Optimal selection from a Big Dataset

Laura Deldossi[1] Chiara Tommasi[2]

[1]Department of Statistical Sciences,
Università Cattolica del Sacro Cuore, Milan, Italy

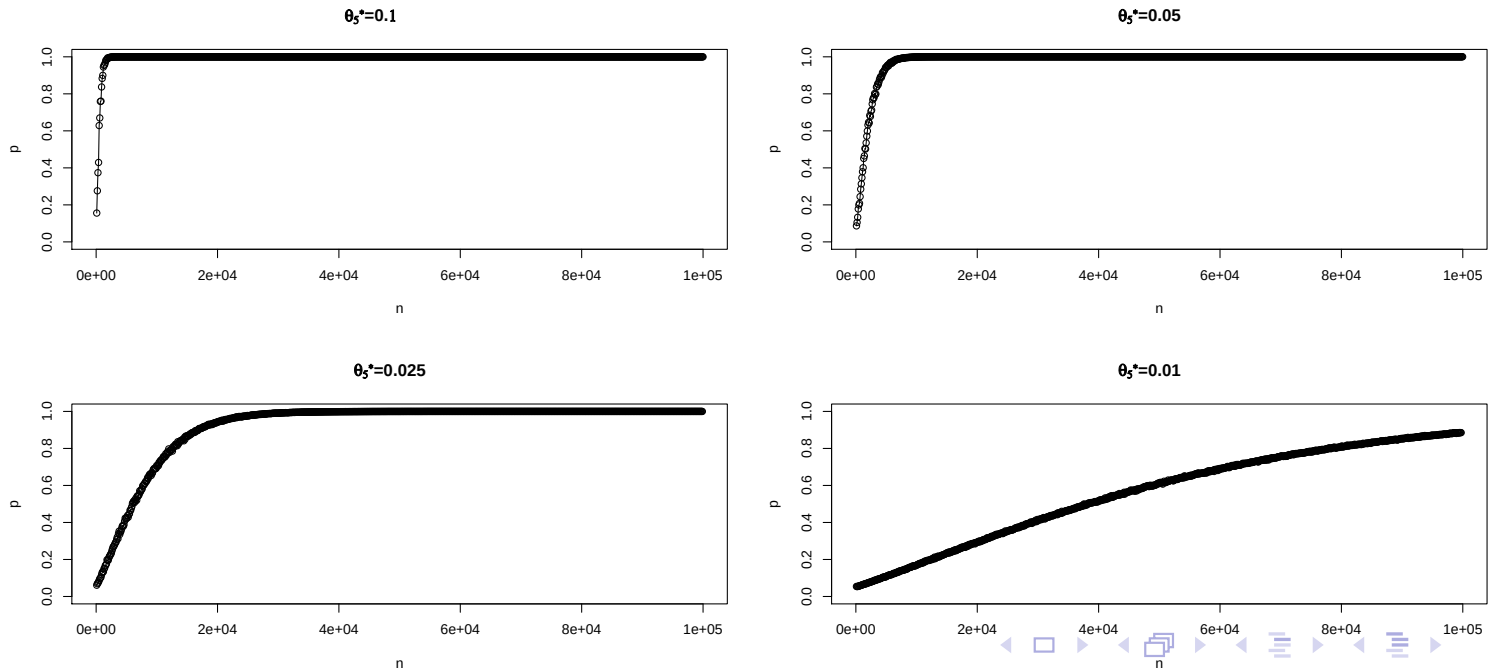
[2]Department of Economics, Management and Quantitative Methods
University of Milan, Italy

GMDS & CEN-IBS
December 2-3, 2020
Online satellite meeting

Motivation of the work

Advances in technology enable us to collect, transfer and store large dataset. The availability of huge quantity of data is a great challenge nowadays, BUT:

- Enormous computational effort is required to analyze an entire Big Dataset.
- Very often a Big Dataset contains redundant data.
- From a practical point of view, some inferential conclusions may be ineffective with large samples.



Goal and outline

CONTEXT

A Big Dataset is conceived as a finite population generated by a super-population model. Goals: to estimate the parameters of this super-population model or to obtain an accurate response prediction.

IDEA

To use the optimal design theory in order to select a sample which contains the majority of information to reach the inferential goal.

