

# Developments and Challenges in Mixture Models, Bump Hunting and Measurement Error Models

June 2-4, Cleveland Ohio

## Overview

Bumps, components, clusters and atypical structures from real data often lead to scientific discoveries or reveal interesting phenomena of a population. They are important in astronomy, biology, data mining, bioinformatics and in applications to virtually all natural and social sciences. The wide interest in such structures has in the last decade led to significant developments in each of these areas: mixture models for component hunting; nonparametric methods for bump or mode hunting; methods for cluster and structure hunting; and Bayesian computational methods for model selection and latent variable mixture models. Additionally, data often come with measurement errors or incomplete information. These problems add additional challenges to component, bump and cluster hunting and lead to another area of active research. Image sharpening can be also considered as an inferential problem involving measurement error models.

**Invited Sessions:** Astronomy, Bayesian Methods, Bioinformatics, Bump Hunting, Classification and Clustering, Image Analysis/Incomplete Data, Measurement Errors, Mixture Models, Neyman Lectures: (a) Physical Sciences (b) Genetics.

---

## Sunday June 2

### [9:55 - 10:00 a.m.] Opening Remarks

Joe Sedransk, Case Western Reserve University

### [10:00 - 11:00 a.m.] Key Note Speaker

*Chair: J. S. Marron*

**Peter Hall, Australian National University**  
**Aspects of modality in density estimation**

We shall address several topics where the number of modes of a probability density has an interesting impact on properties of a density estimator. One will be density estimation under modality constraints. Another will be the effect of kernel type on the estimator. Some kernels are particularly prone to produce estimators where the number of modes is a non-monotone function of bandwidth. For example, in the cases of Epanechnikov and biweight kernels the probability of nonmonotonicity equals 1 for all sample sizes of 2 or more. A third topic will be the effect that modality has on estimator performance. This issue has intriguing aspects when one considers the relative 'efficiencies' of kernel and local log-polynomial esti-

matoms. There, excepting the standard second-order case, log-polynomial local-likelihood estimators can perform relatively poorly, in MISE terms, against conventional kernel estimators when the sampled distribution is multimodal.

---

### [11:00 - 12:30 p.m.] Neyman Lectures: Genetics

*Chair: J. Sunil Rao*

**Genetics in a post-genomics era: lessons from model systems (Geoffrey Duyk, Exelixis)**

The human genome project has reached its midpoint. Its initial goals were to complete genetic and physical maps of the human genome in anticipation of completing the primary nucleotide sequence. The project was initially aimed at providing the infrastructure necessary for the identification of the genetic basis of common disease. This effort spawned sister projects focused on model systems (mouse, rat, zebrafish, arabidopsis, Drosophila, C. elegans, yeast, multiple bacterial species etc.) as well as stimulating the growth of our technology base. An important consequence of this effort was the application of high throughput process technologies to discovery phase of research, specifically the introduction of automation and informatics into the biology work place. The project has also stimulated a paradigm shift in research as the gathering and presentation of information has become an end itself, resulting in the dissociation of data acquisition from classic hypothesis based research.

The goal of this talk will be to review the utility of available genetic systems for target discovery and target validation. Special emphasis will be placed on invertebrate model systems as they offer the opportunity for systematic genetic screening in the context of well established understanding of organismal biology, the availability of high quality genomic/genetic information and tools as well as advanced technology for germline modification. I will also discuss the translation of fundamental approaches, first pioneered in simpler genetic systems, into tools for genetic dissection of vertebrate models.

**Building hearts, getting fat and breaking bones: computation, genetics and systems biology (Joe Nadeau, Case Western Reserve University, School of Medicine)**

A key challenge in biology and medicine is understanding how components of complex biological systems act together to provide functionality at higher levels. Can we infer from assays of component traits how higher level systems function? Can we predict emergent properties?

Functional genomic studies are providing large data sets of many biological features; the problem is developing and testing computational methods for using these data in systems analysis. I will discuss our work on perturbation analysis, component trait assays and computational methods to study three biomedically important complex systems - heart, bones and metabolism. Computational analysis of cardiovascular traits shows that the function of the heart can be correctly predicted, that these computational methods correctly the functional relations among diverse physiological features in obesity, diabetes, dyslipidemia and hypertension, and finally that studies of bone biology as a complex system is more instructive than studies of single features alone. Together these kinds of studies begin to establish the empirical and analytical framework for connecting single genes and complex traits in health and disease in humans and model organisms.

