Yale SCHOOL OF PUBLIC HEALTH

Biostatistics

Scalable inference on complex dependency through multi-scale divide-and-conquer

Li Ma, PhD
Associate Professor
Department of Statistical Science
Duke University

ABSTRACT

A fundamental inference task is the testing and learning of dependency structures between random variables. Existing approaches to quantifying complex (i.e., nonlinear or local) dependencies require polynomially complex algorithms and in addition resampling to evaluate statistical significance, and hence their applicability to modern massive data is limited. We introduce a multi-scale divide-and-conquer framework for testing and identifying complex dependency between random variables that overcomes these limitations. Under this strategy, dependency is characterized by the odds ratios on a collection of 2 by 2 contingency tables formed by sequentially partitioning the sample space, and statistical inference proceeds through scanning over these 2 by 2 tables from coarse to fine resolutions and carry out a simple test on the odds ratio for each table. The evidence from the tables is combined through multiple testing adjustment. We show that due to a factorization of Fisher's multivariate hypergeometric (MHG) likelihood into the product of the univariate hypergeometric likelihoods, when Fisher's exact test is adopted on the 2 by 2 tables, a sequential generative representation for the MHG model implies the mutual independence (up to deviation due to discreteness) among the individual tests in the absence of dependence. This leads to an exact characterization of the joint null distribution of the p-values and gives rise to an effective inference recipe that achieves finite-sample guarantees through simple multiple testing procedures such as Šidák and Bonferroni corrections. The computational complexity of the inference algorithm is approximately linear in the sample size, which along with the avoidance of resampling makes it ideal for analyzing massive datasets. If time permits, I will also present two further generalizations of the framework: (i) testing independence between random vectors possibly of large dimensions and (ii) testing conditional independence between two random variables conditional on a third.

12:00 Noon, Tuesday, November 5, 2019
47 College Street, Room 106B
11:45 AM - Lunch served outside Room 106B