Efficient subsampling for exponential family models

Dr. Subhadra Dasgupta
(Ruhr-Universität Bochum)

Joint work with Prof. Dr. Holger Dette

July 2023

Outline

In the age of big data, technical advances have enabled exponential growth in data collection.

$$\mathcal{D} = \{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$$
$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$$

Examples

- Sensor response time data- $n \approx 4 * 10^6$ and $p = 14$
- Flight arrival and departure data- $n \approx 10^8$ and $p = 29$
- Cross-Continental square kilometer array data generated by an Astronomical telescope- 700 TB/sec

Techniques- To deal with the data size

1. Divide and conquer- Takes advantage of parallel computing technologies
2. Dimensionality reduction- when $n << p$
3. Subsampling- when $n >> p$

Subsampling

Sample: $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$ and

Subsample $\mathcal{D}_k = \{(\boldsymbol{x}_{s_i}, y_{s_i}) : i = 1, \ldots, k\}$ such that $\mathcal{D}_k \subset \mathcal{D}$.

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ is the parameter corresponding to a model.

- $\hat{\boldsymbol{\beta}}$ denote the estimator based on the **full sample** $\mathcal{D}$
- $\hat{\boldsymbol{\beta}}_{\mathcal{D}_k}$ denote the estimator based on the **subsample** $\mathcal{D}_k$

Aims

1. $\hat{\boldsymbol{\beta}}_{\mathcal{D}_k}$ is very close to $\hat{\boldsymbol{\beta}}$
2. The subsampling algorithm should be computationally cheaper

Existing Subsampling Techniques

- Generalized linear models- [Ai et al., 2021](non-deterministic, based on L-optimality and A-optimality), [Deldossi and Tommasi, 2022] (deterministic and design based on approximate optimal design)
- Logistic regression [Wang et al., 2018], [Cheng et al., 2020]
- Linear regression [Wang et al., 2019],[Ma et al., 2015],[Ren and Zhao, 2021], [Wang et al., 2021]

Our Goal

To find a subsampling algorithm that addresses

- Applicability to a wide class of models
- Provides good estimation accuracy
- Reasonable time complexity

Proposed Subsampling Algorithm

$\mathcal{D} = \{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$ is the sample such that $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^p \times \mathbb{R}$.

Model

$$f(y|\boldsymbol{x}, \boldsymbol{\beta}) = h(y)\ \exp\{\eta^\top(\boldsymbol{x}, \boldsymbol{\beta}) T(y) - A(\boldsymbol{x}, \boldsymbol{\beta})\},$$

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top \in \Theta \subset \mathbb{R}^{p+1}$, $h(y)$ is assumed to be a positive measurable function, $\eta : \mathcal{X} \times \Theta \to \mathbb{R}^l$, $A : \mathcal{X} \times \Theta \to \mathbb{R}$, and T denote a l-dimensional statistic.

Proposed Subsampling Algorithm

$\mathcal{D} = \{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$ is the sample such that $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^p \times \mathbb{R}$.

Model

$$f(y|\boldsymbol{x}, \boldsymbol{\beta}) = h(y)\ \exp\{\eta^\top(\boldsymbol{x}, \boldsymbol{\beta}) T(y) - A(\boldsymbol{x}, \boldsymbol{\beta})\},$$

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top \in \Theta \subset \mathbb{R}^{p+1}$, $h(y)$ is assumed to be a positive measurable function, $\eta : \mathcal{X} \times \Theta \to \mathbb{R}^l$, $A : \mathcal{X} \times \Theta \to \mathbb{R}$, and T denote a l-dimensional statistic.