---

**[12:30 - 2:30 p.m.]**

**Lunch in Pavillion Dining Room (Free)**

**Poster Session**

*Chairs: Nidhan Choudhuri and Steve Ganocy*

Roxana Alexandridis\*, Ohio State University  
 Swati Biswas\*, Ohio State University  
 Hsien Wern Chan\*, CWRU and U. of Western Australia  
 Richard J. Charnigo, Case Western Reserve University  
 Nidhan Choudhuri, Case Western Reserve University  
 Eloisa Diaz-Frances, C. de Investigacion en Matematicas  
 Ryan Thomas Elmore\*, Pennsylvania State University  
 George Z. Fan, Case Western Reserve University  
 Jeffrey D. Hart, Texas A&M University  
 Berta Ibanez\*, Public University of Navarra  
 Hemant Ishwaran, Cleveland Clinic Foundation  
 and Glen Takahara, Queen's University  
 Woncheol Jang\*, Carnegie Mellon University  
 Jiashun Jin, Stanford University  
 Murray A. Jorgensen, U. of Waikato and U. of Victoria  
 Mary Lesperance, University of Victoria  
 Walter Stewart Liggett, NIST  
 Miguel Nakamura, C. de Investigacion en Matematicas  
 Surajit Ray\*, Pennsylvania State University  
 David Todem\*, University of Wisconsin-Madison  
 Bin Wang, Case Western Reserve University  
 Kai Wang, University of Iowa  
 Liquin Wang, Univeristy of Manitoba  
 Yanzhong Wang\*, University of Glasgow

\*Student Award

---

**[2:30 - 4:00 p.m.] Astronomy**

*Chair: Jiayang Sun*

**Star streams: bump hunting in the Milky Way's halo (Heather Morrison, Case Western Reserve University)**

The outer regions of the Milky Way are remarkably uncharted, given the advances in knowledge of our own and other galaxies. There are two competing scenarios for the formation of these outer regions. The early classical work described a smooth, rapid collapse, which would result in a smooth spatial distribution and a unimodal velocity distribution. More recent work, backed up by our increasing understanding of galaxy formation from a cosmological perspective, suggests a more extended and chaotic formation which would produce a less smooth spatial distribution and substructure (modes or bumps) in velocity distributions. This relates to the important statistics areas of bump hunting, mixture modeling and inference, and cluster analysis.

I will discuss the "Spaghetti" survey of the Galactic halo, focusing on new discoveries of star streams in its outer regions, and what the streams will tell us about the history of the Galaxy and its structure.

*Acknowledgement: This is joint work with the "Spaghetti" project: Robbie Dohm-Palmer, Ken Freeman, Amina Helmi, Paul Harding, Mario Mateo, Ed Olszewski, Stephen Sheckman and Jiayang Sun.*

**Distant cluster hunting: a review of methods for finding clusters of galaxies (Megan Donahue, Space Telescope Science Institute)**

I will review the various methods and techniques that astronomers use to discover and identify clusters of galaxies, particularly distant clusters of galaxies. I will discuss briefly how the properties of clusters allow us to find them, including their optical and X-ray appearances and their affect on the cosmic microwave background. I will compare search efforts with optical and X-ray telescopes, in particular the ROSAT Optical X-ray Survey (ROXS). I will also review various techniques used in the optical to identify clusters in the Sloan Digital Sky Survey. I will mention some up-coming space and ground-based missions that will use these techniques to discover the most distant clusters in the universe.

**Mining a virtual observatory: mixture models, classification and the need for new statistical tools (Andrew Connolly, University of Pittsburgh)**

With recent technological advances in survey Astrophysics it has now become possible to map the distribution of galaxies within the local and distant Universe across a wide spectral range (from X-rays through to the radio). Combining these diverse data sets will provide the first panchromatic view of the Universe. The goal of such a Virtual Observatory is to enable astrophysicists to seam-

lessly interact with and analyze the data. In this talk I will discuss some of the questions that we wish to address with these new data sets and the challenges that they will provide. I will show that many of these questions relate directly to the statistics of density estimation and structure finding. As an example, the application of mixture models to studying the multidimensional properties of galaxies (whether these distributions are a 3D representation of galaxies across the sky or the distributions of the properties of galaxies such as their colors) will be used to show how we can extract new scientific results through the application of these statistical techniques. I will focus on the questions we can currently answer with mixture models, the need for new tools that can account for the heterogeneous errors within astronomical data sets and the computational challenges that these current and future techniques present when analysing massive, multi-dimensional data sets.

