Keynote Abstracts

"Mixtures at the Interface"
David Scott, Rice University

Mixture modeling provides an effective framework for complex, high-dimensional data. The potential of mixture models is still largely untapped. This talk considers a number of applications and results for gaussian mixtures.

"The Practice of Cluster Analysis"
Jon Kettenring, Drew University

Cluster analysis is one of the main methodologies for analyzing multivariate data. While not as neat and tidy as other methods, such as principal components analysis or discriminant analysis, its use is widespread and growing rapidly. The goal of the talk is to document this growth, characterize current usage, provide examples of recent applications, and highlight both good and risky practices. The presentation is based on evidence extracted from database searches and literature reviews covering the years 1995-2004. It will conclude with discussion of several areas where advances in research could have a large impact on the practice of clustering.

"Bayesian Perspectives on Combining Models"
Merlise Clyde, Duke University

Consideration of multiple models is routine in statistical practice. With computational advances over the past decade, there has been increased interest in methods for making inferences based on combining models. Examples include boosting, bagging, stacking, and Bayesian Model Averaging (BMA), which often lead to improved performance over methods based on selecting a single model. Bernardo and Smith have described two Bayesian frameworks for model selection known as the M-closed and M-open perspectives. The standard formulation of Bayesian Model Averaging arises as an optimal solution for combining models in the M-closed perspective where one believes that the 'true' model is included in the list of models under consideration.

In the M-open perspectives the 'true' model is outside the space of models to be combined, so that model averaging using posterior model probabilities is no longer applicable. Using a decision theoretic approach, we present optimal Bayesian solutions for combining models in both frameworks. We illustrate the methodology with an example of combining models representing two distinct classes, prospective classification trees and retrospective multivariate discriminant models applied to gene expression data in advanced stage serous ovarian cancers.

"Bayesian Causal Inference from Observational Data"
Siddhartha Chib, Washington University in St. Louis

In many inferential problems, the main objective of the analysis is to isolate the "causal effect" of a binary indicator variable (the treatment) on a response. The determination of this effect is a complex problem in practice because of observed and, importantly, unobserved confounders. Such confounders are the norm in observational settings. Interestingly, randomized clinical trials

are, in general, not impervious to this problem when, for example, compliance to the assigned treatment is less than perfect, or when participants in the trial are subject to attrition. Unfortunately, in each such case, it is not possible to point identify the treatment effect without auxiliary, non-testable assumptions. A large literature dealing with these questions has now emerged. The purpose of this talk is to describe a new model-based Bayesian framework for calculating causal effects that is based on the assumption of a continuous confounder. Unlike previous Bayesian approaches, the modeling does not require the unknowable joint distribution of the potential outcomes (the outcomes for each level of the treatment). Analysis requires a joint model of the observed outcome and the treatment intake, while inferences about the causal effect are obtained by comparing the marginal predictive distribution of the potential outcomes. Calculations require Markov chain Monte Carlo methods that are tuned to the specifics of this problem. We illustrate the ideas with data arising from an eligibility trial (a trial in which subjects randomized into the treatment arm do not necessarily take the treatment whereas those randomized into the control arm are prevented from doing otherwise). By way of contrast, we also develop an alternative Bayesian approach for eligibility trials that is based on the assumption of a discrete unobserved confounder. The latter approach is simpler, because it does not entail the modeling of the treatment intake mechanism, but inferences can be vulnerable to the identifying assumptions. Nonetheless, as we discuss, each approach can be compared via marginal likelihoods and Bayes factors.

"Predictive Learning via Rule Ensembles"
Jerome Friedman, Stanford University

General regression and classification models are constructed as linear combinations of simple rules derived from the data. Each rule consists of a conjunction of a small number of simple statements concerning the values of individual input variables. These rule ensembles are shown to produce predictive accuracy comparable to the best methods. However their principal advantage lies in interpretation. Because of its simple form, each rule is easy to understand, as is its influence on individual predictions, selected subsets of predictions, or globally over the entire space of joint input variable values. Similarly, the degree of relevance of the respective input variables can be assessed globally, locally in different regions of the input space, or at individual prediction points. Techniques are presented for automatically identifying those variables that are involved in interactions with other variables, the strength and degree of those interactions, as well as the identities of the other variables with which they interact. Graphical representations are used to visualize both main and interaction effects. (Joint work with Bogdan Popescu)

"Extracting Biological Meaning from High-Dimensional Datasets"
John Quackenbush, Dana-Farber Cancer Institute and Harvard School of Public Health

The revolution of genomics has come not from the "completed" genome sequences of human, mouse, rat, and other species. Nor has it come from the preliminary catalogues of genes that have been produced in these species. Rather, the genomic revolution has been in the creation of technologies – transcriptomics, proteomics, metabolomics – that allow us to rapidly assemble data on large numbers of samples that provide information on the state of tens of thousands of biological entities. Although the gene-by-gene hypothesis testing approach remains the standard for dissecting biological function, 'omic technologies have become a standard laboratory tool for generating new, testable hypotheses. The challenge is now no longer generating the data, but rather in analyzing and interpreting it. Although new statistical and data mining techniques are being developed, they continue to wrestle with the problem of having far fewer samples than

necessary to constrain the analysis. One way to deal with this problem is to use the existing body of biological data, including genotype, phenotype, the genome, its annotation and the vast body of biological literature. Through examples, we will demonstrate show how diverse datasets can be used in conjunction with computational tools to constrain ‘omics datasets and extract meaningful results that reveal new features of the underlying biology.