Working principle- Proposed Algorithm

- Subsample is close to the **approximate optimal design** corresponding to the underlying model

Optimal Design Based Subsampling (ODBSS)

Aim

Accurate estimation of the maximum likelihood estimate of $\boldsymbol{\beta}$ using a subsample of $\mathcal{D}$

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^\top \in \mathbb{R}^{p+1}$$

Optimal Design Based Subsampling (ODBSS)

Aim

Accurate estimation of the maximum likelihood estimate of $\boldsymbol{\beta}$ using a subsample of $\mathcal{D}$

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^\top \in \mathbb{R}^{p+1}$$

Fisher information matrix at the point $\boldsymbol{x} \in \mathcal{X}$

$$\mathcal{I}(\boldsymbol{\beta}, \boldsymbol{x}) = \mathbb{E}\left[\left\{\frac{\partial}{\partial \boldsymbol{\beta}} \log f(y|\boldsymbol{x}, \boldsymbol{\beta})\right\} \left\{\frac{\partial}{\partial \boldsymbol{\beta}} \log f(y|\boldsymbol{x}, \boldsymbol{\beta})\right\}^\top\right]$$

Optimal Design Based Subsampling (ODBSS)

Aim

Accurate estimation of the maximum likelihood estimate of $\boldsymbol{\beta}$ using a subsample of $\mathcal{D}$

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^\top \in \mathbb{R}^{p+1}$$

Fisher information matrix at the point $\boldsymbol{x} \in \mathcal{X}$

$$\mathcal{I}(\boldsymbol{\beta}, \boldsymbol{x}) = \mathbb{E}\Big[\Big\{\frac{\partial}{\partial \boldsymbol{\beta}} \log f(y|\boldsymbol{x}, \boldsymbol{\beta})\Big\} \ \Big\{\frac{\partial}{\partial \boldsymbol{\beta}} \log f(y|\boldsymbol{x}, \boldsymbol{\beta})\Big\}^\top\Big]$$

An **approximate design**

$$\xi(\mathcal{X}, \boldsymbol{\beta}) = \begin{Bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \ldots & \boldsymbol{x}_d \\ w_1 & w_2 & & w_d \end{Bmatrix},$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d \in \mathcal{X}$ and $w_1 + w_2 + \ldots + w_d = 1$.

$$M(\xi, \boldsymbol{\beta}) := \sum_{i=1}^{d} w_i \mathcal{I}(\boldsymbol{x}_i, \boldsymbol{\beta}),$$

Covariance matrix of the maximum likelihood estimator $\sqrt{n}\hat{\boldsymbol{\beta}}$ converges to the matrix $M^{-1}(\xi, \boldsymbol{\beta})$

Optimal Design Based Subsampling (ODBSS)

An **approximate optimal design**

$$\xi^*(\mathcal{X}, \boldsymbol{\beta}) = \left\{ \begin{array}{cccc} \boldsymbol{x}_1^* & \boldsymbol{x}_2^* & \dots & \boldsymbol{x}_d^* \\ w_1^* & w_2^* & & w_d^* \end{array} \right\},$$

where $\boldsymbol{x}_1^*, \dots, \boldsymbol{x}_d^* \in \mathcal{X}$ and $w_1^* + w_2^* + \dots + w_d^* = 1$ is obtained by maximizing $\Phi(\boldsymbol{\mathcal{M}}(\xi))$ for $\boldsymbol{x}_i^*$ and w_i^*, where $\Phi(\cdot)$ concave function.

Example, for D-optimality $\Phi(\cdot) = log(det(M(\xi, \boldsymbol{\beta})))$.

Optimal Design Based Subsampling (ODBSS)

Input: The sample $\mathcal{D}$ of size n
Output: The subsample $\mathcal{D}_k$ a of size k

Step 1: Initial sampling

(1.1) Take a uniform subsample of size k_0 denoted by $\mathcal{D}_{k_0}$

(1.2) Find an estimate of the design space $\mathcal{X}_{k_0}$ based on $\mathcal{D}_{k_0}$

(1.3) Calculate an initial parameter estimate $\hat{\boldsymbol{\beta}}_{\mathcal{D}_{k_0}}$ based on $\mathcal{D}_{k_0}$

Optimal Design Based Subsampling (ODBSS)

Input: The sample $\mathcal{D}$ of size n
Output: The subsample $\mathcal{D}_k$ a of size k

Step 1: Initial sampling

(1.1) Take a uniform subsample of size k_0 denoted by $\mathcal{D}_{k_0}$

(1.2) Find an estimate of the design space $\mathcal{X}_{k_0}$ based on $\mathcal{D}_{k_0}$

(1.3) Calculate an initial parameter estimate $\hat{\boldsymbol{\beta}}_{\mathcal{D}_{k_0}}$ based on $\mathcal{D}_{k_0}$

