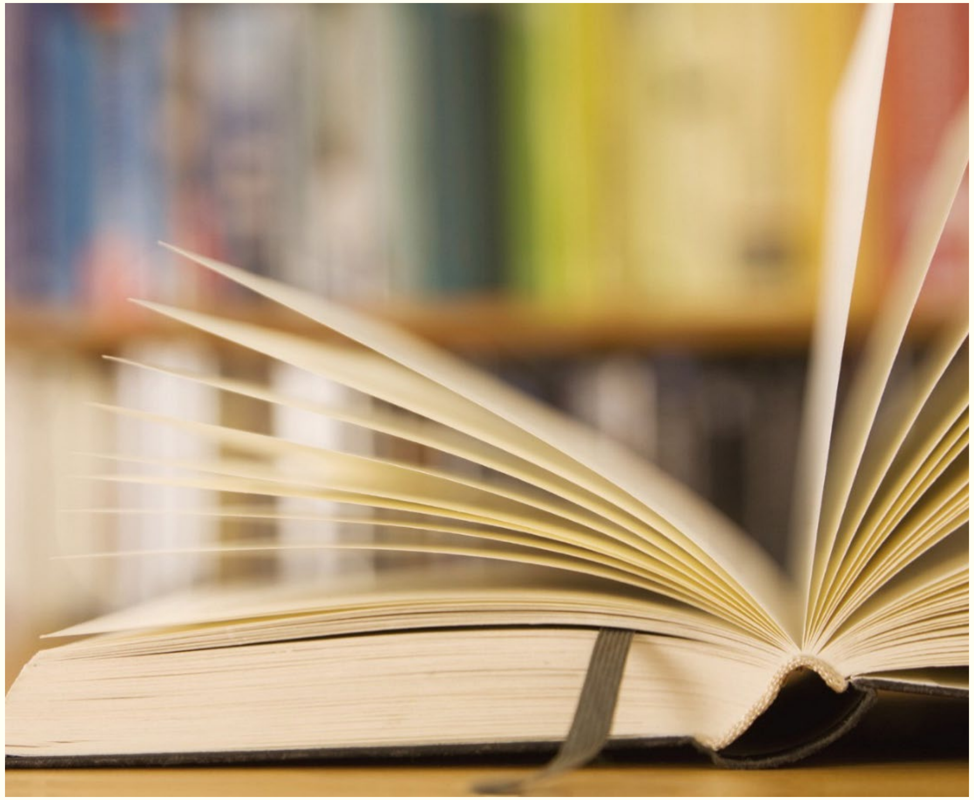
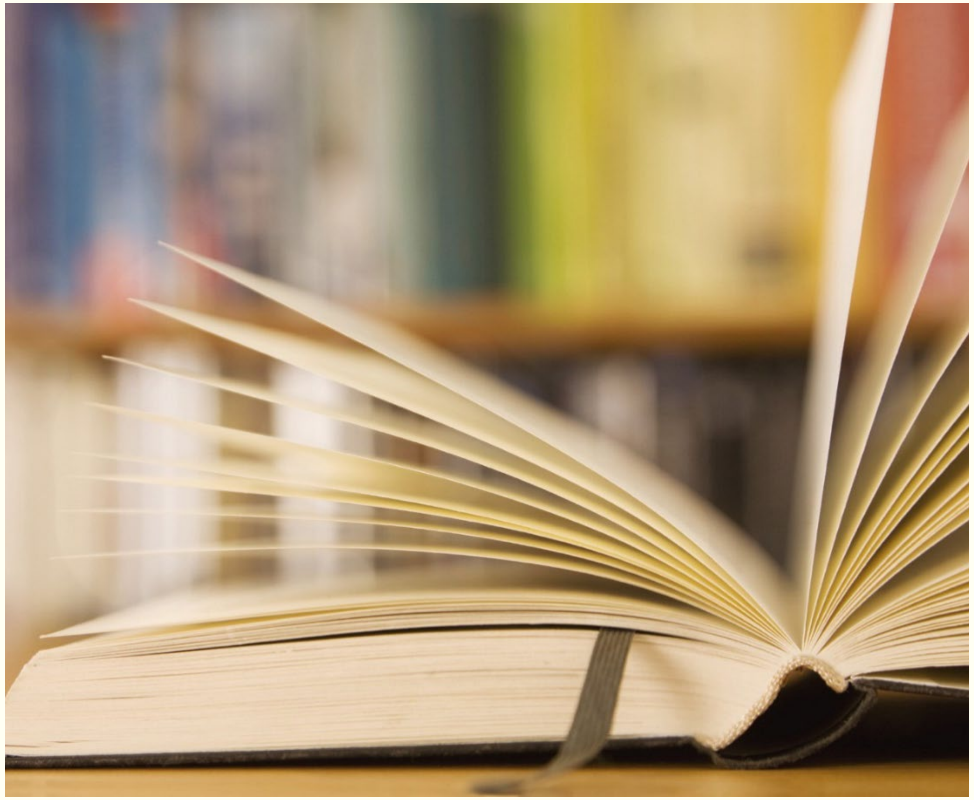
OPTIMAL SUBDATA SELECTION FOR LARGE-SCALE MULTI-CLASS LOGISTIC REGRESSION

