NSF Workshop '97

Bumps, Jumps, Clustering and Discrimination

May 11-14, 1997
Houston, Texas
Holiday Inn – Houston Medical Center

PROGRAM

Sponsored by:
THE NATIONAL SCIENCE FOUNDATION
WILLIAM MARSH RICE UNIVERSITY, SCHOOL OF ENGINEERING

Hosted by:

WILLIAM MARSH RICE UNIVERSITY
David W. Scott, Program Chair

Program Committee:
DAVID W. SCOTT, RICE UNIVERSITY
STEVE MARRON, UNIVERSITY OF NORTH CAROLINA
WENDY POSTON, NSWC
KATHRYN ROEDER, CARNEGIE MELLON UNIVERSITY
JOAN STANISWALIS, UNIVERSITY OF TEXAS, EL PASO

NSF Workshop 1997

Program Chair

David W. Scott
Department of Statistics, MS-138
Rice University
6100 Main Street
Houston, TX 77005-1892
(713) 527-6037
scottdw@rice.edu

Program Committee

David W. Scott, Rice University
Steve Marron, University of North Carolina
Wendy Poston, NSWC
Kathryn Roeder, Carnegie Mellon University
Joan Staniswalis, University of Texas, El Paso

Conference Arrangements

Diane Brown, David Scott, John Salch, David Andrews, Peter Thall, Luciano Batista, Michael Pearlman, Farrell Gerbode, the Rice University Statistics Department

Workshop Support

Support provided by:
The National Science Foundation
The George R. Brown School of Engineering, Rice University

Other Support

Other support provided by:
SGI, Inc. (Computers)
Southwestern Bell (ISDN Line)
Farrell Gerbode (ISDN Modem)

General Information

Conference Facilities
The Holiday Inn – Houston Medical Center Hotel conference facilities have been recently remodeled and are located on the first and second floors. The registration desk is located off the lobby on the second floor.

Sessions
All technical sessions will be held on the second floor. A map of the meeting space may be found on the inside cover of this program guide.

Snacks
Breaks will be held in the area adjacent to the registration desk outside Houston II. There will be a mid-morning and mid-afternoon break for thirty minutes each. Please check the program guide for exact times each day.

Banquet
Monday evening's banquet will be held at the Faculty Club at the Rice University Cohen House. An outstanding menu of Beef Tenderloin and Achiote Marinated Chicken and wine has been selected for your enjoyment. If you wish to have a vegetarian meal, please indicate your preference at the registration desk as early as possible. After dinner, there will be a twenty minute banquet address by Emanuel F. Parzen, Professor of Statistics at Texas A&M University. Afterwards, you are invited to remain and continue your conversation.

Dining
Other meals will be on your own. Houston has an extensive selection of high quality and ethnic dining experiences within two miles of the hotel. The hotel shuttle bus can take you to the Rice Village or the Galleria. An annotated list (courtesy of Peter Thall) is available at the registration counter.

Internet Access
The Dallas III meeting room will be equipped with a number of computer workstations (courtesy of **SGI, Inc and Rice University**) connected to an ISDN line (courtesy of **Southwestern Bell Company**). This arrangement should allow adequate access to electronic mail via the internet. The 128 Kb capacity means that users should not expect to run anything more than telnet without affecting other users. The room will be locked after hours. Check with registration for details.

Emergency Access
If it is necessary for someone to reach you while you are attending sessions, ask at the Registration Area for our cellular phone number. We request that you limit the use of this number to emergencies, as the phone charges can be expensive. Non-emergency messages can be left at the Hotel (713-797-1110). If it is necessary for you to receive a fax, the hotel maintains a guest fax number (713-797-1034).

Keynote: John Hartigan
The keynote address will be given by John Hartigan, Eugene Higgins Professor of Statistics at Yale University. Professor Hartigan received his Ph.D. in 1962 from Princeton. His research interests include foundations of probability and statistics, Bayes' theorem, classification, statistical computing and graphical methods. Professor Hartigan is a Fellow of the IMS, ASA, and ISI.

Some features of this program

To make finding sessions convenient, we have marked sessions with filled circles (●●). Also note that in the Program Schedule, only the first author's affiliation is given. The complete affiliation information is given in the Abstracts section.