1. Notation: exact and continuous designs
2. Big Data, super-population model and DOE
3. Efficiency as a measure of the "quality" of a Big-Data
4. Selection of the most informative observations through optimum design
5. Illustrative examples

Notation

- $\boldsymbol{x} \in \mathcal{X}$: **experimental condition** chosen by the experimenter.
- $y = y(\boldsymbol{x})$: **response** variable.
- Responses and experimental conditions are related (at least approximatively) through a **regression model**:

$$y_i = \eta(\boldsymbol{x}_i, \boldsymbol{\theta}) + \varepsilon_i \cong \boldsymbol{f}(\boldsymbol{x}_i)^T \boldsymbol{\theta} + \varepsilon_i = \sum_{j=1}^{m+1} f_j(\boldsymbol{x}_i)\, \theta_j + \varepsilon_i,$$

$$\mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{Var}(\varepsilon_i) = \sigma^2, \quad \mathrm{Cov}(\varepsilon_i, \varepsilon_l) = 0.$$

Exact designs

- **Exact design**: $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \quad \longmapsto \quad \{y(\boldsymbol{x}_1), \ldots, y(\boldsymbol{x}_n)\}$

  $$\xi_n = \left\{ \begin{array}{ccc} \boldsymbol{x}_1 & \ldots & \boldsymbol{x}_k \\ \dfrac{n_1}{n} & & \dfrac{n_k}{n} \end{array} \right\}, \quad k < n$$

- **Precision matrix** of the BLUE for $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^{m+1}$:

$$\frac{1}{\sigma^2} \boldsymbol{F}_n^T \boldsymbol{F}_n \quad = \quad \frac{1}{\sigma^2} \sum_{i=1}^{n} \boldsymbol{f}(\boldsymbol{x}_i) \boldsymbol{f}(\boldsymbol{x}_i)^T = \frac{n}{\sigma^2} \sum_{j=1}^{k} \boldsymbol{f}(\boldsymbol{x}_j) \boldsymbol{f}(\boldsymbol{x}_j)^T \frac{n_j}{n}$$

- **Information matrix** of an exact design ξ_n:

$$\boldsymbol{M}(\xi_n) = \sum_{j=1}^{k} \boldsymbol{f}(\boldsymbol{x}_j) \boldsymbol{f}(\boldsymbol{x}_j)^T \frac{n_j}{n}; \quad \boldsymbol{F}_n = \begin{bmatrix} 1 & f_1(\boldsymbol{x}_1) & \ldots & f_m(\boldsymbol{x}_1) \\ \vdots & \vdots & \ldots & \vdots \\ 1 & f_1(\boldsymbol{x}_n) & \ldots & f_m(\boldsymbol{x}_n) \end{bmatrix}$$

Continuous designs

- **A continuous design** ξ is a discrete probability measure on $\mathcal{X}$ with a finite number of support points:

$$\xi = \begin{Bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_k \\ \omega_1 & \cdots & \omega_k \end{Bmatrix}, \quad 0 \leq \omega_j \leq 1, \sum_{j=1}^{k} \omega_j = 1.$$

- **Information matrix** for an approximate design ξ:

$$\boldsymbol{M}(\xi) = \sum_{j=1}^{k} \boldsymbol{f}(\boldsymbol{x}_j)\boldsymbol{f}(\boldsymbol{x}_j)^T \omega_j.$$

The variance of the BLUE for θ is proportional to $\boldsymbol{M}(\xi)^{-1}$, thus it makes sense to "maximize" some criterion function $\Phi[\boldsymbol{M}(\xi)]$.

Big Data and super-population model

- Let the Big Dataset be a $N \times (1+p)$ matrix with $p << N$:
  1. First column: N observations for a response variable Y;
  2. Remaining $N \times p$ matrix (denoted by $\mathcal{X}$): N **observed** values for an auxiliary variable $X \in \mathbb{R}^{\mathrm{p}}$.
- Super-population model:

$$\mathbf{y}_U \cong \mathbf{F}_U \boldsymbol{\theta} + \boldsymbol{\varepsilon}_U,$$

- $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_m)$ is a $(m+1) \times 1$ vector of unknown parameters of interest,
- $\boldsymbol{\varepsilon}_U = (\varepsilon_1, \ldots, \varepsilon_N)$ is a vector of homoschedastic independent errors such that $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$

$$\mathbf{F}_U = \begin{bmatrix} \mathbf{f}(\boldsymbol{x}_1)^T \\ \vdots \\ \mathbf{f}(\boldsymbol{x}_N)^T \end{bmatrix} = \begin{bmatrix} 1 & f_1(\boldsymbol{x}_1) & \ldots & f_m(\boldsymbol{x}_1) \\ \vdots & \vdots & \ldots & \vdots \\ 1 & f_1(\boldsymbol{x}_N) & \ldots & f_m(\boldsymbol{x}_N) \end{bmatrix}$$

is an **observed** design matrix.

Big Data and DOE: precise estimation

"Nature" as an experimenter

If Nature had been a "wise" experimenter then it would have chosen the N values for explanatory vector variable X according to an optimality criterion $\Phi[\cdot]$.

Inferential goal: to estimate the vector parameter θ as precisely as possible.