MIN YANG

UNIVERSITY OF ILLINOIS AT CHICAGO



SCALABLE METHODOLOGIES FOR BIG DATA ANALYSIS: INTEGRATING FLEXIBLE STATISTICAL MODELS AND OPTIMAL DESIGNS



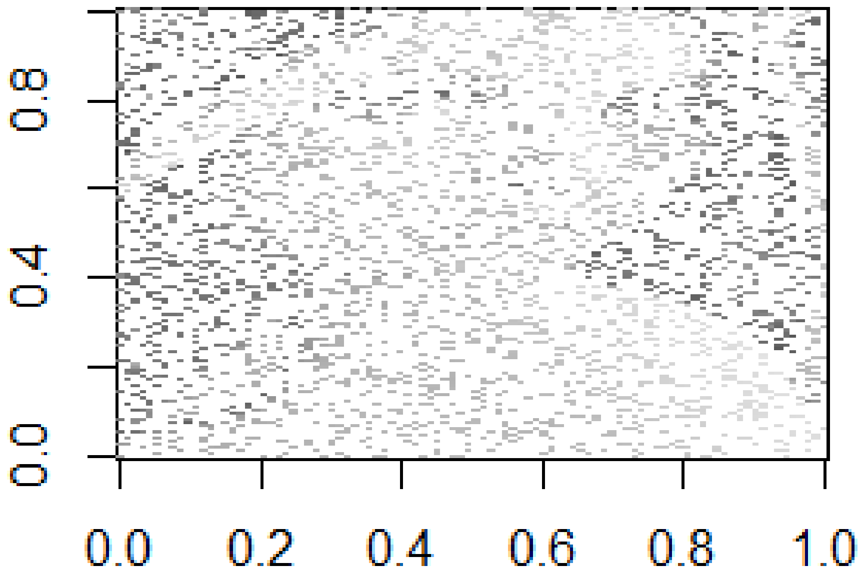
MIN YANG

UNIVERSITY OF ILLINOIS AT CHICAGO

Phenomenon of data science

- Scientific and evidence-based decision
  - Policy making, marketing strategy, …
  - Interpretable
- How to detect relationships
  - Size
  - Complex
  - Statistical models?

A motivated example



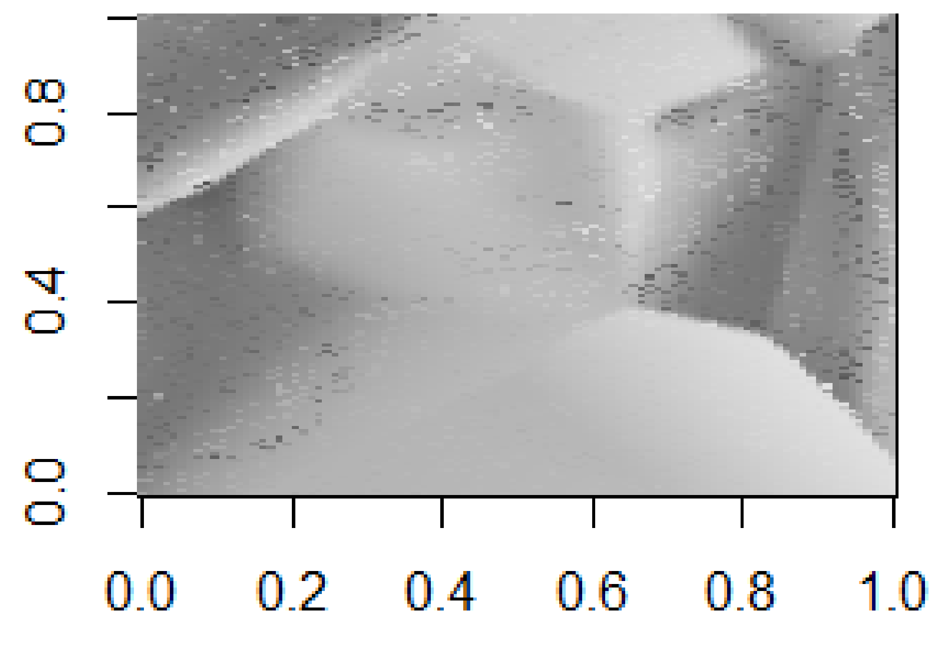
X	Y	Gray
1	3	118
1	10	127
1	17	127
1	21	127
1	25	130
1	26	134
1	28	135
1	35	135
1	36	141
1	37	137
.	.	.
.	.	.
.	.	.

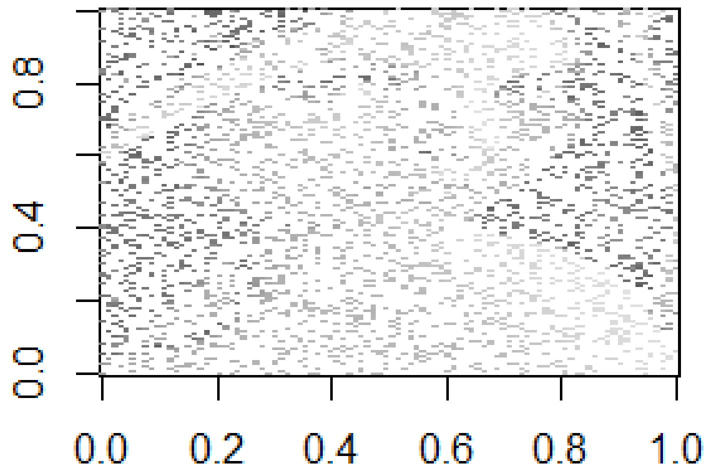
Fundamental question

Given $(\boldsymbol{X_1}, Y_1), \cdots, (\boldsymbol{X_N}, Y_N)$,

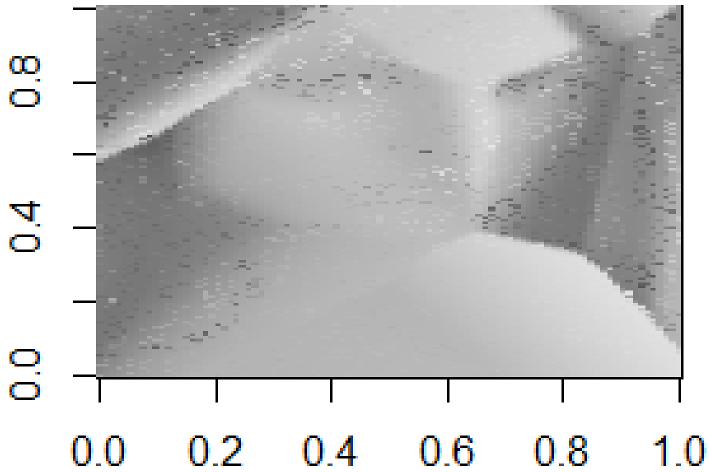
$$Y = f(\boldsymbol{X}, \theta)$$

A motivated example





0.0
0.4
0.8
0.0 0.2 0.4 0.6 0.8 1.0



0.0
0.4
0.8
0.0 0.2 0.4 0.6 0.8 1.0

Mixture-of-Experts modeling

- Proposed by Jacobs et al. (1991)
- Discover the hidden clusters
- Striking a balance between flexibility and interpretability