Sunday

Sunday, 6:00 p.m. - 8:30 p.m.

Registration Gallery

Sunday, 7:30 p.m. - 10:00 p.m.

Informal Mixer El Paso
Light Refreshments and Cash Bar.

Monday

Monday, 8:00 a.m. - 9:00 a.m.

Registration Gallery

Monday, 8:45 a.m. - 9:00 a.m.

Opening and Welcome Houston II
Michael Carroll, Dean of Engineering, Rice University
David W. Scott, Rice University

Monday, 9:00 a.m. - 10:00 a.m.

•• **Keynote Address** Houston II
Chaired by David W. Scott, Rice University
9:00 *Looking for Modes — An Overview*, John Hartigan, Yale University [1]

Monday, 10:00 a.m. - 10:30 p.m.

BREAK

Monday, 10:30 a.m. - Noon

•• **Mode and Bump Visualization and Estimation** Houston II
Chaired by Michael Minnotte, Utah State University

10:30 *Visualizing Structure with Trees and Forests*, David J. Marchette, Naval Surface Warfare Center

11:00 *Interactive Mode Testing*, Enno Mammen, J. S. Marron, and Frederic Udina, University of Heidelberg

11:30 *Least-Squares Mixture Decomposition*, Donggeon Kim and George R. Terrell(*), Virginia Polytechnic Institute

Monday, Noon - 2:00 p.m.

LUNCH

Monday, 2:00 p.m. - 3:30

•• **Jumps: Overview and Identification** Houston II
Chaired by Nancy Heckman, University of British Columbia

[1] Only the affiliation of the first author is given. Complete affiliation information is given in the abstracts.

2:00 *Emergence of Non-Smoothness*, Hans-Georg Müller, University of California, Davis

3:00 *Using Local Polynomials to Identify Jumps, Troughs, and Shoulders: An Application to Cold Tolerance Features in Plants*, William R. Schucany and Patrick D. Gerard, Southern Methodist University

Monday, 3:00 p.m. - 3:30 p.m.

BREAK

Monday, 4:00 p.m. - 5:05 p.m.

Structure and Kernels Houston II ••

Chaired by Frederic Udina, Universitat Pompeu Fabra

4:00 *Structure Exploration in Smoothing and Scale Space Theory from Computer Vision*, P. Chaudhuri and J. S. Marron, University of North Carolina at Chapel Hill

4:45 *On the Practical Advantage of Higher Order Kernels*, Shean-Tsong Chiu, Colorado State University

Monday, 5:05 p.m. - 5:45 p.m.

Small Group Discussions Houston II/Dallas I/Dallas II

Monday, 6:30 p.m. - 9:30 p.m.

Banquet Rice University Faculty Club - Cohen House

Music: Bluegrass Band

Banquet Address Cohen House

8:15 *Teaching Smoothing in Introductory Statistics*, E. Parzen, Texas A&M University

Tuesday

Tuesday, 9:00 a.m. - 10:10 a.m.

Local Level Sets and Modes Houston II ••

Chaired by Stephan R. Sain, Southern Methodist University

9:00 *Local Testing of Mode Existence*, Michael C. Minnotte, Utah State University

9:30 *Using Level Set Estimators in Statistical Analysis*, Wolfgang Polonik, Universität Heidelberg

Tuesday, 10:10 a.m. - 10:40 a.m.

BREAK

Tuesday, 10:40 a.m. - Noon

Testing and Applications Houston II ••

Chaired by Bill Schucany, Southern Methodist University

10:40 *Testing Monotonicity of Regression*, Adrian Bowman, M.C. Jones, and I. Gijbels, University of Glasgow

11:10 *Medical Applications of Cluster Analysis*, Alfred Lin(*), Catherine A. Sugar(*), and Richard A. Olshen, Stanford University

11:40 *Testing Modes in the Population Distribution of Beliefs from Experimental Games*, Ernan Haruvy, Department of Economics, University of Texas

Tuesday, Noon - 2:00 p.m.