Step 2: Optimal design determination

(2.1) Find a (locally) approximate optimal design $\xi^*(\mathcal{X}_{k_0}, \hat{\boldsymbol{\beta}}_{\mathcal{D}_{k_0}}) = \begin{Bmatrix} \boldsymbol{x}_1^* & \boldsymbol{x}_2^* & \dots & \boldsymbol{x}_d^* \\ w_1^* & w_2^* & & w_d^* \end{Bmatrix}$

Optimal Design Based Subsampling (ODBSS)

Input: The sample $\mathcal{D}$ of size n
Output: The subsample $\mathcal{D}_k$ a of size k

Step 1: Initial sampling

(1.1) Take a uniform subsample of size k_0 denoted by $\mathcal{D}_{k_0}$

(1.2) Find an estimate of the design space $\mathcal{X}_{k_0}$ based on $\mathcal{D}_{k_0}$

(1.3) Calculate an initial parameter estimate $\hat{\boldsymbol{\beta}}_{\mathcal{D}_{k_0}}$ based on $\mathcal{D}_{k_0}$

Step 2: Optimal design determination

(2.1) Find a (locally) approximate optimal design $\xi^*(\mathcal{X}_{k_0}, \hat{\boldsymbol{\beta}}_{\mathcal{D}_{k_0}}) = \left\{ \begin{matrix} \boldsymbol{x}_1^* & \boldsymbol{x}_2^* & \dots & \boldsymbol{x}_d^* \\ w_1^* & w_2^* & & w_d^* \end{matrix} \right\}$

Step 3: Optimal design based subsampling

(3.1) Determine the remaining subsample $\mathcal{D}_{k_1}$ ($k_1 = k - k_0$), such that, $\lfloor w_i^* \ k_1 \rfloor$ observations are "close" to the support points $\boldsymbol{x}_i^*$ of the optimal design $\xi^*(\mathcal{X}_{k_0}, \hat{\boldsymbol{\beta}}_{\mathcal{D}_{k_0}})$ ($i = 1, \dots, d$).

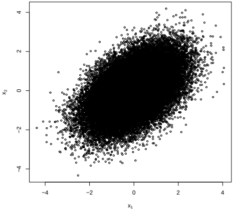
(3.2) The final subsample $\mathcal{D}_k = \mathcal{D}_{k_0} \cup \mathcal{D}_{k_1}$

The points $\boldsymbol{x}_i^* \in \mathcal{X}$ but might not be a part of the original sample

Optimal Design Based Subsampling (ODBSS)

Logistic regression with two covariates and $\boldsymbol{\beta} = (.1, .5, .5)$

$(\boldsymbol{x}_1, \boldsymbol{x}_2) \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}$

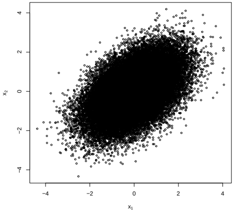


(a) full data ($n = 50000$)

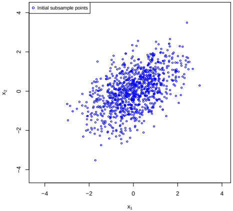
Optimal Design Based Subsampling (ODBSS)

Logistic regression with two covariates and $\boldsymbol{\beta} = (.1, .5, .5)$

$(\boldsymbol{x}_1, \boldsymbol{x}_2) \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}$



(a) full data ($n = 50000$)

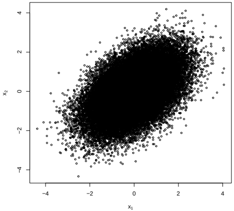


(b) $\mathcal{D}_{k_0}$ with $k_0 = 1000$

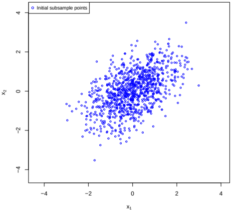
Optimal Design Based Subsampling (ODBSS)

Logistic regression with two covariates and $\boldsymbol{\beta} = (.1, .5, .5)$

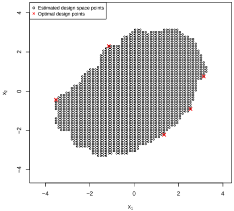
$(\boldsymbol{x}_1, \boldsymbol{x}_2) \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}$



