THE UNIVERSITY OF GEORGIA
DEPARTMENT OF STATISTICS

Colloquium Series

Yingying Wei, Ph.D.

Department of Statistics

The Chinese University of Hong Kong

Friday, March 29, 2019

3:30pm in room 102, Caldwell Building

Meta-Clustering of Genomic Data

Like traditional meta-analysis which pools effect sizes across studies to improve statistical power, it is of increasing interest to conduct clustering jointly across datasets to identify disease subtypes for bulk genomic data and discover cell types for single-cell RNA-sequencing (scRNA-seq) data. Unfortunately, due to the prevalence of technical batch effects among high-throughput experiments, directly clustering samples from multiple datasets can lead to wrong results. The recent emerging meta-clustering approaches require all datasets to contain all subtypes, which is not feasible for many experimental designs.

In this talk, I will present our Batch-effects-correction-with-Unknown-Subtypes (BUS) framework. BUS is capable of correcting batch effects explicitly, grouping samples that share similar characteristics into subtypes, identifying features that distinguish subtypes, and enjoying a linear-order computational complexity. We prove the identifiability of BUS for not only bulk data but also scRNA-seq data whose dropout events suffer from missing not at random. We mathematically show that under two very flexible and realistic experimental designs--the "reference panel" and the "chain-type" designs--true biological variability can also be separated from batch effects. BUS outperforms the state-of-the-art methods on real data.