LUNCH

Tuesday, 2:00 p.m. - 3:00

•• **Local Methods** Houston II

Chaired by Adrian Bowman, University of Glasgow

2:00 *Local Maximum Likelihood Estimation and Inference*, Jianqing Fan, Mark Farmen(*), and Irène Gijbels, The University of North Carolina at Chapel Hill, Chapel Hill, NC

2:30 *Exploring the Potential of Locally Adaptive Density Estimation*, Stephan R. Sain, Southern Methodist University

Tuesday, 3:00 p.m. - 3:30 p.m.

BREAK

Tuesday, 3:30 p.m. - 5:00 p.m.

○○ **Mixture Models and Clustering** Houston II

Chaired by George R. Terrell, Virginia Polytechnic Institute

3:30 *Bayesian Restoration Maximization Algorithm for Estimation in Mixture Models*, Olga I. Cordero Brana and Gilles Celeux, American University

4:00 *Identification of Outliers and Clusters using Weighted Likelihood Methodology*, Marianthi Markatou, Columbia University

4:30 *Mixtures in Supervised Classification*, Brian D. Ripley, Oxford University

Tuesday, 5:00 p.m. - 5:45 p.m.

Small Group Discussions Houston II/Dallas I/Dallas II

Wednesday

Wednesday, 9:00 a.m. - 10:00 a.m.

•• **Modelets and Wavelets** Houston II

Chaired by William Szewczyk, Department of Defense

9:00 *Modelets*, John Hartigan, Yale University

9:30 *Wavelets in Bayesian Change-point Analysis*, R. Todd Ogden, University of South Carolina

Wednesday, 10:00 a.m. - 10:30 a.m.

BREAK

Wednesday, 10:30 a.m. - Noon

•• **Unimodal Densities and Bump Hunting** Houston II

Chaired by Wendy Poston, Naval Surface Warfare Center

10:30 *An Empirical Approach to Estimate Unimodal Density Functions*, Ming-Yen Cheng and Peter Hall, Australian National University, Australia

11:00 *Bump Hunting Using Model-Based Penalized Least Squares*, Nancy E. Heckman, University of British Columbia

11:30 *Bump Hunting Problems: A New Test*, Jiayang Sun, Case Western Reserve University

Wednesday, Noon - 2:00 p.m.

LUNCH

Wednesday, 2:00 p.m. - 3:20

Wavelets, Mixtures, and Mode Trees Houston II ••

Chaired by Mark Farmen, Ohio State University

2:00 *Rough Function Estimation via Wavelets*, Yazhen Wang, University of Missouri-Columbia

2:30 *Adaptive Mixture Techniques for Model Complexity Estimation*, Jeffrey L. Solka and George Rogers, Naval Surface Warfare Center

3:00 *Bumps Along the Road Towards Multivariate Mode Trees*, David W. Scott and William F. Szewczyk, Rice University

Wednesday, 3:20 p.m. - 4:00 p.m.

Wrap-Up Houston II

Monday

Monday, 9:00 a.m. - 10:00 a.m.

Keynote Address: Looking for Modes — An Overview

Looking for Modes — An Overview

John Hartigan
Yale University

Modes and clusters have a variety of uses in data analysis and inference. One is the discovery of new phenomena. Another is in making predictions. We need to discover modes that are present in subsets of variables of the data, so that mode searching requires selecting sets of variables as well as localities in space of surprisingly high densities. Some general techniques including density estimates, mixture models, mini minimum spanning trees, excess mass methods are described for identifying modes. Examples are given from astronomy and parliamentary voting.

Monday, 10:30 a.m. - Noon

Invited Session: Mode and Bump Visualization and Estimation

Visualizing Structure with Trees and Forests

David J. Marchette
Naval Surface Warfare Center

Mode trees, which have been used to investigate modal structure, are probably familiar to most researchers in the area. Modifications to the basic mode tree have been proposed, and will be briefly described. A new idea, the mode forest, uses resampling to reduce the clutter in the mode tree and produces a more appealing picture. The ”forest” idea is applied to the problem of determining the number of components in a mixture.

Interactive Mode Testing

Enno Mammen
University of Heidelberg
J. S. Marron
University of North Carolina. Chapel Hill
Frederic Udina
Universitat Pompeu Fabra

In a kernel density estimation setting, we develop and analyze some tests to decide if a user selected mode can be really from the underlying density. A multivariate extension is also considered. The test are implemented in a fully interactive environment based on Xlisp-Stat, where the user can select the mode or modes to be tested using point and click devices. We will show it in action and discuss some interaction issues.

