香港中文大學統計學系
Department of Statistics
THE CHINESE UNIVERSITY OF HONG KONG

SEMINAR

DEPARTMENT OF STATISTICS
THE CHINESE UNIVERSITY OF HONG KONG

Statistical Significance of Clustering for High Dimensional Data

INVITED SPEAKER

Yufeng Liu
Professor
Department of Statistics and Operations Research
The University of North Carolina at Chapel Hill

TIME

June 25, 2024 (Tue) · 2:30 pm - 3:30 pm

VENUE

LSB LT2 · Lady Shaw Building - LT2 · CUHK

ABSTRACT

Clustering serves as a fundamental tool for exploratory data analysis, but a key challenge lies in determining the reliability of the clusters identified by these methods, differentiating them from artifacts resulting from natural sampling variations. In this talk, I will present statistical significance of clustering (SigClust) as a cluster evaluation tool for high dimensional data. To begin, we define a cluster as data originating from a single Gaussian distribution and frame the assessment of statistical significance of clustering as a formal testing procedure. Addressing the challenge of high-dimensional covariance estimation in SigClust, we employ a combination of invariance principles and a factor analysis model. I'll also discuss an enhanced SigClust using multidimensional scaling (MDS) on dissimilarity matrices. SigClust for hierarchical clustering will be presented as well. Simulations and real data, including cancer subtype analysis, validate SigClust's effectiveness in assessing clustering significance.