---

#### [4:00 - 4:15 p.m.] Break

---

#### [4:15 - 5:45 p.m.] Bump Hunting

Chair: J. A. Hartigan

#### A SiZer analysis of IP flow start times (J. S. Marron, University of North Carolina)

The SiZer technique is used to study the homogeneity of a point process of Internet traffic flow start times. It is seen that a homogenous Poisson process is an inappropriate model, because it does not yield observed statistically significant burstiness. Some Weibull waiting processes gives better, but still inadequate performance. A clustered Poisson process gives the best fit.

*Acknowledgement: This is joint work with Felix Hernandez-Campos and F. D. Smith. Partially supported by NSF Grant DMS-9971649. Most of the data analysis in this paper was done in the stimulating environment of the course OR 778, taught by J. S. Marron in the Fall of 2001 at Cornell University.*

#### Kernel oscillation analysis for the mixture complexity (Guenther Walther, Stanford University)

The problem under consideration is to determine the number of components in a location mixture, in the case where one does not want to make parametric assumptions on the component distribution. It turns out that whenever mode- or bump-hunting works, it is as well possible to do a much more sensitive analysis that can resolve mixtures that are unimodal (say). Further, a simple criterion can be derived for this analysis. Moreover, this more sensitive analysis comes without penalty in simpler situations: Even if one knew a priori that the mixture were multimodal and hence detectable by mode-hunting, the refined analysis will detect it essentially as well as any mode-hunting procedure possibly could. I will explain this heuristically and also give a precise theoretical result.

#### A new approach to finding and testing clusters (David Scott, Rice University)

Finding clusters in multivariate data is a fundamental problem underlying many investigations. The most successful algorithm is hierarchical clustering, which forms clumps by finding points that are close together. A more sophisticated algorithm is  $k$ -means, which iteratively reassigns points to the closest cluster center. Perhaps the most sophisticated algorithm is mixture modeling, in which the entire dataset is fit by a complicated combination of multivariate Normal distributions.

In practice, these methods can be fooled. The most difficult problem is determining the correct number of clusters. The second problem is that all of these methods work best when the clusters have the same shape (spherical, for example) and the same numerosity.

We examine some new research that aims at handling the difficult yet practical case: multivariate data, different cluster numerosity, different cluster shapes, and an unknown number of clusters. Our solution relies upon novel fitting technology and interactive graphical visualization.

*Acknowledgement: This work was supported in part by NSF grant DMS-9971797 and contract EIA-9983459.*

---

### Monday June 3

#### [8:45 - 10:15 a.m.] Bayesian Methods

Chair: Mary Lesperance

#### Bayesian hierarchical models in meta-analysis of diagnostic test accuracy studies (Vanja Dukic, University of Chicago)

Bayesian hierarchical models have recently received much attention in the area of meta-analysis, where they can provide a natural framework for dealing with multiple sources of data heterogeneity. Meta-analytic methods based on Bayesian hierarchical models can produce study-specific effect estimates while also yielding a combined estimate of the overall effect. Due to "borrowing of strength" across studies in the estimation of the study-specific effects, these effects are estimated with greater precision than if using data from each study separately. This borrowing of information occurs because all the study-specific estimates are based on the posterior distributions of the study-specific parameters, which are in fact mixtures over the posterior densities of other model parameters given the data. In addition, the values of study-specific estimates are shrunk toward the overall population effect.

Current meta-analytic methods for diagnostic test accuracy assessments studies are generally applicable to selection of studies reporting only estimates of sensitivity and specificity, or at most to studies whose results are reported using equal number of ordered categories. In this talk we propose a new Bayesian hierarchical model for the evaluation of an overall test accuracy. We derive a summary ROC curve for a collection of studies evaluating diagnostic tests, when test results are possibly reported in

an unequal number of non-nested ordered categories. We propose several ways to construct ROC credible intervals. We illustrate our approach with data from a recently published meta-analysis assessing accuracy of a single serum progesterone test in diagnosing pregnancy failure.

