Department of Biostatistics and Medical Informatics Seminar

Heping Zhang, PhD

Susan Dwight Bliss Professor of Biostatistics
Professor of Child Study
Professor of Statistics and Data Science
Yale University

Friday, September 30, 2022
12:00-1:00 pm
Biotech Center Auditorium *or* via Zoom Link
https://uwmadison.zoom.us/j/97419332700

Genes, Brain, and Us

Abstract: Many human conditions, including cognition, are complex and depend on both genetic and environmental factors. After the completion of the Human Genome Project, genome-wide association studies have associated genetic markers such as single-nucleotide polymorphisms with many human conditions and diseases. Despite the progress, it remains difficult to identify genes and environmental factors for complex diseases - the so-called geneticist's nightmare. Furthermore, although the impact of these discoveries on human health is not shock and awe, “drugs with support from human genetic studies for related effects succeed from phase I trials to final approval twice as often as those without such evidence.” Therefore, it is important and promising, while challenging, to identify genetic variants for complex human health-related conditions.

This talk is not intended to provide a comprehensive review of massive progress of related methods and discoveries. Instead, I will focus on some of the work that many of my students assisted me in over the past several years. The first area is the identification of super-variants. A super-variant is a set of alleles in multiple loci of human genome although unlike the loci in a gene, contributing loci to a super-variant can be anywhere in the genome. The concept of super-variant follows a common practice in genetic studies by the means of collapsing a set of variants, specifically single nucleotide polymorphisms. The novelty and challenge lie in how to find, replicate, interpret, and eventually make use of the super-variants. Our work has been mainly based on the use of tree- and forest-based methods, and a data analytic flow that we proposed in 2007, which in retrospect resembles the spirit of “deep learning” that Hinton coined in 2006. The second area is our progress in conducting statistical inference for high dimensional and structured data objects. Such data objects not only more and more commonly appear in imaging genetic studies, but also in other areas of data science including artificial intelligence. They do not belong to a Euclidean space for which most of the statistical theory and methods such as the distribution function are developed. How do we analyze data objects in non-Euclidean spaces?

School of Medicine
and Public Health
UNIVERSITY OF WISCONSIN–MADISON