General Framework of Mixture of Experts

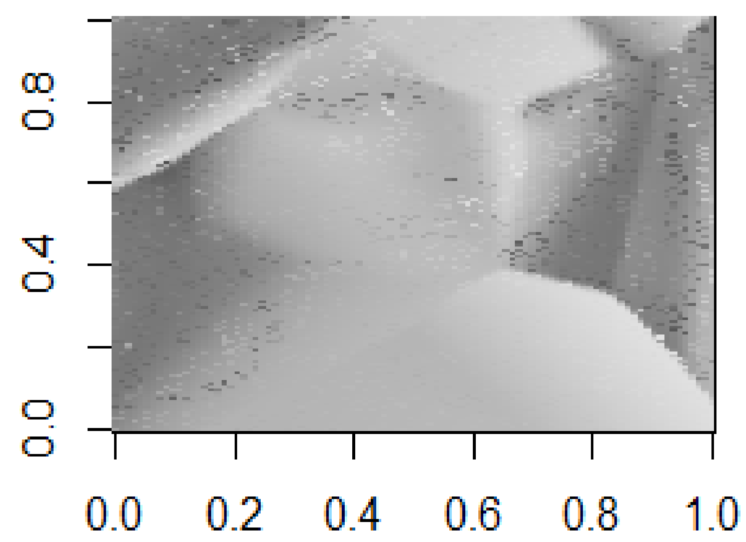
- $(\boldsymbol{X_1}, Y_1), \cdots, (\boldsymbol{X_N}, Y_N)$

- K gate functions and K regression models (experts)
  - Y_i is modeled by $\boldsymbol{X_i}$ through one of the experts
  - It is unknown which expert is employed

- $g_k(\boldsymbol{X_i}, \boldsymbol{\gamma}) = \frac{\exp(\boldsymbol{\gamma}_k' \boldsymbol{X_i})}{1+\sum_{j=1}^{K-1} \exp(\boldsymbol{\gamma}_j' \boldsymbol{X_i})}$

- Experts
  - Depends on the nature of the responses: linear, GLMS, $\cdots$

A motivated example



$$\text{Gray} = \sum_k \frac{\exp(\boldsymbol{\gamma}'_k \boldsymbol{X_i})}{1+\sum_{j=1}^{K-1}\exp(\boldsymbol{\gamma}'_j \boldsymbol{X_i})}(\boldsymbol{\theta}'_k \boldsymbol{X_i})$$

γ_1	110.43	-1.63	3.47
γ_2	-57.56	0.55	-0.79
γ_3	-117.84	-0.49	5.16
γ_4	-15.35	-0.48	2.95
.	.	.	.
.	.	.	.
.	.	.	.
γ_{29}	-362.35	-1.90	10.31

θ_1	-266.27	1.83	4.98
θ_2	-59.82	0.95	-1.40
θ_3	-38.51	1.01	1.49
θ_4	448.81	-1.97	-0.73
.	.	.	.
.	.	.	.
.	.	.	.
θ_{30}	49.93	-0.06	0.09

Ability to capture complex relationships

- Compare with functional data analysis (Chen, Hall, and Müller, 2011)

Table 1: RASE comparison between CHM and ME

Model	N	$R=0.1$		$R=0.5$		Model	N	$R=0.1$		$R=0.5$	
		CHM	ME	CHM	ME			CHM	ME	CHM	ME
(i)	50	0.0464	0.0222	0.1096	0.0242	(ii)	50	0.0970	0.0298	0.2562	0.0829
	200	0.0279	0.0216	0.0577	0.0221		200	0.0486	0.0132	0.1122	0.0392
	800	0.0156	0.0201	0.0315	0.0206		800	0.0226	0.0069	0.0526	0.0184

- Compare with Reproducing Kernel Hilbert Space (RKHS) Approach (Xiong, Qian, and Wu, 2013; Sauer, Gramacy, and Higdon,2023)

Table 3: MSPE comparison between RKHS and ME

Case (i)	$N=1000$		$N=2000$	
	RKHS	ME	RKHS	ME
MSPE	0.0703	0.0324	0.0436	0.0271

Case (ii)	$N=200$		$N=500$	
	RKHS	ME	RKHS	ME
MSPE	0.9725	0.3924	0.4088	0.2365

Computation issue

- No closed form solution

$$L(\beta|\boldsymbol{X},\boldsymbol{Y}) = \prod_{i=1}^{N}\left(\sum_{k=1}^{S} g_k(\boldsymbol{X}_i,\gamma)L(\beta|\boldsymbol{X}_i,y_i)\right)$$

  - EM algorithm or Bayesian approach

- Computation expensive for large dataset
  - Depends on # of clusters, # of start values
  - For the motivated example (n≈3000)
    - One hour
  - Take weeks for n=10^6
    - Two weeks

Computation complexity and statistical efficiency

- With size of data and number of clusters increase, the computation cost increase dramatically
- The tradeoff between the computation complexity and statistical efficiency?
- One of six suggested core research topics of theoretical foundations of data science (NSF)

Two main approaches

- Subsampling with sampling probability
  - Pro: Robustness, outliers
  - Con: Limited by subsize

- Information-based subdata selection
  - Based on optimal design theory

  - Fixed n, the information increases with N

A TOY EXAMPLE ABOUT OPTIMAL DESIGN

Rationale

- Matrix form: $Y = X\beta + \epsilon$
- BLUE: $\hat{\beta} = (X'X)^{-1}X'Y$ and $Var(\hat{\beta}) = (X'X)^{-1}\sigma^2$
- How to select X such that $(X'X)^{-1}$ is “minimized"?