(a) full data ($n = 50000$)



(b) $\mathcal{D}_{k_0}$ with $k_0 = 1000$

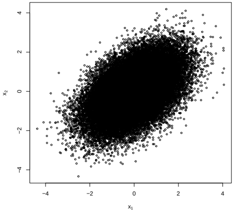


(c) $\mathcal{X}_{k_0}$ and $\xi^*(\mathcal{X}_{k_0}, \hat{\boldsymbol{\beta}}_{\mathcal{D}_{k_0}})$

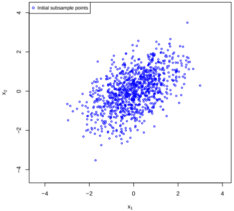
Optimal Design Based Subsampling (ODBSS)

Logistic regression with two covariates and $\boldsymbol{\beta} = (.1, .5, .5)$

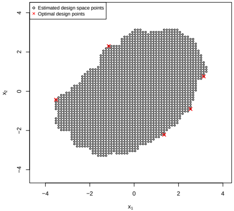
$(\boldsymbol{x}_1, \boldsymbol{x}_2) \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}$



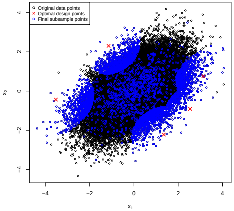
(a) full data ($n = 50000$)



(b) $\mathcal{D}_{k_0}$ with $k_0 = 1000$



(c) $\mathcal{X}_{k_0}$ and $\xi^*(\mathcal{X}_{k_0}, \hat{\boldsymbol{\beta}}_{\mathcal{D}_{k_0}})$



(d) $\mathcal{D}_k$ ($k = 5000$)

Optimal Design Based Subsampling (ODBSS)- Details

Step 1: Initial sampling

(1.1) Take a uniform subsample of size k_0 denoted by $\mathcal{D}_{k_0}$

(1.2) Find an estimate of the design space $\mathcal{X}_{k_0}$ based on $\mathcal{D}_{k_0}$

Reasons

- In real-world problems the design space is not known
- The optimal design depends upon the design space
- This also ensures reduced time complexity

Technique

- Done using density-based clustering [Ester et al., 1996]
- Used the DBSCAN package in $R - software$ [Hahsler et al., 2022]

(1.3) Calculate an initial parameter estimate $\hat{\beta}_{\mathcal{D}_{k_0}}$ based on $\mathcal{D}_{k_0}$

Reasons

- In non-linear models the optimal design depends on the parameter

Optimal Design Based Subsampling (ODBSS)- Details

Step 2: Optimal design determination

(2.1) Find a (locally) approximate optimal design $\xi^*(\mathcal{X}_{k_0}, \hat{\boldsymbol{\beta}}_{\mathcal{D}_{k_0}}) = \left\{ \begin{matrix} \boldsymbol{x}_1^* & \boldsymbol{x}_2^* & \dots & \boldsymbol{x}_d^* \\ w_1^* & w_2^* & & w_d^* \end{matrix} \right\}$

Technique

- Approximate optimal designs are determined numerically using *OptimalDesign* in $R - software$ [Harman and Lenka, 2019] for our simulation studies

Optimal Design Based Subsampling (ODBSS)- Details

Step 3: Optimal design based subsampling

(3.1) Determine the remaining subsample $\mathcal{D}_{k_1}$ ($k_1 = k - k_0$), such that, $\lfloor w_i^* \, k_1 \rfloor$ observations are "close" to the support points $\boldsymbol{x}_i^*$ of the optimal design $\xi^*(\mathcal{X}_{k_0}, \hat{\boldsymbol{\beta}}_{\mathcal{D}_{k_0}})$ ($i = 1, \ldots, d$).

(3.2) The final subsample $\mathcal{D}_k = \mathcal{D}_{k_0} \cup \mathcal{D}_{k_1}$

Optimal Design Based Subsampling (ODBSS)- Details

Step 3: Optimal design based subsampling

(3.1) Determine the remaining subsample $\mathcal{D}_{k_1}$ $(k_1 = k - k_0)$, such that, $\lfloor w_i^* \; k_1 \rfloor$ observations are "close" to the support points $\boldsymbol{x}_i^*$ of the optimal design $\xi^*(\mathcal{X}_{k_0}, \hat{\boldsymbol{\beta}}_{\mathcal{D}_{k_0}})$ $(i = 1, \ldots, d)$.

