子集抽樣：大數據分析的省力神器

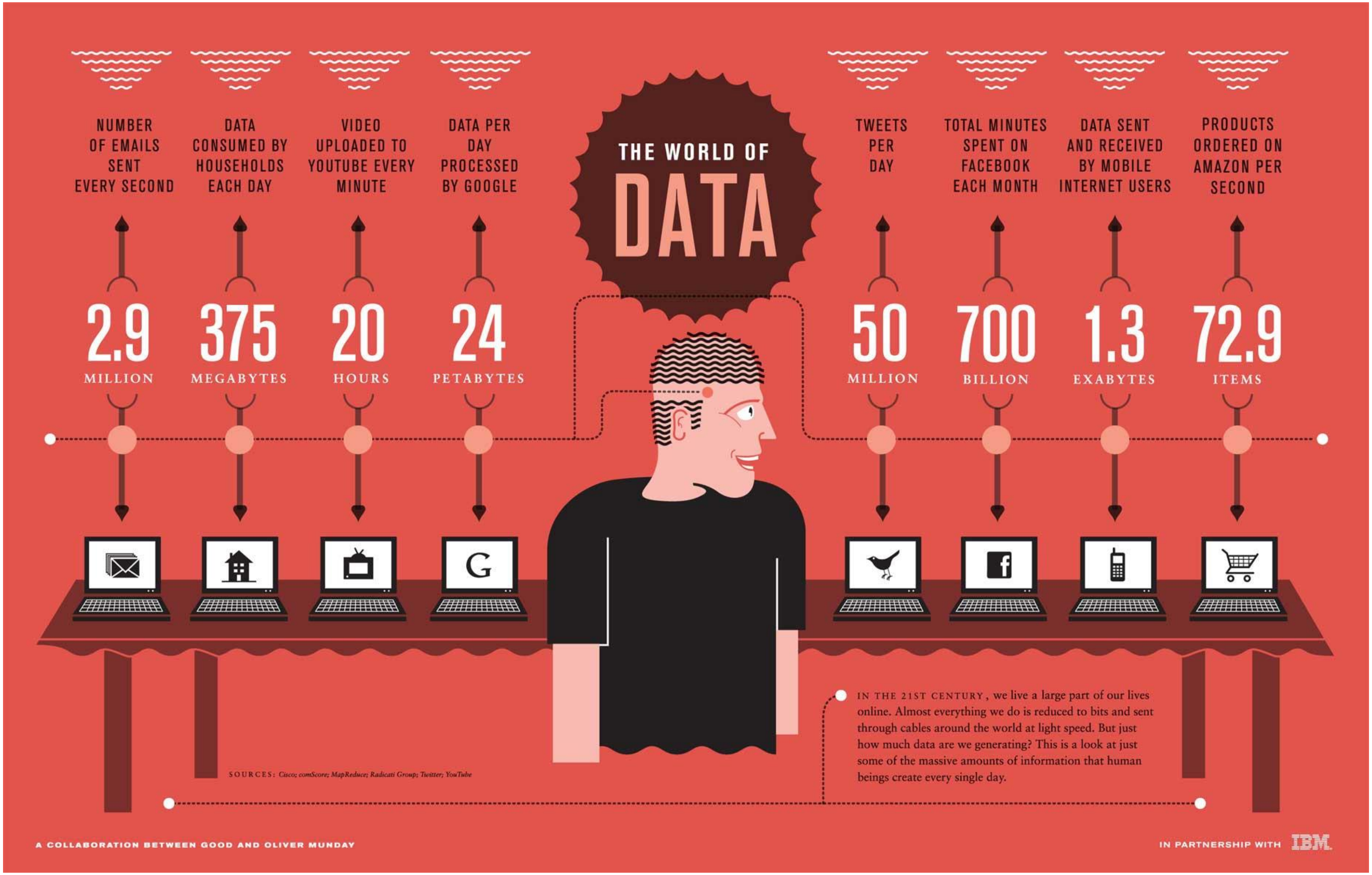
張明中

中央研究院統計科學研究所

STATISTICAL SCIENCE CAMP

統計科學營

2024.8.27(二)-8.28(三)


THE WORLD OF DATA
NUMBER OF EMAILS SENT EVERY SECOND
2.9 MILLION
DATA CONSUMED BY HOUSEHOLDS EACH DAY
375 MEGABYTES
VIDEO UPLOADED TO YOUTUBE EVERY MINUTE
20 HOURS
DATA PER DAY PROCESSED BY GOOGLE
24 PETABYTES
TWEETS PER DAY
50 MILLION
TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH
700 BILLION
DATA SENT AND RECEIVED BY MOBILE INTERNET USERS
1.3 EXABYTES
PRODUCTS ORDERED ON AMAZON PER SECOND
72.9 ITEMS
G
SOURCES: Cisco; comScore; MapReduce; Radicati Group; Twitter; YouTube
A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY
IN PARTNERSHIP WITH IBM

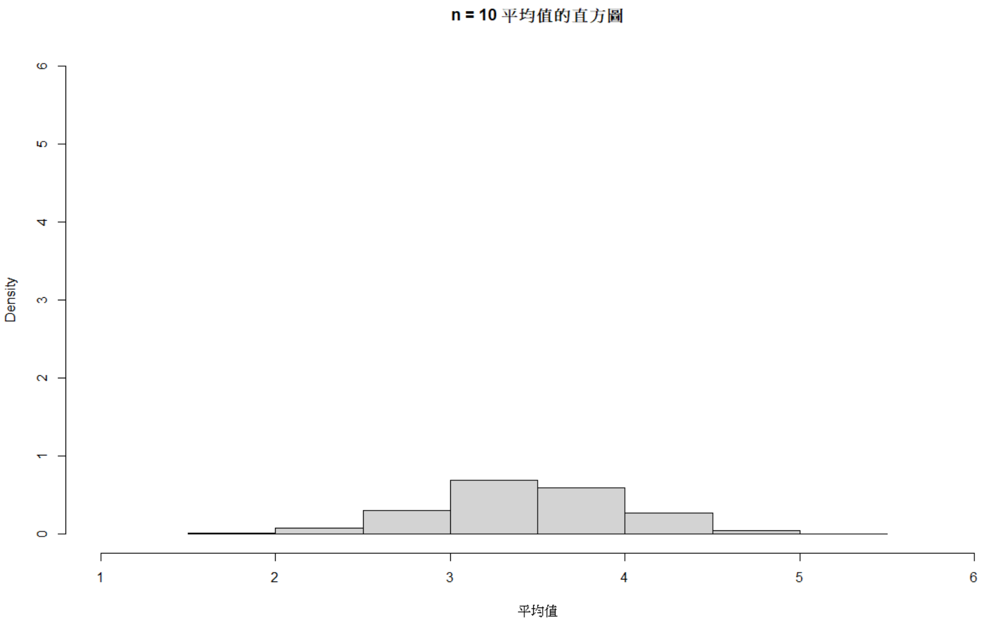
IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.

https://floyden.home.blog/2019/04/28/definition-of-big-data/

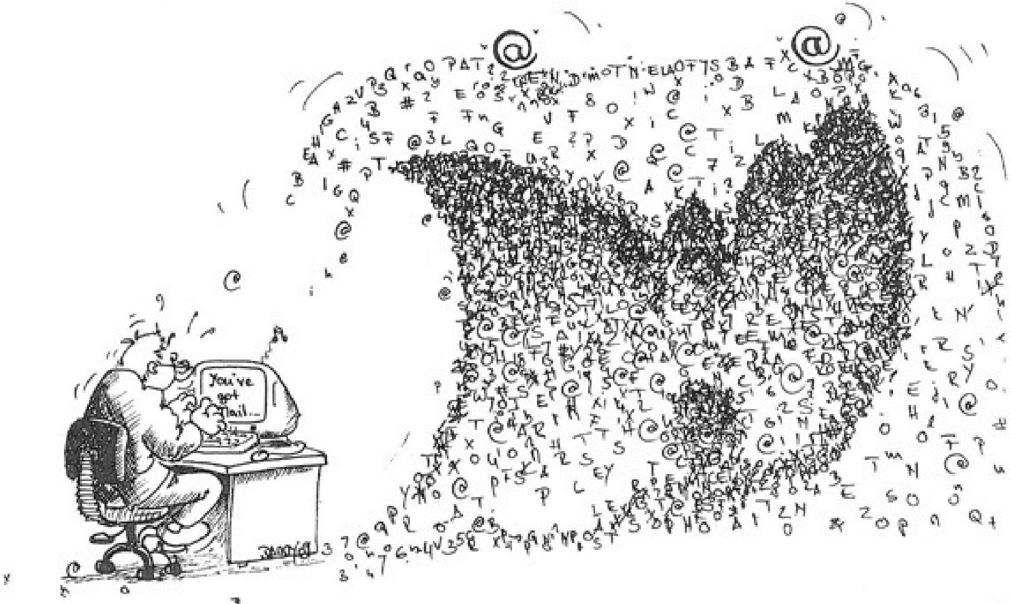
中央極限定理 https://youtu.be/SbRUtf-b3DA

Theorem 1 (Lindeberg–Lévy Central Limit Theorem). Let $\{X_n\}$ be a sequence of iid RVs with $0 < \mathrm{var}(X_n) = \sigma^2 < \infty$ and common mean μ. Let $S_n = \sum_{j=1}^{n} X_j$, $n = 1, 2, \ldots$. Then for every $x \in \mathcal{R}$

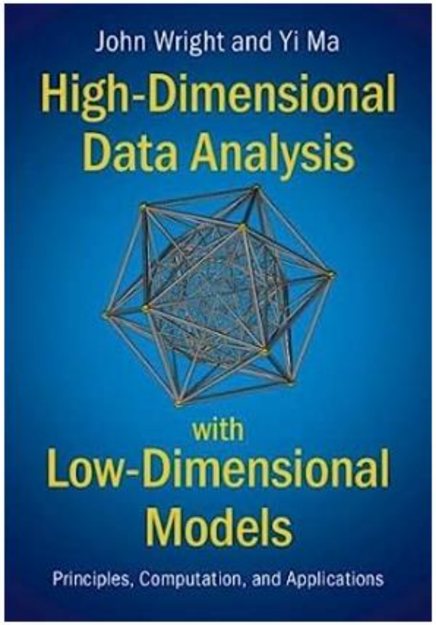
$$\lim_{n\to\infty} P\left\{\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right\} = \lim_{n\to\infty} P\left\{\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le x\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-u^2/2}du.$$



1



https://openviewpartners.com/blog/7-inconvenient-truths-about-big-data/


John Wright and Yi Ma

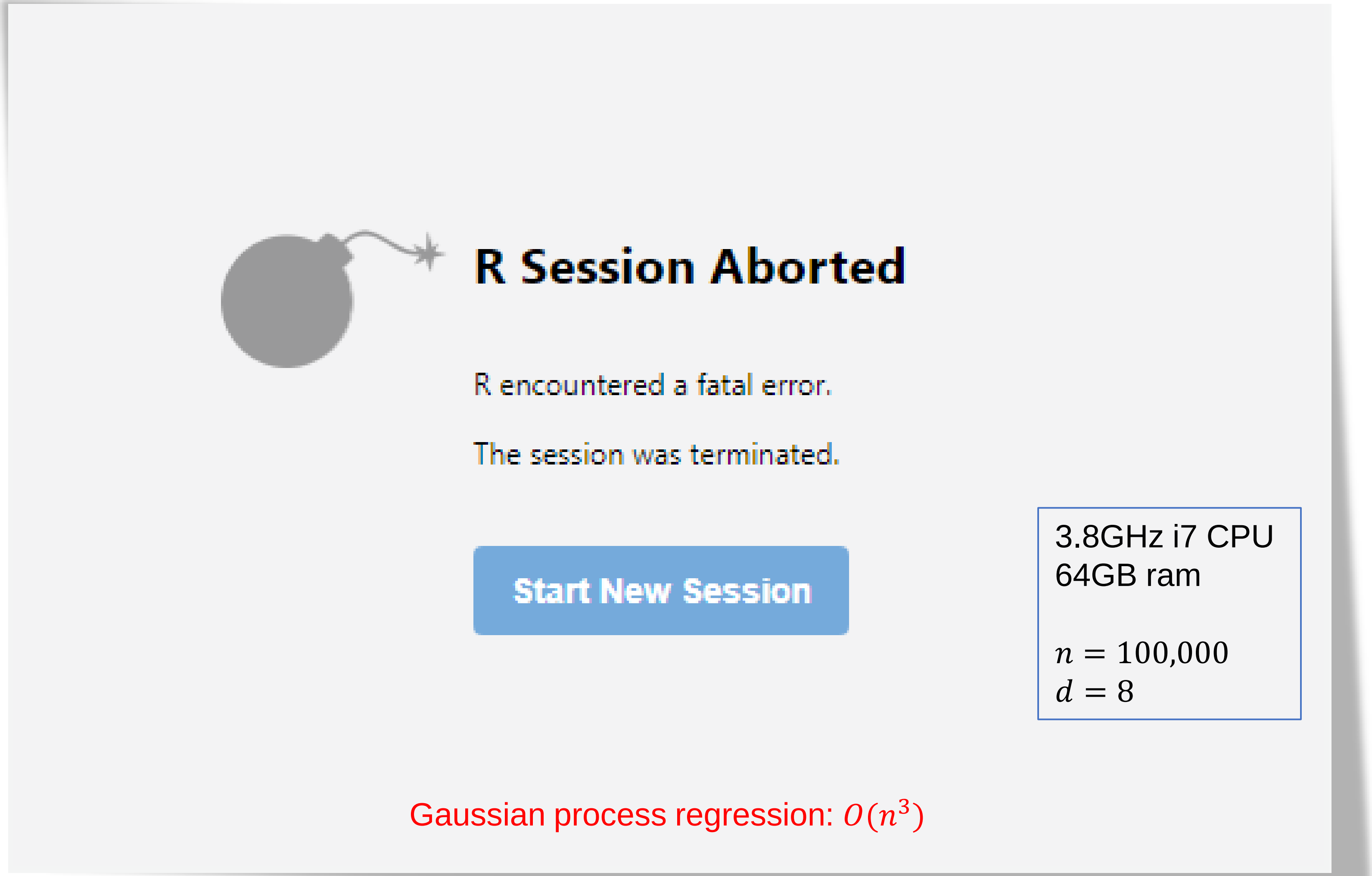
High-Dimensional Data Analysis with Low-Dimensional Models

Principles, Computation, and Applications

Nevertheless, data-rich does not necessarily imply "information-rich," at least not for free. Massive amounts of data are being collected, sometimes without

$$\textbf{Data} \quad = \quad \textbf{Information} \; + \; \textbf{Irrelevant Data}.$$

2



R Session Aborted

R encountered a fatal error.

The session was terminated.