Least-Squares Mixture Decomposition

Donggeon Kim and George R. Terrell
Virginia Polytechnic Institute

We often find ourselves with large random samples which we would like to interpret as coming from mixtures of homogeneous subpopulations. Traditionally, this is done by maximum likelihood, or by Bayesian methods. These have two disadvantages: they are computationally challenging (often the algorithms do not converge), and they do not tell you (without strong prior information) how many subpopulations are present. Least- squares mixture decomposition (a relative of kernel density estimation) solves the first problem and makes promising progress on the second.

Monday, 2:00 p.m. - 3:30

Invited Session: Jumps: Overview and Identification

Emergence of Non-Smoothness

Hans-Georg Müller
University of California, Davis

A basic assumption in nonparametric curve and surface estimation is smoothness of the function to be estimated. This smoothness assumption is violated at known endpoints of the support of the function, leading to the well known boundary effects when estimating compactly supported functions. The problem gets more interesting when the location of discontinuities is unknown. The issue of non-smoothness in otherwise smooth curves goes back at least to the early years of this century. A review is given on methods, results and applications when estimating smooth functions with isolated jump discontinuities. Of particular interest are multidimensional extensions (break curves, edges), the case of an unknown number of discontinuities, and the question whether discontinuities exist to begin with (smooth versus discontinuous curve fitting).

Using Local Polynomials to Identify Jumps, Troughs, and Shoulders: An Application to Cold Tolerance Features in Plants

William R. Schucany
Southern Methodist University
Patrick D. Gerard
Mississippi State University

We examine some specifics of an implementation of local polynomial fitting to identify a jump, local minima, and certain inflection points. The physical problem that motivates this involves controlled experiments in freezing plant tissues. For the single changepoint we search all differences between one-sided local linear fits. For each of these we use an estimated local bandwidth for each candidate extreme boundary point. Subsequent searches beyond the jump examine both first and second derivative estimators with plug-in estimates of the global bandwidths that minimize the respective asymptotic mean integrated squared error. Both of these require simultaneous inference for the significance of the relevant axis crossings. We report on some limited experience with estimated standard errors. For this particular application the plant scientists prefer a sensitive automated procedure. In other words, for these less dramatic freezing episodes it is better to identify extra minima and shoulders, than to declare them insignificant before they are inspected visually.

Monday, 4:00 p.m. - 5:05 p.m.

Invited Session: Structure and Kernels

Structure Exploration in Smoothing and Scale Space Theory from Computer Vision

P. Chaudhuri and J. S. Marron
University of North Carolina at Chapel Hill

A useful statistical viewpoint for mode testing, and other types of feature identification, has been to study families of smooths indexed by the smoothing parameter. This same family has been well developed as an important model for "computer vision" by computer scientists, who have carried the mathematics very far. An overview of these ideas is given, with an emphasis on the relation to statistics. An application of these ideas, SiZer, is proposed, which studies SIgnificant ZERo crossings of the derivative smooths, from a scale space viewpoint. This gives visual insight into features that are statistically significant at various levels of resolution.

On the Practical Advantage of Higher Order Kernels

Shean-Tsong Chiu
Colorado State University

It is well known that the density estimate based on higher order kernels have a faster convergence rate. Since the difference in the rate is very small, higher order kernels may not have a practical advantage over the second order kernel. In particular, when the bandwidth has to selected based on the data. In the paper, the problem of bandwidth selection is studied. Asymptotic properties of the cross-validation and stablized bandwidth selection for higher order kernel are presented. The simulate results indicate that the higher order kernel has a practical advantage only for very smooth densities. The simulation also shows the importance of having a good bandwidth selector when using higher order kernels.