(3.2) The final subsample $\mathcal{D}_k = \mathcal{D}_{k_0} \cup \mathcal{D}_{k_1}$

Distance between points

- **Frobenius distance**
  $d_F(\boldsymbol{x}, \boldsymbol{x}') := \|\mathcal{I}(\boldsymbol{x}, \beta) - \mathcal{I}(\boldsymbol{x}', \beta)\|_F :=$
  $tr\Big\{ \big(\mathcal{I}(\boldsymbol{x}, \beta) - \mathcal{I}(\boldsymbol{x}', \beta)\big)^\top \; \big(\mathcal{I}(\boldsymbol{x}, \beta) - \mathcal{I}(\boldsymbol{x}', \beta)\big) \Big\}^{1/2}$
- **Square root distance**
  $d_s(\boldsymbol{x}, \boldsymbol{x}') := \|\mathcal{I}(\boldsymbol{x}, \beta)^{1/2} - \mathcal{I}(\boldsymbol{x}', \beta)^{1/2}\|_F$
- **Procrustes distance**
  $d_p(\boldsymbol{x}, \boldsymbol{x}') := \inf_{\boldsymbol{K} \in O(\mathbb{R}^{(p+1)\times(p+1)})} \Big\{ \|\mathcal{I}(\boldsymbol{x}, \beta) - \mathcal{I}(\boldsymbol{x}', \beta)\boldsymbol{K}\|_F \Big\}^{1/2}$, where
  $O(\mathbb{R}^{(p+1)\times(p+1)})$ is set of orthogonal matrices

Simulation study

- The subsampling algorithms are compared by the $MSE = \mathbb{E}[\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\mathcal{D}_k}\|^2]$ by performing 100 simulation runs
- Logistic regression model $p = 7$, no intercept, $\boldsymbol{\beta} = (0.5, 0.5, \ldots, 0.5)$, and n=100000
- Covariates $\boldsymbol{x} = (x_1, x_2, \ldots, x_7)$ follows multivariate a centered normal distribution with covariance $\boldsymbol{\Sigma}$
  (1) $\boldsymbol{\Sigma}_1 = (0.5^{|i-j|})_{i,j=1,\ldots,7}$.
  (2) $\boldsymbol{\Sigma}_2 = 2\, \boldsymbol{e}_1\boldsymbol{e}_1^\top + 1.8\, \boldsymbol{e}_2\boldsymbol{e}_2^\top + 1.6\, \boldsymbol{e}_3\boldsymbol{e}_3^\top + 1.4\, \boldsymbol{e}_4\boldsymbol{e}_4^\top + 1.2\, \boldsymbol{e}_5\boldsymbol{e}_5^\top + 0.1\, \boldsymbol{\Sigma}_1$ where, $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, $\boldsymbol{e}_3$, $\boldsymbol{e}_4$, and $\boldsymbol{e}_5 \in$ on $\mathbb{S}_6 \subset \mathbb{R}^7$ are mutually orthogonal and chosen randomly in each simulation
  (3) Similarly to (2) $\boldsymbol{\Sigma}_3 = 3\, \boldsymbol{e}_1\boldsymbol{e}_1^\top + 2\, \boldsymbol{e}_2\boldsymbol{e}_2^\top + 1\, \boldsymbol{e}_3\boldsymbol{e}_3^\top + 0.1\, \boldsymbol{\Sigma}_1$
- In the simulation studies we consider approximate A-optimal designs

Subsampling Matrix distances for Logistic regression

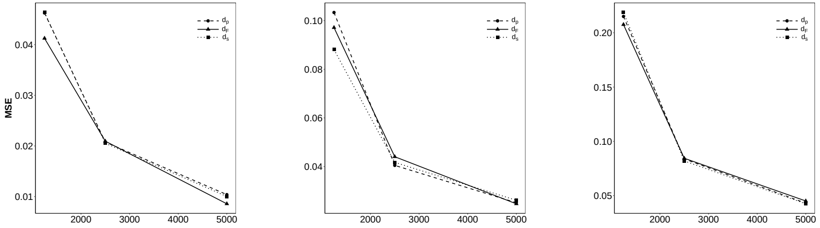
Distance between points- When information matrix is rank 1