$$X_I = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}; X_{II} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{pmatrix};$$

$$X_{III} = \begin{pmatrix} 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix};$$

$(X_I'X_I)^{-1} = I;$

$(X_{II}'X_{II})^{-1} = I/2;$

$(X_{III}'X_{III})^{-1} = I/4$

Selecting an optimal subset

- Difference between optimal design and subdata selection
  - Perfect points may not exist

- Large N and n

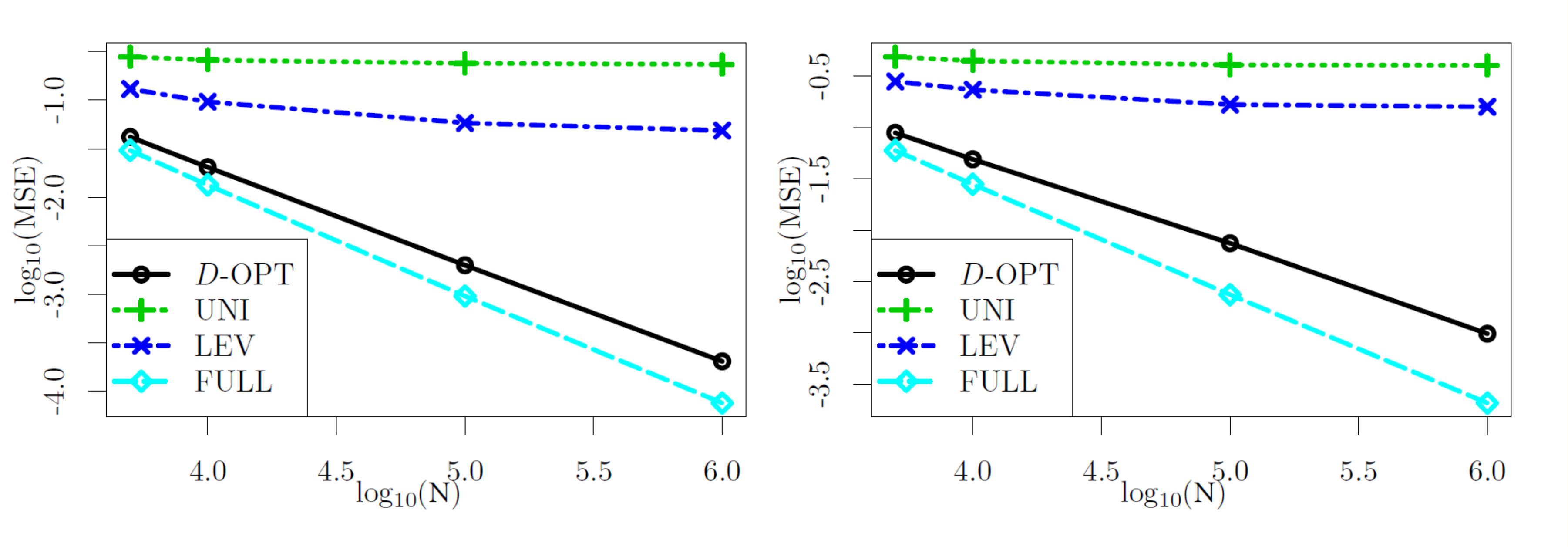
- Intractable
  - Discrete nature
  - No tool
  - N-P hard problem

Available approaches

- Information-Based Optimal Subdata Selection (IBOSS) (Wang, Yang, and Stufken, 2019)
- Linear model $E(\boldsymbol{Y}) = \boldsymbol{X\beta}$
- Characterizing the design maximizing information matrix
- An algorithm of selecting the subset based on the characterization
  - Fixed n, the information increases with N

- Subsampling with sampling probability
  - Limited by subsize

IBOSS approach



- Builds the theoretical foundation
- Extends to nonlinear models: Logistic regression model
- Extends to variable selection: LASSO

Limitations

- Simple model
  - Unlikely suitable for large dataset with complexity structure
  - Possible solution: Mixture of experts

- Efficiency of the algorithm?
  - Based on the characterization of optimal design
  - May not be efficient

Challenges for Mixture of Experts

- Information matrix
  - No explicit form
- Charactering optimal designs
- Algorithm?

Strategy

$$\boldsymbol{I}(\boldsymbol{\delta}) = \sum_{i\in\delta} (\boldsymbol{I}_{C_i} - \boldsymbol{I}_{M_i}) \le \sum_{i\in\delta} \boldsymbol{I}_{C_i}, \text{ so that}$$

$$det(\boldsymbol{I}(\boldsymbol{\delta})) \le det(\sum_{i\in\delta} \boldsymbol{I}_{C_i}).$$

- Choosing a subset δ
  - Maximizing $\sum_{i\in\delta} I_{C_i}$
  - Minimizing $\sum_{i\in\delta} I_{M_i}$

Asymptotic result

- Under clusterwise linear regression model, where gate functions are constants and experts are linear

Let $\boldsymbol{\mu}_z = (\mu_{z1}, ..., \mu_{zp})^T$ and $\boldsymbol{\Sigma}_z = \boldsymbol{\Phi}_z \boldsymbol{\rho} \boldsymbol{\Phi}_z$ be a full rank covariance matrix, where $\boldsymbol{\Phi}_z = blkdiag(\sigma_{z1}, \ldots, \sigma_{zp})$ is a diagonal matrix of standard deviations and $\boldsymbol{\rho} = (\rho_{jj'})_{p\times p}$ is a correlation matrix.

