Analysis of Big Data

Description
The amount of data in our world is exploding and analyzing large datasets is becoming central in science, technology, and our society. This course introduces new statistical methods and computational tools for analyzing Big Data. This course covers the following topics:

1. Statistical challenges of Big Data Analysis
2. Handling massive data: divide-and-conquer statistical inference
3. Handling high dimensional data: regularization method
4. Handling complex and noisy data: robust semiparametric method
5. Large-scale statistical optimization: model-based optimization

The main focus of this course is to introduce new methodology and explain theoretical justification of these methods at an intuitive level.

Prerequisite: introductory statistical inference and basic knowledge of convex optimization. Targeted audience include graduate students, statistical researchers, and data analytical professionals.

Reference
Lecture notes and supplementary reading materials will be provided.

Details
1. Statistical challenges of analyzing Big Data
   • Heterogeneity vs. Commonality
   • Challenges of massive data
   • Challenges of high dimensional data
   • Challenges of complex and noisy data

2. Analyzing massive data
   • Introduction of MapReduce framework
• Divide-and-conquer approach for statistical estimation
• Divide-and-conquer approach for asymptotic statistical inference
• Divide-and-conquer approach for bootstrapping based statistical inference

3. Analyzing high dimensional data
• Introduction of regularization framework
• Sparsity in predictive analysis: sparse regression and classification
• Sparsity in exploratory analysis: sparse covariance estimation, graphical model and PCA
• Sparse multivariate time series analysis
• Multi-task learning and calibrated inference

4. Analyzing complex and noisy data
• A new perspective on the roles of Gaussianity and linearity in multivariate analysis
• Relaxing Gaussianity: transelliptical modeling and regularized rank based methods
• Relaxing Linearity: sparse additive modeling and semiparametric GLMs
• New applications: Semiparametric topic models, missing values in GLMs, semiparametric portfolio selection

5. Large-scale statistical optimization
• Parametric simplex method
• Douglas-Rachford operator splitting method (or ADMM method)
• Peaceman-Rachford operator splitting method
• Statistical optimization for large-scale nonconvex optimization
• Tradeoff computation and statistics

---------------------------

Computing

If time permits, related R packages will be introduced.