\ p = 7$, 500 simulations



k=28

k=42

k=56

Figure 4: The mean execution time (in seconds) of Alg1.

Execution time of VAlg1

- ◇ $\mathbf{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$
- ◇ $\boldsymbol{\Sigma} = (\Sigma_{ij}),\ i, j = 1, 2, \ldots, p,\ \Sigma_{ij} = 1,\ i = j$ and $\Sigma_{ij} = 0.5,\ i \neq j$
- ◇ $n = 1000,\ p = 7$
- ◇ 500 simulations
- ◇ 1 iteration

k	28	28	28	42	42	42	56	56	56
K	20	40	60	20	40	60	20	40	60
Time	0.2244	0.39534	0.5611	0.3361	0.5989	0.8321	0.4593	0.8226	1.1368

Table 1: The mean execution time (in seconds) of VAlg1.

About iterations of Alg1

◇ $\mathbf{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$

◇ $\boldsymbol{\Sigma} = (\Sigma_{ij}),\ i, j = 1, 2, \ldots, p,\ \Sigma_{ij} = 1,\ i = j$ and $\Sigma_{ij} = 0.5,\ i \neq j$

◇ $n = 1000,\ p = 7$, 500 simulations

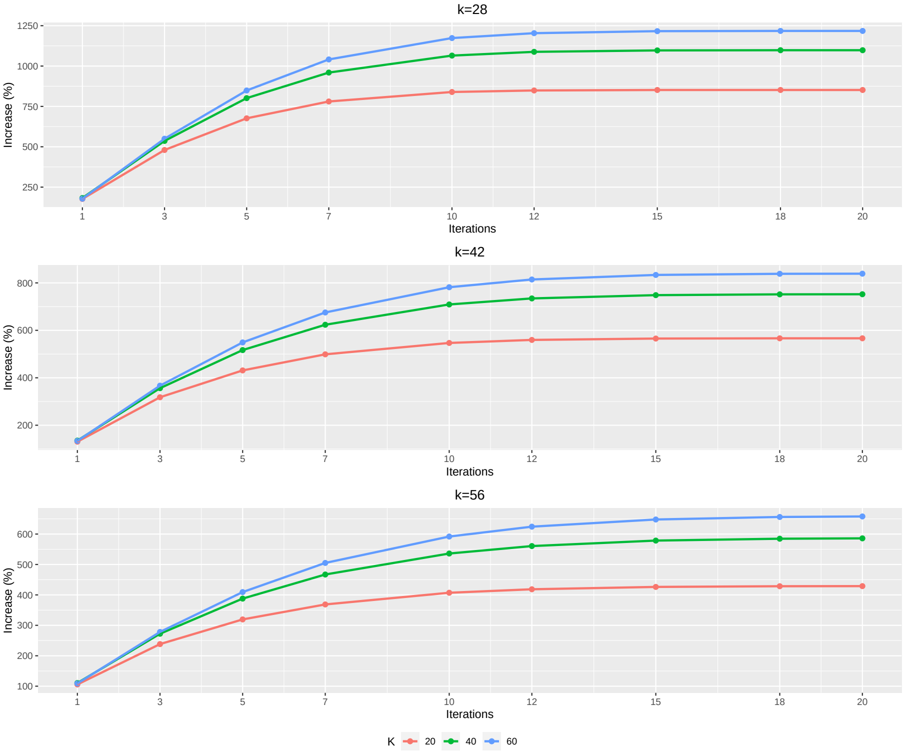


Figure 5: The mean percent increase in the generalized variance by Alg1.

Power consumption data (Salam and Hibaoui, 2018)

◇ y_i: power consumption of the 2^{nd} zone of Tetouan city (north Morocco)
◇ $n = 52,417$ data points
◇ $p = 5$: temperature, humidity, wind speed, diffuse flows and general diffuse flows
◇ 1000 bootstrap samples
◇ $K = 10$, Alg1: 5 iterations - VAlg1: 1 iteration