- **D-optimality:** $\Phi_D[\boldsymbol{M}(\xi)] = |\boldsymbol{M}(\xi)|,\ \boldsymbol{M}(\xi) \propto \mathbf{F}^T\mathbf{F}$

  A D-optimum design minimizes the **generalized variance** of $\hat{\theta}$:

  $$\xi_D^* = \arg\max_{\xi} \Phi_D[\boldsymbol{M}(\xi)] = \arg\min_{\xi} |\boldsymbol{M}(\xi)^{-1}|.$$

- **A-optimality:** $\Phi_A[\boldsymbol{M}(\xi)] = -\mathrm{Tr}\left[\boldsymbol{M}(\xi)^{-1}\right]$

  An A-optimum design minimizes the **total variation** of $\hat{\theta}$:

  $$\xi_A^* = \arg\max_{\xi} \Phi_A[\boldsymbol{M}(\xi)] = \arg\min_{\xi} \mathrm{Tr}\left[\boldsymbol{M}(\xi)^{-1}\right].$$

Big Data and DOE: prediction goal

In Big Data context a common goal is **prediction accuracy**, the criterion which reflects this aim is:

- Integrated **I-optimality** (Atkinson, 2014):

$$I[\boldsymbol{M}(\xi)] = \int_{\mathcal{X}} \mathbf{f}(\boldsymbol{x})^T \boldsymbol{M}(\xi)^{-1} \mathbf{f}(\boldsymbol{x})\, d\boldsymbol{x} = n\mathrm{Tr}\left[(\boldsymbol{F}^T\boldsymbol{F})^{-1} \int_{\mathcal{X}} \mathbf{f}(\boldsymbol{x})\mathbf{f}(\boldsymbol{x})^T d\boldsymbol{x}\right]$$

An I-optimum design is $\xi_I^* = \arg\min_\xi \Phi_I[\mathbf{M}(\xi)]$

A measure for the quality of the Big Data

- Given a super-population model it is always possible to compute the (continuous) Φ-optimum design:

$$\xi_{\Phi}^{*} = \begin{Bmatrix} \boldsymbol{x}_1^* & \cdots & \boldsymbol{x}_j^* & \cdots & \boldsymbol{x}_k^* \\ \omega_1^* & \cdots & \omega_j^* & \cdots & \omega_k^* \end{Bmatrix}.$$

- A wise Nature would have generated $N\omega_j^*$ responses at $\boldsymbol{x}_j^*$ (with $j = 1, \ldots, k$) providing the **"ideal" Big Data**.
- $\mathbf{F}_U^*$ denotes "ideal" design matrix and $\mathbf{M}_U^* = \mathbf{F}_U^{*T}\mathbf{F}_U^*/N$ is the corresponding "ideal" information matrix.
- $\mathbf{M}_U = \mathbf{F}_U^T\mathbf{F}_U/N$ measures the per-unit information contained in the **observed** Big Dataset

Φ-efficiency as a measure of the quality of the Big Dataset

$$0 \leq \mathrm{Eff}_{\Phi}(\mathbf{M}_U) = \frac{\Phi[\mathbf{F}_U^T\mathbf{F}_U]}{\Phi[\mathbf{F}_U^{*T}\mathbf{F}_U^*]} \leq 1$$

To extract the most informative observations

We aim at selecting only the most informative observations.

- $\mathbf{F_s}$ denotes the design matrix of a sample s of n observations selected from the Big Dataset; $\boldsymbol{M}_s = \mathbf{F}_s^T \mathbf{F}_s / n$ is the **sample information matrix** which measures the per-unit information contained in s.
- If we select all the $\binom{N}{n}$ samples of size n from the Big Data, we can compute all the corresponding

$$\mathrm{Eff}_\Phi[\mathbf{M}_s] = \frac{\Phi[\mathbf{F}_s^T \mathbf{F}_s / n]}{\Phi[\mathbf{F}_U^{*T} \mathbf{F}_U^* / N]}.$$

- Let s^* be sample of size n with the **largest** value of $\mathrm{Eff}_\Phi[\mathbf{M}_s]$.
- **Drawback:** s^* cannot be computed if N and n are large.
- **Optimal design theory** to provide approximations of s^*.

Exchange Algorithm sampling method: EXCH

An approximation of s^* is provided by an **exchange algorithm**; see for instance Atkinson et al. (2007), §12.6.

- An initial sample $s_0 = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is chosen at random from $\mathcal{X}$, i.e. the N rows of the Big Data.
- s_0 is improved by adding that point $x_{n+1} \in \mathcal{X}$ which **most improves** the Φ-criterion, followed by removing that point from $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x}_{n+1}\}$, which gives the **smallest reduction** in the Φ-criterion.
- This add/remove procedure is continued until it converges, with the same point being added and then removed.
- We have applied the R package *Optimal design* by Harman and Filova (2019).