3.8GHz i7 CPU
64GB ram

$n = 100{,}000$
$d = 8$

Gaussian process regression: $O(n^3)$

3

B. HAALAND AND P. Z. G. QIAN

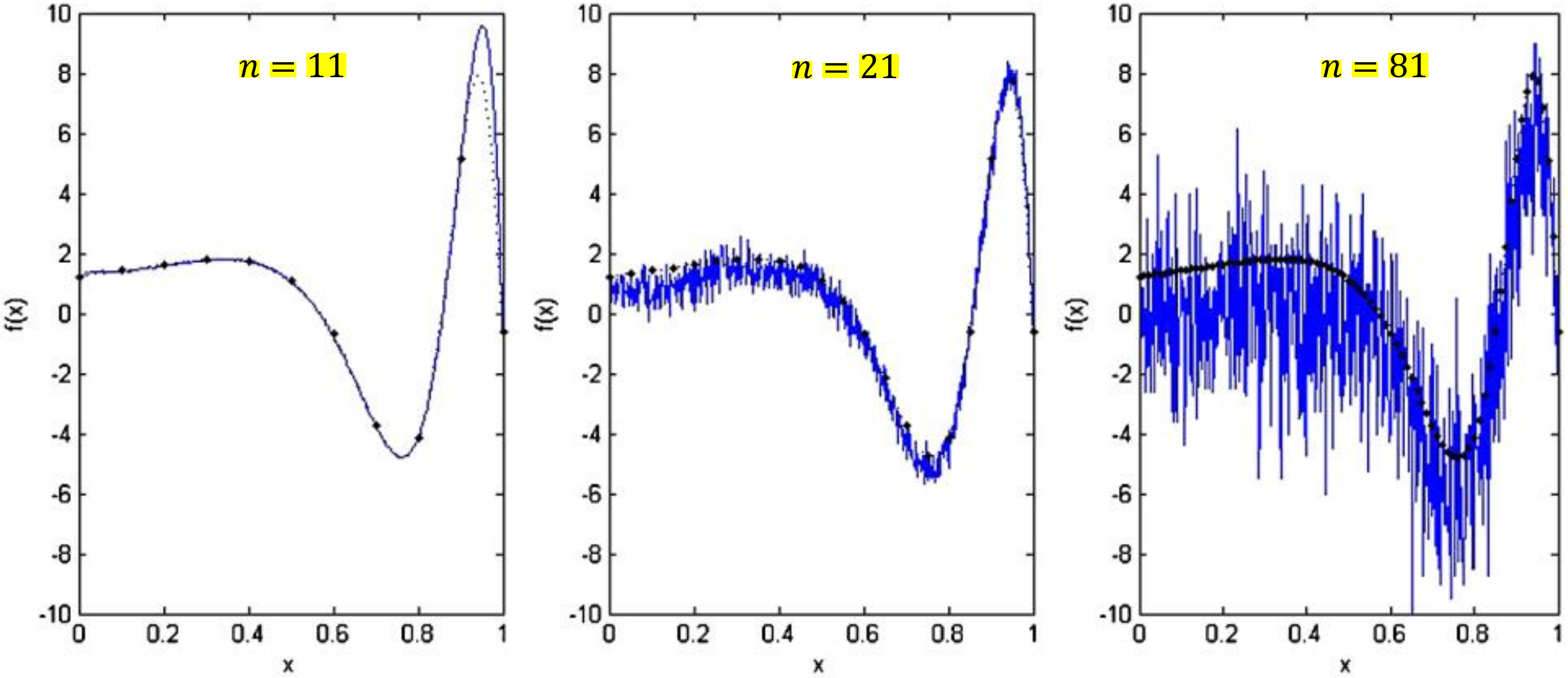


FIG. 1. *Panels 1–3: interpolator in solid blue and actual function in dotted black with collected data indicated by black dots.*

$$f(x) = \exp\{(x + 1/2)^2\} \sin(\exp\{(x + 1/2)^2\})$$

$$\Phi(x - y) = \exp\{-(x - y)^2\}$$

Ben Haaland. Peter Z. G. Qian. "Accurate emulators for large-scale computer experiments." Ann. Statist. 39 (6) 2974 - 3002, December 2011. https://doi.org/10.1214/11-AOS929

The more the better?

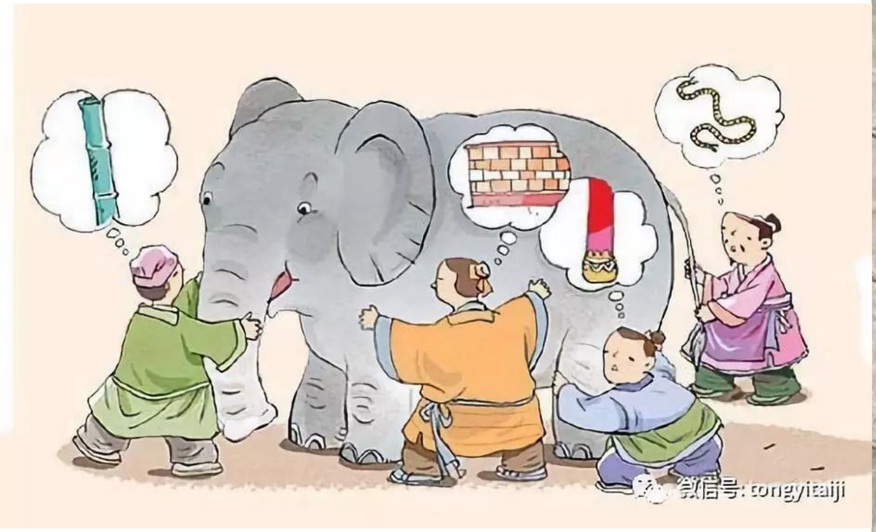
- Variable selection

- Sufficient statistics

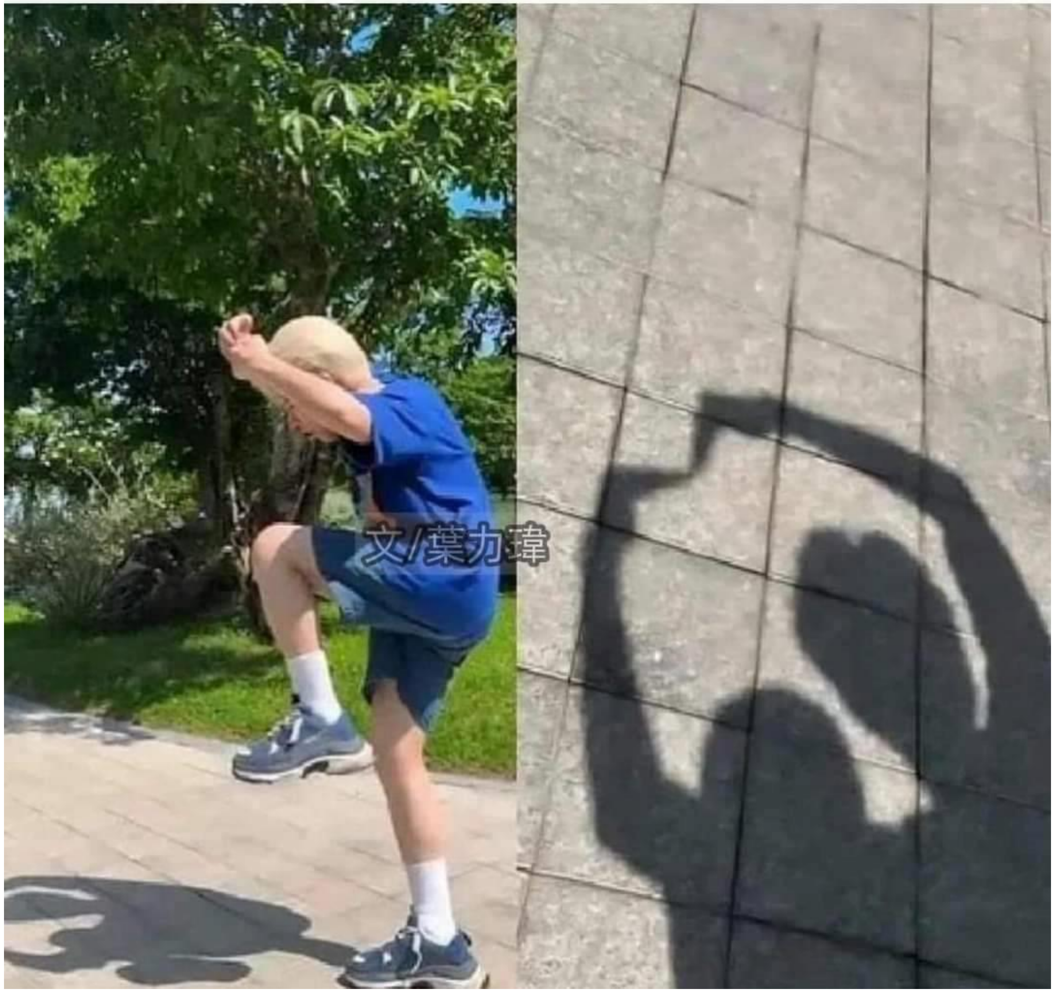
- Max pooling layer

- Stochastic gradient descent

Representative



如何一個人完成兩個人的約會



她的左臉
她的右臉
：MRTV：香日台

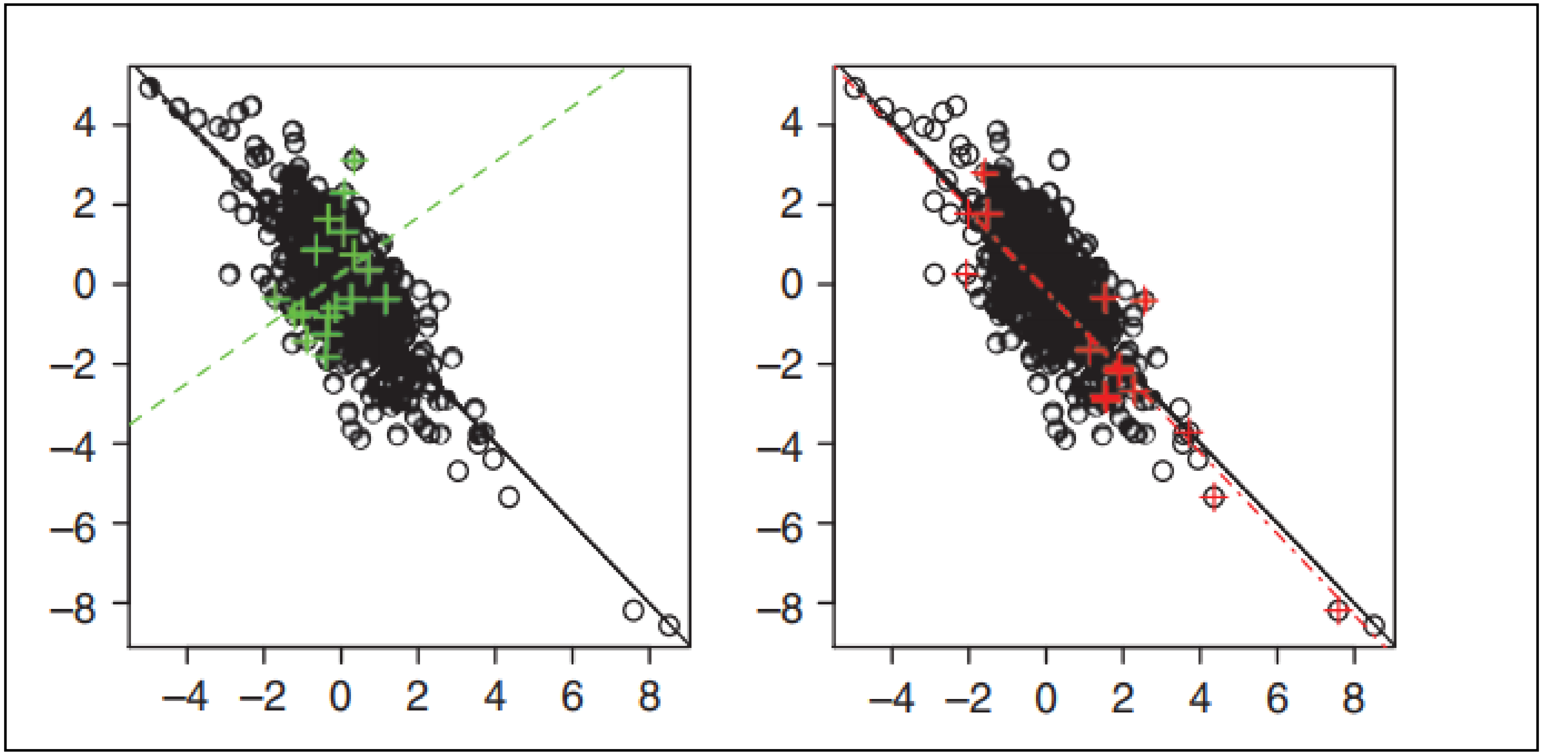
她的左臉
她的右臉
MRTV 香日台
她的正臉

What Information?

- Data-rich ≠ Information-rich
- Information is objective-dependent
- Here, two types of information from the data:
  - **Conditional distribution**
  - **Joint distribution**

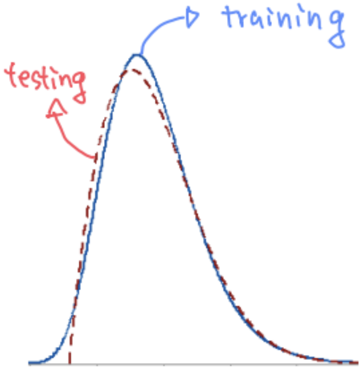
Conditional Distribution of $X_1|X_2,\ldots,X_p$

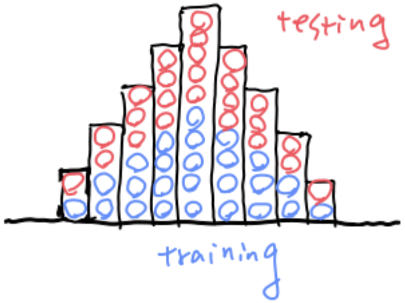
Ma et al. (2015)

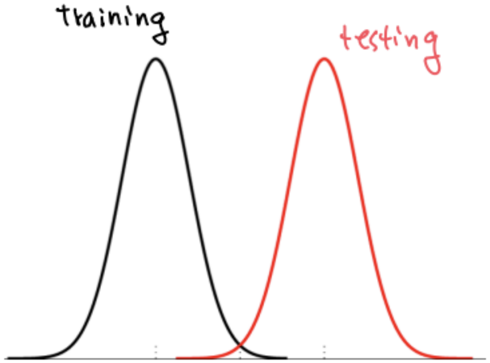


Joint Distribution of $X_1, \dots, X_p$

- What is similar?







