中央研究院統計科學研究所

學 術 演 講

講 題：Boosting Data Analytics with Synthetic Volume Expansion

演 講 人：Prof. Xiaotong Shen

School of Statistics, University of Minnesota

時 間： 2024-05-20(Mon) 10:30-12:00

地 點：Auditorium, B1F, Institute of Statistical Science；The tea reception will be held at 10:10.

備 註：Online live streaming through Cisco Webex will be available.

Abstract

Synthetic data generation heralds a paradigm shift in data science, addressing the challenges of data scarcity and privacy and enabling unprecedented performance. As synthetic data gains prominence, questions arise regarding the accuracy of statistical methods compared to their application on raw data alone. Addressing this, we introduce the Synthetic Data Generation for Analytics framework, which applies statistical methods to high-fidelity synthetic data produced by advanced generative models like tabular diffusion models through knowledge transfer. These models, trained using raw data, are enriched with insights from relevant studies. A significant finding within this framework is the generational effect: the error of a statistical method initially decreases with the integration of synthetic data but may subsequently increase. This phenomenon, rooted in the complexities of replicating raw data distributions, introduces the "reflection point," an optimal threshold of synthetic data defined by specific error metrics. Through one data example, we demonstrate the effectiveness of this framework. This work is joint with Y. Liu and R. Shen. Please click here for participating the talk online.

中 央 研 究 院

統計科學研究所