Exchange algorithms depend on an initial design and sometimes may not reach the convergence.

Information Based Optimal Subdata Selection (IBOSS)

IBOSS sampling algorithm is a novel approach to data selection proposed by Wang, Yang and Stufken (2018) for linear models.

Suppose that $r = \frac{n}{2\,m}$ is an integer, where n is the sample size and m is the number of slope parameters in a linear model: $\theta_1, \ldots, \theta_m$ (without the intercept).

1. for $f_1(\boldsymbol{x}_i)$ with $1 \leq i \leq N$, include r data points with the r smallest $f_1(\boldsymbol{x}_i)$ values and r data points with the largest $f_1(\boldsymbol{x}_i)$ values;
2. for $j = 2, \cdots, m$, exclude data points that were previously selected and from the remainder ones, select r data points with the smallest $f_j(\boldsymbol{x}_i)$ values and r data points with the largest $f_j(\boldsymbol{x}_i)$ values.

ODB: Optimal design based sample

The (continuous) Φ-optimal design

$$\xi_{\Phi}^{*} = \begin{Bmatrix} \boldsymbol{x}_1^* & \cdots & \boldsymbol{x}_j^* & \cdots & \boldsymbol{x}_k^* \\ \omega_1^* & \cdots & \omega_j^* & \cdots & \omega_k^* \end{Bmatrix}.$$

suggests the following sampling rule:

select from the Big Data the $n\omega_j^*$ rows of $\mathbf{F}_U$ which are closest to $\mathbf{f}(\mathbf{x}_j^*)^T$ for $j = 1, \ldots, k$.

- Given a super-population model the Φ-optimal design is easily found (we have applied the R package *Optimal design* by Harman and Filova (2019)).
- If $n\omega_j^*$ is not integer than a suitable rounding-off rule can be applied (see Pulkesheim and Rieder, 1992).
- As a measure of closeness we have applied Euclidean distance but any other distance can be used.

Theoretical justification for using optimal design sampling strategies

- OLS estimator based on a sample s:

$$\hat{\boldsymbol{\theta}}_s = (\mathbf{F}_s^T \mathbf{F}_s)^{-1} \mathbf{F}_s \mathbf{y}_s = \left(\sum_{l=1}^{N} \mathbf{f}(\boldsymbol{x}_l) \mathbf{f}(\boldsymbol{x}_l)^T i_l \right)^{-1} \sum_{l=1}^{N} \mathbf{f}(\boldsymbol{x}_l) y_l \, i_l$$

  where $i_l = \begin{cases} 1 & \text{if } l \in s \\ 0 & \text{otherwise} \end{cases}$ is the inclusion indicator.

- $\mathrm{E}_{\mathbf{I}_U, \mathbf{Y}}(\hat{\theta}_s | \mathbf{X}_U) = \boldsymbol{\theta}$ and $\mathrm{Var}_{\mathbf{I}_U, \mathbf{Y}}(\hat{\boldsymbol{\theta}}_s | \mathbf{X}_U) = \sigma^2 \, \mathrm{E}_{\mathbf{I}_U} \left[\left(\mathbf{F}_s^T \mathbf{F}_s \right)^{-1} \right]$
- If the criterion Φ is concave or linear, then

$$\Phi \left\{ \mathrm{E}_{\mathbf{I}_U} \left[\left(\mathbf{F}_s^T \mathbf{F}_s \right)^{-1} \right] \right\} \geq \mathrm{E}_{\mathbf{I}_U} \left\{ \Phi \left[\left(\mathbf{F}_s^T \mathbf{F}_s \right)^{-1} \right] \right\} \geq \Phi \left[\left(\mathbf{F}_{s^*}^T \mathbf{F}_{s^*} \right)^{-1} \right]$$

- Optimal design sampling strategies approximate s^*; hence, we expect they work better than "standard" random samples.