*Acknowledgement: This is joint work with Constantine Gatsonis, Center for Statistical Sciences, Brown University.*

### **Finding outliers in density space with Bayesian nonparametrics (Michael Escobar, University of Toronto)**

This talk will discuss methods of finding individuals with unusual distributions of values. For example, with modern diagnostic equipment, one can measure the individual cell sizes from a sample of ones blood. It is believed that different illness can be characterised by the shape of the distribution of sizes of the red blood cells.

In this talk, a hierarchical model for the distribution of the blood cells is developed. This model is somewhat equivalent to put a distribution on the family of kernel density estimates. Escobar and West (1995) showed how the kernel density estimator approximates a Bayesian method of estimating densities based on a mixture of Dirichlet processes (MDP). However, since the MDP method is a proper Bayesian model, one can use hierarchical priors and calculate posterior distributions of functionals of interest.

Using these techniques, we develop a highly flexible hierarchical model in the space of distributions. This model allows us to model samples of densities and to find outliers in the space of distributions. This talk will discuss the methods used to compute these models and to assess outliers. These techniques will be used to identify diseased subjects based on the distribution of the size of the subject's red blood cells.

*Acknowledgement: This is joint work with George Tomlinson and Christine McLaren.*

### **Bayesian mixtures: missing data models and their generalizations (Hemant Ishwaran, Cleveland Clinic Foundation)**

Bayesian missing data models incorporating discrete random measures are a powerful and flexible tool for inference in many types of semiparametric and nonparametric problems. I will discuss the role discreteness plays in clustering missing information, illustrating the idea by several examples. Specific computational details related to inference in finite mixture models will be given.

---

### **[10:15 - 10:30 a.m.] Break**

---



---

### **[10:30 a.m. - 12:00 a.m.] Neyman Lectures: Physical Sciences**

*Chair: Michael Woodroofe*

#### **Some statistical issues in particle physics experiments (Byron Roe, University of Michigan)**

Current topics in Particle Physics analysis will be discussed. There are a number of current and recent experiments setting limits on possible particle production (bumps) in the presence of background. For these experiments questions have arisen about setting limits for lower than expected counting rates. A number of other experiments involve the classification of events, and distinguishing between hypotheses. Some recent methods for doing this will also be discussed.

*Acknowledgement: I wish to acknowledge the support of the National Science Foundation for work on the mini-BooNE experiment.*

#### **Problems in the separation of superposed complex valued signals (Mark Stuff, Veridian Systems)**

In some remote sensing problems, it is common to receive the sum (superposition) of a finite set of complex valued sinusoidal signals. Typically, these signals have unknown amplitudes, frequencies, and phases, which one would like to estimate. When only a single realization of such a signal is available, methods from the theory of spectral estimation for time series and/or nonlinear regression methods are typically employed. For that single realization case, such methods probably include the best available. But, there are situations in which a sequence of realizations is collected, for which we know that the unknown amplitudes, frequencies, and phases change in a continuously differentiable way. This situation permits one to borrow power from the adjacent realizations, and to construct what seem to be reasonable estimates, even for those realizations in which the some of the signals have been essentially annihilated by destructive interference. We will present approaches to this estimation problem, empirical evidence in favor of the estimates, and raise questions concerning how the statistics of such estimates should be studied.

#### **Bumps and clumps in experimental particle physics (B. Paul Padley, Rice University)**

Experimental Particle Physics attempts to understand the most basic constituents of matter and the forces that act upon them. The research is carried out at National and Multinational Laboratories such as Fermilab and CERN by large international collaborations. Petabytes of data are produced in our experiments, which must be analyzed to look for signals of interesting physics. Often these signals are in the form of bumps or clusters in the data. This talk will describe some of the problems faced in these searches and some of the methods being applied.