Theorem 3. *Let $\mathbf{z}_1, ..., \mathbf{z}_N$ be iid, where $\mathbf{z}_i = (z_{i1}, z_{i2}, ..., z_{ip})^T$. Assuming that $y_i \sim \sum_{g=1}^{G} \pi_g \phi(\mathbf{x}_i^T \beta_g, \sigma_g^2)$, where $\mathbf{x}_i^T = (1, \mathbf{z}_i^T)^T$, and $\boldsymbol{\delta}^*$ corresponds to subdata selected by Algorithm $\boxed{1}$, then $\sum_{i\in\delta^*} \mathbf{I}_{M_i} \xrightarrow{\mathbb{P}} \mathbf{0}_{(Gp+3G-1)\times(Gp+3G-1)}$ when $N \to \infty$ under one of the following conditions:*

(a) $\mathbf{z}_i \sim \mathrm{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ and for any triplet (g, g', j) with $g, g' \in \{1, ..., G\}, g \neq g'$ and $j \in \{1, ..., p\}$, it holds that $\sum_{l=1}^{p} \rho_{lj}\sigma_{zj}(\beta_{g,l} - \beta_{g',l}) \neq 0$;

(b) $\mathbf{z}_i \sim \mathrm{LN}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ and for any triplet (g, g', j) with $g, g' \in \{1, ..., G\}, g \neq g'$ and $j \in \{1, ..., p\}$, it holds that $\beta_{g,j} - \beta_{g',j} \neq 0$ and $\sum_{l\in\mathcal{L}_{\min,j}} (\beta_{g,l} - \beta_{g',l}) \neq 0$, where $\mathcal{L}_{min,j} = \{l \mid \rho_{lj} = \rho_{\min,j}\ ;\ l = 1, ..., p\}$ and $\rho_{\min,j} = \min_l \rho_{lj} < 0$.

How to derive an efficient algorithm?

- Algorithm based on characterization of an optimal design?
  - Pro: very fast
  - Con:
    - characterization may not be feasible
    - May not be efficient

- New strategy
  - approximate bounded optimal design approach

Rationale

- Approximate deign context
  - Equivalence theorem
- Subdata selection
  - Be selected at most once: $\omega_i = 0$ or $\frac{1}{n}$
- Bounded approximate optimal design
  - $x_1, \dots, x_n \Leftrightarrow \left(x_1, \frac{1}{n}\right), \dots, \left(x_n, \frac{1}{n}\right)$
  - $\Xi = \{\xi | \xi = (x_i, \omega_i), i = 1, \dots, k, 0 \le \omega_i \le \frac{1}{n}\}$

Rationale

- $\xi^* = argmin_{\xi \in \Xi} \Phi(I_\xi)$
- ξ^{exact}: $\omega_i = \frac{1}{n}$, $i = 1, \dots, n$ based on ξ^*
- $\xi^{opt-exact}$: the optimal exact subdata (projected on Ξ)
- For a selected subdata ξ^{sub}(projected on Ξ), its efficiency is $\frac{\Phi(\xi^{opt-exact})}{\Phi(\xi^{sub})}$, and

$$\frac{\Phi(\xi^*)}{\Phi(\xi^{sub})} \leq \frac{\Phi(\xi^{opt-exact})}{\Phi(\xi^{sub})} \leq \frac{\Phi(\xi^{exact})}{\Phi(\xi^{sub})}$$

- To make this strategy work
  - Derive ξ^*
  - $|\Phi(\xi^*) - \Phi(\xi^{exact})| < \varepsilon$

General Equivalence Theorem

Theorem

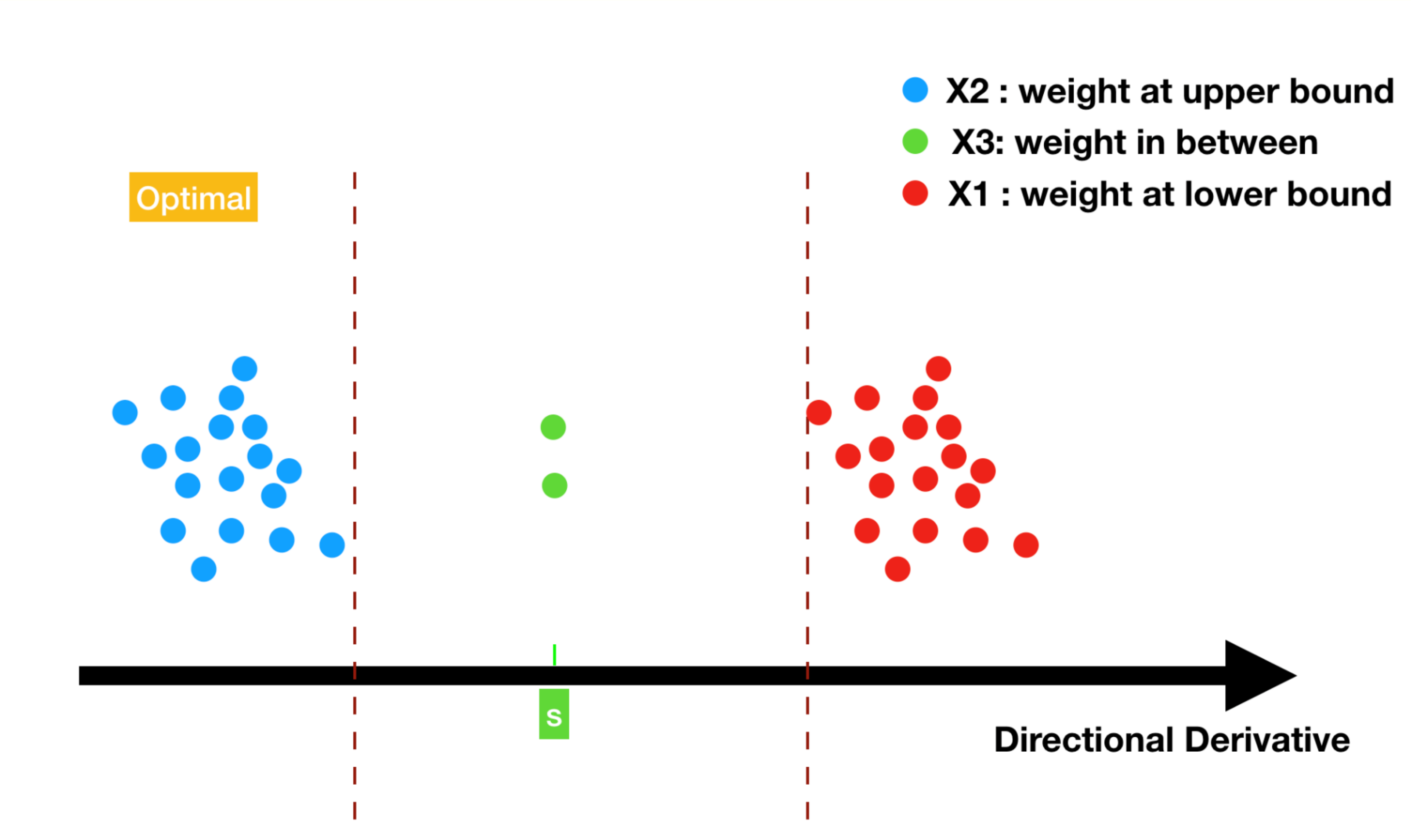
The following two statements are equivalent:

