Data-driven Space-filling Design

Dr. Aijun Zhang

The University of Hong Kong

[Joint work with M. Zhang (Sichuan U.) and Y.-D. Zhou (Nankai U.)]

10 November 2018

2018 Workshop on Experimental Design, Nankai University

Outline of the presentation

Small Data Representation

The quest for a small data to represent a big data is important for big data analytics and large-scale machine learning:

- Data compression: save storage, snapshot for big data (small data proxy)
- Data exploration: descriptive statistics, visualization, clustering ...
- Subsampled modeling: supervised learning (regression and classification)

Emerging literature on subsampling or subdata selection, including:

- randomized algorithms (Mahoney, 2011)
- algorithmic leveraging (Ma, Mahoney and Yu, 2015)
- iterative Hessian sketch (Pilanci and Wainright, 2016)
- optimal subsampled linear regression (Wang, Yang and Stufken, 2018)

Low-Discrepancy Points

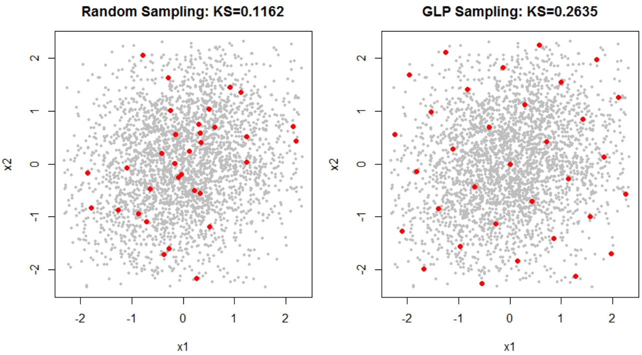

Figure: Toy example of synthetic 3,000 observations. Simple random sampling and good lattice point sampling with KS statistics evaluated.

Empirical $F_{\mathcal{X}}$-discrepancy

- Denote by $F_{\mathcal{X}}(\boldsymbol{z})$ the empirical distribution of big data $\mathcal{X}$ (sample size N):

$$F_{\mathcal{X}}(\boldsymbol{z}) = \frac{1}{N}\sum_{i=1}^{N} I\{\boldsymbol{x}_i \leq \boldsymbol{z}\},\ \boldsymbol{z} \in \mathbb{R}^s \tag{1}$$

- For a small data $\mathcal{P}$ with sample size $n \ll N$, Kolmogorov-Smirnov statistic in the two sample test can be used to measure the degree of representation:

$$D_{\ell_\infty}(\mathcal{P};\mathcal{X}) = \sup_{\boldsymbol{z}\in\mathbb{R}^s} \left|F_{\mathcal{P}}(\boldsymbol{z}) - F_{\mathcal{X}}(\boldsymbol{z})\right| \tag{2}$$

- This is an empirical version of F-discrepancy in Fang and Wang (1994).
- It becomes the classical star discrepancy $D_*(\mathcal{P})$ if $F_{\mathcal{X}}$ is replaced by the uniform distribution on the unit cube C^s.
- $\mathcal{P}$ is a good **data-driven space-filling design** if it has low $D_{\ell_\infty}(\mathcal{P};\mathcal{X})$.

Outline of the presentation

Connection with Star Discrepancy

- Given big data $\mathcal{X} \subset \mathbb{R}^s$, denote $F_{\mathcal{X}_{(j)}}(x) = N^{-1}\sum_{i=1}^{N} I(x_{ij} \leq x)$ as the jth marginal empirical distribution. Define for $\boldsymbol{x} \in \mathbb{R}^s$ a multivariate mapping:

$$T_{\mathcal{X}}(\boldsymbol{x}) = \left(F_{\mathcal{X}_{(1)}}(x_1), \ldots, F_{\mathcal{X}_{(s)}}(x_s)\right) \tag{3}$$

- Joint independence assumption:

$$F_{\mathcal{X}}(\boldsymbol{x}) = \prod_{j=1}^{s} F_{\mathcal{X}_{(j)}}(x_j), \ \boldsymbol{x} \in \mathbb{R}^s \tag{4}$$

Theorem (Connection with Star Discrepancy)

Let the number of repeated observations within $\mathcal{X}$ is upper bounded, then

$$D_{l_\infty}(\mathcal{P};\mathcal{X}) = D_*(T_{\mathcal{X}}(\mathcal{P})) + O(1/N^*) \tag{5}$$

where $N^* = \min_{j\in[s]} N_j$ with N_j denoting the number of distinct values in $\mathcal{X}_{(j)}$.