1

Two Objectives

- Conditional distribution of $X_1 | X_2, \ldots, X_p$
  - Model-dependent subsampling
  - Model-free subsampling

- Joint distribution of $X_1, \ldots, X_p$
  - Model-free subsampling

Conditional Distribution

Model-dependent

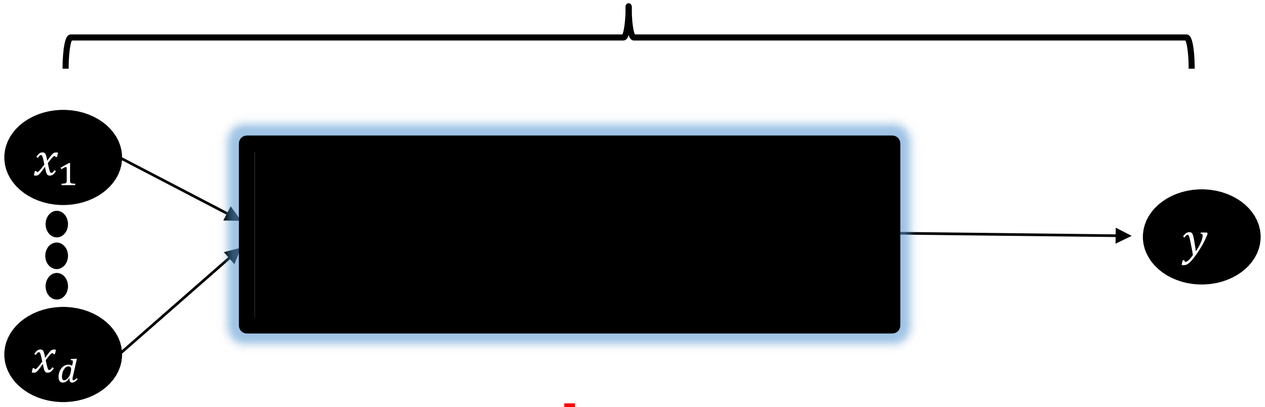
- Wang, H., Yang, M., and Stufken, J. (2019), “Information-based Optimal Subdata Selection for Big Data Linear Regression,” *Journal of the American Statistical Association*, 114, 393–405
- M.-C. Chang (2023). ``Predictive Subdata Selection for Computer Models’’, *Journal of Computational and Graphical Statistics*, 32, 613-630. DOI: 10.1080/10618600.2022.2097247. [SPEC科學推展中心報導 https://spec.ntu.edu.tw/20230901-research-math/]

Model-free

- Joseph, V. R., and Mak, S. (2021), “Supervised Compression of Big Data,” *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 14, 217–229.
- M.-C. Chang (2024). ``Supervised Stratified Subsampling for Predictive Analytics’’, *Journal of Computational and Graphical Statistics*, to appear. DOI:10.1080/10618600.2024.2304075.

Regression problem / Supervised learning

Unveil the **blackbox** among features/variables



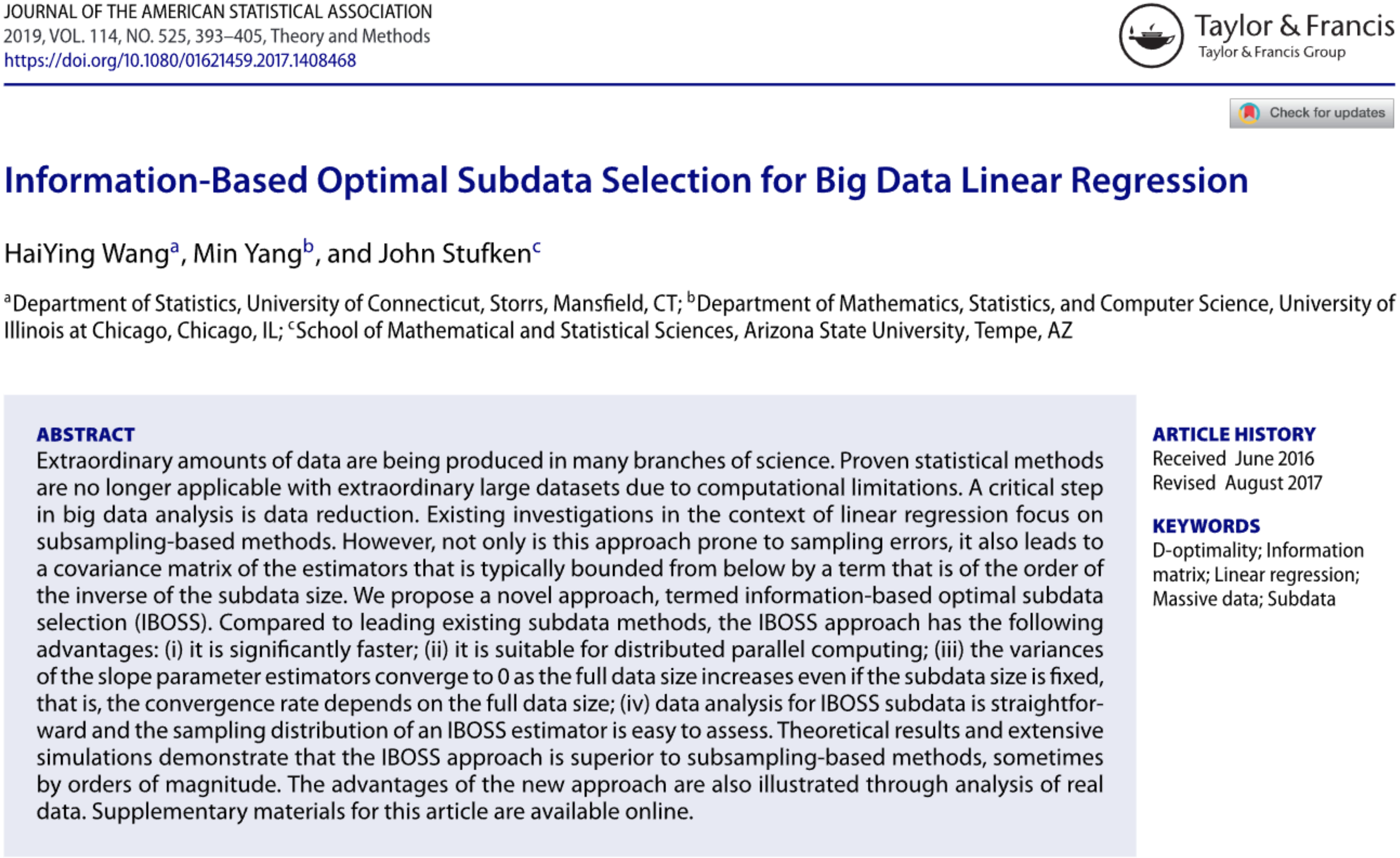
THE WORLD OF DATA

Statistical model fitting
(*Linear model*, *Gaussian process regression*, etc.)

$$f(\mathbf{x}) = -\sum_{i=1}^{d} \sin(x_i)\sin^{2m}\left(\frac{ix_i^2}{\pi}\right)$$

Conditional Distribution

Model-Dependent Subsampling



JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 114, NO. 525, 393–405, Theory and Methods
https://doi.org/10.1080/01621459.2017.1408468

Taylor & Francis
Taylor & Francis Group

Check for updates

Information-Based Optimal Subdata Selection for Big Data Linear Regression

HaiYing Wang[a], Min Yang[b], and John Stufken[c]

[a]Department of Statistics, University of Connecticut, Storrs, Mansfield, CT; [b]Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, IL; [c]School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ

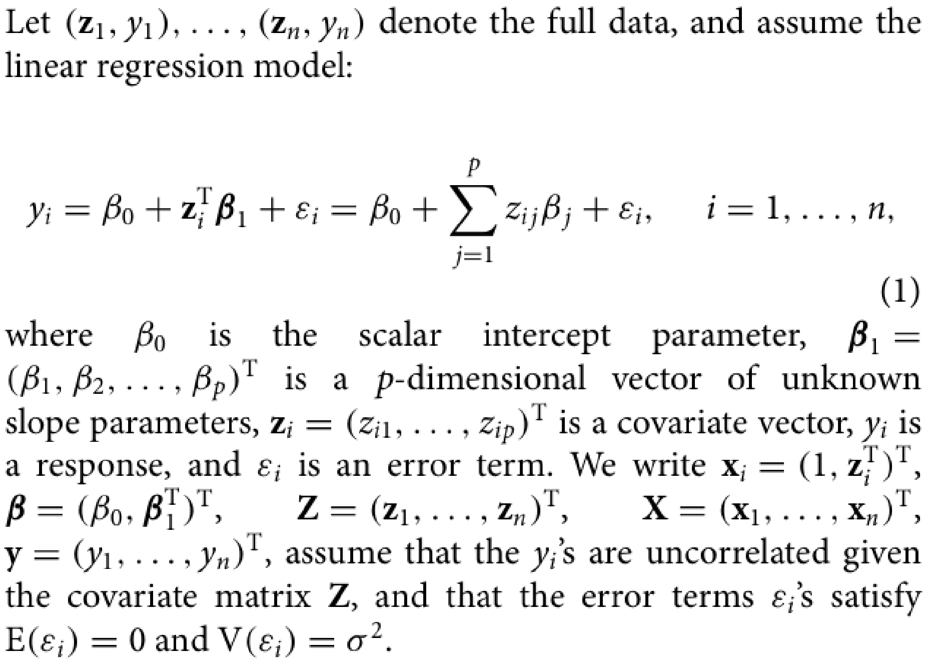
ABSTRACT

Extraordinary amounts of data are being produced in many branches of science. Proven statistical methods are no longer applicable with extraordinary large datasets due to computational limitations. A critical step in big data analysis is data reduction. Existing investigations in the context of linear regression focus on subsampling-based methods. However, not only is this approach prone to sampling errors, it also leads to a covariance matrix of the estimators that is typically bounded from below by a term that is of the order of the inverse of the subdata size. We propose a novel approach, termed information-based optimal subdata selection (IBOSS). Compared to leading existing subdata methods, the IBOSS approach has the following advantages: (i) it is significantly faster; (ii) it is suitable for distributed parallel computing; (iii) the variances of the slope parameter estimators converge to 0 as the full data size increases even if the subdata size is fixed, that is, the convergence rate depends on the full data size; (iv) data analysis for IBOSS subdata is straightforward and the sampling distribution of an IBOSS estimator is easy to assess. Theoretical results and extensive simulations demonstrate that the IBOSS approach is superior to subsampling-based methods, sometimes by orders of magnitude. The advantages of the new approach are also illustrated through analysis of real data. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2016
Revised August 2017

KEYWORDS
D-optimality; Information matrix; Linear regression; Massive data; Subdata

2. The Framework



Let $(\mathbf{z}_1, y_1), \ldots, (\mathbf{z}_n, y_n)$ denote the full data, and assume the linear regression model:

$$y_i = \beta_0 + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \varepsilon_i = \beta_0 + \sum_{j=1}^{p} z_{ij}\beta_j + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where β_0 is the scalar intercept parameter, $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$ is a p-dimensional vector of unknown slope parameters, $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})^{\mathrm{T}}$ is a covariate vector, y_i is a response, and ε_i is an error term. We write $\mathbf{x}_i = (1, \mathbf{z}_i^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^{\mathrm{T}}$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$, $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, assume that the y_i's are uncorrelated given the covariate matrix $\mathbf{Z}$, and that the error terms ε_i's satisfy $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{V}(\varepsilon_i) = \sigma^2$.

$$\Downarrow$$

$$\hat{\boldsymbol{\beta}}_f = \left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i$$

$$\Downarrow$$

$$\mathbf{M}_f = \frac{1}{\sigma^2}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$$

In our framework, for given full data of size n, the D-optimality criterion suggests the selection of subdata of size k so that

$$\boldsymbol{\delta}_{\mathrm{D}}^{\mathrm{opt}} = \arg\max_{\boldsymbol{\delta}} \left| \sum_{i=1}^{n} \delta_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \right|, \quad \sum_{i=1}^{n} \delta_i = k.$$

Obtaining an exact solution is computationally far too expensive. In working toward an approximate solution, we first derive an upper bound for $|\mathbf{M}(\boldsymbol{\delta})|$ which, while only attainable for very special cases, will guide our later algorithm.

Theorem 2 (D-optimality). For subdata of size k represented by $\boldsymbol{\delta}$,

$$|\mathbf{M}(\boldsymbol{\delta})| \leq \frac{k^{p+1}}{4^p \sigma^{2(p+1)}} \prod_{j=1}^{p} (z_{(n)j} - z_{(1)j})^2, \qquad (17)$$

where $z_{(n)j} = \max\{z_{1j}, z_{2j}, \ldots, z_{nj}\}$ and $z_{(1)j} = \min\{z_{1j}, z_{2j}, \ldots, z_{nj}\}$ are the nth and first-order statistics of $z_{1j}, z_{2j}, \ldots, z_{nj}$. If the subdata consists of the 2^p points $(a_1, \ldots, a_p)^{\mathrm{T}}$ where $a_j = z_{(n)j}$ or $z_{(1)j}$, $j = 1, 2, \ldots, p$, each occurring equally often, then equality holds in (17).

For the result in (22), it can be shown from the proof that

$$\mathrm{V}(\hat{\beta}_j^{\mathrm{D}}|\mathbf{Z}) \asymp_P \frac{p}{k\left\{F_j^{-1}(1-n^{-1}) - F_j^{-1}(n^{-1})\right\}^2}, \qquad j = 1, \ldots, p.$$

What we find more interesting is the fact that, when k is fixed, from Theorems 2.8.1 and 2.8.2 in Galambos (1987), $z_{(n-r+1)j} - z_{(r)j}$ goes to infinity with the same rate as that of $z_{(n)j} - z_{(1)j}$. Thus, the order of the variance of a slope estimator is the inverse of the squared full data sample range for the corresponding covariate. If the sample range goes to ∞ as $n \to \infty$, then the variance converges to 0 even when the subdata size k is fixed. This suggests that subdata may preserve information at a scale related to the full data size. We will return to this for specific cases with more details. In the remainder of this section, we focus on the case that both p and k are fixed.

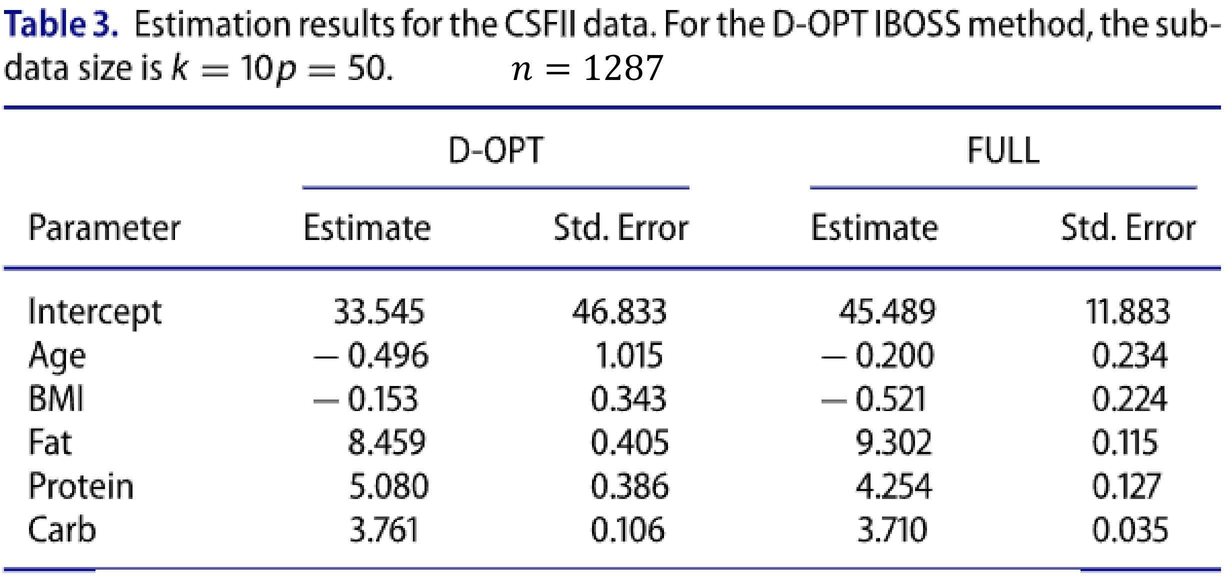


Table 3. Estimation results for the CSFII data. For the D-OPT IBOSS method, the subdata size is $k = 10p = 50$. $n = 1287$

	D-OPT		FULL	
Parameter	Estimate	Std. Error	Estimate	Std. Error
Intercept	33.545	46.833	45.489	11.883
Age	− 0.496	1.015	− 0.200	0.234
BMI	− 0.153	0.343	− 0.521	0.224
Fat	8.459	0.405	9.302	0.115
Protein	5.080	0.386	4.254	0.127
Carb	3.761	0.106	3.710	0.035

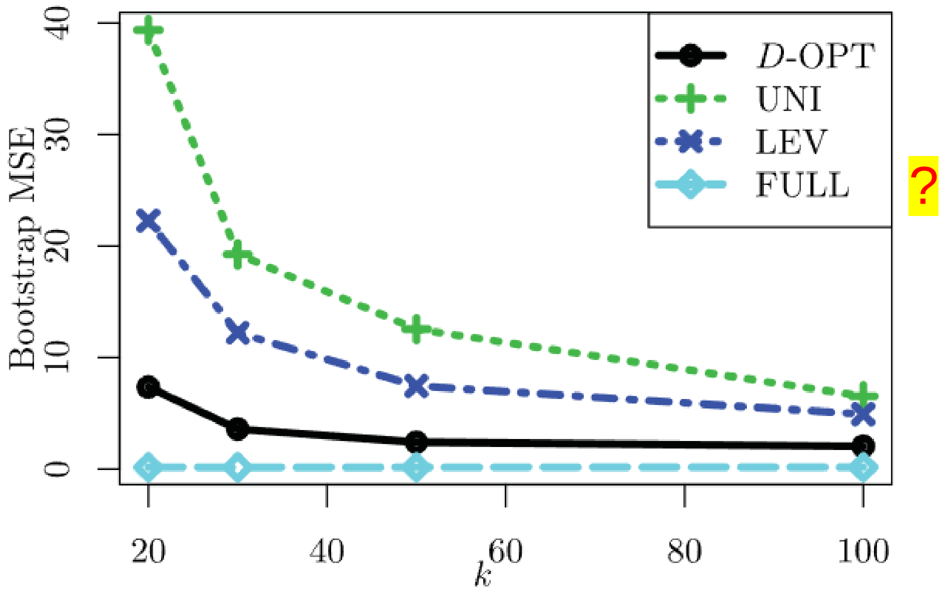


Figure 6. MSEs for estimating slope parameters for the CSFII data. They are computed from 1000 bootstrap samples.

