

STANFORD UNIVERSITY
DEPARTMENT OF STATISTICS
DEPARTMENTAL SEMINAR

4:15 p.m., Tuesday, May 27, 2008
Sequoia Hall Room 200
(Cookies at 3:45 in 1st Floor Lounge)

Andrew B. Nobel
Department of Statistics and Operations Research
University of North Carolina, Chapel Hill

Finding Significant Large-Average Submatrices in High Dimensional Data

Exploratory analysis of gene expression and other high dimensional data often begins with row and column clustering, which yields a partition of the data matrix into disjoint sample-variable blocks (submatrices). Of particular interest in practice are submatrices whose entries are large on average. In conjunction with clinical and functional annotation, large average submatrices are frequently the starting point for subsequent analyses, such as the identification of genetic pathways and new disease subtypes.

We describe a simple algorithm, belonging to the general category of biclustering methods, for identifying large average submatrices in high dimensional data. Like other biclustering methods, the algorithm improves on independent sample variable clustering in several respects: the submatrices it identifies can overlap and they need not cover the entire data matrix (features that better reflect underlying biology), and the inclusion of samples and variables in a submatrix does not depend on their expression values outside the submatrix. The algorithm seeks to maximize a simple measure of statistical significance, which also provides an objective basis for comparing and selecting among submatrices of different sizes and average intensities. We will discuss the applications of the algorithm to a recent gene-expression based cancer studies, and will compare its performance with several other biclustering methods. The talk should be accessible to statisticians, computer scientists and computational biologists.

Joint work with Andrey Shabalin, Vic Weigman, and Charles Perou.