Third International Conference on STATISTICS FOR TWENTY-FIRST CENTURY

Trivandrum, India
December 14-16, 2017

Invited Session

Statistical Methods for Big Data

All Participants form:

Division of Biostatistics and Data Science
Department of Population Health Sciences
Medical College of Georgia at Augusta University
Augusta, USA

Organizer & Chair
Varghese George

1. Deepak Ayyala: Discrete Multivariate Models for Genomic Count Data

In gene expression experiments such as bulk and single cell RNA-Seq, the data collected is generally converted to discrete count data. There are two main problems of interest in single cell RNA-seq studies: cell-type detection and gene selection. Univariate models ignore the inherent dependence structure in the data, making them inefficient for such data. To address the two problems simultaneously, we explored the application of discrete multivariate models to single cell gene expression data, particularly the Dirichlet-multinomial distribution. In this talk, we will present a penalized Dirichlet-multinomial model for analyzing single cell RNA-Seq data. An ℓ_2-penalty is imposed to the likelihood to regularize differential abundance of genes between the different cell types. We developed an approximate Newton's method that is fast and computationally efficient even when using a large set of genes. We applied our method to several data sets to detect cell types and identify differentially expressed genes in different human tissues.

2. Santu Ghosh: A Two-Sample Test for Data with Large Dimension and Small Sample Size

With the rapid development of modern computing techniques, scientists are encountering data with much higher dimension, i.e., “large-p-small-n” setting. Consequently, due to their loss of accuracy or power, some classical statistical inferences are being challenged by non-exact approaches. This work is concerned with the two-sample test for population mean vector of non-normal high-dimensional multivariate data. Several tests for high dimensional mean vector, based on modifying the classical Hotelling T^2 test, have been proposed in the literature. For the above testing problem, we develop an alternative approach. The proposed procedure, known as the generalized component test (GCT), is free from the estimation of the covariance matrix. Each component of GCT is constructed by transforming a univariate two-sample studentized pivot into another one based on a symmetrizing transformation. The test statistic has a standard normal distribution as its limiting distribution, and the asymptotic result holds for divergent p. We compare the performance of the proposed test to its counterparts using simulations. The usefulness of our method is further illustrated through an example.

3. Daniel Linder: Network Topology Inference via Synthetic Likelihood with Variable Selection Priors

Systems biologists seek to understand the higher-level organizational properties that a biological system exhibits from the interactions of its many lower level components. It is then typically beneficial to learn, in a statistical sense, the nature of these interactions from experimental data; this is sometimes referred to as reverse engineering of the biological system. This is known to be a challenging problem due to high-dimensionality of the data and intractability of the exact likelihood. In this talk, we will discuss methods for learning the kinetic parameters from trajectories of stochastic systems that are measured at discrete time points. The focus will be on estimating the system topology, or network structure, with a method based on the notion of a synthetic likelihood coupled with Bayesian variable selection priors.