1. ξ^* *is* Φ *optimal in* Ξ_μ^ν
2. *There are subsets* $\mathcal{X}_1$, $\mathcal{X}_2 \subset \mathcal{X}$ *and a number* s *such that*
   (a) $\omega_i = \nu_i$ *for* $x_i \in \mathcal{X}_1$ *and* $\omega_i = \mu_i$ *for* $x_i \in \mathcal{X}_2$,
   (b) $\max\limits_{x_i \in \mathcal{X}_2} F_\Phi(\xi^*; x_i) \leq s \leq \min\limits_{x_i \in \mathcal{X}_1} F_\Phi(\xi^*; x_i)$,
   (c) $F_\Phi(\xi^*; x) = s$ *on* $\mathcal{X} \setminus (\mathcal{X}_1 \cup \mathcal{X}_2)$ *if* $\mathcal{X} \setminus (\mathcal{X}_1 \cup \mathcal{X}_2) \neq \emptyset$ *with*

$$s = \frac{-\sum_{x_i \in \mathcal{X}_1} F_\Phi(\xi^*; x_i)\nu_i - \sum_{x_i \in \mathcal{X}_2} F_\Phi(\xi^*; x_i)\mu_i}{1 - \left(\sum_{x_i \in \mathcal{X}_1} \nu_i + \sum_{x_i \in \mathcal{X}_2} \mu_i\right)} \tag{1}$$

where $\xi^* = \{(x_i, \omega_i)\} \in \Xi_\mu^\nu$, $\mathcal{X}_1 = \{x_i | x_i \in \mathcal{X}, \omega_i = \nu_i\}$ *and* $\mathcal{X}_2 = \{x_i | x_i \in \mathcal{X}, \omega_i = \mu_i\}$; $F_\Phi(\xi^*; x)$ *be the directional derivative of* Φ *in the direction of* x.

General Equivalence Theorem



Algorithm

The Main Group Exchange Algorithm

Input $\mathcal{X}$, k, μ and tol and execute following steps:

1. Initialize $\boldsymbol{\omega}$ such that $\xi = (\mathcal{X}, \boldsymbol{\omega})$ is a valid bounded design.
2. Let $\mathcal{X}_1 = \{\mathbf{x}_i | \mathbf{x}_i \in \mathcal{X}, \omega_i = 0\}$, $\mathcal{X}_2 = \{\mathbf{x}_i | \mathbf{x}_i \in \mathcal{X}, \omega_i = \mu_i\}$, $\mathcal{X}_3 = \{\mathbf{x}_i | \mathbf{x}_i \in \mathcal{X}, 0 < \omega_i < \mu_i\}$.
3. If
   (a) $F_{\Phi_p}(\xi, \mathbf{x}_1^{(minF)}) - F_{\Phi_p}(\xi, \mathbf{x}_3^{(maxF)}) > (-tol)$, and
   (b) $F_{\Phi_p}(\xi, \mathbf{x}_3^{(minF)}) - F_{\Phi_p}(\xi, \mathbf{x}_2^{(maxF)}) > (-tol)$

   then output ξ. Otherwise, go to step 4.
4. (a) If step 3 (a) is not satisfied, move $\mathbf{x}_1^{(minF)}$ to $\mathcal{X}_3$.
   (b) If step 3 (b) is not satisfied, move $\mathbf{x}_2^{(maxF)}$ to $\mathcal{X}_3$.
5. Derive optimal weights $\forall \mathbf{x}_i \in \mathcal{X}_3$ through Newton's method. Then, go to step 3.

where $\mathcal{X}_1$, $\mathcal{X}_2$ and $\mathcal{X}_3$ are defined as a set of points from design space with weights equal to lower bounds, equal to upper bounds and in between lower and upper bounds respectively.(continued...)

Algorithm

(...continued)

Let $\mathbf{x}_1^{(minF)}$, $\mathbf{x}_3^{(maxF)}$, $\mathbf{x}_3^{(minF)}$ and $\mathbf{x}_2^{(maxF)}$ be defined as:

1. $\mathbf{x}_1^{(minF)} = \arg\min_{\mathbf{x}_i \in \mathcal{X}_1} F_{\Phi_p}(\xi, \mathbf{x}_i)$
2. $\mathbf{x}_3^{(maxF)} = \mathit{ifelse}(\mathcal{X}_3 = \emptyset, \mathbf{x}_1^{(minF)}, \arg\max_{\mathbf{x}_i \in \mathcal{X}_3} F_{\Phi_p}(\xi, \mathbf{x}_i))$
3. $\mathbf{x}_3^{(minF)} = \mathit{ifelse}(\mathcal{X}_3 = \emptyset, \mathbf{x}_1^{(minF)}, \arg\min_{\mathbf{x}_i \in \mathcal{X}_3} F_{\Phi_p}(\xi, \mathbf{x}_i))$
4. $\mathbf{x}_2^{(maxF)} = \mathit{ifelse}(\mathcal{X}_2 = \emptyset, \mathbf{x}_3^{(minF)}, \arg\min_{\mathbf{x}_i \in \mathcal{X}_2} F_{\Phi_p}(\xi, \mathbf{x}_i))$

where ifelse(a,b,c) is a function which returns b if condition a is satisfied, otherwise returns c.

