Nuclear/Astro Seminar

Tuesday, October 30th - 3:45 p.m. - Corcoran Hall 404 A

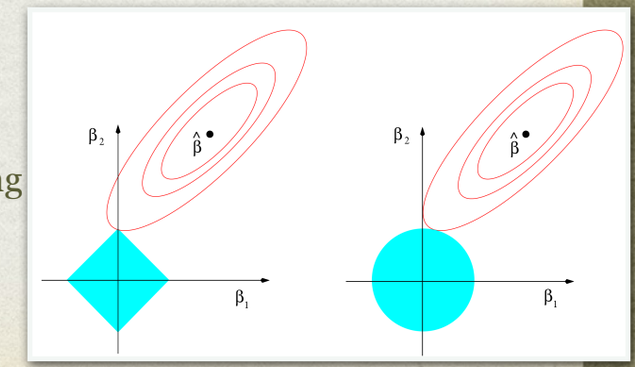
EMRE BARUT

DEPARTMENT OF STATISTICS - GWU

The LASSO and His Friends: A Non-Comprehensive Tour Through the Recent Advances in High Dimensional Statistics

Variable selection, the process of identifying "useful" or "important" covariates, is one of the most fundamental problems in statistics. This problem has become even more critical in the modern data age, where one has to perform variable selection using a dataset with a small sample size and a very large number of variables. For instance, in genome-wide association studies, one seeks to find SNPs associated with a specific outcome, when the sample size, given by the number of people in the study, is usually two orders of magnitudes smaller compared to the number of analyzed SNPs. In statistics, this is referred to as the "large p, small n" paradigm, where p and n refer to the number of variables, and the sample size, respectively. This paradigm, also called "High Dimensional Statistics" by the more sophisticated, has seen massive interest in the last twenty years, and tens of thousands of various statistical methodologies have been proposed to address countless variations of high dimensional problems.

In this talk, we present a general overview of the recent developments from the field of high dimensional statistics. We start with the method that gave rise to it all: Tibshirani's Lasso from 1996, which has been referred to as the "linear regression of the 21st century". The rest of the first half of the talk is dedicated to more recent developments in the Lasso-verse; including concave penalties such as the SCAD and the MCP, and generalizations of the Lasso penalty such as the group Lasso and fused Lasso penalties.



In the second half of the talk, we focus on developments made in the last ten years. More specifically, we briefly cover (i) screening methods that reduce the dimensionality of the problem; (ii) provably efficient optimization algorithms for fitting Lasso (and its variations) to big datasets, including LARS, FISTA and other proximal based approaches such as ADMM; and (iii) post-selection inference methods that can be used to build confidence intervals for parameters fit with Lasso. We conclude with applications of the Lasso framework to numerous statistical problems ranging from high dimensional classification to graphical model estimation.

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

DEPARTMENT OF PHYSICS
COLUMBIAN COLLEGE OF ARTS AND SCIENCES