When rank of $\mathcal{I}(\boldsymbol{x},\boldsymbol{\beta})$ is 1, then $\mathcal{I}(\boldsymbol{x},\boldsymbol{\beta}) = \Phi(\boldsymbol{x},\boldsymbol{\beta})\Phi(\boldsymbol{x},\boldsymbol{\beta})^\top$ where $\Phi(\boldsymbol{x},\boldsymbol{\beta}) \in \mathbb{R}^{p+1}$.

- $d_F(\boldsymbol{x},\boldsymbol{x}') = \left\{\|\Phi(\boldsymbol{x},\boldsymbol{\beta})\|^4 + \|\Phi(\boldsymbol{x}',\boldsymbol{\beta})\|^4 - 2(\Phi(\boldsymbol{x}',\boldsymbol{\beta})^\top\Phi(\boldsymbol{x},\boldsymbol{\beta}))^2\right\}^{1/2}$
- $d_S(\boldsymbol{x},\boldsymbol{x}') = \left\{\|\Phi(\boldsymbol{x},\boldsymbol{\beta})\|^2 + \|\Phi(\boldsymbol{x}',\boldsymbol{\beta})\|^2 - 2\dfrac{(\Phi(\boldsymbol{x}',\boldsymbol{\beta})^\top\Phi(\boldsymbol{x},\boldsymbol{\beta}))^2}{\|\Phi(\boldsymbol{x},\boldsymbol{\beta})\|\;\|\Phi(\boldsymbol{x}',\boldsymbol{\beta})\|}\right\}^{1/2}$
- $d_p(\boldsymbol{x},\boldsymbol{x}') = \|\Phi(\boldsymbol{x},\boldsymbol{\beta}) - \Phi(\boldsymbol{x}',\boldsymbol{\beta})\| = \left\{\|\Phi(\boldsymbol{x},\boldsymbol{\beta})\|^2 + \|\Phi(\boldsymbol{x}',\boldsymbol{\beta})\|^2 - 2(\Phi(\boldsymbol{x}',\boldsymbol{\beta})^\top\Phi(\boldsymbol{x},\boldsymbol{\beta}))\right\}^{1/2}$,
  where $\|\cdot\|$ is the Euclidean norm.

Simulation study

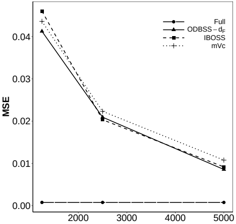
- ODBSS based on the distances d_F, d_s, and d_p are comparable
- Some more simulations indicated d_F would be a better choice among the three distances



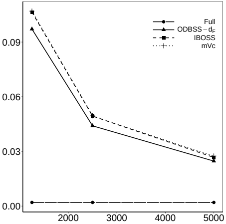
$\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_1)$ $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_2)$ $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_3)$

Mean squared error of the parameter estimate using OBDSS subsampling with the metric $d_F()$, $d_S()$, and $d_p()$ at different subsample sizes

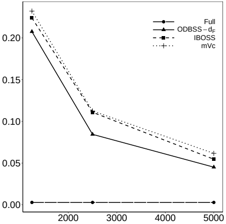
Simulation study- Comparison with existing algorithms



$\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_1)$



$\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_2)$



$\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_3)$

Computational complexity of ODBSS

Time Components

(1) **Area estimation:** Complexity of DBSCAN algorithm with p-dimensional k_0 points is $\mathcal{O}(k_0^2 p)$

(2) **Calculation of optimal design** over $\mathcal{X}_{k_0}$: $\mathcal{O}((sp)^3)$, where $s = |\mathcal{X}_{k_0}|$ (s is controlled by the experimenter)

(3) **Subsample allocation:** $\mathcal{O}(dnp) + \mathcal{O}(dn)$.

