RUTGERS
School of Arts and Sciences

Department of Statistics and Biostatistics
Rutgers, The State University of New Jersey
Busch Campus
501 Hill Center
101 Frelinghuysen Road
Piscataway, NJ 08854

Web: stat.rutgers.edu
Email: office@stat.rutgers.edu
Phone: 848-445-2690
Fax: 848-445-3428

RUTGERS UNIVERSITY
DEPARTMENT OF STATISTICS AND BIOSTATISTICS
www.stat.rutgers.edu

Seminar

Speaker: **Professor Peter Z. G. Qian**
University of Wisconsin-Madison

Title: **OEM for Big Data**

Time: **3:20 – 4:20pm, *Wednesday*, November 20, 2013**

Place: **552 Hill Center**

Abstract

Big data now arise in internet, marketing, engineering and many other fields. We propose a new statistical algorithm, called OEM (a.k.a. orthogonalizing EM), intended for fitting various least squares problems. The first step, named active orthogonization, orthogonalizes an arbitrary regression matrix by elaborately adding more rows. The second step imputes the responses of the new rows. The third step solves the least squares problem of interest for the complete orthogonal design. The second and third steps have simple closed forms, and iterate until convergence. The maximum number of points required in active orthogonalization is bounded by the number of columns of the original matrix, which makes OEM particularly appealing for Big Data with large sample size. The algorithm works for ordinary least squares and regularized least squares with the lasso, SCAD, MCP and other penalties. It has several attractive theoretical properties. For the ordinary least squares with a singular regression matrix, an OEM sequence converges to the Moore-Penrose generalized inverse-based least squares estimator. For the SCAD and MCP, an OEM sequence can achieve the oracle property. For ordinary and regularized least squares with various penalties, an OEM sequence converges to a point having grouping coherence for fully aliased regression matrices. Results on convergence rate of OEM show that for the same data set, OEM converges faster for regularized least squares than ordinary least squares. This provides a theoretical comparison between these methods. The underlying idea of OEM can be extended to other Big Data problems with singularity issues like fitting massive Gaussian process models. Numerical examples are provided to illustrate the proposed algorithm. This talk is based on joint work with Shifeng Xiong at Chinese Academy of Science and Bin Dai at Tower Research Capital.

***** Refreshments will be served @2:50pm in Room 502 Hill Center *****