

THE CHINESE UNIVERSITY OF HONG KONG

Department of Statistics

will present a seminar entitled

Impact of dimensionality and correlation learning

by

**Professor Jianqing Fan
Princeton University**

on

**Wednesday, 17 December 2008
11:00am – 12:00pm**

in

**Lady Shaw Building C1
The Chinese University of Hong Kong**

Abstract:

Model selection and classification using high-dimensional features arise frequently in many contemporary statistical studies such as tumor classification using microarray or other high-throughput data. The impact of dimensionality on classifications is largely poorly understood. We first demonstrate that even for the independence classification rule, classification using all the features can be as bad as the random guessing due to noise accumulation in estimating population centroids in high-dimensional feature space. In fact, we demonstrate further that almost all linear discriminants can perform as bad as the random guessing. Thus, it is paramountly important to select a subset of important features for high-dimensional classification, resulting in Features Annealed Independence Rules (FAIR). The connections with the sure independent screening (SIS) and iterative SIS (ISIS) of Fan and Lv (2008) in model selection will be elucidated and extended. Further extension of the correlation learning results in independence learning for feature selection in general loss functions. The choice of the optimal number of features, or equivalently, the threshold value of the test statistics are proposed based on an upper bound of the classification error. Simulation studies and real data analysis support our theoretical results and demonstrate convincingly the advantage of our new classification procedure.

All are Welcome