---

**[12:00 - 1:15 p.m.] Lunch in Pavillion Dining Room (free)**

---

**[1:15 - 2:45 p.m.] Bioinformatics**

*Chair: Guenther Walther*

**Inferring informative genes from regression parameters (Jenny Bryan, University of British Columbia)**

In genomic datasets produced by various high-throughput techniques, such as DNA microarrays and deletion set studies, we often make inferences on hundreds or thousands of (possibly multidimensional) parameters. We will present a method of selecting informative genes from such datasets, with particular attention to the case in which each gene contributes data from a range of conditions. Examples of gene-specific response include expression level (DNA arrays) or the size of a yeast colony characterized by the deletion of the DNA for an individual gene (deletion set studies); examples of covariates include time and the concentration of a compound of interest in the growth medium. The common theme is that the relationship between the response and the covariates can be summarized in a regression parameter. Sets and clusters of informative genes can be defined in terms of these gene-specific regression parameters. The statistical performance of these subsets and clusters will then depend on the family-wise performance of thousands of regression parameter estimates. We will provide closed form results to control the probability of detecting 'extremely' false positives in certain settings. A bootstrap approach is employed to study other aspects of family-wide performance, such as crude and cluster-specific sensitivity and positive predictive value. The method will be motivated and illustrated with an analysis of yeast deletion set data, in which collaborators seek to uncover the mechanism of a compound known to inhibit angiogenesis in various human cancers.

*Acknowledgement: This is joint work with Mark van der Laan of UC Berkeley (statistical approach) and Kristin Baetz and Michel Roberge of the University of British Columbia (yeast deletion set studies).*

**Mammographic CAD using bootstrap ensembles (J. Sunil Rao, Case Western Reserve University)**

We present a method for recognizing suspicious masses in digital mammograms. Mammography is an effective tool for the early detection of possibly irregular breast masses. Computer-aided diagnosis (CAD) using digital mammograms can be a very effective complement to a visual diagnosis.

Our approach classifies each pixel indirectly into mass or non-mass via one-split classification stump derived from a broken line regression model using the pixel intensity profile with respect to a given origin. A bootstrap aggregation method (bagging) combining multiple versions of these simple circle approximations can be an effective

method for detecting more complex shapes.

Two data resampling schemes are implemented: re-sampling broken line residuals and resampling candidate pixel centers. Both approaches were tested on simulated "masses" of circular and elliptical shapes under different noise conditions. Results follow the theory, with bootstrapping residuals having a beneficial effect for generally circular masses, and bootstrapping centers providing large decreases in misclassification error for more general shapes.

Among the major advantages of the method are its simplicity, adaptability and lack of need for supervised training.

*Acknowledgement: This is joint work with Mireya Diaz of Case Western Reserve University.*

**Statistical problems of genetic mapping (David Siegmund, Stanford University)**

The goal of genetic mapping is to locate genes affecting particular traits (e.g., genes that affect human susceptibility to particular diseases or genes that affect productivity of agriculturally important species) by comparing the phenotypes and genotypes of related individuals. Changes in experimental technique that provide large numbers of informative genetic markers at known locations throughout a genome suggest new statistical problems concerned with the design and analysis of gene mapping experiments. I will discuss three such problems arising from genome scans to detect anonymous genes: (i) multiple comparisons arising from the simultaneous testing of many markers for linkage to the trait of interest; (ii) statistical power to map genes as a function of the true genetic model, especially when there is gene-gene or gene-environment interaction; and (iii) confidence bounds for estimation of genetic effects.

*Acknowledgement: This is joint work with H.-K. Tang.*

---

**[2:45 - 3:00 p.m.] Break**

---

**[3:00 - 4:30 p.m.] Classification and Clustering**

*Chair: David Scott*

**Robust clustering and outlier detection (David M. Rocke, University of California, Davis)**

We examine relationships between the problem of robust estimation of multivariate location and shape and the problem of maximum likelihood assignment of multivariate data to clusters and we offer a synthesis and generalization of computational methods reported in the literature. These connections are important because they can be exploited to support effective robust analysis of large data sets. Recognition of the connections between estimators for clusters and outliers immediately yields one important result that we demonstrate in this paper; namely, the ability to detect outliers can be improved a great deal using a combined perspective from outlier detection and cluster identification. One can achieve practical breakdown values that approach the theoretical limits by us-

ing algorithms for both problems. Computational results are reported that demonstrate the effectiveness of this approach.

*Acknowledgement: The research reported in this paper was supported by grants from the National Science Foundation (DMS 95-10511, DMS 96-26843, ACI 96-19020, and DMS 98-70172) and the National Institute of Environmental Health Sciences, National Institutes of Health (P42 ES04699). The author is grateful to Torsten Reinert for the use of his computer code and his help with computational experiments.*

