Title: Robust, scalable and fast bootstrap method for analyzing large scale data

Speaker: Professor Visa Koiuvnen, Department of Signal Processing and Acoustics, Aalto University, Finland

Date: Tuesday, June 21, 2016

Abstract:

This talk address the problem of performing statistical inference for large scale data sets i.e., Big Data. The volume and dimensionality of the data may be so high that it cannot be processed or stored in a single computing node. We propose a scalable, statistically robust and computationally efficient bootstrap method, compatible with distributed processing and storage systems. The proposed method combines distributed bootstrap with computationally efficient fixed point equations. Many statistically robust and highly efficient estimators lend themselves to such computation. Bootstrap resamples are constructed from a smaller number of distinct data points that correspond to multiple disjoint subsets of data, similarly to the bag of little bootstrap method (BLB) by Kleiner et al. This facilitates distributed storage and computation in inference. Significant saving in computation is achieved by avoiding the re-computation of the estimator for each bootstrap sample. Instead, an initial estimate is improved by using an efficient fixed-point estimation equation. An analytically found correction term compensating for underestimated variability is applied. Our proposed bootstrap method facilitates the use of highly robust statistical methods in analyzing large scale data sets. The favorable statistical properties of the method are established analytically. Numerical examples on finding confidence intervals in parameter estimation and hypothesis testing problems demonstrate scalability, low complexity and robust statistical performance of the method in analyzing large data sets.