For favour of Posting

THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

Department of Mathematics

SEMINAR ON STATISTICS

False Discovery Rate Control via Data Splitting for Testing-after-Clustering

By

Prof. Lijun WANG

Zhejiang University

Abstract

Testing for differences in features between clusters in various applications often leads to inflated false positives when practitioners use the same dataset to identify clusters and then test features, an issue commonly known as ``double dipping". To address this challenge, inspired by data-splitting strategies for controlling the false discovery rate (FDR) in regressions (Dai et al., 2023), we present a novel method that applies data-splitting to control FDR while maintaining high power in unsupervised clustering. We first divide the dataset into two halves, then apply the conventional testing-after-clustering procedure to each half separately and combine the resulting test statistics to form a new statistic for each feature. The new statistic can control the FDR due to its property of having a sampling distribution that is symmetric around zero for any null feature. To further enhance stability and power, we suggest multiple data splitting, which involves repeatedly splitting the data and combining results. Our proposed data-splitting methods are mathematically proven to asymptotically control FDR in Gaussian settings. Through extensive simulations and analyses of single-cell RNA sequencing (scRNA-seq) datasets, we demonstrate that the data-splitting methods are easy to implement, adaptable to existing single-cell data analysis pipelines, and often outperform other approaches when dealing with weak signals and high correlations among features.

Date : 22 January 2026 (Thursday)

Time : 2:00p.m.-3:00p.m.

Venue : Room 2463 (near Lift 25/26)

All are welcome!