### **Analysis of massive aviation inspection reports using text classification (Regina Y. Liu, Rutgers University)**

To ensure compliance with the Federal Aviation Regulation, the FAA regularly conducts surveillance inspections on aviation entities such as air carriers, airports, repair stations and flight training schools. Reports of findings from these inspections are collected in several FAA databases. Properly analyzing these data can help the FAA identify the areas of greater aviation risk, and oversee more effectively aviation operations and safety activities. These databases are massive, and they contain numerical data of various measurements as well as textual data in the form of fixed-format word reports or free-style write-ups. Many existing statistical methods have been applied to aviation safety analysis, but the applications so far have been mostly focused on the analysis of numerical data. The textual data of millions of reports detailing inspection activities and findings are another important source of aviation safety information, and they should be systematically extracted, analyzed, and then used for conducting a more effective aviation safety analysis. This talk discusses applications of various text classification approaches to develop a systematic analysis of the FAA inspection report database, and to show how the text classification procedure can be a critical element of the aviation safety decision support system. Further breakdowns of the misclassification errors and related findings from the report data also suggest ways to assess data quality and to gather information which are more pertinent to the intended goals for filing inspection reports.

*Acknowledgement: This is joint work with David Madigan and Susana Eyheramendy. We acknowledge the support from the National Science Foundation and the National Security Agency, and in particular the support from the Federal Aviation Administration through grant #00-G007.*

### **Listening Post (Mark Hansen, Bell Laboratories, Lucent Technologies)**

In this talk we describe a multi-media installation that attempted to characterize the "collective voice of the Internet." Over the last few years, the Web has become a massive communications channel: Conservatively, 35,000 sites advertise some form of chat or online forum, and recent IRC (Internet Relay Chat) statistics suggest a user pool of over half a million people. Through Listening Post,

we offered the public a view of this vast world of on-line activity, distilling content from tens of thousands of online exchanges in real time and revealing the patterns and rhythms of people communicating with each other. At the center of this uniquely designed space was a large array of 110 vacuum fluorescent displays (VFDs), each screen holding up to four lines of 20 characters. This grid displayed portions of online conversations, and cycled through several modalities in an attempt to convey scale (the number and diversity of Web discussions) and immediacy (that the displayed pieces of conversations are generated in realtime), while preserving the character (the individual exchanges mediated through various chat programs) and content (common topics of discussion and their frequency) of the underlying sources. In addition to the grid of VFDs, Listening Post also made use of a multi-layered soundscape projected through a series of speakers distributed throughout the installation space. Lucent's text-to-speech engine, the Articulator, was a large component of the audio, giving voice to many of the chat fragments displayed on the VFDs.

The technical challenges implied here are considerable; from scalable monitoring agents that continually recognized and culled new content on the Web, to statistical natural language processing and dynamic clustering schemes that allowed us to track topics and extract representative phrases. Toward this end, we constructed a network of computers that continually monitored 10,000 chat rooms and bulletin boards. To be efficient, the monitoring agents adapted their polling rates to the activity levels of the sources. As discussions took place, the content was categorized into topics based on (learned) key words, the participant's previous posts, and other recent or concurrent discussions. In upcoming incarnations of the installation, we are also considering covariates drawn from the Web site sponsoring the forum, as well as live news feeds. This categorization is by necessity dynamic, changing over the course of a day, and is highly influenced by current events. The installation and its supporting monitoring components are described fully in Hansen and Rubin (2001, 2002). Listening Post was part of the Brooklyn Academy of Music's Next Wave Festival and was open to the public December 6-20, 2001.

---

### **[4:30 - 4:45 p.m.] Group Photo**

---

### **[4:45 - 6:00 p.m.] Breakout sessions: Current Trends, Challenges, Discussions**

Astronomy/Physical Sciences<sup>1</sup>

Genetics/Bioinformatics<sup>2</sup>

Bump hunting/Classification and Clustering<sup>3</sup>

Measurement errors/Image analysis, Incomplete Data<sup>4</sup>

Mixture models/Bayesian methods<sup>5</sup>

<sup>1</sup>Heather Morrison, Michael Woodroffe

<sup>2</sup>J. Sunil Rao

<sup>3</sup>David Scott, Regina Y. Liu

<sup>4</sup>Jiahua Chen, Carey E. Priebe

<sup>5</sup>Bruce Lindsay

---

**[6:00 - 7:00 p.m.] Reception (cash bar)**

---

**[7:00 - 8:30 p.m.] Banquet Dinner, Severance Hall (free with registration)**

## Opening Remarks:

Lynne Singer, Provost CWRU

Samuel Savin, Dean CWRU

## Special Speaker: J. A. Hartigan, Yale University

The normal mixture model dates from Karl Pearson (1899) and standard likelihood methods are available for estimating parameters and determining the numbers of components. These methods are not appropriate for discovering modes, because there is no need for  $k$  different components to produce  $k$  different modes. I will present some methods for determining modes by fitting normal mixture models in which the  $k$  fitted components are constrained to produce  $k$  different modes.

