# Interesting Genes, Informative Genes and Microarray Classification

*Hui Zou*
*University of Minnesota*

This talks focuses on classification with microarray gene expression data which is one of the leading examples behind the recent surge of interests in high dimensional data analysis. It is now known that feature selection is crucial and often necessary in high dimensional classification problems. In the current literature there are two main approaches to gene (feature) selection in microarray classification. The first approach is basically a two-stage procedure. First, one performs large-scale multiple hypothesis testing to find the “interesting genes". After reporting a subset of “interesting genes", one then builds a discriminative function only using these “interesting genes". The second approach integrates feature selection and discriminant analysis so that one can directly target on finding the “informative genes" that are responsible to classifying the binary response.

Consider feature selection under the linear discriminant analysis model. We carefully examine the fundamental difference between the “interesting genes" and “informative genes" and show that the first approach could lead to a fraud classification rule. On the other hand, the computational challenges in deriving sparse discriminant analysis rules force people to use the independent rules that basically ignore the correlations among features in the classification procedure. In this talk we argue that such compromise is not necessary and present a new sparse discriminant analysis method that has the same computation complexity as sparse regularized least squares. We further provide theoretical justifications for our method. Our analysis concerns the scenario where both n (sample size) and p (dimension) can go to infinity and p can grow faster than any polynomial order of n. We show that, under reasonable regularity conditions, the proposed sparse LDA classifier can simultaneously achieve two goals: (1) consistently identify the subset of informative predictors; (2) consistently estimate the Bayes classification direction. Numerical examples will be presented as well.