Banquet Address: Teaching Smoothing in Introductory Statistics

Teaching Smoothing in Introductory Statistics

E. Parzen
Texas A&M University

A 20 minute banquet talk aimed at stimulating discussion about the following topics. (0) What is role of jokes in teaching? How many smoothers does it take to change a light bulb? One or a crew? When bulbs fail or when bulbs are scheduled for replacement? (1) What do smoothers talk about at all those (exclusive European and North American) workshops being held in 1997? How to measure progress? Do smoothers (like other workshop addicted researchers) ask why does not everyone think like us? (2) Why and how should smoothing be taught in an introductory statistics course? Because statistical methods are a combination of statistical inference, probability plug in, numerical analysis/ approximation theory, and computing algorithms? (3) Why should statistical methods mining be taught in an introductory statistics course? (4) How to explain how and why kernel smoothing works? How central are ideas of bias- variance trade-off and mean square error rate? (5) What data sets provide good examples to illustrate practical issues in smoothing? (6) Should we have a Smoothing Teaching Hot line where teachers of introductory statistics can ask questions about how to teach smoothing in introductory statistics courses? Can smoothing become a statistical consumer product?

Tuesday

Tuesday, 9:00 a.m. - 10:10 a.m.

Invited Session: Local Level Sets and Modes

Local Testing of Mode Existence

Michael C. Minnotte
Utah State University

A test is proposed for the weight of evidence of individual observed modes. The test statistic used is a measure of the size of the mode, the absolute integrated difference between the estimated density and the same density with the mode in question excised at the level of the higher of its two surrounding antimodes. Samples are simulated from a conservative member of the composite null hypothesis to estimate p-values within a Monte Carlo setting. Such a test can be used with the graphical "mode tree" of Minnotte and Scott (1993) to examine, in a locally adaptive fashion, not only the reality of individual modes, but also the overall number of modes of the density.

Using Level Set Estimators in Statistical Analysis

Wolfgang Polonik
Universität Heidelberg

Level set estimation is intrinsically connected to cluster analysis, since level sets (of densities) are sets which carry high mass concentration. This has long been recognized in the literature. For instance, Hartigan (1975) used the notion λ-cluster for connected components of level sets.

There are different approaches to (direct) level set estimation. To mention some keywords: minimum volume sets, modal sets, excess mass. These estimators can be used directly or indirectly in statistical analysis and inference. After a brief review of some recent work in this area we will concentrate on the minimum volume approach. A new method will be presented to extend classical goodness-of-fit techniques to arbitrary (finite) dimensions. More precisely, a generalized Kolmogorov-Smirnov-test and corresponding generalized P-P-plots, named C-C-plots, will be presented. They are shown to have similar properties as classical techniques: for simple hypothesis the test statistics are (asymptotically) distribution-free with same well-known asymptotic distribution, and C-C-plots show straight lines, if and only if the hypothesis is true. This holds for any finite dimension. The underlying key observation is that minimum volume sets naturally lead to a certain genralized quantile transformation, as will be discussed.

Tuesday, 10:40 a.m. - Noon

Invited Session: Testing and Applications

Testing Monotonicity of Regression

Adrian Bowman
University of Glasgow
M.C. Jones
The Open University
I. Gijbels
Catholic University of Louvain

This paper provides a test of monotonicity of a regression function. The test is based on the size of a 'critical' bandwidth, the amount of smoothing necessary to force a nonparametric regression estimate to be monotone. It is analogous to Silverman's test of multimodality in density estimation. Bootstrapping is used to provide a null distribution for the test statistic. The methodology is particularly simple in regression models in which the variance is a specified function of the mean, but we also discuss in detail the homoscedastic case with unknown variance. Simulation evidence indicates the usefulness of the method. Two examples are given.

Medical Applications of Cluster Analysis

Alfred Lin, Catherine A. Sugar, and Richard A. Olshen
Stanford University

At present, an area of great interest is the optimal allocation of available resources for maintaining or improving the public's health. To facilitate these decisions, some have devoted energy to quantifying the *quality of life* of patients in different *health states*. A collection of health states for a specified population is a partition of the space of features that enable description of the health of an individual member of the population. Policy can then be based on an attempt to improve quality of life by moving people from one health state to a better one. We will discuss how we have applied k-means clustering to several data sets that inform on quality of life issues and the insights we have gleaned about cluster analysis in the process.