---

---

**Tuesday June 4****[8:45 - 10:15 a.m.] Mixtures***Chair: Marianthi Markatou***Fitting sequence data with mixtures of self-reproducing kernels (Bruce Lindsay, Penn State University)**

Self-reproducing (SR) kernels make an attractive building block for fitting high-dimensional sequence data, as they provide highly flexible mixture models as well as simple and effective model selection mechanisms. By sequence data, we mean strings of correlated data points whose dimensionality is challenging, such as sequences of binary digits of length 64, for which the sample space is enormous. An example of a self-reproducing kernel is the normal, which satisfies the rule "a normal mixture of normals is also normal." We will introduce two other such kernels, the mutation kernel for binary sequences and the Poisson kernel for spherical data.

*Acknowledgement: This work is being carried out with students Shu-Chuan Chen, Surajit Ray, and Ke Yang.*

**Fitting mixtures of multivariate distributions via penalized dual method (Ramani S. Pilla, University of Illinois at Chicago and Case Western Reserve University)**

Mixture models have been used extensively for modeling heterogeneous data. There is an increasing need for efficient estimation of mixture distributions, especially following the explosion in their use in many applied fields. Focus of this research is to develop theory and an efficient method for the multivariate mixtures based on the fact that there is a dual problem connected with the mixture or primal problem (Titterton, 1975; Lesperance &

Kalbfleisch, 1992; Lindsay, 1995). The goal is to create a *penalized dual method* based on the dual concept and in turn use the penalized dual estimators to solve the primal problem. In particular, the constrained dual optimization problem will be turned into an unconstrained one via a *penalty function* that keeps the solution near the feasible region. It is shown that as the penalty is tightened, the solution is forced towards the optimal dual point. Proposed method is tested on several data sets including genomic data. Multivariate mixtures have applications in unsupervised pattern recognition, disease mapping, genetics, medical imaging and minefield detection.

*Acknowledgement: This is joint work with Francesco Bartolucci and Bruce Lindsay. This research is supported in part by the Probability and Statistics Program, Office of Naval Research Grant N00014-02-1-0316.*

**Comparing parametric and nonparametric mixed normal regression models (Murray Aitkin, Education Statistics Services Institute and University of Newcastle UK)**

This talk compares parametric and nonparametric mixed normal regression models for the Brownlee stack-loss data. These data, of 21 observations on an outcome and three explanatory variables, are well-known and frequently used to evaluate outlier detection methods in regression – there are four, or sometimes five, outliers which are clearly identifiable. The parametric approach uses  $t$ -regression (Little, Lange and Taylor JASA 1989) which identifies these observations and downweights them smoothly. The nonparametric approach uses a general overdispersion model (Aitkin, Statistics and Computing 1996) and estimates the mixing distribution nonparametrically. This approach gives a much higher likelihood and identifies a four-component mixture which can be attributed to the omission of a time-related factor in the data, leading to quite different conclusions from the analysis.

**Discussant: Marianthi Markatou, Columbia University**

---

**[10:15 - 10:30 a.m.] Break**

---

**[10:30 - 12:00 a.m.] Measurement Errors***Chair: Jiahua Chen***Measurement errors: connections, solutions and beyond (Andrey Feuerverger and Jiayang Sun, University of Toronto and Case Western Reserve University)**

Many interesting scientific problems are related to measurement error models. For example, imaging data analysis, astronomy problems and even some genetic data applications can be framed under a general framework involving measurement errors. In this talk, we'll make connections to various interesting problems, study solutions

from fundamental side - deconvolutions involving characteristic functions, to their modifications and then to non-deconvolution solutions (no characteristic function estimation). We'll show why the plain deconvolution estimator fail sometimes in practice. New results linking to restricted spline estimates will be discussed. Performances of estimates will be illustrated via simulations and analyses.