Data-driven Space-filling Design Construction

- For one-dimensional $\mathcal{X}$, DSD construction by the inversion method:

Corollary (1-D Construction)

Given the big data $\mathcal{X} \subset \mathbb{R}$ with sample size N, the n-run design $\mathcal{P}$ given by

$$\xi_i = F_{\mathcal{X}}^{-1}\left(\frac{2i-1}{2n}\right), \; i = 1, \ldots, n$$

is asymptotically optimal (as $N \to \infty$) under the empirical $F_{\mathcal{X}}$-discrepancy.

- For multi-dimensional $\mathcal{X}$, the joint independence is a needed condition. We propose two preprocessing methods:
  (1) SVD rotation: a simple approach for homogeneous data;
  (2) Rosenblatt transform: an advanced approach for manifold data.

Rotation-Inversion Construction

Input: Big data $\mathcal{X} \in \mathbb{R}^s$, Uniform design $\mathcal{D} \in C^s$.

1) Perform SVD for $\mathcal{X}$ to obtain the rotation $\boldsymbol{V}$, the singular-valued matrix Λ, and the rotated data $\mathcal{Z}$;
2) For each point $\zeta_i \in \mathcal{D}$, perform the $T_{\mathcal{Z}}^{-1}$ transform

$$\eta_{ij} = F_{\mathcal{Z}_{(j)}}^{-1}(\zeta_{ij}), \; j = 1, \ldots, s.$$

3) Generate the point set $\mathcal{P}$ by $\xi_i = \boldsymbol{V}\Lambda\eta_i$ for each i.

Output: Data-driven space-filling design $\mathcal{P}$.

Rotation-Inversion Construction

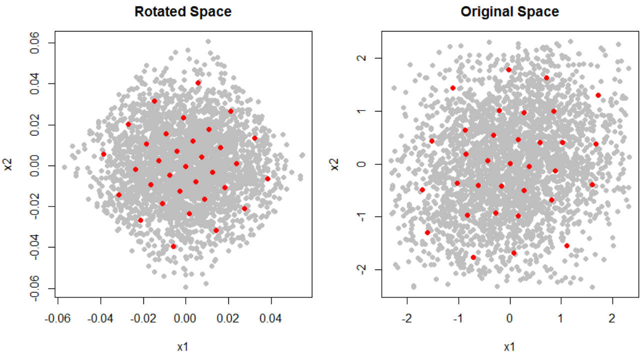

Figure: Toy example with data-driven space-filling design, by the rotation and inversion method based on the leave-one-out Fibonacci lattice in 2D.

Outline of the presentation

Subdata Selection

- **Purpose:** to select a subdata $\mathcal{P}^{\dagger} \subset \mathcal{X}$ with the preserved distribution.
- We develop an effective data-driven space-filling sampling algorithm based on low empirical $F_{\mathcal{X}}$-discrepancy design.
- For each design point, its nearest neighbor within the non-uniform grid (in the rotated space) is sampled.
- Parallel processing strategy is used for speeding up the computation.

Non-uniform Stratification

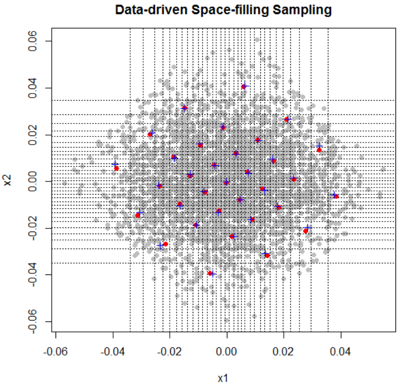

Figure: Data-driven space-filling sampling by non-uniform stratification and nearest neighbor in the rotated space.