The current state of the art by which health state descriptions are devised is to choose some set of features and divide each of them into equally spaced levels. This leads to a factorial design, with one health state for each possible combination of levels, the centerpoint of each square grid representing a prototypical member of the state. There are problems with this approach, the most serious of which are that it produces far too many health states, and that there is no *a priori* reason why the data would cluster in such a symmetric fashion. To circumvent these problems, we propose a data-driven approach; use k-means cluster analysis (or some other form of clustering) to divide the data into a parsimonious set of (possibly) asymmetrically spaced clusters. This approach, in our studies, has led to models that are far more efficient than factorial designs (in terms of having smaller mean squared distance to cluster center for any number of clusters), and that at the same time give clinically meaningful characterizations of the health states.

Many statistical issues arise when implementing this methodology. On what dimensions should the clustering be based, and how should these dimensions be scaled? When a standard routine such as the Lloyd algorithm is being used, how should one initialize and run it so as to have high probability of obtaining the optimal clustering? How does one decide how many clusters are "right" for the data? How should health state descriptions be defined from the output of the clustering algorithm? How can one determine whether the clusters represent real health states or are an artifact of the data–i.e., how stable are the clusters? What sort of imputation should be done in the case of missing data? How can cluster analysis be used to check for longitudinal trends? We will explain briefly how we dealt with each of these issues and the discoveries we made in the process about cluster analysis in general.

Since the meeting is devoted to clustering, very broadly construed, it may be of interest to the group if we spend a very few minutes at the end summarizing different aspects of clustering in which a group of us are involved. This includes adapting k-means in a tree-structured fashion, not only for long standing applications to lossy image compression, but also to molecular genetics. As well, it includes an approach to ascertaining from data "what mediates" a relationship found in data, for example the relationship between bone mineral density and stroke in elderly women. Website: http://www-stat.stanford.edu/projects/Cluster_Analysis*

Testing Modes in the Population Distribution of Beliefs from Experimental Games

Ernan Haruvy
Department of Economics, University of Texas

In recent years, many attempts have been made in game theory literature to characterize the diversity of behavior in a population in a parsimonious way, particularly the way hypotheses are formed. Recent experimental work provides data sets with ideal properties for characterizing this diversity. The data consist of vectors of hypotheses in the form of probabilities on the percentage of other players choosing each available strategy. Using nonparametric density estimates we look for modes in the data corresponding to various behavioral types and test for the existence of these modes in the underlying population distribution using a version of the smoothed bootstrap test. However, linear constraints unique to the probability metric that are not present in other types of data and the dependence nature of this type of behavioral data make standard approaches to density estimation of bounded data unsuitable. A possible solution to this problem is outlined, empirical results are presented, and potential problems not yet resolved are discussed.

Tuesday, 2:00 p.m. - 3:00

Invited Session: Local Methods

Local Maximum Likelihood Estimation and Inference

Jianqing Fan
The University of North Carolina at Chapel Hill, Chapel Hill, NC
Mark Farmen
Ohio State University, Columbus, OH
Irène Gijbels
Université Catholique de Louvain, Louvain-la-Neuve, Belgium

Local maximum likelihood estimation is a nonparametric counterpart of the widely-used parametric maximum likelihood techniques. It extends the scope of the parametric maximum likelihood methods to a much wider class of parametric spaces. Associated with this nonparametric estimation scheme is the issue of bandwidth selection and bias and variance assessment. A unified approach to selecting a bandwidth and constructing confidence intervals for local maximum likelihood estimation will be presented. The confidence intervals for the curve and its derivative can prove useful for assessing the presents of modes. The approach has been used for least-squares nonparametric regression. Computational issues and an encouraging simulation study in the case of nonparametric logistic regression will be presented. Our experience in these two settings show that the general idea outlined here is powerful and encouraging.

Exploring the Potential of Locally Adaptive Density Estimation

Stephan R. Sain
Southern Methodist University

Locally adaptive density estimation has great potential for accurately discerning structure in data from complex distributions. In this talk, a fully automatic, data-driven locally adaptive procedure based on binning (Sain and Scott, 1996) will be presented. Simulation studies will also be shown that compare this procedure with fixed bandwidth methods (plug-in and least-squares cross-validation) in an attempt to better understand the behavior and exploring the potential of each procedure with respect to capturing the true structure of the underlying density.

Tuesday, 3:30 p.m. - 5:00 p.m.

Contributed Session: Mixture Models and Clustering

Bayesian Restoration Maximization Algorithm for Estimation in Mixture Models

Olga I. Cordero Brana
American University
Gilles Celeux
I.N.R.I.A.

