
Recent Advances in Data Science with Applications to Epidemiology and Genetics
Avancées récentes en science des données et applications en épidémiologie et en génétique
(Org: Candemir Cigsar and/et Yildiz Yilmaz (Memorial University))

LAURENT BRIOLLAIS, Lunenfeld-Tanenbaum Research Institute

The Scalable Birth-Death MCMC Algorithm for Mixed Graphical Model Learning with Application to Genomic Data Integration

Recent advances in biological research have seen the emergence of high-throughput technologies with numerous applications. In cancer research, the challenge is now to perform integrative analyses of high-dimensional multi-omic data with the goal to better understand genomic processes that correlate with cancer outcomes. We propose here a novel mixed graphical model approach to analyze multi-omic data of different types (continuous, discrete and count) and perform model selection by extending the Birth-Death MCMC (BDMCMC) algorithm. We compare the performance of our method to the LASSO and the standard BDMCMC methods using simulations and found that our method is superior in terms of both computational efficiency and the accuracy of the model selection results. Finally, an application to the TCGA breast cancer data shows that integrating genomic information at different levels (mutation and expression data) leads to better subtyping of breast cancers.

JC LOREDO-OSTI, Memorial University

Stochastic modelling of an infectious disease outbreak

There are many ways to model an infectious disease outbreak. Hawkes processes are a class of self-exting processes that can be used in numerous applications to model event clustering and causal inference. In spite of their simple formulation, this class of processes can model quite complex phenomena. While most literature on Hawkes processes refers to continuous-time processes, there are discrete-time variants that can be viewed as stochastic versions of popular compartmental models used in epidemiology. Due to its flexibility, Hawkes processes are a good alternative to model disease outbreaks with public health interventions and other time-dependent covariates. In this presentation, we discuss the link/equivalence between variants of SIR models and Hawkes processes to model Covid-19 in small populations.

BRADY RYAN, University of Michigan

Using External Reference Panel and Meta-Analysis Summary Statistics for Rare-Variant Aggregation Tests

Genome-wide association studies (GWAS) have identified thousands of associations between common genetic variants and a wide range of human diseases and traits. These studies are often underpowered to identify associations with rare genetic variants, which are thought to contribute to the heritability of many common diseases and traits. Aggregation tests pool the genetic signal across multiple variants in a region of the genome to test the cumulative effect of these variants on a disease or trait. These aggregation tests can increase the power to detect rare variant genetic association in these regions. To further increase power, meta-analysis is employed to pool information across studies via summary statistics such as effect sizes and p-values. To perform proper aggregation test meta-analysis, accurate estimates of the covariances for the single-variant test statistics are also needed. Covariance files are often too large to be shared and estimation requires access to individual level data for each of the participating studies. Unfortunately, individual-level genetic data is often unable to be shared due to privacy concerns. In this study, we apply a previously proposed method of estimating single-variant test statistic covariance from an external reference panel to perform aggregation tests on a variety of traits from the UK Biobank. We propose a two-stage approach by first filtering genes using a null covariance to perform aggregation tests and in stage two testing only those genes passing a p-value threshold. We find this to be an efficient strategy which can lead to significant computational improvements.