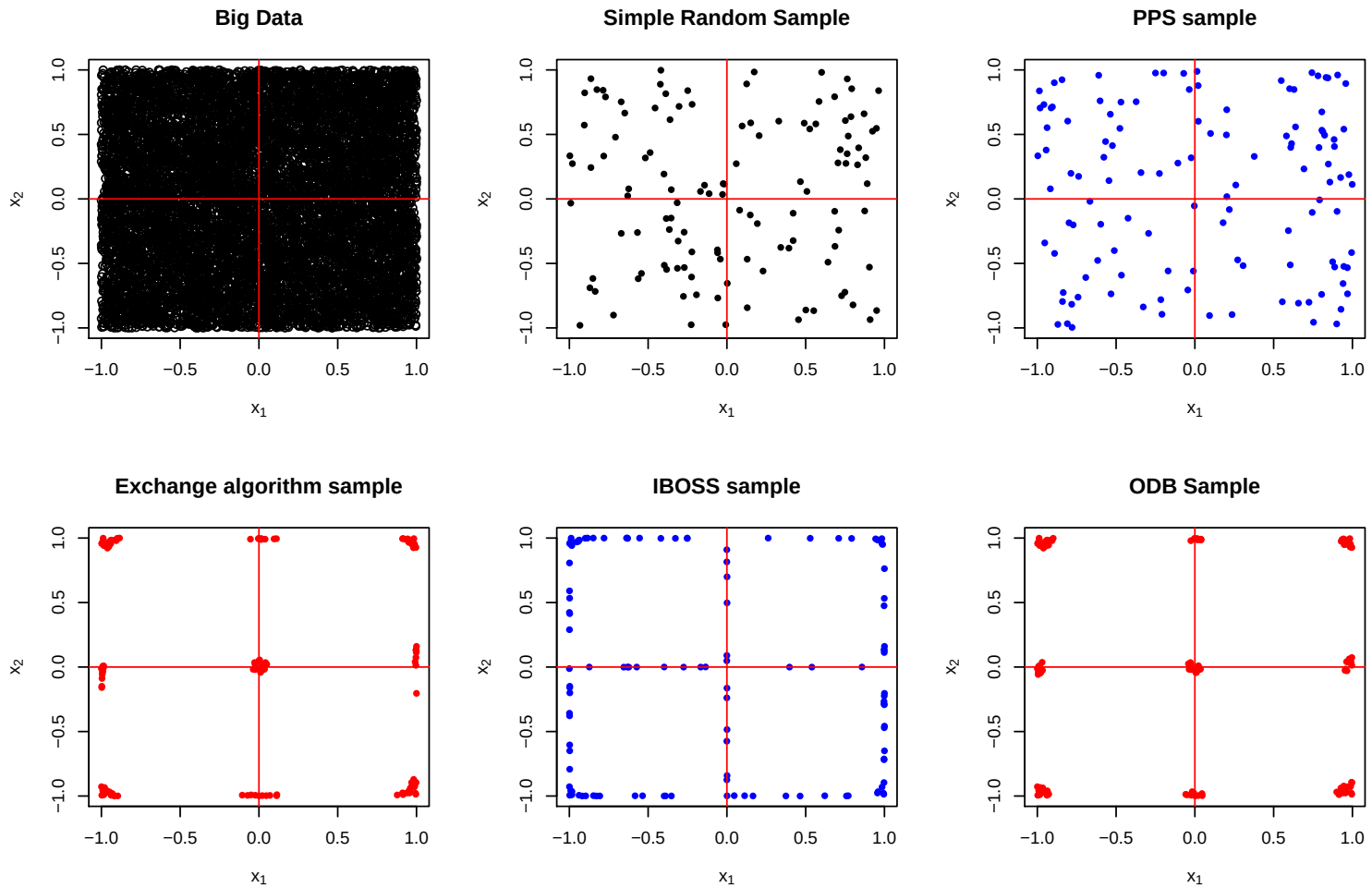
Comparison with two classical sampling strategies

- The simple random sampling without replacement (SRS);
- The probability proportional to size (PPS) sampling with selection probabilities given by

$$p_i = \frac{\mathbf{f}(\mathbf{x}_i)^T (\mathbf{F}_\mathbf{U}^\mathbf{T} \mathbf{F}_\mathbf{U})^{-1} \mathbf{f}(\mathbf{x}_i)}{m+1}, \quad i = 1, \cdots, N$$