We consider estimation of mixture models using BRM algorithm. This is a variation of the Weighted Likelihood Bootstrap algorithm of Newton and Raftery (1994) to approximate posterior distributions

Identification of Outliers and Clusters using Weighted Likelihood Methodology

Marianthi Markatou
Columbia University

In this talk we will discuss the application of the weighted likelihood estimating equations methodology to identify clusters and outliers. First we will introduce the method of weighted likelihood in the case of a random sample from a probability mass or density function and will discuss a bootstrap root search algorithm which enables us to identify structure in the data. We will then show how to use the method to identify outliers and clusters.

Mixtures in Supervised Classification

Brian D. Ripley
Oxford University

Mixtures, often of Gaussian densities, have frequently been proposed as flexible models in discriminant analysis. The talk will view some of the ways that mixtures can be fitted, some approximations to full Bayesian learning, and some new ideas on specifying hierarchical priors.

Wednesday

Wednesday, 9:00 a.m. - 10:00 a.m.

Invited Session: Modelets and Wavelets

Modelets

John Hartigan
Yale University

A modelet is an empirically defined set that consists of a region of high density surrounded by a region of low density. In one dimension, the modelets consist of intervals of locally maximal density in the stretched string fit that passes between the graph of the empirical distribution function, and the same graph shifted vertically by D, as the gap D varies. The modelets form a hierarchical tree that may be thought of as an estimate of the population mode tree. Tests for the number of modes in one dimension are constructed by considering for each emerging modelet whether or the c corresponding gap D is sufficiently large, referred to the shoulder in the distribution from which the mode emerges. Similar techniques in higher dimensions define modelets on the minimum spanning tree.

Wavelets in Bayesian Change-point Analysis

R. Todd Ogden
University of South Carolina

Wavelet methods have become popular tools in statistical function estimation because of their ability to adapt well to discontinuities and other irregular features of the underlying function. This ability makes wavelets an attractive alternative for use in various versions of the change-point problem. A Bayesian analysis based on the empirical wavelet coefficients of a set of data is developed for the standard change-point problem, in which the prior information on the location of the change-point is expressed in terms of the wavelet domain. This general methodology, developed first for the standard single change-point problem, is easily extended to the problem of detecting jumps or other unusual features in functions that are otherwise smooth by tailoring the prior information to the appropriate type of feature and by considering only higher-level coefficients in the computation of the posterior. This procedure is illustrated by simulated data examples.

Wednesday, 10:30 a.m. - Noon

Invited Session: Unimodal Densities and Bump Hunting

An Empirical Approach to Estimate Unimodal Density Functions

Ming-Yen Cheng and Peter Hall
Australian National University, Australia

Suppose that we are given a random sample drawn from a population following a unimodal density function. For simplicity, assume that the density has support $[0, 1]$. To estimate the unimodal density function, take a known unimodal density (which we call a template) supported on $[0, 1]$ and match (by a monotone and increasing transformation) its distribution function with the empirical distribution function based on the sample. Then our estimator of the unknown density is taken to be the template combined with the transformation. This gives a unimodal estimator since the template is unimodal and the transformation is monotone and increasing. The desired transformation basically can be achieved using the curve alignment method of Cheng, Gasser and Hall (1997). However, some sort of smoothing is required since the empirical distribution is discontinuous while the alignment technique applies only to continuous curves. Also, it is crucial to choose the heights of the template at the points zero and one in an appropriate proportion. We will demonstrate these problems and discuss ways to overcome them. In addition, the amount of smoothing and the proportion of heights can be chosen adaptively by the data. Numerical properties of the method will be discussed and simulation results show that it has very good performance.

