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 Lawrence Brown**
Statistics Department
Wharton School, University of Pennsylvania

Title: **Semi-supervised Inference with numerical dependent variable Means, Covariances and Linear Prediction; General Theory**

Time: **3:20 – 4:20pm, *Wednesday*, October 29, 2014**

Place: **552 Hill Center**

Abstract

Assume $\left(Y,X\right)\sim F$. Let be an iid sample. This is the “labeled” portion of the sample. Here, *Y* is a numerical variable, and the covariates in *X* comprise a p-dimensional vector. No special assumptions are made about the distribution, *F*, other than non-singularity of $E\left(XX'\right)$ and the existence of low-order moments. Suppose there is also available an independent sample of unlabeled X-values, $\left\{X_i : i=n+1,..,n+m\right\}$.

Interest lies in linear estimates of properties related to *Y* – about its mean, $E_F\left(Y\right)$, or about the covariances, $\operatorname{cov}\left(X,Y\right)$, or about predictions of a future value of Y, say Y^*. The estimates (or predictions) are linear in the usual sense of being linear functions of the X-vector, X^*; they need not be linear in the data $\left\{\left(Y_i,X_i\right): i=1,..,n\right\}$. The theory in this talk concerns asymptotics that hold as $m,\, n\to\infty$ with $\liminf\left(m/n\right)>0$ and *p* remains fixed, or at least grows only slowly with *n*. More precisely, key parts of the theory apply when $p^k << n$. Under reasonable assumptions the best value of *k* is *k* =2.

For all the problems considered we describe simply implementable estimates and predictions that are improved alternates to the naïve standard procedures when there is no unlabeled data. Our procedures have easily described asymptotic distributions, and asymptotically dominate the usual standard procedures whenever there is any useful information in X about Y; otherwise they are asymptotically equivalent. The current talk is mainly about mathematical theory that we believe will prove useful but there will also be some rudimentary simulation results.

This is joint work with several others at Wharton – especially Anru Zhang and Tony Cai as well as our POSI/Conspiracy research group (Berk, Buja, George, McCarthy, Pitkin, Zhang, Zhao).

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