REVISED

For favour of posting

DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE
THE UNIVERSITY OF HONG KONG

Seminar

Professor K.W. NG

Patrick S C Poon Professor in Statistics and Actuarial Science
The University of Hong Kong

will give a talk

entitled

THE TRUTH OF SPURIOUS CORRELATIONS

Abstract

In a read paper of Royal Statistical Society which was commented by 41 discussants, Fan and Lv (2008) considered the problem of screening a set of random predictors $(X_1, X_2, \ldots, X_p)$ for response Y, based on realistically small sample size $n \ll p$. To illustrate that "When p is large, some of the intuition might not be accurate" in the Introduction, the authors reported the so-called "spurious correlations" or "noise accumulation" from simulations. The phenomena without theoretical explanation were again reported in *Journal of Machine Learning Research* by Fan, Samsworth and Wu (2009, p.2014), in an Invited Review Article in *Statistica Sinica* by Fan and Lv (2010, p.102), in *Annu. Rev. Econ.* by Fan, Lv and Qi (2011, p.239, p.294), and in *J. R. Statist. Soc., B* by Fan, Guo and Hao (2012, p.39). It is about the seemingly excessive values of the largest absolute correlation coefficient in their simulations, even if the $p+1$ variables are known to be independently standard normal. Such intuitively surprising phenomena have important consequences, particularly in biostatistics and genetics where the variables (features) in scope of study are usually in thousands or tens of thousands while the sample size in tens or hundreds. The topic and its discussions have been cited in nearly a thousand publications according to Google Scholar.

This seminar shows the theoretical genesis of such "spurious correlations" in a framework more general than normality, where the p independent samples, respectively $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$, are such that each of the vectors $\{(\boldsymbol{x}_j - \mu_j \mathbf{1}) : j = 1, \ldots, p\}$ has a spherically symmetric density function of dimension n from a possibly different family with a possibly different scaling parameter, while the independent vector $\boldsymbol{y}$ has an arbitrary density function of dimension n. Under these assumptions, the squared largest sample correlation coefficient is distributed as the maximum of p random variables which are independently and identically distributed as Beta$(1/2, (n-2)/2)$ distribution. And after controlling q of the X-variables where $q < n - 2$, the result for the remaining $p - q$ sample partial correlation coefficients is as in the simple correlation case but with sample size n reduced to $n - q$. The proof assuming normality can be incorporated into undergraduate curriculum without difficulty. The results can assist the sample-size design to reduce spurious correlations' confounding effects and facilitate the Benjamini and Hochberg procedures of protecting false discovery rate of multiple tests for zero-slope in all of the simple regressions of Y on individual X_j.

on

Friday, February 22, 2013

2:30 p.m. — 3:30 p.m.

at

Room 524, Meng Wah Complex
(behind the Chong Yuet Ming Amenities Centre)

Visitors Please Note that the University has limited parking space. If you are driving please call the Department at 2859 2466 for parking arrangement.

All interested are welcome