Bump Hunting Using Model-Based Penalized Least Squares

Nancy E. Heckman
University of British Columbia

A standard method for locating modes of a regression function μ is to first estimate μ (or μ') using a local polynomial or smoothing spline estimate and then to find the maxima of $\hat{\mu}$ (or the zeroes of $\hat{\mu'}$). The usual smoothing spline estimate of μ (or μ') is based on minimizing the penalized sum of squares $\sum (Y_i - \mu(x_i))^2 + \lambda \int \left(\mu^{(m)}\right)^2$ where x_i, Y_i are the observed data and, typically $m = 2$ or 3. The penalty $\int \left(\mu^{(m)}\right)^2$ in some sense defines a default model, the set of μ's with zero penalty, that is the default model is a polynomial of degree $m - 1$. In some cases, one may actually have in mind a different default parametric model. For instance, if one thinks that $\mu(x) = \beta_1 + \beta_2 \exp(\gamma x)$ is an approximate model, then one could use the penalty $\int (\mu'' - \gamma\mu')^2$. I'll compare the estimation of modes using several penalties.

Bump Hunting Problems: A New Test

Jiayang Sun
Case Western Reserve University

Is a small galaxy falling into our Galaxy? Is a part of the galaxy been torn apart? Or, how do we detect if there is an extended moving group near the galactic disk? These questions are related to the bump hunting problems, especially, the test of unimodality versus bimodality. Our test for testing unimodality vs bimodality and its application to radial velocity data will be presented. Some computational issues and challenging questions from the astronomical data will also be discussed.

Wednesday, 2:00 p.m. - 3:20

Invited Session: Wavelets, Mixtures, and Mode Trees

Rough Function Estimation via Wavelets

Yazhen Wang
University of Missouri-Columbia

In scientific studies, signals and objects are often far from smooth or piecewise smooth. These rough signals and objects are mathematically modeled by fractal functions. Fractal dimension is used to measure and characterize their roughness and quantitatively describe their irregularity. Fractal functions can't be well approximated by smooth or piecewise smooth functions and hence it may not be appropriate to estimate them by classic smoothing methods. The present paper investigates estimation of fractal functions by wavelet methods. It is shown that wavelet shrinkage can remove noise from a fractal function and virtually preserve its fractal dimension.

Adaptive Mixture Techniques for Model Complexity Estimation

Jeffrey L. Solka and George Rogers
Naval Surface Warfare Center

This talk will focus on some of our recent work for the estimation of appropriate mixture model complexity. These techniques focus on the use of adaptive mixture type estimators. In particular I will discuss a method for the estimation of model complexity using adaptive mixtures estimates of bootstrap resamples and several new iterative adaptive mixture type estimators.

Bumps Along the Road Towards Multivariate Mode Trees

David W. Scott
Rice University
William F. Szewczyk
Department of Defense

Minnotte and Scott proposed the mode tree as a means of capturing features in a kernel density estimator as the bandwidth varies. They noted the tree-like structure could replace the dendogram as a clustering summary. Silverman's work emphasizes the importance of the Normal kernel in this application. Here we illustrate some of the aspects of multivariate kernel estimation that are relevant to use of the mode tree as a clustering tool.

Index

Bowman, A., 5, 6, 11

Carroll, M., 4
Celeux, G., 6, 14
Chaudhuri, P., 5, 10
Cheng, M., 7, 15
Chiu, S., 5, 10
Cordero Brana, O. I., 6, 14

Fan, J., 6, 13
Farmen, M., 6, 7, 13

Gerard, P. D., 5, 9
Gijbels, I., 5, 6, 11, 13

Hall, P., 7, 15
Hartigan, J., 4, 6, 8, 15
Haruvy, E., 6, 13
Heckman, N. E., 4, 7, 16

Jones, M., 5, 11

Kim, D., 4, 8

Lin, A., 5, 12

Müller, H., 5, 9
Mammen, E., 4, 8
Marchette, D. J., 4, 8
Markatou, M., 6, 14
Marron, J. S., 4, 5, 8, 10
Minnotte, M. C., 4, 5, 11

Ogden, R. T., 6, 15
Olshen, R. A., 5, 12

Parzen, E., 5, 10
Polonik, W., 5, 11
Poston, W., 6

Ripley, B. D., 6, 14
Rogers, G., 7, 17

Sain, S. R., 5, 6, 13
Schucany, W. R., 5, 9
Scott, D. W., 4, 7, 17
Solka, J. L., 7, 17
Sugar, C. A., 5, 12
Sun, J., 7, 16
Szewczyk, W. F., 6, 7, 17

Terrell, G. R., 4, 6, 8

Udina, F., 4, 5, 8

Wang, Y., 7, 16