Convergence Theorem

$S^{(t)} = (\mathcal{X}_1^{(t)}, \mathcal{X}_2^{(t)}, \mathcal{X}_3^{(t)})$.

Theorem

Let $\frac{\partial f}{\partial \boldsymbol{\theta}^T}$ *be a matrix of full row rank and the initial sets* $S^{(0)}$ *satisfies* $M_{\xi_{S^{(0)}}} > 0$. *Then sequence of designs* $\{\xi_{S^{(t)}}; t \geq 0\}$, *converges to an optimal design which minimizes* $\Phi_p(\boldsymbol{\Sigma}_\xi(f))$, *as* $t \to \infty$.

Bounded optimal design to subdata

Subdata selection method with GE
Input $\mathcal{X}$, n.

1. Let $\mu = \frac{1}{n}$
2. Find optimal design ξ through GE algorithm with upper bound μ and lower bound 0.
3. Output n points of ξ which have largest n weights.

Simulation setup

- N=100,000
- n=10,000
- $X \sim Normal\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, 0.5I + 0.5J\right)$
- $Prob(Y = i|X,k) = \frac{\exp(\boldsymbol{\theta}'_{ki}\boldsymbol{X})}{1+\sum_{l=1}^{2}\exp(\boldsymbol{\theta}'_{kl}\boldsymbol{X})}$
- $Prob(k|X) = \frac{\exp(\boldsymbol{\gamma}'_{k}\boldsymbol{X})}{1+\sum_{j=1}^{2}\exp(\boldsymbol{\gamma}'_{j}\boldsymbol{X})}$

β_1	β2
-5	-5
-5	6
10	-11
-17	18

θ_{11}	θ_{12}	θ_{21}	θ_{22}	θ_{31}	θ_{32}
0	0	0	0	0	0
9	-12	-15	18	27	-30
12	-15	-18	21	30	-33
15	-18	-21	24	33	-36

Competing methods

- SRS1 – 10,000
- SRS2 – 20,000
- Full data: 100,000
- Optimal subdata
  - SRS: 3,000
  - Optimal subdata: 7,000
  - Combined: 3,000+7,000

Criteria

- Computation time
- Efficiency
  - $Prob(Y|X)$
  - Root-mean-squared error of prediction (RMSEP)

- Size of test data: 100,000
- Repeat: 100 times

Comparisons

	SRS1	SRS2	Full	OPT
RMSEP	0.0555	0.0513	0.0497	0.0394
Time(s)	20.08	56.43	328.13	24.85*

24.85 = 9.38 + 1.27 + 14.20

Multinomial logistic regression model: 0.6385

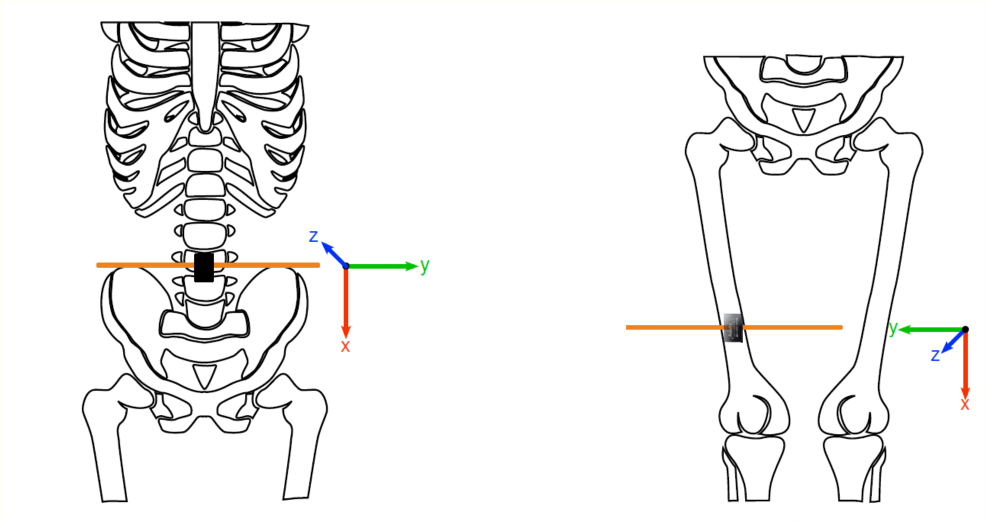
HARTH: A Human Activity Recognition Dataset for Machine Learning

- HARTH - UCI Machine Learning Repository

- A benchmark dataset for researchers to develop innovative machine learning approaches for precise human activity recognition in free living.

HARTH

- A professionally-annotated dataset containing 22 subjects wearing two 3-axial accelerometers for around 2 hours in a free-living setting. The sensors were attached to the right thigh and lower back.
- Video recordings of a chest-mounted camera were used to annotate the performed activities frame-by-frame.