*Acknowledgement: Work is partially supported by NSF grants.*

### **Nonparametric regression with measurement error (David Ruppert, Cornell University)**

In nonparametric regression we estimate the conditional expectation of a response  $Y$  given a covariate  $X$  without assuming a parametric structure for this function. In regression with measurement error, we do not observe  $X$  itself but rather a proxy  $W$  which is commonly assumed to be  $X$  plus measurement error. Although both parametric measurement error models and nonparametric regression without measurement error have been intensively studied for many years, there has been surprisingly little work on nonparametric regression with measurement error. This lack of research may indicate the difficulty of the problem.

In recent work with colleagues, several effective methodologies have been developed for nonparametric regression with mismeasured  $X$ 's: SIMEX and local polynomial regression, SIMEX with penalized splines, a flexible structural spline method, and a fully Bayesian spline method. The Bayesian approach seems most promising and will be discussed in this talk. The Bayesian approach allows for inference that takes into account the measurement error and uncertainty about the smoothing parameter. This approach enables simultaneous global confidence bands on the regression function and its derivatives. These can be used for bump hunting. Instrumental variable estimators have also been developed.

*Acknowledgement: This is joint work with Raymond Carroll, Tor Tosteson, and Scott Berry. Work supported by NSF Grant DMS-9804058 and NCI Grant CA50597.*

### **Empirical likelihood inference in the presence of measurement error (Jiahua Chen, University of Waterloo)**

We consider the case where several different imperfect instruments and one perfect instrument are used independently to measure a characteristic of interest of a target population. We wish to combine the information from these independent samples to make statistical inference on parameters of interest, in particular the population mean and the population cumulative distribution function. We develop maximum empirical likelihood (MEL) estimators and study their asymptotic properties. We also present simulation results on the finite sample efficiency of MEL estimators.

*Acknowledgement: This is joint work with Bob Zhong and J. N. K. Rao, and the research was supported by NSERC.*

---

### **[12:00 - 1:00 p.m.] Lunch in Pavillion Dining Room (free)**

---

### **[1:00 - 2:30 p.m.] Image Analysis/Incomplete Data**

*Chair: Carey E. Priebe*

#### **A directed graph approach to locally adaptive sensing (David Marchette, Naval Surface Warfare Center)**

The Cluster Catch Digraph (CCD) covers the support of a random variable with balls, using random digraph techniques to select the minimal number of covering balls. This models the support as a mixture of compact components, where the number of components is chosen via the dominating set of a random digraph. Using the resultant cover, one can perform local dimensionality reduction, local classification, or in the case of an adaptive sensor, suggest a new sensor setting to collect information optimized for the task at hand. This is illustrated on a hyperspectral pixel classification problem.

*Acknowledgement: This is joint work with Carey Priebe and Jason DeVinney of The Johns Hopkins University, and Diego Socolinsky of Equinox Corporation.*

#### **Estimating a unimodal density (Michael Woodroffe, University of Michigan)**

The connection between biased sampling and the estimation of a non-increasing density will be reviewed and illustrated by examples. Then techniques for estimating unimodal and non-increasing densities will be described and a new one proposed. The new approach requires the estimator to be concave on an interval about the mode. The interval may be specified by a user or computed as part of estimator and may shrink to zero as the sample size increases. An algorithm for computing the new estimator will be presented, and properties of the new estimator will be described using a combination of simulation and asymptotic analysis.

#### **Consistent Estimation of Mixture Complexity (Carey E. Priebe, Johns Hopkins University)**

The consistent estimation of mixture complexity is of fundamental importance in many applications of finite mixture models. An enormous body of literature exists regarding the application, computational issues, and theoretical aspects of mixture models when the number of components is known, but estimating the unknown number of components remains an area of intense research effort. We present a semiparametric methodology yielding almost sure convergence of the estimated number of components to the true but unknown number of components. The scope of application is vast, as mixture models are routinely employed across the entire diverse application range of statistics, including nearly all of the social and experimental sciences. We consider application to image

analysis; specifically, we employ the methodology to determine the number of mixture components for subsequent borrowed strength spatial scan analysis.

*Acknowledgement: This presentation will include aspects of joint work with D. Chen, L.F. James, D.J. Marchette, T. Olson, J.S. Pang, R.S. Pilla, G.W. Rogers, J.L. Solka, P. Tao. This work is supported in part by Office of Naval Research Grants N00014-95-1-0777 and N00014-01-1-0011.*

---

---

**[2:30 - 5:00 p.m.] NSF Grantsmanship**

Marianthi Markatou, Columbia University