Gaussian Process Regression

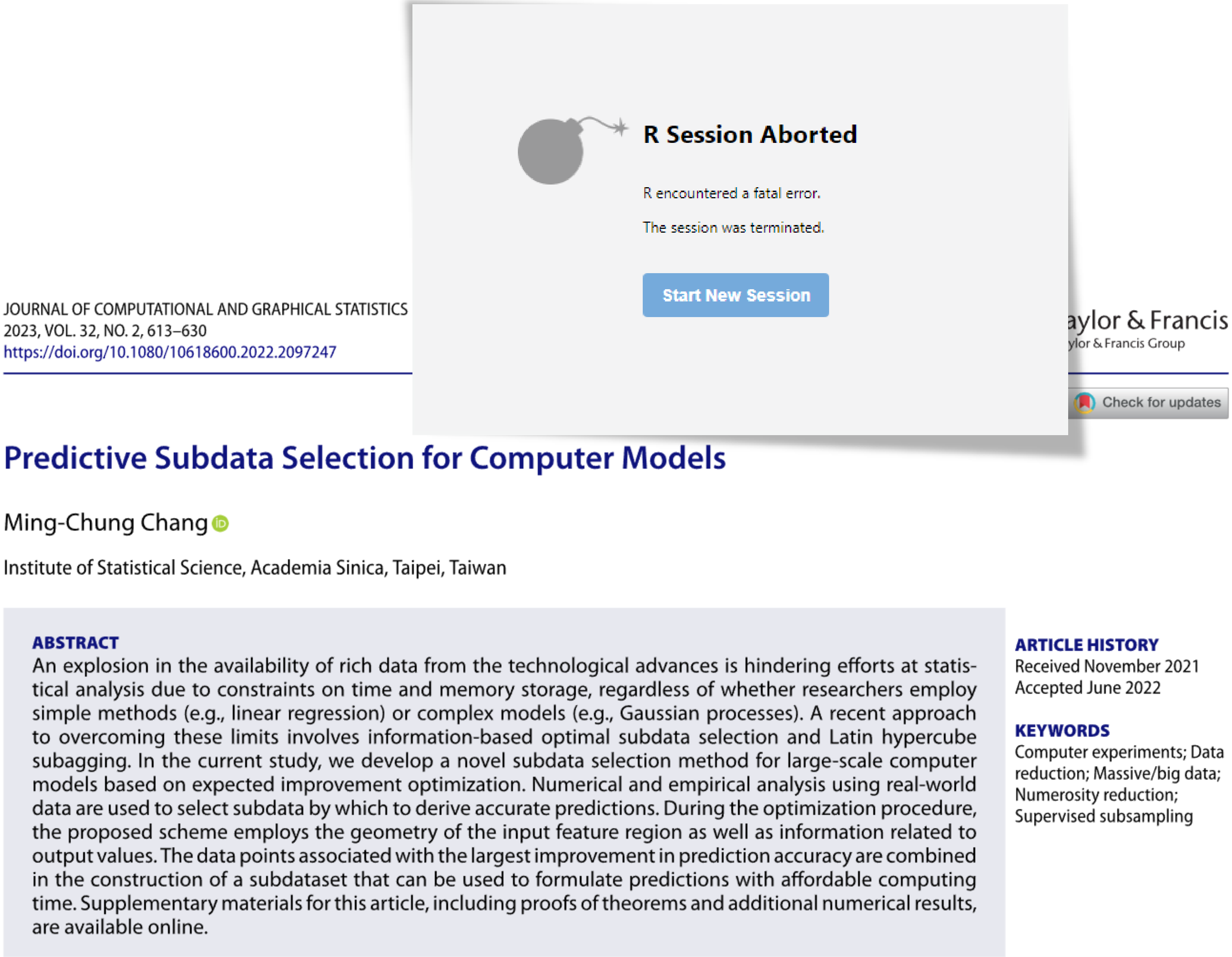
- Gaussian process regression (GPR):

$$y = \sum_{j=1}^{p} \beta_j g_j(\boldsymbol{x}) + Z(\boldsymbol{x}) + \varepsilon$$

- $Z(\boldsymbol{x})$ is a Gaussian process (GP)
  - For any $L \geq 1$, any choice of $\boldsymbol{x}_1, \dots, \boldsymbol{x}_L$, the vector $(Z(\boldsymbol{x}_1), \dots, Z(\boldsymbol{x}_L))$ has a multivariate normal distribution
  - Determined by its *mean function* and *covariance function*

- $\mu(\boldsymbol{x}) = \mathrm{E}[Z(\boldsymbol{x})]$ and $C(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{cov}(Z(\boldsymbol{x}), Z(\boldsymbol{x}'))$

- **Stationary**: the joint distribution of $Z(\boldsymbol{x}_1), \dots, Z(\boldsymbol{x}_L)$ is identical to that of $Z(\boldsymbol{x}_1 + \boldsymbol{h}), \dots, Z(\boldsymbol{x}_L + \boldsymbol{h})$ for any L, any $\boldsymbol{x}_1, \dots, \boldsymbol{x}_L$, any $\boldsymbol{h}$



R Session Aborted

R encountered a fatal error.

The session was terminated.

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2023, VOL. 32, NO. 2, 613–630
https://doi.org/10.1080/10618600.2022.2097247

Predictive Subdata Selection for Computer Models

Ming-Chung Chang

Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

ABSTRACT

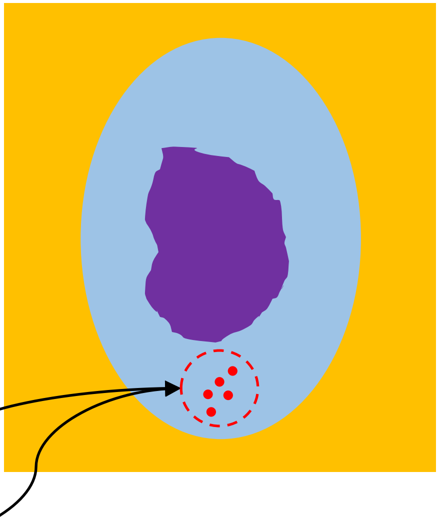
An explosion in the availability of rich data from the technological advances is hindering efforts at statistical analysis due to constraints on time and memory storage, regardless of whether researchers employ simple methods (e.g., linear regression) or complex models (e.g., Gaussian processes). A recent approach to overcoming these limits involves information-based optimal subdata selection and Latin hypercube subagging. In the current study, we develop a novel subdata selection method for large-scale computer models based on expected improvement optimization. Numerical and empirical analysis using real-world data are used to select subdata by which to derive accurate predictions. During the optimization procedure, the proposed scheme employs the geometry of the input feature region as well as information related to output values. The data points associated with the largest improvement in prediction accuracy are combined in the construction of a subdataset that can be used to formulate predictions with affordable computing time. Supplementary materials for this article, including proofs of theorems and additional numerical results, are available online.

ARTICLE HISTORY
Received November 2021
Accepted June 2022

KEYWORDS
Computer experiments; Data reduction; Massive/big data; Numerosity reduction; Supervised subsampling

Methodology

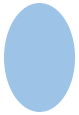
- Expected improvement (EI) optimization
- Different from the traditional EI
  - Purpose → **Prediction**
  - **Output values available** outside the subdata



- The proposed criterion for subdata selection
  - $\boldsymbol{\eta}_0$: GPR outputs at $\boldsymbol{x}_0 = (x_{i_1}, \dots, x_{i_k})$
  - $\boldsymbol{t}_0$: true outputs (known) at $\boldsymbol{x}_0 = (x_{i_1}, \dots, x_{i_k})$
  - Discrepancy function: $I(\boldsymbol{x}_0) = (\boldsymbol{\eta}_0 - \boldsymbol{t}_0)^T(\boldsymbol{\eta}_0 - \boldsymbol{t}_0)$
  - Conditional expectation:

$$\begin{aligned} \mathrm{E}[I(\boldsymbol{x}_0)|\boldsymbol{y}] &= \mathrm{tr}\left(\boldsymbol{\Sigma}_{\eta}\right) + (\mathrm{E}[\boldsymbol{\eta}_0|\boldsymbol{y}] - \boldsymbol{t}_0)^T(\mathrm{E}[\boldsymbol{\eta}_0|\boldsymbol{y}] - \boldsymbol{t}_0) \\ &= \mathrm{tr}\left(\boldsymbol{\Sigma}_{\eta}\right) + (\mathrm{E}[\boldsymbol{\eta}_0|\boldsymbol{y}] - \boldsymbol{y}_0)^T(\mathrm{E}[\boldsymbol{\eta}_0|\boldsymbol{y}] - \boldsymbol{y}_0) + \mathrm{tr}(\boldsymbol{\Sigma}_{\epsilon}) \end{aligned}$$

Methodology

$\boldsymbol{x}_0 = \left(x_{i_1}, \dots, x_{i_k}\right) \subset$  $-$

$$\mathrm{E}[I(\boldsymbol{x}_0)|\boldsymbol{y}] = \mathrm{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\right) + (\mathrm{E}[\boldsymbol{\eta}_0|\boldsymbol{y}] - \boldsymbol{y}_0)^T(\mathrm{E}[\boldsymbol{\eta}_0|\boldsymbol{y}] - \boldsymbol{y}_0) + \mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$$

⬇

$$\sum_{j=1}^{k} \{\text{prediction var.} + \text{prediction error}^2 + \text{noise var.}\}_{\text{at } x_{i_j}}$$

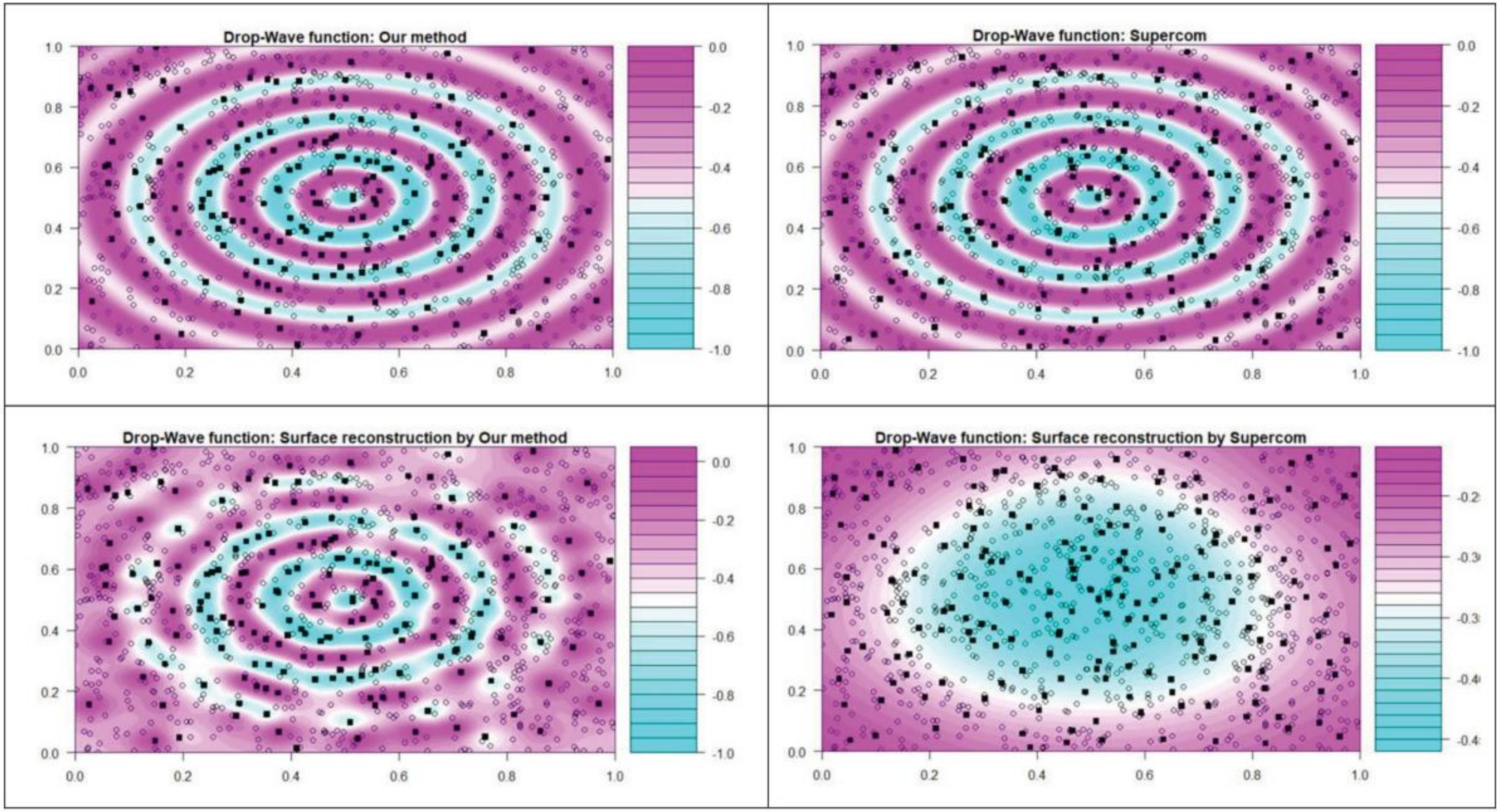
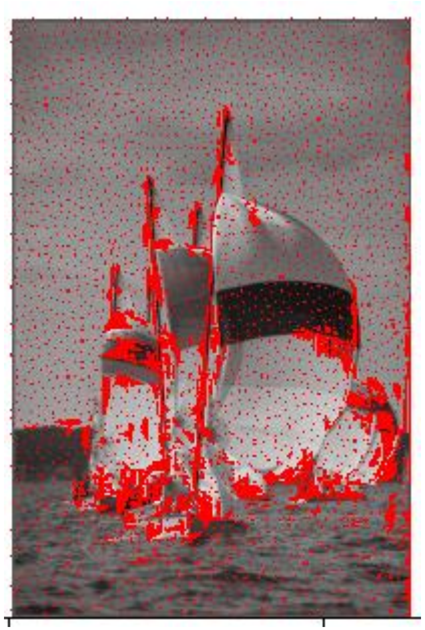


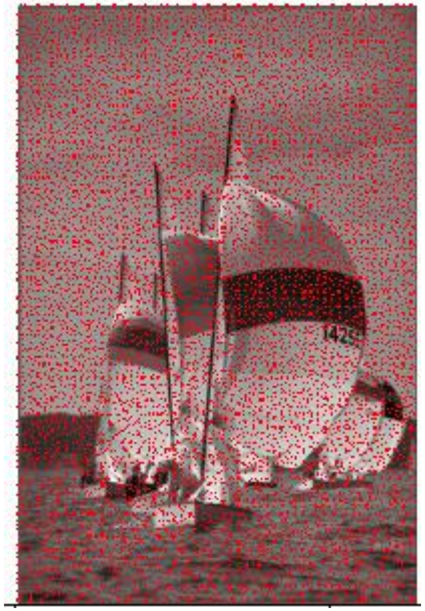
Figure 9. Contour of the Drop-Wave function: black squares for subdata. Upper: true contour. Bottom: fitted contour.