Space-filling Subdata Selection

Input: Big data $\mathcal{X} \in \mathbb{R}^s$, data-driven space-filling design $\mathcal{P}$

1) Perform SVD $\mathcal{X} = \mathcal{Z}\Lambda \boldsymbol{V}^T$;
2) Parallel for each point z_i in $\mathcal{Z}$, label its cell index

$$I_j(i) = \left\lceil nF_{\mathcal{Z}_{(j)}}(z_{ij}) \right\rceil, \; j = 1, \ldots, s.$$

3) Obtain $\eta_k = \Lambda^{-1}\boldsymbol{V}^T\xi_k$ for each ξ_k in $\mathcal{P}$;
4) Parallel for each η_k, identify its neighboring cells and find the nearest sample with index i_k^*.

Output: Space-filling subdata $\mathcal{P}^\dagger$ with sample indexes $\{i_k^*, k = 1, \ldots, n\}$.

Numerical Examples

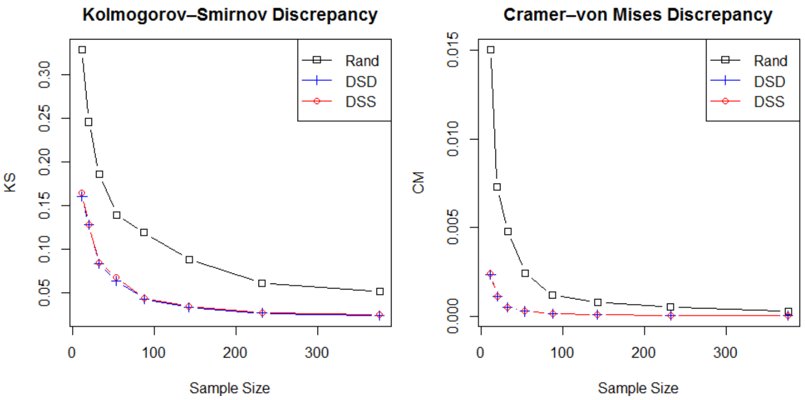

Figure: Toy example with data-driven space-filling design and sampling, as compared to simple random sampling.

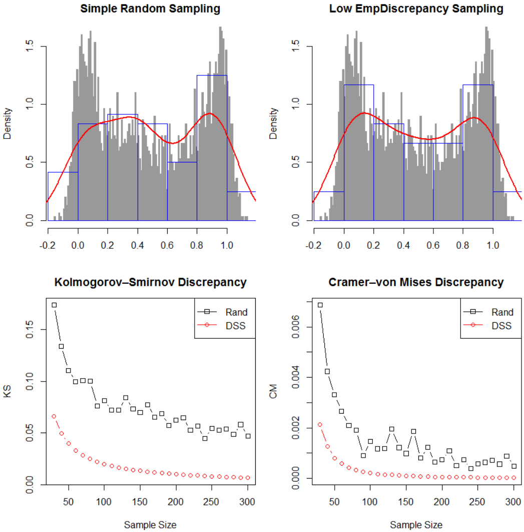

Figure: Simulated data from the contaminated Beta distribution.

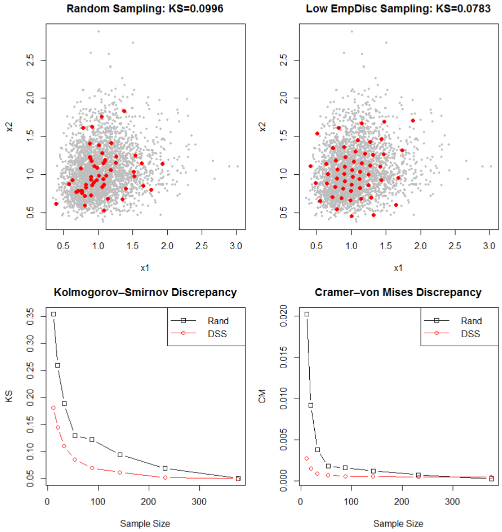

Figure: Simulated data from the bivariate lognormal distribution.

Outline of the presentation

Application: Big Data Exploration

- Real data: proptein tertiary structure dataset from the UCI ML-repository. It contains 45,730 samples with 9 continuous attributes.
- Directly exploring such big data is sometimes cumbersome for graphical visualization tasks (e.g. pairwise scatter plots).
- We instead perform large-scale unsupervised learning based on the proposed space-filling subdata selection method.
- By PCA with the scaling option, PC1 and PC2 altogether explain 83.7% of total variation in the original data.
- The subdata selection is performed on these two PC coordinates based on the leave-one-out Fibonacci lattice with $n = 986$.