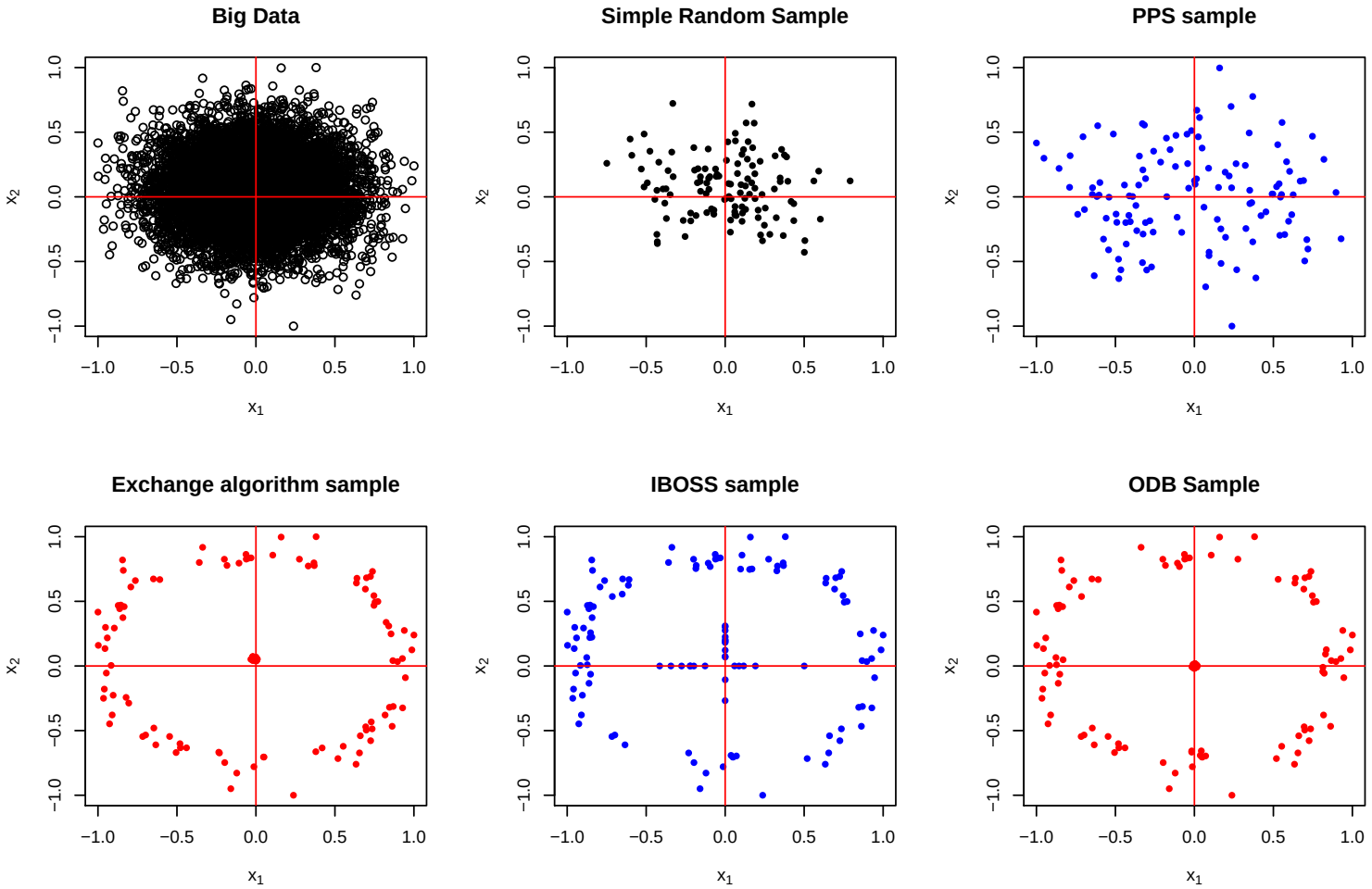
to select more frequently the rows (the units) with largest prediction variance (see also Ma and Sun, 2015).

N=10000, n=120; linear model $f(x)^T=(1, x_1, x_2, x_1 \cdot x_2, x_1^2, x_2^2)$, $x \sim U_2(-1, 1)$



N=10000, n=120; Linear model $f(x)^T=(1, x_1, x_2, x_1 \cdot x_2, x_1^2, x_2^2)$

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 16 & 0 \\ 0 & 16 \end{bmatrix} \right), \quad x_j = \frac{z_j - \left(z_{j(1)} + z_{j(N)}\right)/2}{\left(z_{j(N)} - z_{j(1)}\right)/2}$$



Example 1: Estimation accuracy

- We generate a $N \times (p+1)$ design matrix $\mathbf{X}_U$, with $N = 10^6$ and $p = 10$:

$$X_i \sim U(0,5) \text{ (independent uniform r.v.), for } i = 1,2,3;$$

$$(X_4, X_5, X_6, X_7)^T \sim N(\mathbf{0}, \mathbf{\Sigma}_4);\ Cov(X_i, X_j) = -1,\ Var(X_i) = 9;$$

$$(X_8, X_9)^T \sim t_3(\mathbf{0}, \mathbf{\Sigma}_2);\ Cov(X_i, X_j) = 0.5;\ Var(X_i) = 4, \text{ for } i, j = 8, 9;$$

$$X_{10} \sim \mathcal{P}(5) \text{ (Poisson distribution).}$$

- A $Nx1$ response vector $\mathbf{y}_U$ is simulated $R = 1000$ times from

$$\mathbf{y}_U = \mathbf{F}_U \boldsymbol{\theta} + \boldsymbol{\varepsilon}_U,$$

$\boldsymbol{f}(\mathbf{x})^T = (1, x_1, x_2, \cdots, x_{10})$, $\boldsymbol{\theta} = (1,1,1,1,1,1,1,1,1,1,1)^T$, $Var(\epsilon_i) = 9$,

- At each step r, with $r = 1, 2, \cdots, R = 1000$:
  a sample of size $n = 1000$ is drawn from the Dataset according to the five different sampling schemes.