Image Data

98304 data points



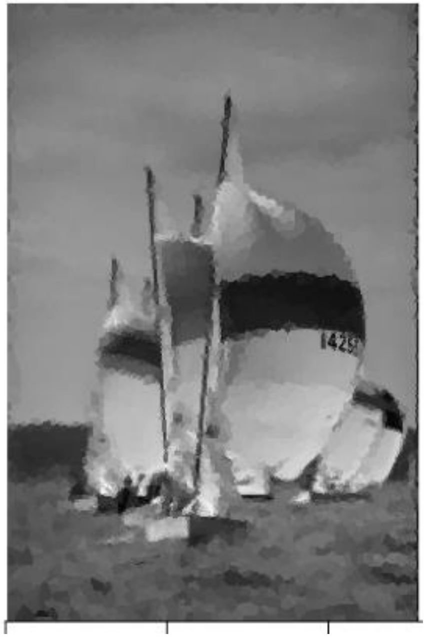
10000 subdata points
~ Proposed



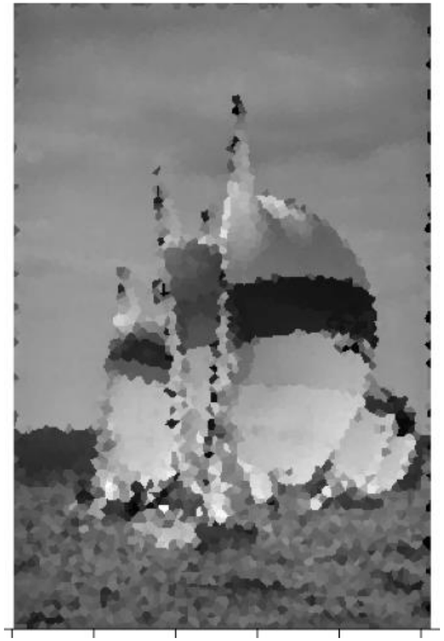
10000 subdata points
~ SPlit

Image Data

98304 data points



Reconstruction using 10000 (Proposed) subdata points



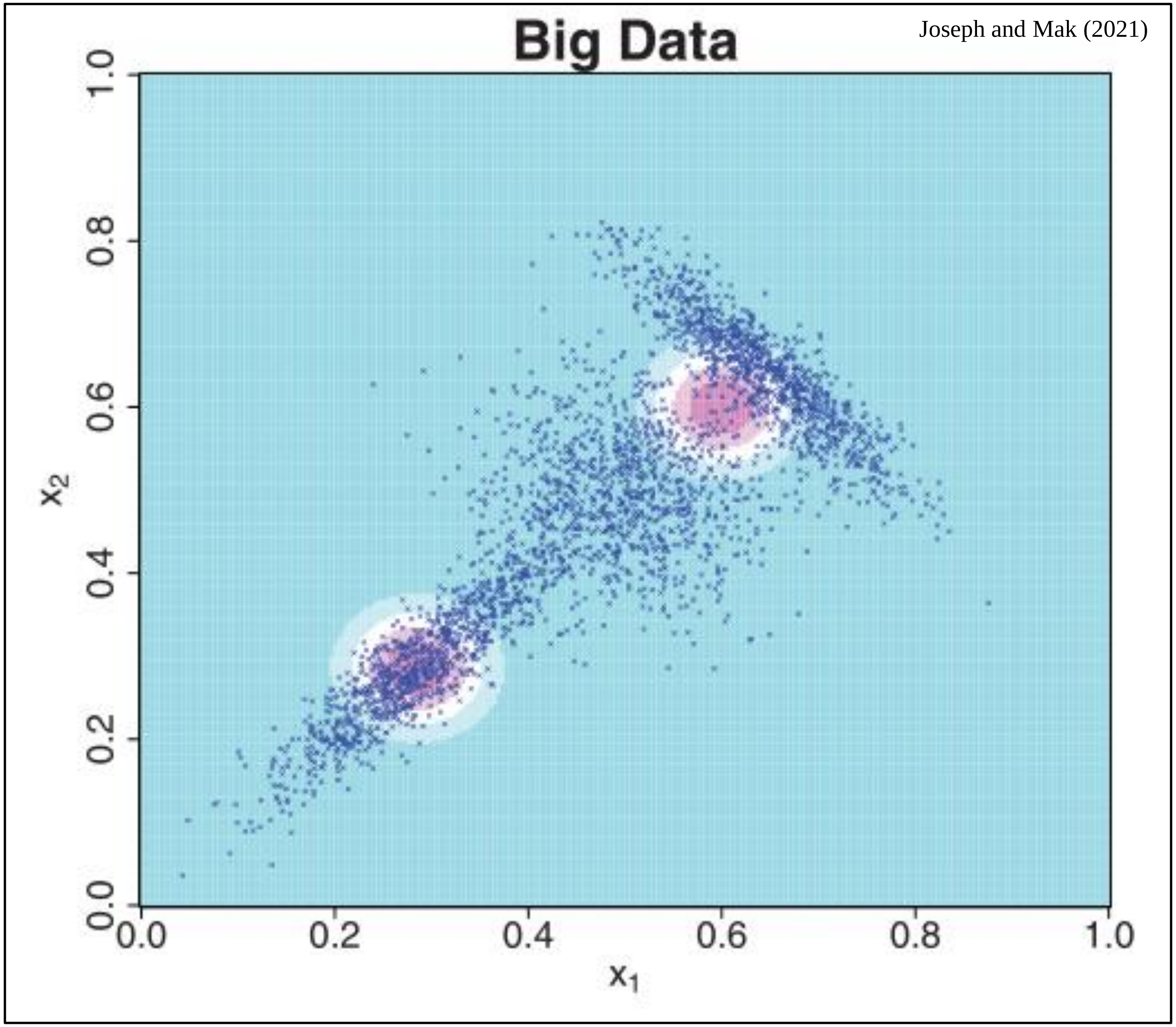
Reconstruction using 10000 (SPlit) subdata points

Fitted by ***partitioning estimate***

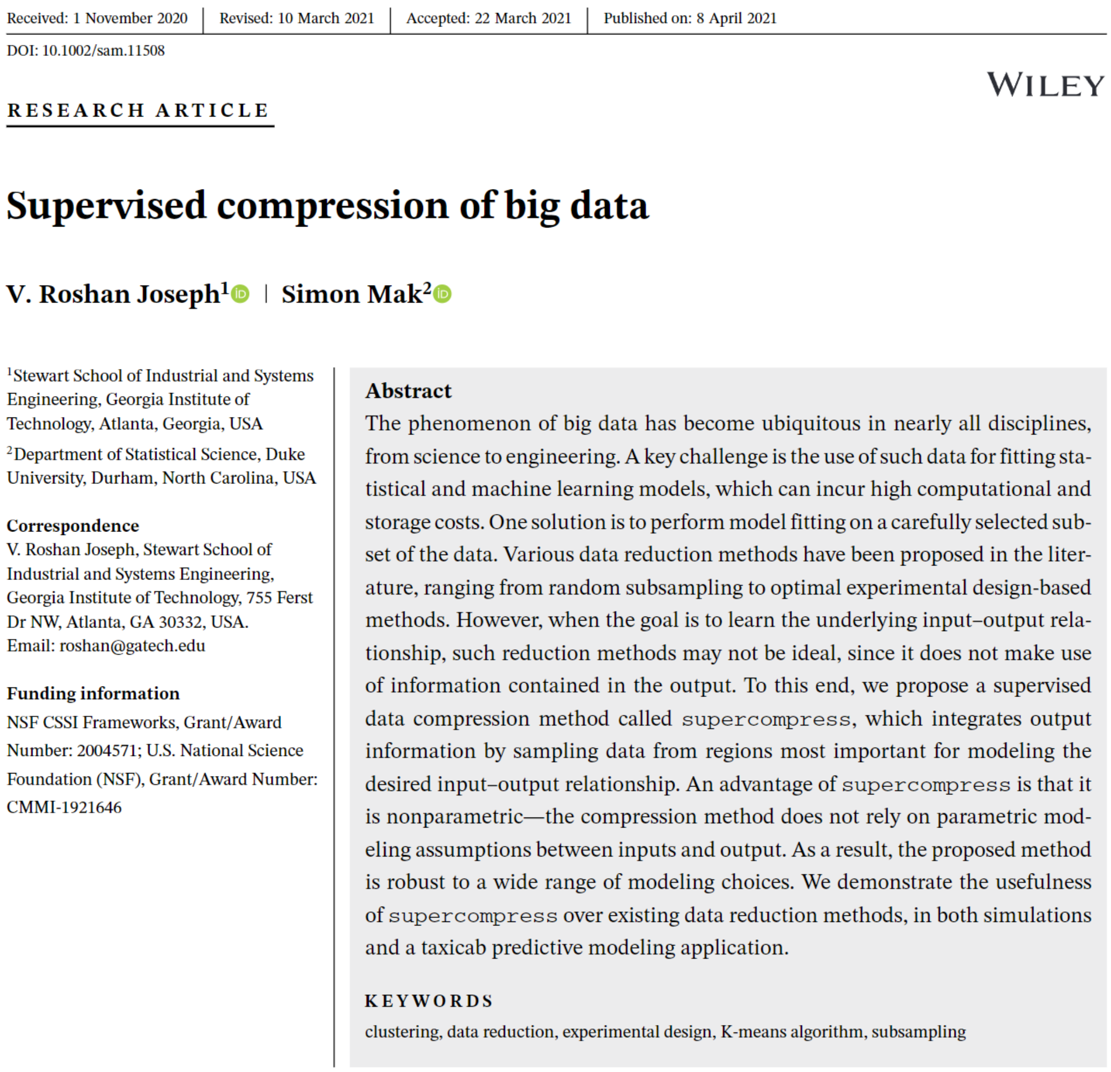
(aka regressogram or regression histogram)

Conditional Distribution

Model-Free Subsampling



Joseph and Mak (2021)



Received: 1 November 2020 | Revised: 10 March 2021 | Accepted: 22 March 2021 | Published on: 8 April 2021

DOI: 10.1002/sam.11508

WILEY

RESEARCH ARTICLE

Supervised compression of big data

V. Roshan Joseph[1] | Simon Mak[2]

[1]Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA

[2]Department of Statistical Science, Duke University, Durham, North Carolina, USA

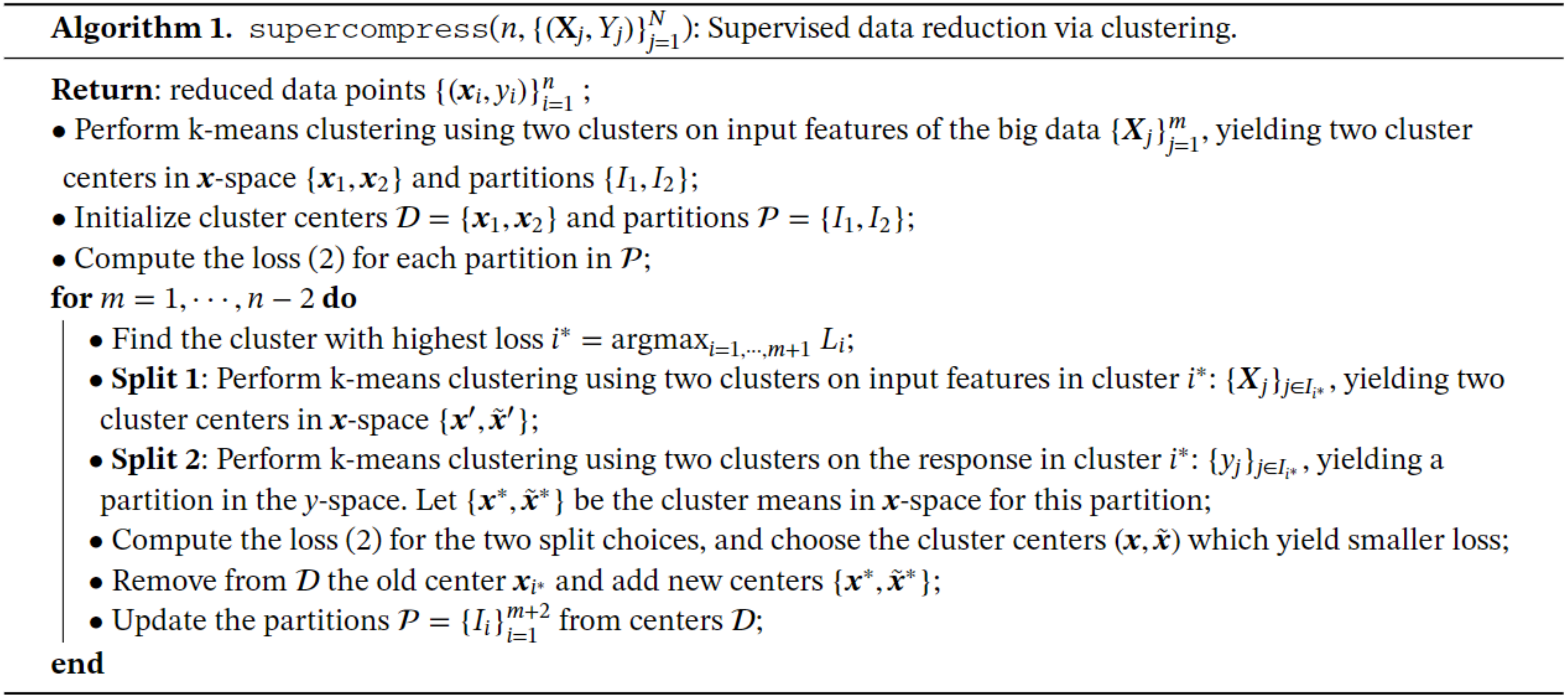
Correspondence
V. Roshan Joseph, Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Dr NW, Atlanta, GA 30332, USA.
Email: roshan@gatech.edu

Funding information
NSF CSSI Frameworks, Grant/Award Number: 2004571; U.S. National Science Foundation (NSF), Grant/Award Number: CMMI-1921646

Abstract
The phenomenon of big data has become ubiquitous in nearly all disciplines, from science to engineering. A key challenge is the use of such data for fitting statistical and machine learning models, which can incur high computational and storage costs. One solution is to perform model fitting on a carefully selected subset of the data. Various data reduction methods have been proposed in the literature, ranging from random subsampling to optimal experimental design-based methods. However, when the goal is to learn the underlying input–output relationship, such reduction methods may not be ideal, since it does not make use of information contained in the output. To this end, we propose a supervised data compression method called `supercompress`, which integrates output information by sampling data from regions most important for modeling the desired input–output relationship. An advantage of `supercompress` is that it is nonparametric—the compression method does not rely on parametric modeling assumptions between inputs and output. As a result, the proposed method is robust to a wide range of modeling choices. We demonstrate the usefulness of `supercompress` over existing data reduction methods, in both simulations and a taxicab predictive modeling application.