Computational complexity of ODBSS

(2) Run-time for finding optimal design is low - $\mathcal{O}((sp)^3)$

- If area approximation is not done, then $\mathcal{D}$ serves as the approximation of the design space.
- In the above case, the time complexity for finding optimal design is $\mathcal{O}((np)^3)$
- Simulation studies show that area approximation does not have any negative impact on parameter estimation (performs well with respect to mVc, IBOSS)
- Area estimation reduced the time for the subsampling algorithm significantly

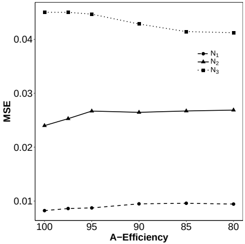
		n			
		100000	200000	300000	400000
$\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_1)$	ODBSS	7.05	8.66	7.79	8.43
	ODBSS-2	5.63	8.95	9.60	13.05
$\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_2)$	ODBSS	6.52	6.01	6.69	8.33
	ODBSS-2	4.17	7.11	10.29	12.51
$\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_3)$	ODBSS	8.34	7.33	8.72	8.05
	ODBSS-2	5.11	8.09	11.07	11.51

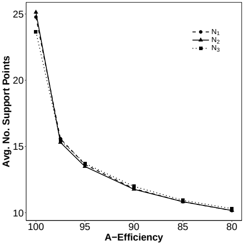
Table: *Comparison of run times (in seconds) of ODBSS is with area estimation and ODBSS-2 is with and without area approximation*

Computational complexity of ODBSS

(3) Reduce run-time for subsample allocation- $\mathcal{O}(dnp) + \mathcal{O}(dn)$

- Number of support points d of the approximate optimal design is quite high (although bounded by p(p+1)/2
- Efficiency of a design ξ:
$$\text{eff}(\xi, \boldsymbol{\beta}) = \frac{\Phi(M(\xi, \boldsymbol{\beta}))}{\Phi(M(\xi^*(\boldsymbol{\beta}, \mathcal{X}), \boldsymbol{\beta})} \in [0, 1]$$
- Use a design with reduced efficiency





Conclusion and Future Directions

Summary

- Provides a universal subsampling framework for any model
- Propose ways to minimize the run-time of the subsampling algorithm without compromising the quality of estimation
- Simulation studies show that ODBSS outperforms the existing algorithms for linear and logistic regression

Conclusion and Future Directions

Summary

- Provides a universal subsampling framework for any model
- Propose ways to minimize the run-time of the subsampling algorithm without compromising the quality of estimation
- Simulation studies show that ODBSS outperforms the existing algorithms for linear and logistic regression

Future Directions

- Need to investigate if there is a theoretical justification as to why the proposed approach performs better
- To investigate the statistical properties of the ODBSS estimators (with various matrix distances)
- The optimal design determination is computationally expensive and we need to find if this could be reduced

References

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

Ben-Tal, A. and Nemirovski, A. (2001).
Lectures on modern convex optimization: analysis, algorithms, and engineering applications.
SIAM.

Cheng, Q., Wang, H., and Yang, M. (2020).
Information-based optimal subdata selection for big data logistic regression.
Journal of Statistical Planning and Inference, 209:112–122.

Deldossi, L. and Tommasi, C. (2022).
Optimal design subsampling from big datasets.
Journal of Quality Technology, 54(1):93–101.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996).
A density-based algorithm for discovering clusters in large spatial databases with noise.
In *kdd*, volume 96, pages 226–231.

Hahsler, M., Piekenbrock, M., Arya, S., and Mount, D. (2022).
dbscan: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Related Algorithms.
R package version 1.1-11.

Harman, R. and Lenka, F. (2019).

OptimalDesign: A Toolbox for Computing Efficient Designs of Experiments.
R package version 1.0.1.

Ma, P., Mahoney, M., and Yu, B. (2015).
A statistical perspective on algorithmic leveraging.
Journal of Machine Learning Research, 16(27):861–911.

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

Ren, M. and Zhao, S.-L. (2021).
Subdata selection based on orthogonal array for big data.
Communications in Statistics-Theory and Methods, pages 1–19.

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.