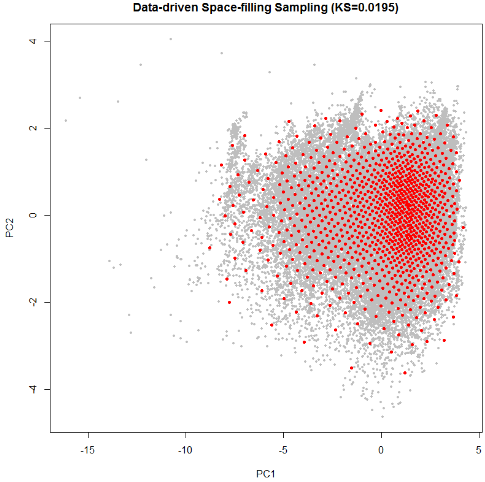

Figure: Data-driven space-filling sampling for the protein structure data in the principal component space: KS $= 0.0195$, sampling ratio 2.16%.

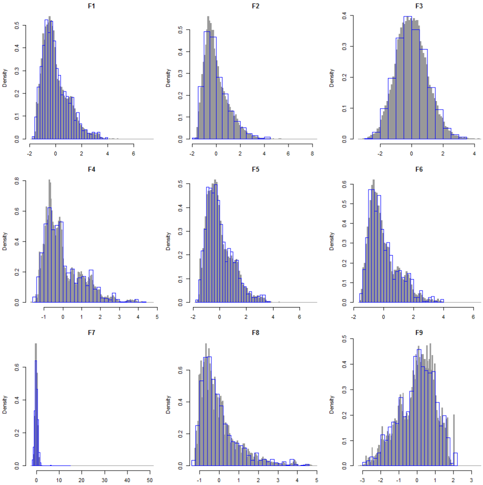

Figure: Histograms of each scaled attribute by the original data (background in gray color) and the selected subdata (foreground in blue).

	F1	F2	F3	F4	F5	F6	F7	F8	F9
F1	0.00	0.00	0.10	0.00	0.00	0.01	0.28	0.04	0.01
F2	0.00	0.00	0.03	0.01	0.00	0.01	0.30	0.05	0.01
F3	0.10	0.03	0.00	0.15	0.10	0.00	0.26	0.29	0.18
F4	0.00	0.01	0.15	0.00	0.00	0.01	0.24	0.03	0.02
F5	0.00	0.00	0.10	0.00	0.00	0.01	0.28	0.05	0.01
F6	0.01	0.01	0.00	0.01	0.01	0.00	0.28	0.03	0.03
F7	0.28	0.30	0.26	0.24	0.28	0.28	0.00	0.19	0.19
F8	0.04	0.05	0.29	0.03	0.05	0.03	0.19	0.00	0.06
F9	0.01	0.01	0.18	0.02	0.01	0.03	0.19	0.06	0.00

Table: Relative errors of pairwise correlation approximation by the subdata.

Outline of the presentation

Conclusion

- We have proposed the notion of data-driven space-filling design (DSD) under the empirical $F_{\mathcal{X}}$-discrepancy;
- By an established asymptotic equivalence between discrepancy measures, we propose a simple rotation-inversion method for DSD construction;
- Then, an efficient and effective DSD-based subsampling algorithm is developed for the purpose of subdata selection;
- By numerical examples and real data analysis, the proposed subsampling method is demonstrated to be useful for big data exploration;
- The proposed data-driven space-filling design has great potential for big data analytics and large-scale machine learning.

Ongoing Work

Following the proposed notion of data-driven space-filling design, we are currently investigating the following interesting problems:

1. Empirical Rosenblatt transformation for manifold data analytics with heterogeneous distribution and/or irregular experimental domain.
2. Asymptotic equivalence between the generalized $F_{\mathcal{X}}$-discrepancy and RKHS-induced ℓ_2-discrepancy (e.g. CD2, WD2, MD2);
3. Modified Koksma-Hlawka inequality based on empirical $F_{\mathcal{X}}$-discrepancy, an important theory for small data functional approximation;
4. Applications to large-scale machine machine under the ERM (empirical risk minimization) paradigm.

Thank You !

Q&A or Email ajzhang@hku.hk。