KEYWORDS
clustering, data reduction, experimental design, K-means algorithm, subsampling

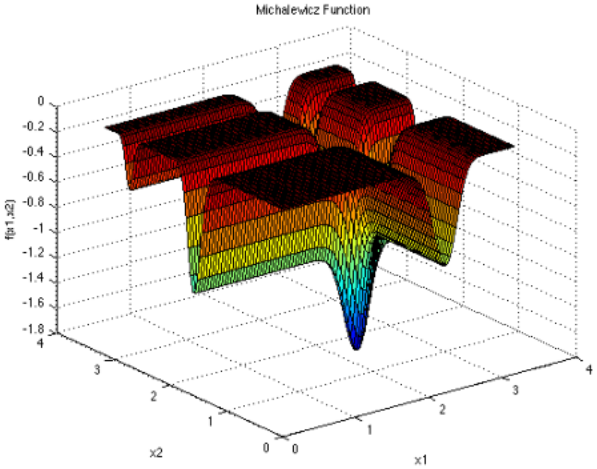
https://cran.r-project.org/web/packages/supercompress/index.html



Algorithm 1. `supercompress`$(n, \{(\mathbf{X}_j, Y_j)\}_{j=1}^{N})$: Supervised data reduction via clustering.

```
Return: reduced data points $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ ;
• Perform k-means clustering using two clusters on input features of the big data $\{\boldsymbol{X}_j\}_{j=1}^{m}$, yielding two cluster
  centers in $\boldsymbol{x}$-space $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ and partitions $\{I_1, I_2\}$;
• Initialize cluster centers $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ and partitions $\mathcal{P} = \{I_1, I_2\}$;
• Compute the loss (2) for each partition in $\mathcal{P}$;
for $m = 1, \cdots, n-2$ do
    • Find the cluster with highest loss $i^* = \text{argmax}_{i=1,\cdots,m+1} L_i$;
    • Split 1: Perform k-means clustering using two clusters on input features in cluster $i^*$: $\{\boldsymbol{X}_j\}_{j \in I_{i^*}}$, yielding two
      cluster centers in $\boldsymbol{x}$-space $\{\boldsymbol{x}', \tilde{\boldsymbol{x}}'\}$;
    • Split 2: Perform k-means clustering using two clusters on the response in cluster $i^*$: $\{y_j\}_{j \in I_{i^*}}$, yielding a
      partition in the $y$-space. Let $\{\boldsymbol{x}^*, \tilde{\boldsymbol{x}}^*\}$ be the cluster means in $\boldsymbol{x}$-space for this partition;
    • Compute the loss (2) for the two split choices, and choose the cluster centers $(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ which yield smaller loss;
    • Remove from $D$ the old center $\boldsymbol{x}_{i^*}$ and add new centers $\{\boldsymbol{x}^*, \tilde{\boldsymbol{x}}^*\}$;
    • Update the partitions $\mathcal{P} = \{I_i\}_{i=1}^{m+2}$ from centers $D$;
end
```

MICHALEWICZ FUNCTION



$$f(\mathbf{x}) = -\sum_{i=1}^{d} \sin(x_i) \sin^{2m}\left(\frac{ix_i^2}{\pi}\right)$$

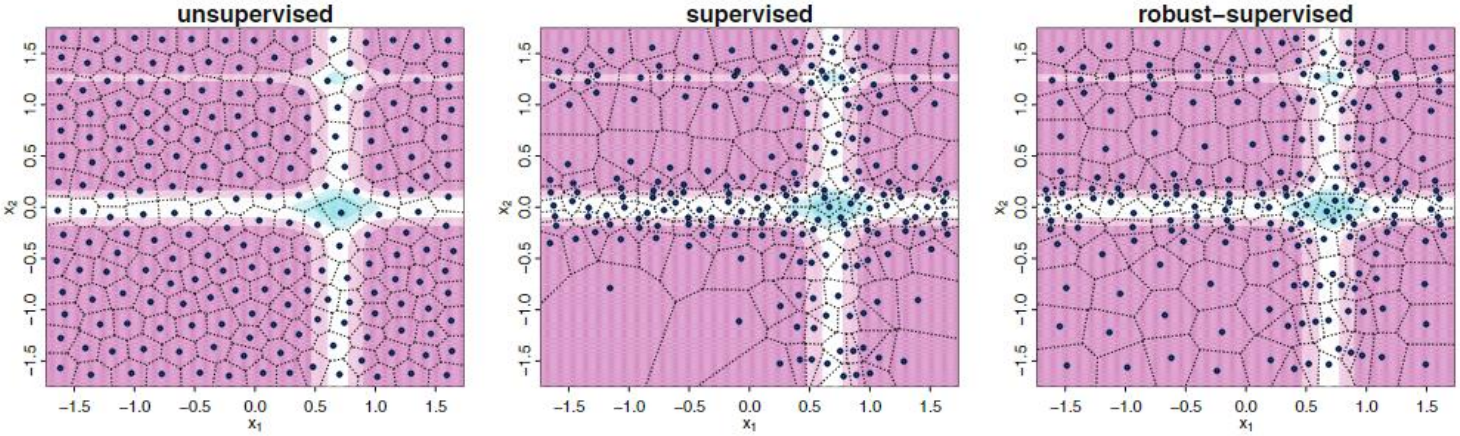
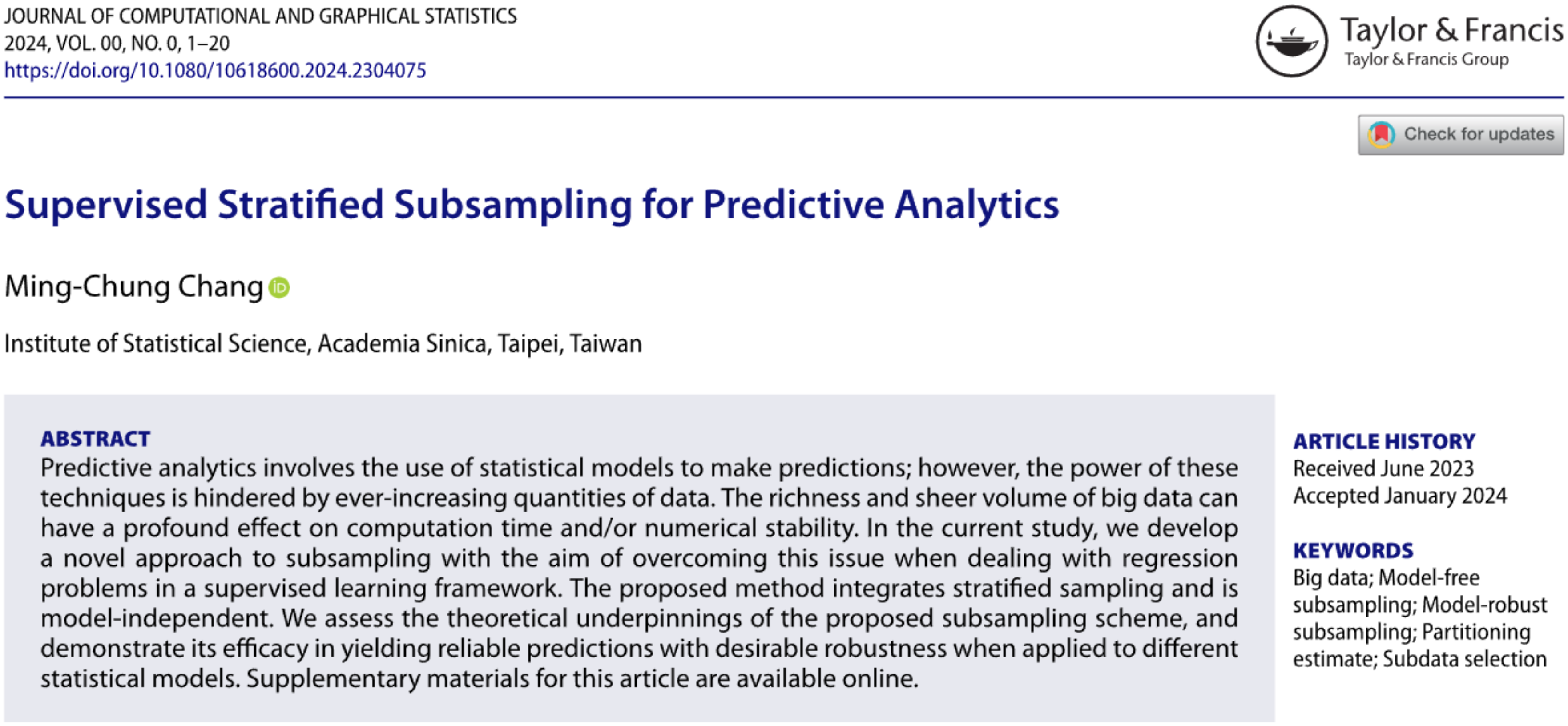


FIGURE 8 Two-dimensional example using Michaelwicz function



JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2024, VOL. 00, NO. 0, 1–20
https://doi.org/10.1080/10618600.2024.2304075

Taylor & Francis
Taylor & Francis Group
Check for updates

Supervised Stratified Subsampling for Predictive Analytics

Ming-Chung Chang

Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

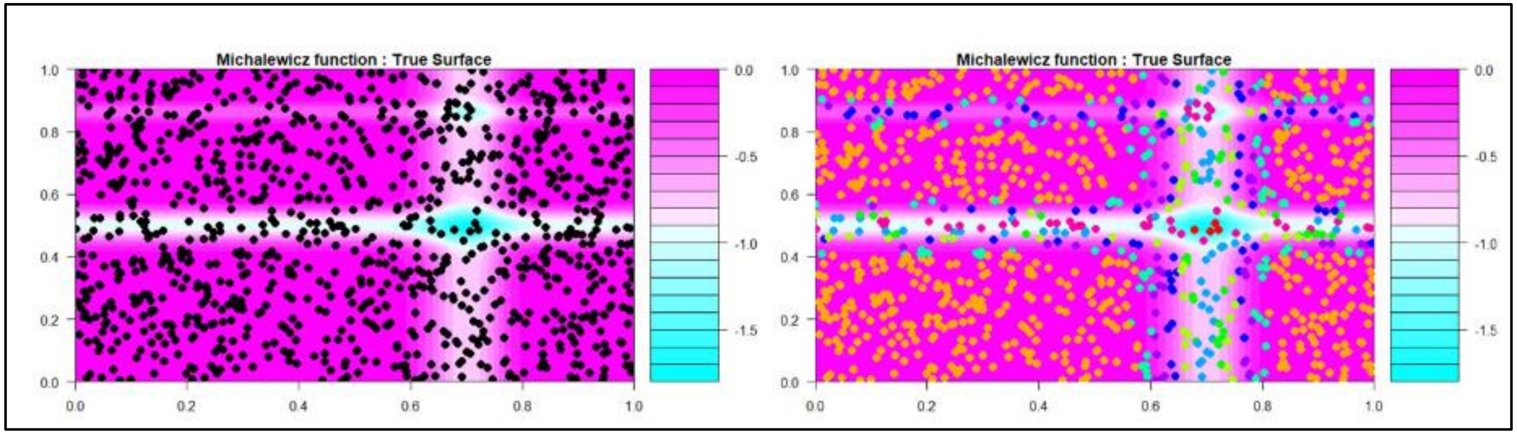
ABSTRACT

Predictive analytics involves the use of statistical models to make predictions; however, the power of these techniques is hindered by ever-increasing quantities of data. The richness and sheer volume of big data can have a profound effect on computation time and/or numerical stability. In the current study, we develop a novel approach to subsampling with the aim of overcoming this issue when dealing with regression problems in a supervised learning framework. The proposed method integrates stratified sampling and is model-independent. We assess the theoretical underpinnings of the proposed subsampling scheme, and demonstrate its efficacy in yielding reliable predictions with desirable robustness when applied to different statistical models. Supplementary materials for this article are available online.

