UNC
GILLINGS SCHOOL OF
GLOBAL PUBLIC HEALTH

Department of Biostatistics

Bernard G. Greenberg Distinguished Lecture Series

May 28-29, 2014

Presented by

Jianqing Fan, PhD
Frederick L. Moore Professor in Finance and Chair
Department of Operations Research & Financial Engineering
Princeton University

All lectures will be held in the Blue Cross Blue Shield Memorial Auditorium 0001 Michael Hooker Research Center

---

Lecture I: May 28 - WEDNESDAY
Time: 10:00-11:00 AM
Title: Are assumptions in high-dimensional inference verifiable?

Abstract: The fundamental assumption in high-dimensional statistics is that the noise variable and covariates are uncorrelated. It has been taken for granted in all statistical modeling. Can this exogeneity assumption be validated by the data? If not, what are the consequences? We will argue that this exogeneity assumption is typically violated due to the large collections of covariates, a phenomenon we call incidental endogeneity prominently featured in Big Data. How to prove the existence of such incidental endogeneity in big data? What are the null distributions when the covariates are indeed uncorrelated with the noise variable, namely what are the distributions of spurious correlations? In what way they depend on the covariance of covariates themselves? How can we do right statistical inferences in presence of incidental endogeneity? This talk will provide some insights and preliminary results to these fundamentally important issues. Our solution has also relation to the random geometric graphs and significant implications on modern statistical learning.

Lecture II: May 28 - WEDNESDAY
Time: 2:00-3:00 PM
Title: FDR control under dependence – Joint work with Xu Han

Abstract: Large-scale multiple testing with highly correlated test statistics arises frequently in many scientific research. Incorporating correlation information in estimating false discovery proportion has attracted increasing attention in recent years. The covariance of the test statistics can assist in obtaining better false discovery control. The talk first introduces the principal factor approximation method of Fan, Han \& Gu (2012), which is a consistent estimate of False Discovery Proportion (FDP) under arbitrary dependence structure, when the covariance of test statistics is known. We then provide methodological modification and theoretical investigations for estimation of FDP with unknown covariance. When data are sampled from an approximate factor model, which encompasses most practical situations, we provide a consistent estimate of FDP via exploiting this specific dependence structure. Our results are demonstrated by simulation studies and some real data applications.

Lecture III: May 29 - THURSDAY
Time: 10:00-11:00 AM
Title: Robust Sparse Quadratic Discrimination – Joint work with Tracy Ke, Han Liu and Lucy Xia

Abstract: We propose a novel Rayleigh quotient based sparse quadratic dimension reduction method -- named QUADRO -- for analyzing high dimensional data. Unlike in the linear setting where Rayleigh quotient optimization coincides with classification, these two problems are very different under nonlinear settings. One major challenge of Rayleigh quotient optimization is that the variance of quadratic statistics involves all fourth cross-moments of predictors, which are infeasible to compute for high-dimensional applications and may accumulate too many stochastic errors. This issue is resolved by considering a family of elliptical models. Moreover, for heavy-tail distributions, robust estimates of mean vectors and covariance matrices are employed to guarantee uniform convergence in estimating nonpolynomially many parameters, even though the fourth moments are assumed. Computationally, we propose an efficient linearized augmented Lagrangian method to solve the constrained optimization problem. Theoretically, we provide explicit rates of convergence in terms of Rayleigh quotient under both Gaussian and general elliptical models. Thorough numerical results on both synthetic and real datasets are also provided to back up our theoretical results.