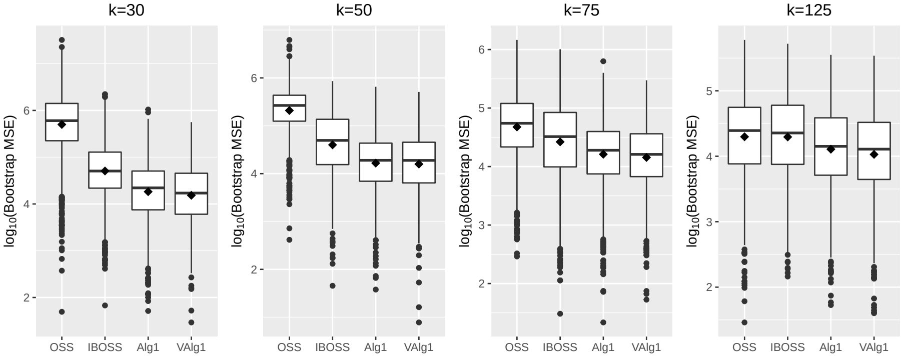


Figure 6: The bootstrap MSEs for estimating slope parameters by different approaches.

Chemical sensors data (Fonollosa *et al.*, 2015)

◇ y_i: readings of a chemical sensor exposed to the mixture of Ethylene and CO at varying concentrations in air

◇ $n = 4,188,261$ data points

◇ $p = 14$: readings of 14 chemical sensors exposed to the mixture of Ethylene and CO at varying concentrations in air

◇ $k = 140$, $K = 10$

◇ Alg1: 5 iterations - VAlg1: 1 iteration

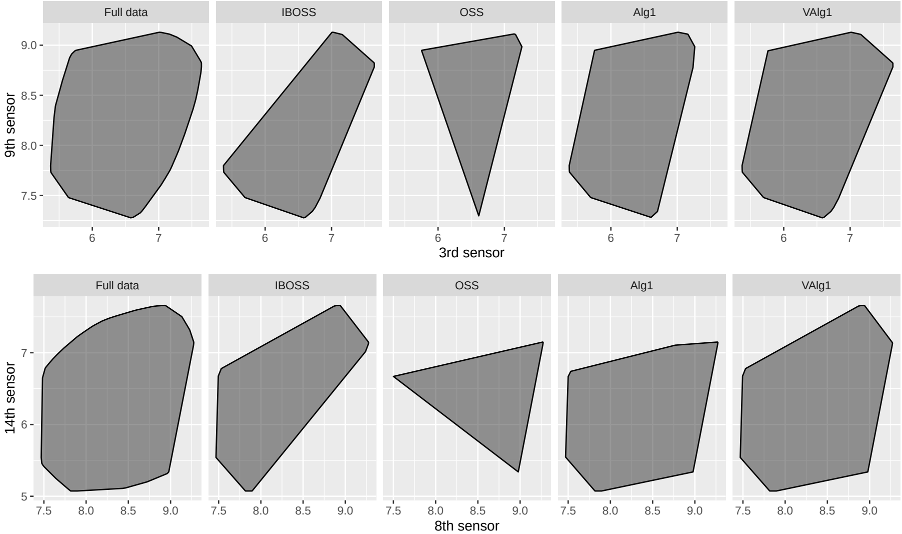


Figure 7: The convex hulls between the 9th and the 3rd sensor, as well as between the 14th and the 8th sensor, for the subdata selected by different approaches.

Discussion

◇ Which approach should one prefer?

References

Drineas, P., Mahoney, M. W., Muthukrishnan, S., and Sarlós, T. (2011). Faster least squares approximation. *Numerische mathematik*, **117**(2): 219–249.

Fonollosa, J., Sheik, S., Huerta, R., and Marco, S. (2015). Reservoir computing compen-sates slow response of chemosensor arrays exposed to fast varying gas concentrationsin continuous monitoring. *Sensors and Actuators B: Chemical*, **215**: 618–629.

Ren, M. and Zhao, S.-L. (2021). Subdata selection based on orthogonal array for bigdata. *Communications in Statistics - Theory and Methods*. doi: `doi.org/10.1080/03610926.2021.2012196`.

Salam, A. and Hibaoui, A. E. (2018). Comparison of machine learning algorithms for the power consumption prediction: - case study of Tetouan city –. In *2018 6th International Renewable and Sustainable Energy Conference (IRSEC)*, pages 1–5.

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

Thank you for your attention!