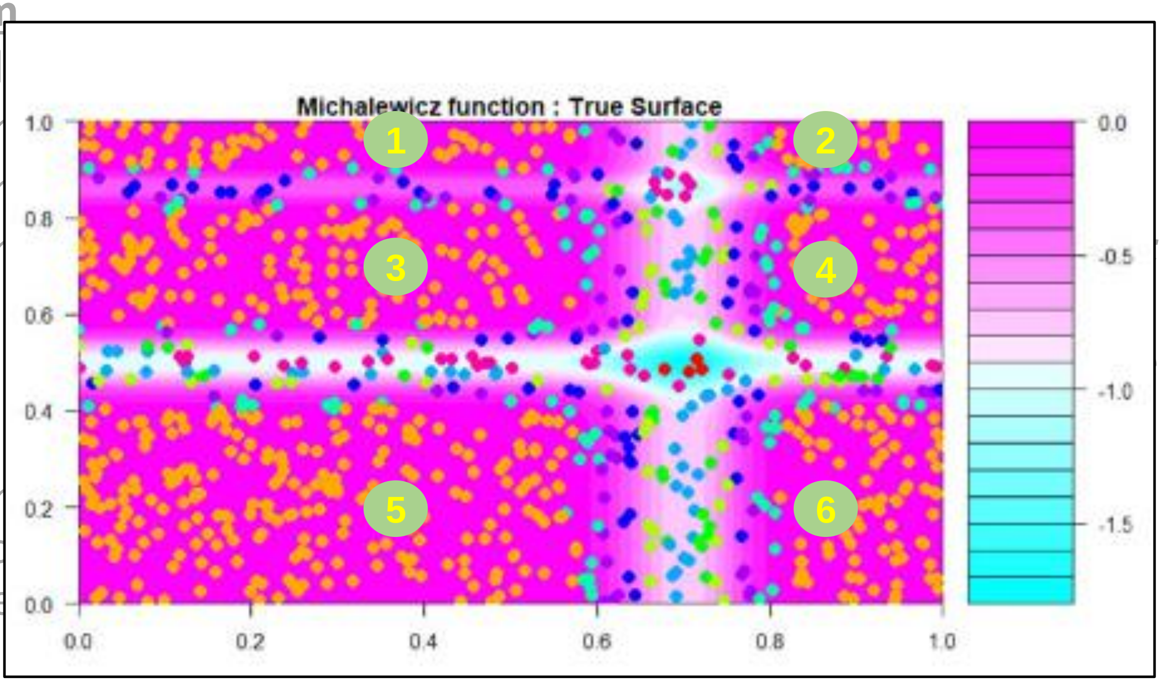
ARTICLE HISTORY
Received June 2023
Accepted January 2024

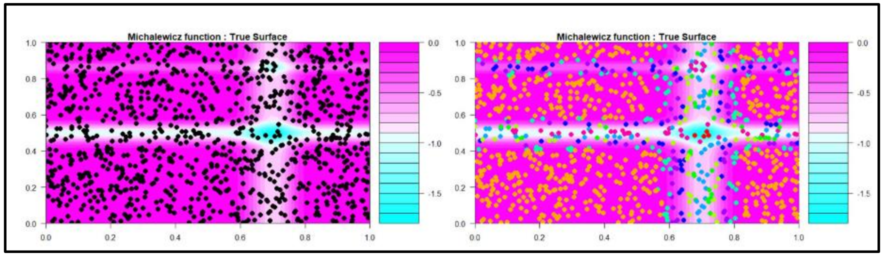
KEYWORDS
Big data; Model-free subsampling; Model-robust subsampling; Partitioning estimate; Subdata selection

- Model-free approach: **Chang (2024)**



- **Algorithm**
    - Appl
    - Form
    - Form
    - Form $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ik_i}$: i $= 1,$
        - gion with k
    - Form
    - Ranc
    - Repe

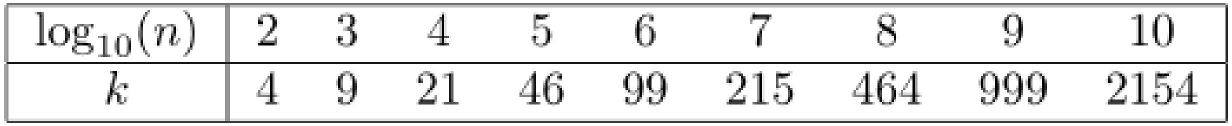




- Assume: **(i)** $f(\boldsymbol{x})$ is bounded; **(ii)** $\mathrm{Var}(Y|x)$ is bounded; **(iii)** $g(y)$ is bounded, defined on a compact support, and has 1st to 4th bounded derivatives. Then, the MISE for the **partitioning estimate** $\hat{f}(\boldsymbol{x})$ is:

$$\mathrm{E}\left\{\int\left(\hat{f}(\boldsymbol{x})-f(\boldsymbol{x})\right)^2\mu(d\boldsymbol{x})\right\}=\mathrm{O}\left(\frac{k}{n}+\frac{1}{k^2}\right)$$

- Suggest $k=n^{\frac{1}{3}}$ ➔ Convergence rate: $\mathrm{O}\left(n^{-\frac{2}{3}}\right)$



$\log_{10}(n)$	2	3	4	5	6	7	8	9	10
k	4	9	21	46	99	215	464	999	2154

WEC Dataset

- The Wave Energy Converters (WEC) dataset, provided by UCI Machine Learning Repository (Dua and Graff, 2019)
  - Y: total power output
  - X: 32 location variables and 16 absorbed power variables ($d = 48$)
  - 288,000 = 252,000 for training + 36,000 for testing (divided by SRS)
  - Subdata size: 1,000
  - $B = 5$

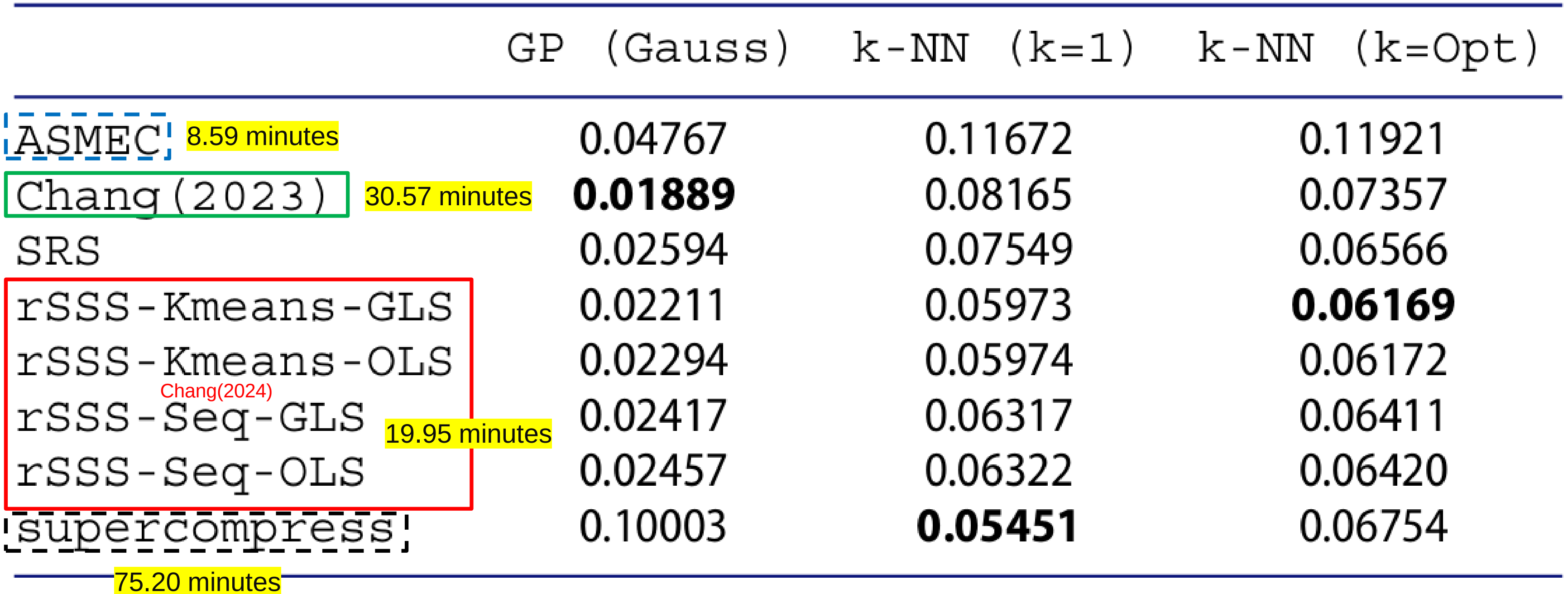
40 replications ➔ 40 RMSPEs

Chang(2024)

- Methods: **rSSS-K-GL/OL**, **rSSS-Seq-GL/OL**, supercompress, ASMEC, SRS,

  Chang(2023)
- Models
  - Gaussian process regression (mleHomGP)
    - Gaussian correlation function
  - k-NN ($k = 1$ and $k = 5$, knn.reg)

Table 8. Medians of the 40 RMSPEs (bold for the minimum): WEC data.



	GP (Gauss)	k-NN (k=1)	k-NN (k=Opt)
ASMEC	0.04767	0.11672	0.11921
Chang(2023)	**0.01889**	0.08165	0.07357
SRS	0.02594	0.07549	0.06566
rSSS-Kmeans-GLS	0.02211	0.05973	**0.06169**
rSSS-Kmeans-OLS	0.02294	0.05974	0.06172
rSSS-Seq-GLS	0.02417	0.06317	0.06411
rSSS-Seq-OLS	0.02457	0.06322	0.06420
supercompress	0.10003	**0.05451**	0.06754

8.59 minutes

30.57 minutes

Chang(2024)

19.95 minutes

75.20 minutes

Desktop computer with a 3.20GHz Intel Corei9 CPU and 128GB of RAM

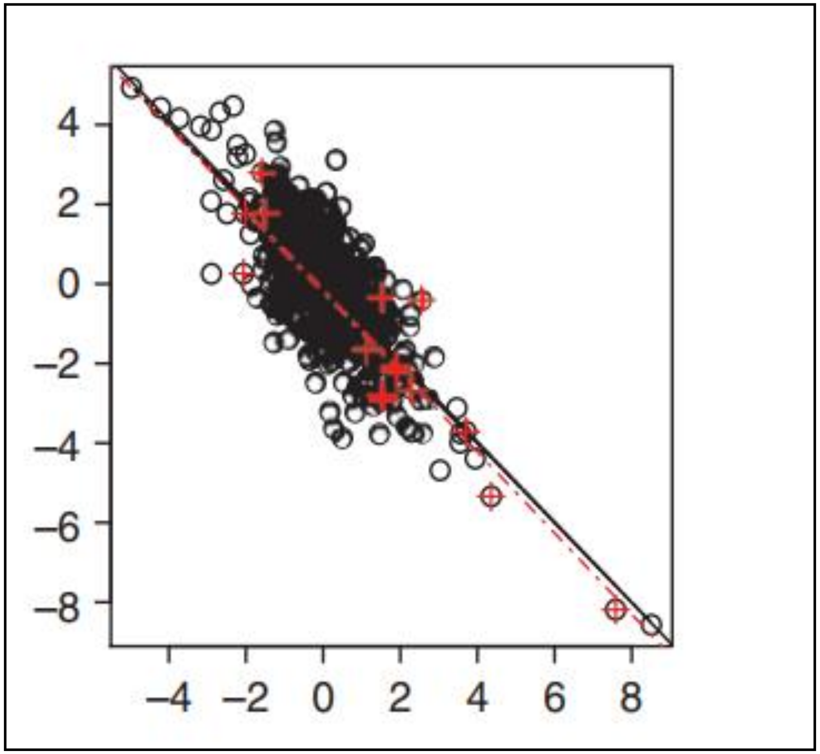
Chang(2023) not good for k-NN

supercompress usually better for 1-NN (non-smooth model)

Chang(2024) seems more robust

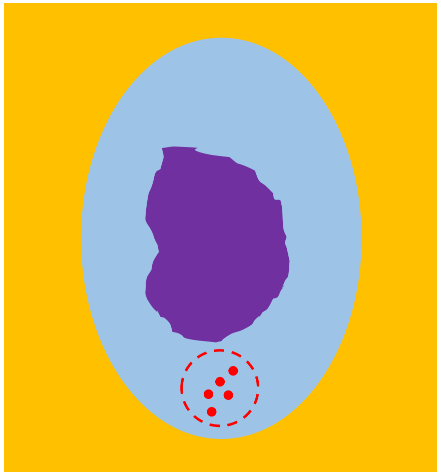
Summary

- **Wang, Yang, and Stufken (2019)**



Summary

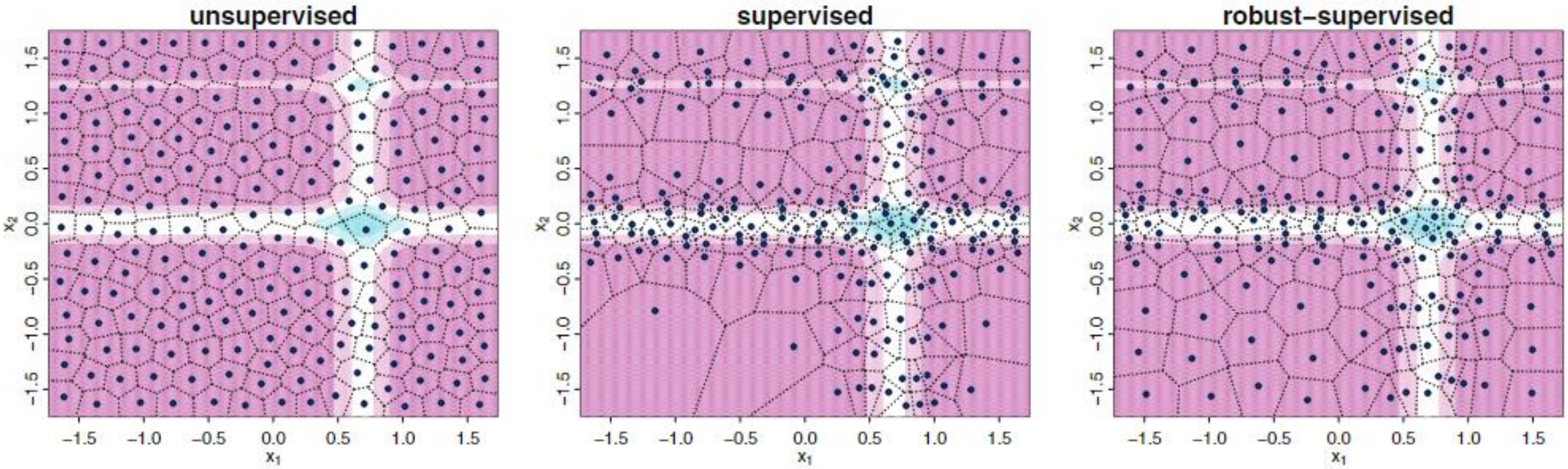
- **Chang (2023)**



$$\sum_{j=1}^{k}\{\text{prediction var.}+\text{prediction error}^2+\text{noise var.}\}_{\text{at } x_{i_j}}$$

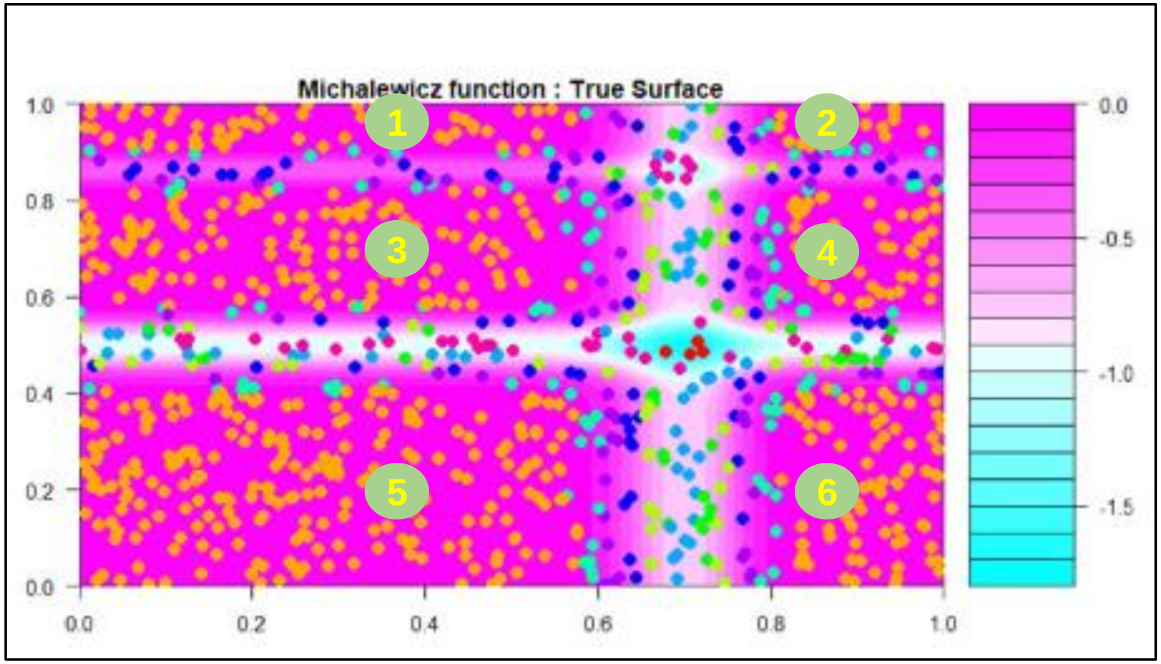
Summary

- Joseph and Mak (2021)



Summary

- **Chang (2024)**



Joint Distribution

Model-Free Subsampling

- Data splitting techniques: SPlit, Twinning, and Optimal ratio (Slides) (Video)

https://sites.google.com/view/mcchang/teaching?authuser=0

- V. Roshan Joseph and Vakayil, A. (2022). “SPlit: An Optimal Method for Data Splitting“. *Technometrics*, 64, 166-176. R package: SPlit. (**Wilcoxon Award**).