EXCH, ODB and IBOSS provide the same sample at each simulation step: they are deterministic sampling schemes.
SRS and PPS strategies are random selection methods and thus we draw 100 different SRS and PPS independent samples at each step r.

Simulation results: $N = 10^6$; $n = 1000$

For each subsample we compute the D- and A-efficiencies, and the OLS estimates of the coefficients in the linear model:

- D- and A-efficiencies of the Big Dataset and the subsamples obtained using ODB, IBOSS, EXCH, SRS and PPS:

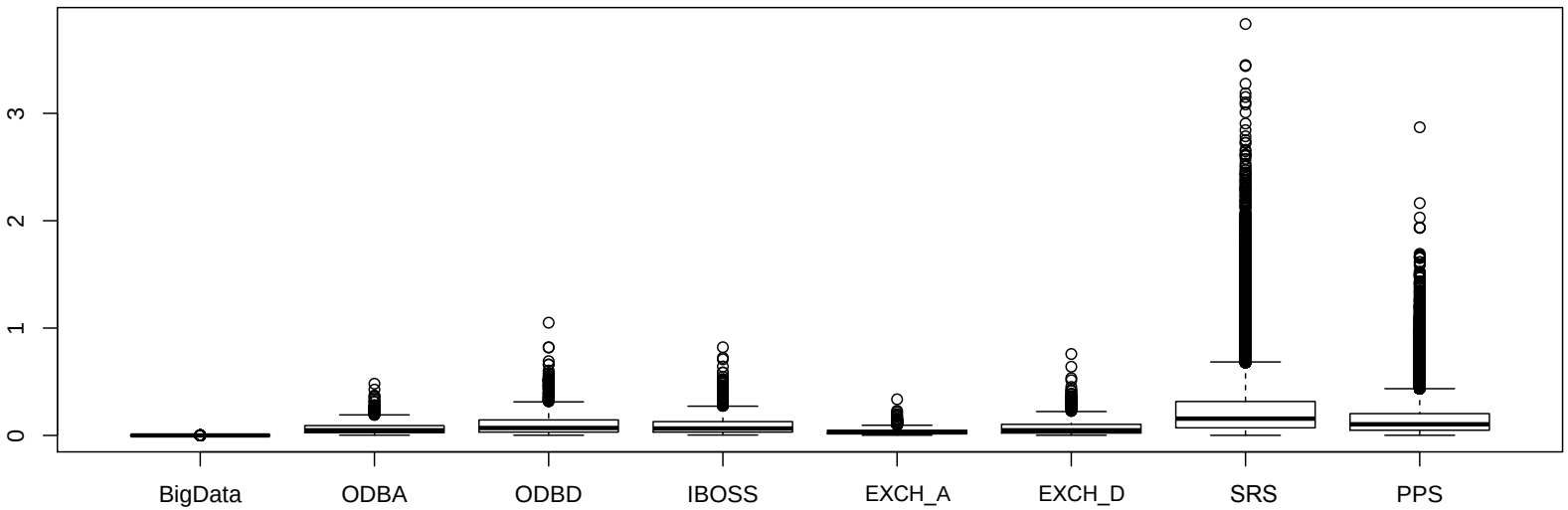
Φ-Efficiency	Big Data	ODB	IBOSS	EXCH	SRS	PPS
D-Efficiency	0.1278	0.4598	0.5042	0.6251	0.1263	0.2214
A-Efficiency	0.1441	0.5014	0.3452	0.8626	0.1416	0.2196

- Determinant and trace of the Monte Carlo covariance matrix of the estimates:

Criterion	Big Data	ODB	IBOSS	EXCH	SRS	PPS
Determinant	1.0e-62	9.2e-36	3.5e-36	3.5e-37	1.6e-29	3.6e-32
Trace	0.0002	0.06755	0.10132	0.03737	0.23987	0.15408

BOXPLOT of $\|\hat{\theta} - \theta\|^2$

Boxplots of the squared Euclidean distance between θ and the OLS estimates based on the full dataset and on the subsamples selected by the different strategies:



Example 2: Prediction accuracy

- If the major goal is prediction accuracy then the I-criterion should guide the subsample selection from the dataset:

$$I[\boldsymbol{M}_s] = n\mathrm{Tr}\left[(\boldsymbol{F}_s^T\boldsymbol{F}_s)^{-1}\int_{\mathcal{X}}\mathbf{f}(\boldsymbol{x})\mathbf{f}(\boldsymbol{x})^T\,d\boldsymbol{x}\right] = n\mathrm{Tr}\left[(\boldsymbol{F}_s^T\boldsymbol{F}_s)^{-1}\boldsymbol{F}_U^T\boldsymbol{F}_U\right].$$