HARTH

- # Instances: 6,461,328
- # Features: 8
  - Time (every 0.02 second)
  - 2×3 sensor signals
  - Label (12 categories)

- Aleksej Logacjov, Kerstin Bach, Atle Kongsvold, H. Bårdstu, P. Mork. 2021
  - Studying 9 categories: walking; running; stairs (ascending); stairs (descending); standing; sitting; lying; cycling (sit); and cycling (stand) (Dataset has 12 categories in total)
  - One-second window
  - X_i: 6×50 matrix
  - 8 competing methods: k-NN, SVM, RF, XGB, BiLSTM, CNN, mCNN
  - leave-one-subject-out cross-validation

HARTH

- Consider 7 categories:
  - walking; running; sitting; lying; cycling (sit); cycling (stand); standing
- N = 115,850
- $\widetilde{X}_i$: 12×1
- Multinomial logistic regression model

HARTH

- Four methods:
  - SRS1 – 10,000
  - SRS2 – 20,000
  - Full data
  - Optimal subdata
    - SRS: 3,000
    - Optimal subdata: 7,000
    - Combined: 3,000+7,000
- Repeat 100 times

Classification accuracy rate

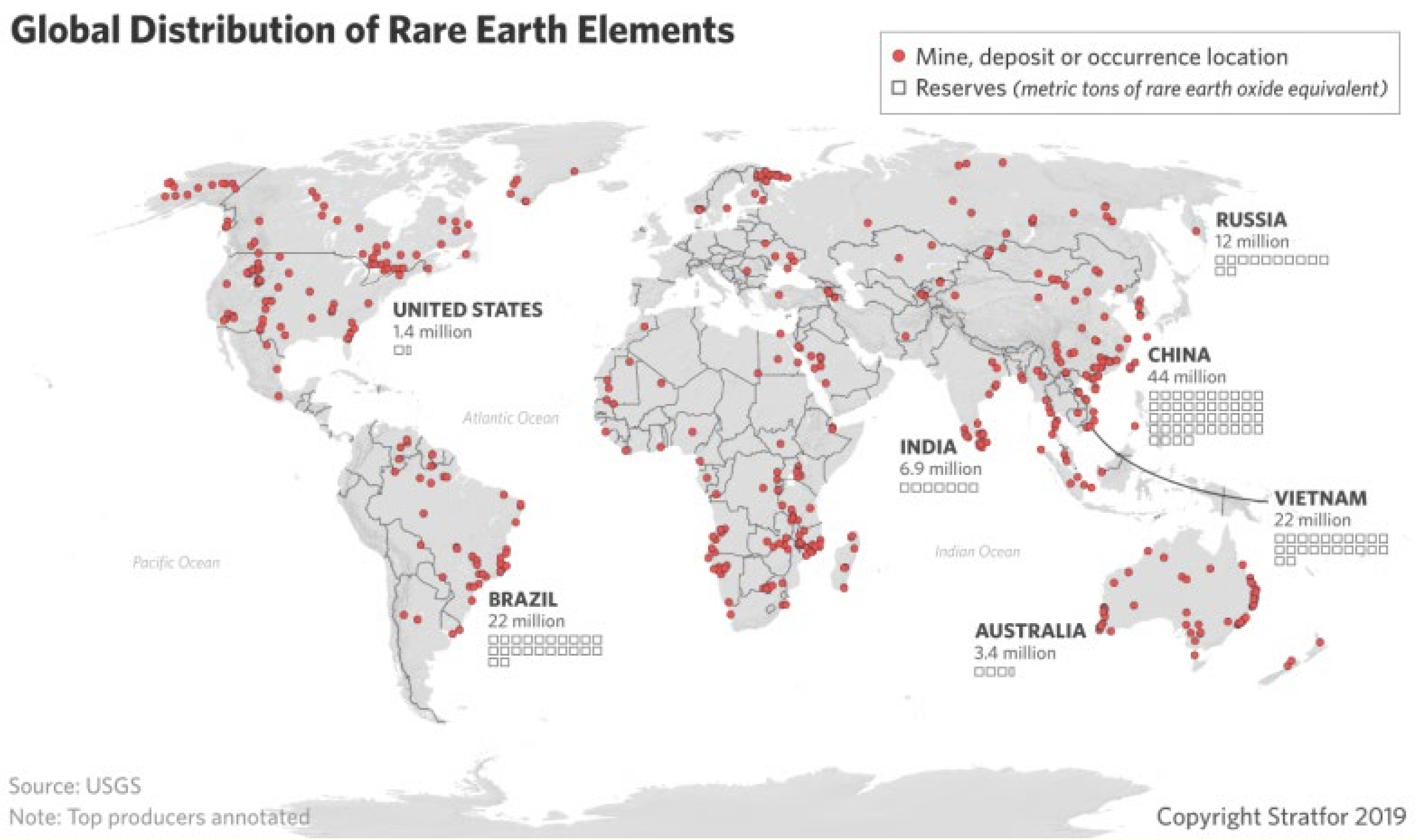
	Walking	Running	Sitting	Lying	Cyling (sit)	Cyling (Stand)	Standing	Average
SRS1	0.91	0.89	0.97	0.98	0.81	0.47	0.91	0.849
SRS2	0.91	0.90	0.98	0.98	0.82	0.47	0.91	0.853
Full	0.92	0.90	0.99	0.98	0.82	0.48	0.91	0.857
OPT	0.92	0.92	0.99	0.99	0.85	0.54	0.91	0.875
SVM	0.90	0.96	0.99	0.95	0.90	0.56	0.86	0.874

HARTH

- Randomly split the data:
  - 2/3 for training and 1/3 for testing
- Four methods:
  - SRS1 – 10,000
  - SRS2 – 20,000
  - Full data
  - Optimal subdata
    - SRS: 3,000
    - Optimal subdata: 7,000
    - Combined: 3,000+7,000
- Repeat 100 times

Classification error rates

	SRS1	SRS2	Full	OPT
Mean	0.0326	0.0315	0.0312	0.0308
Std	0.0032	0.0013	0.0008	0.0013



Global Distribution of Rare Earth Elements
Mine, deposit or occurrence location
Reserves (metric tons of rare earth oxide equivalent)
RUSSIA
12 million
UNITED STATES
1.4 million
CHINA
44 million
INDIA
6.9 million
VIETNAM
22 million
BRAZIL
22 million
AUSTRALIA
3.4 million
Atlantic Ocean
Pacific Ocean
Indian Ocean
Source: USGS
Note: Top producers annotated
Copyright Stratfor 2019

Acknowledgement

- Supported by
  - NSF DMS-2210546