- V. Roshan Joseph. (2022). “Optimal Ratio for Data Splitting“. *Statistical Analysis and Data Mining: The ASA Data Science Journal,* 15, 531-538. R package: SPlit.

- Vakayil, A. and V. Roshan Joseph (2022). “Data Twinning“. *Statistical Analysis and Data Mining: The ASA Data Science Journal,* 15, 598-610. R package. Python package.

- Data structure

Data : 1 2 ⋯ d-1

X_1	Y_1	Z_1
X_2	Y_2	
⋮		
X_N	Y_N	

Let $D = \{\mathbf{Z}_i = [\mathbf{X}_i, Y_i]\}_{i=1}^{N} \in \mathbb{R}^{N\times d}$ be the given dataset, where $\mathbf{X}_i$ is a $d-1$ dimensional vector representing the $d-1$ features in the i^{th} row, and Y_i denotes the corresponding response value. Assume that each row of the dataset is independently drawn from a distribution $\underline{\underline{\mathcal{F}}}$:

e.g., multivariate normal

$$(\mathbf{X}_i, Y_i) \overset{iid}{\sim} \mathcal{F}, \quad i = 1, \ldots, N.$$

$Z_1, Z_2, \ldots, Z_N \overset{iid}{\sim} \mathcal{F}$

- Generalization error

$$\mathcal{E} = E_{\mathbf{X},Y}\left\{L(Y, g(\mathbf{X};\hat{\theta}))|D^2\right\}$$

training data

$\hat{\theta}$ is the estimate of θ using D^2

- Estimate of the generalization error

$$\hat{\mathcal{E}} = \frac{1}{n}\sum_{i=1}^{n} L\left(Y_i^1, g\left(\mathbf{X}_i^1;\hat{\theta}\right)\right)$$

(X_i^1, Y_i^1) from the testing data

$$E(\hat{\mathcal{E}}) = \frac{1}{n}\sum_{i=1}^{n} E\{L(Y_i^1, g(X_i^1;\hat{\theta}))\}$$

$$\overset{?}{=} \frac{1}{n}\cdot n\cdot E\{L(Y_i^1, g(X_i^1;\hat{\theta}))\} = \mathcal{E}$$

- The estimate of the generalization error will be unbiased if $(X_i^1, Y_i^1) \overset{iid}{\sim} \mathcal{F}$
- By LLN, we have $\hat{\mathcal{E}} \xrightarrow{a.s.} \mathcal{E}$ (a.s. : almost surely)

- Mak and Joseph (2018) showed that, under some regularity conditions,

$$\boxed{|\hat{\mathcal{E}} - \mathcal{E}| \le C\sqrt{\mathbb{ED}}}$$

- ED is the energy distance

$$\overline{\mathbb{ED}}_{n,N} = \frac{2}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}\|\mathbf{U}_i - \mathbf{Z}_j\|_2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{U}_i - \mathbf{U}_j\|_2 - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\|\mathbf{Z}_i - \mathbf{Z}_j\|_2,$$

$Z_1,..,Z_N$: full data

$W_1,...,W_n$: a subset of Z_i s

- ***Twinning*** aims to partition a dataset into two disjoint sets such that they have similar statistical properties. We will call the two sets as twins. The twins needn't be of the same size, but they should have similar statistical distributions.

$$\overline{\mathbb{ED}}_{n,N-n} = \frac{2}{n(N-n)}\sum_{i=1}^{n}\sum_{j=1}^{N-n}\|\mathbf{U}_i - \mathbf{V}_j\|_2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{U}_i - \mathbf{U}_j\|_2 - \frac{1}{(N-n)^2}\sum_{i=1}^{N-n}\sum_{j=1}^{N-n}\|\mathbf{V}_i - \mathbf{V}_j\|_2.$$

The twins are obtained by minimizing $\overline{\mathbb{ED}}_{n,N-n}$ with respect to D^1 and D^2, that is,

$\{W_1,..,W_n\} \cup \{V_1,..,V_{N-n}\}$ = full data

$$\left\{\mathbf{U}_i^*\right\}_{i=1}^{n}, \left\{\mathbf{V}_j^*\right\}_{j=1}^{N-n} = \underset{\{\mathbf{U}_i\}_{i=1}^{n}, \{\mathbf{V}_j\}_{j=1}^{N-n}}{\arg\min} \overline{\mathbb{ED}}_{n,N-n}$$

$$\text{subject to}: \quad \{\mathbf{U}_i\}_{i=1}^{n} \cap \left\{\mathbf{V}_j\right\}_{j=1}^{N-n} = \emptyset$$

$$\{\mathbf{U}_i\}_{i=1}^{n} \cup \left\{\mathbf{V}_j\right\}_{j=1}^{N-n} = D. \qquad (9)$$

- Algorithm of Twinning $\gamma = \frac{\text{subdata size}}{\text{full data size}} = \frac{1}{5} = \frac{1}{4+1} \Rightarrow \text{training} : \text{testing} = 4 : 1$

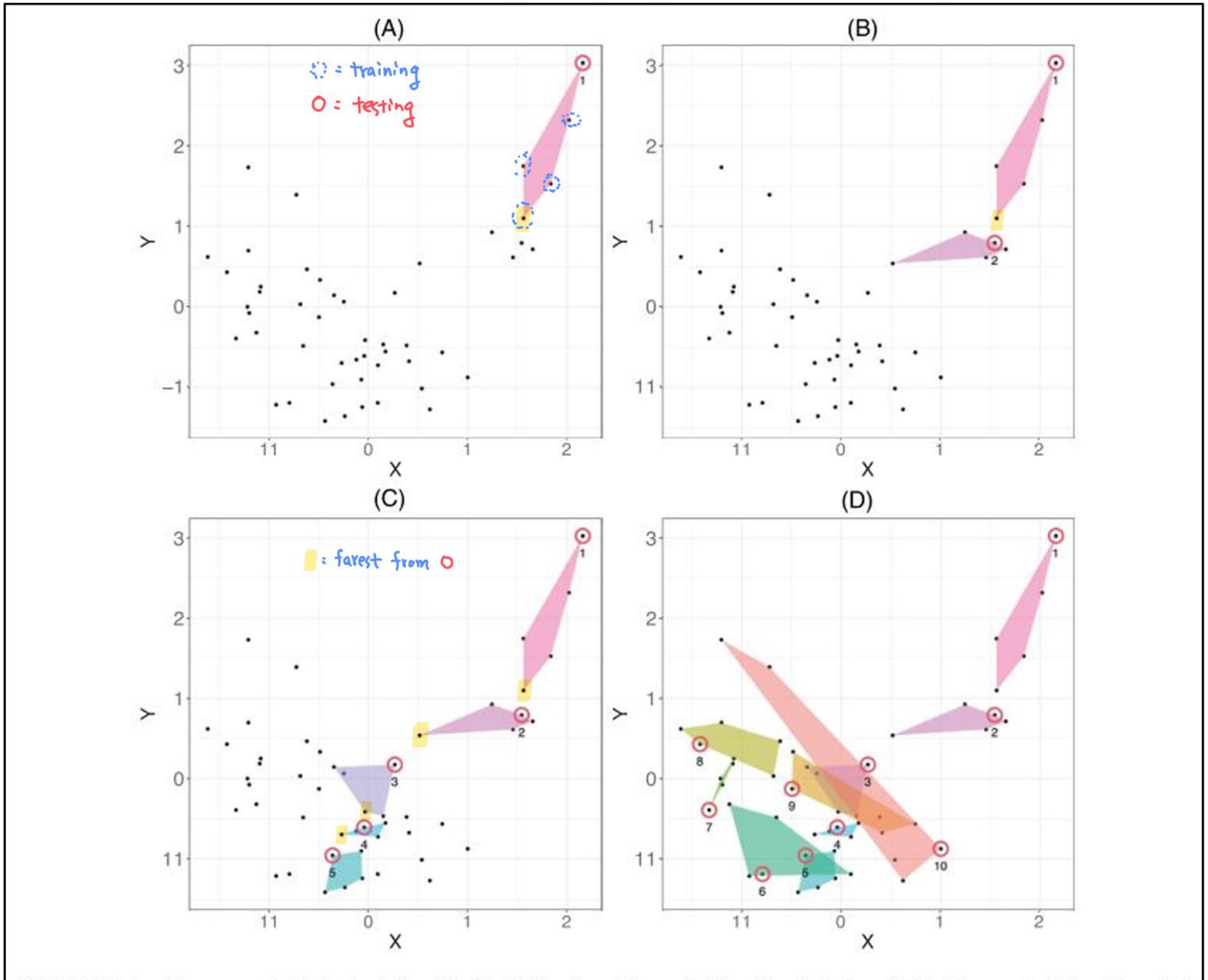
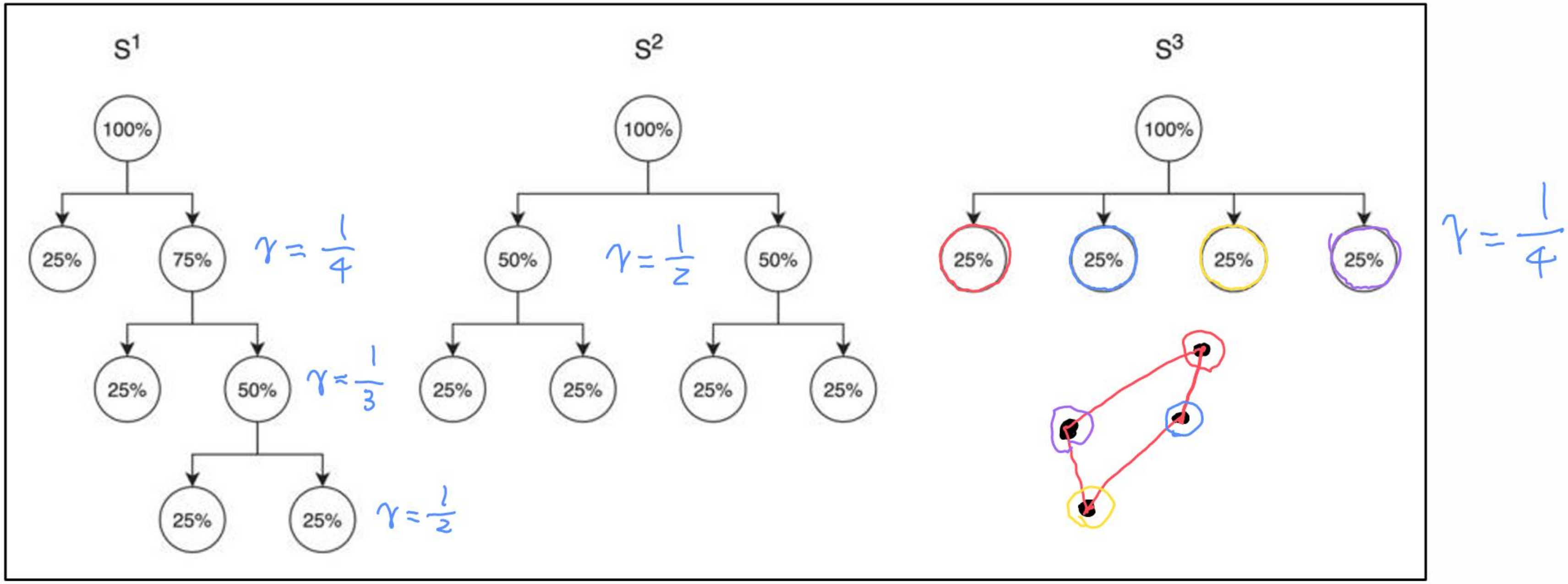
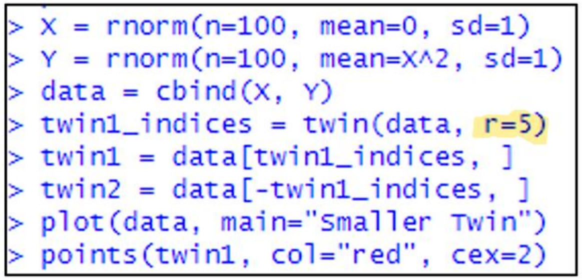


FIGURE 2 The convex hull of subsets identified by Twinning at the end of iterations 1, 2, 5, and 10 for the sample dataset described in Section 3. Points in D^1 are shown as encircled points, and they are numbered in the order they were selected. (A) Iteration 1. (B) Iteration 2. (C) Iteration 5. (D) Iteration 10

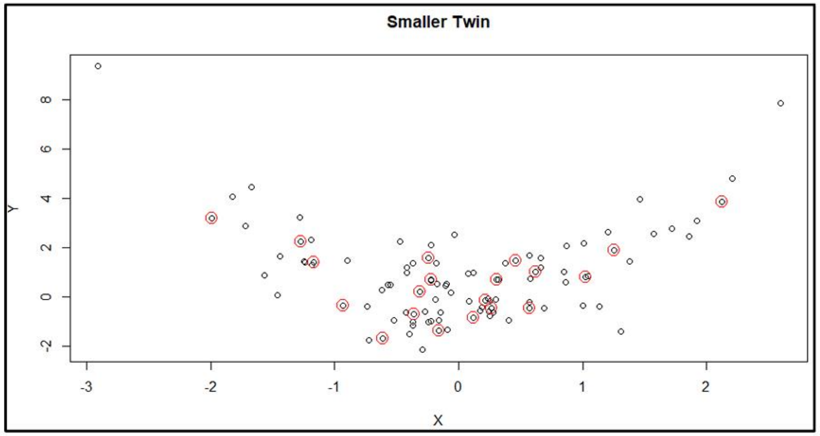
- Three strategies to generate quadruplets



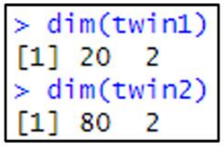
- Example: training&testing



```
> X = rnorm(n=100, mean=0, sd=1)
> Y = rnorm(n=100, mean=X^2, sd=1)
> data = cbind(X, Y)
> twin1_indices = twin(data, r=5)
> twin1 = data[twin1_indices, ]
> twin2 = data[-twin1_indices, ]
> plot(data, main="Smaller Twin")
> points(twin1, col="red", cex=2)
```



$r = \frac{1}{\gamma}$



```
> dim(twin1)
[1] 20  2
> dim(twin2)
[1] 80  2
```

Conclusion

- Why subdata?
  - Data-rich ≠ Information-rich
  - Time consuming
  - Full data beaten by subdata

- Solutions
  - Divide and conquer
  - Subsampling
  - …

- Conditional distribution of $X_1|X_2, \dots, X_p$
  - Model-dependent subsampling
  - Model-free subsampling

- Joint distribution of $X_1, \dots, X_p$
  - Model-free subsampling

Thank you for your attention