STATISTICAL AGGREGATION IN MASSIVE DATA ENVIRONMENT

Nan Lin
Department of Mathematics
Division of Biostatistics
Washington University in St. Louis

Abstract: Due to their size and complexity, massive data sets bring many computational challenges for statistical analysis, such as overcoming the memory limitation and improving computational efficiency of traditional statistical methods. In this talk, I will discuss the statistical aggregation strategy to conquer such challenges posed by massive data sets. Statistical aggregation partitions the entire data set into smaller subsets, compresses each subset into certain low-dimensional summary statistics and aggregates the summary statistics to approximate the desired computation based on the entire data. Results from statistical aggregation are required to be asymptotically equivalent. Statistical aggregation is particularly useful to support sophisticated statistical analyses for online analytical processing in data cubes. We will detail its application to two large families of statistical methods, estimating equation estimation and U-statistics.

DATE: Friday, September 18, 2009
TIME: 2 p.m. - 3 p.m.
PLACE: Bryan Hall, Room 305

Host: Arye Nehorai

Short bio: Nan Lin is an Assistant Professor of Mathematics and Bio-statistics at the Washington University in St. Louis. He received the Ph.D. in Statistics from University of Illinois at Urbana-Champaign in 2003. He has also worked as a Postdoctoral Associate at Yale University School of Medicine from 2003 to 2004. His research interest includes statistical computing, massive data analysis, Bayesian regularization methods, bioinformatics and psychometrics.