

Data-driven Space-filling Design

Aijun Zhang¹, Mei Zhang² and Yong-Dao Zhou³

¹*Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong SAR, China*

²*College of Mathematics, Sichuan University, China*

³*School of Statistics and Data Science, Nankai University, China*

Abstract

The quest for a small data to represent a big data is important for data compression, exploration and subsampled modeling. We consider the data-driven space-filling design with the criterion of empirical F -discrepancy. Asymptotic optimality is established for an inversion construction method based on existing uniform experimental designs. When the small data is required to be a subset of the big data, we develop an effective subdata selection algorithm based on the proposed data-driven space-filling design. Such algorithm has potential applications in large-scale machine learning in both supervised and unsupervised settings.

Keywords

Space-filling design, Empirical F -discrepancy, Big data subsampling, Large-scale machine learning.