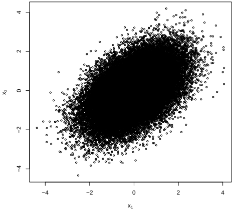
Wang, H., Zhu, R., and Ma, P. (2018).
Optimal subsampling for large sample logistic regression.
Journal of the American Statistical Association, 113(522):829–844.

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

APPENDIX

ODBSS- Area estimation step in details

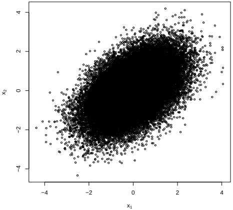
Logistic regression with two covariates and $\boldsymbol{\beta} = (.1, .5, .5)$



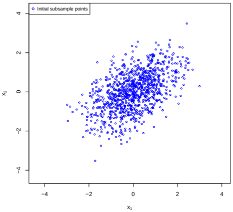
(a) full data ($n = 50000$)

ODBSS- Area estimation step in details

Logistic regression with two covariates and $\boldsymbol{\beta} = (.1, .5, .5)$



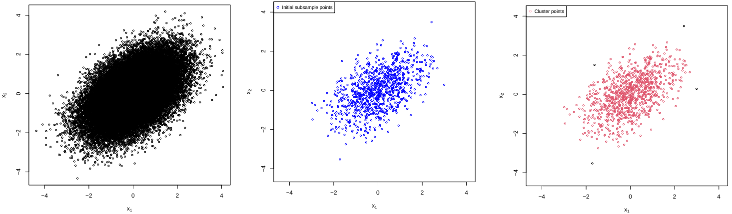
(a) full data ($n = 50000$)



(b) $\mathcal{D}_{k_0}$ with $k_0 = 1000$

ODBSS- Area estimation step in details

Logistic regression with two covariates and $\boldsymbol{\beta} = (.1, .5, .5)$

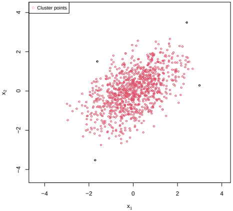


(a) full data ($n = 50000$)

(b) $\mathcal{D}_{k_0}$ with $k_0 = 1000$

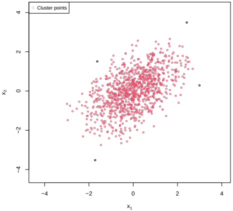
(c) DBSCAN Cluster

ODBSS- Area estimation step in details

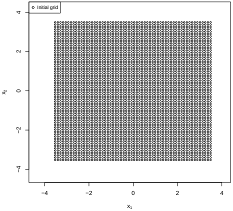


(a) DBSCAN Cluster

ODBSS- Area estimation step in details

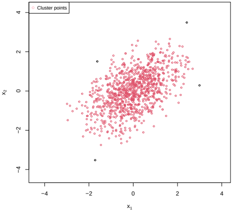


(a) DBSCAN Cluster

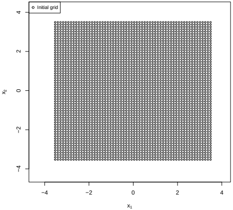


(b) An equispaced grid

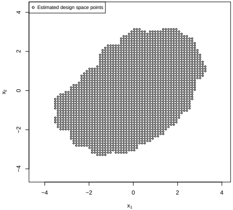
ODBSS- Area estimation step in details



(a) DBSCAN Cluster

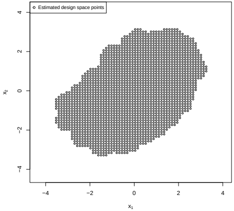


(b) An equispaced grid



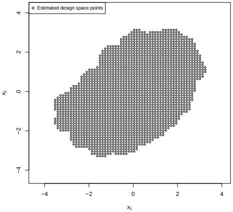
(c) $\mathcal{X}_{k_0}$

ODBSS- Optimal design estimation step in details

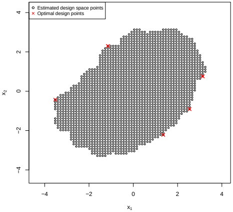


(a) full data ($n = 50000$)

ODBSS- Optimal design estimation step in details



(a) full data ($n = 50000$)



(b) $\mathcal{X}_{k_0}$ and $\xi^*(\hat{\boldsymbol{\beta}}_{\mathcal{D}_{k_0}})$