- Same simulated predictors $X_1, \ldots, X_{10}$ as in Example 1 plus a categorical variable with 3 categories and marginal probabilities $(0.3, 0.4, 0.3)$. Let X_{11} and X_{12} be two dummies for the second and third categories.
- I-efficiency of the subsamples selected using ODB, IBOSS, EXCH, SRS and PPS, for two different sample sizes:

Sample size	ODB	IBOSS	EXCH	SRS	PPS
$n = 100$	0.8587	0.6526	0.9755	0.3608	0.4314
$n = 1000$	0.7262	0.6915	0.9311	0.3941	0.5313

- ODB and EXCH subsamples contain the largest per-unit information for prediction purposes, as they are based on the I-criterion.

Real data: Mortgage default (Drovandi et al., 2017)

1. **Logit Model**: $P(Y=1)=\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2+\theta_3 x_3+\theta_4 x_4)}}$
   Y: a 0/1 binary variable indicating whether or not the mortgage holder defaulted on the loan;
   x_1 : a credit rating;
   x_2 : number of years the mortgage holder has been employed.
   x_3 : the amount of credit card debt.
   x_4 : the age of the house.
   $N = 10^6$; sample size $n = 1000$.
2. It is well known that for nonlinear models the information matrix depends on the unknown parameter:

$$\mathbf{M}(\xi;\boldsymbol{\theta}) = \sum_{j=1}^{k} \tilde{\mathbf{f}}(x_j;\boldsymbol{\theta})\tilde{\mathbf{f}}(x_j;\boldsymbol{\theta})^T \omega_j, \quad \tilde{\mathbf{f}}(x_j;\boldsymbol{\theta}) = \frac{\sqrt{e^{x_j^T\boldsymbol{\theta}}}}{1+e^{x_j^T\boldsymbol{\theta}}}\, x_j.$$

3. Except for SRS all the sampling strategies depend on a guessed value for $\boldsymbol{\theta}$: s_θ.
4. $\boldsymbol{\theta}_0$ is the MLE of θ based on the full dataset: s_{θ_0}.
5. A- and D-efficiencies of for different guessed values of $\boldsymbol{\theta}$:

Guessed value	ODB	IBOSS	EXCH	SRS	PPS
	A-Efficiency				
$\boldsymbol{\theta}_0$	0.1212	0.0617	0.0076	0.0008	0.1013
$2\,\boldsymbol{\theta}_0$	0.1268	0.0655	0.0067	0.0008	0.1655
$1.5\,\boldsymbol{\theta}_0$	0.0831	0.0634	0.0067	0.0008	0.1556
$0.5\,\boldsymbol{\theta}_0$	0.1040	0.0542	0.0046	0.0008	0.0297
	D-Efficiency				
$\boldsymbol{\theta}_0$	0.4187	0.2374	0.5002	0.0019	0.2164
$2\,\boldsymbol{\theta}_0$	0.4163	0.2312	0.4538	0.0019	0.3909
$1.5\,\boldsymbol{\theta}_0$	0.4178	0.2332	0.4816	0.0019	0.3429
$0.5\,\boldsymbol{\theta}_0$	0.3488	0.2306	0.4222	0.0019	0.0510

Conclusions

- The optimal design theory allows to measure the quality of the Big Dataset;
- When $N >> p$, optimal design selection strategies provide "informative" samples to obtain precise estimates or accurate predictions. They work better than SRS and PPS sample.
- IBOSS method is computationally superior to EXCH and ODB strategies;
- Through the local linearization approach EXCH and ODB methods can be applied also to GLM models;
- EXCH and ODB methods can be implemented for different optimality criteria.

References

- Campbell and Broderick (2019), *Automated scalable Bayesian inference via Hilbert coresets*, Journal of Machine Learning Research
- Drovandi C., Holmes C. McGree J.M., Mengersen K., Richardson S. and Ryan E. (2017). Principles of experimental design for Big Data analysis. Statistical Science.
- Harman R. and Filová L. (2019). Optimal Design: A toolbox for computing efficient designs of experiments, R package version 0.0.1, URL https://CRAN.R-project.org/package= OptimalDesign.
- Harman R., Filová L. and Richtárik P. (2020). A Randomized Exchange Algorithm for Computing Optimal Approximate Designs of Experiments, JASA.
- Ma, P. and Sun, X. (2015). Leveraging for big data regression. Wiley Interdisciplinary Review: Computational Statistics.
- Wang, H., Yang, M. and Stufken, J. (2018). Information-Based Optimal Subdata Selection for Big Data Linear Regression, JASA.
- Wang, Zhu and Ma (2018), *Optimal Subsampling for Large Sample Logistic Regression*, JASA.

THANKS FOR YOUR ATTENTION!