An IMS Mini-Meeting on
Imaging, Classification, and Clustering

Department of Statistics
University of Florida

January 11-12, 2002

Contents

Organizing Committee

Jim Booth, George Casella, Alan Hutson, Jim Kepner, Brett Presneli, Dave Wilson, Rongling Wu and Samuel Wu

Conference Schedule

Thursday, January 10

7:00-10:00 P.M. Reception (Keene Faculty Center in Dauer Hall)

Friday, January 11

8:30-9:00 A.M.	Breakfast (Room 2205, New Physics Building)
9:00-10:25 A.M .	Session 1 OPENING SESSION (Room 2205, New Physics Building) Chair: George Casella Neil Sullivan, Dean of the CLAS *Opening Remarks* Peter Hall *Nonparametric Estimation Of Component Distributions In A Multivariate Mixture* Christopher G. Small *The Statistical Analysis Of Dynamic Curves And Sections*
10:25-10:40 A.M.	Break
10:40-12:00 Noon	Session 2 IMAGE ANALYSIS (Room 2205, New Physics Building) Chair: Jim Booth Anuj Srivastava *Probability Models for Statistical Image Understanding: Are We There Yet?* Nicholas Lange *Microscopic Brain Tissue Imaging To Identify And Map Neuronal Distributions In Molecular Pharmacology And Functional Neuroanatomy*
12:00-1:30 P.M.	Lunch (The Micanopy Room, Reitz LTnion)
1:30-2:50 P.M.	Session 3 CLUSTERING AND CLASSIFICATION (Room 2205, NPB) Chair: Brett Presneli Mark C. K. Yang *Cluster identification in Time series and its applications to prediction* Art. B. Owen *Plaid Models and DNA Microarrays*
2:50-3:20 P.M.	Group Picture/Break
3:20-4:00 P.M.	Session 4 PHYNOGENETICS TREES (Room 2205, New Physics Building) Chair: Jim Hobert. Dennis Pearl *Simultaneous Estimation of a Phylogenetic Tree and its Evolutionary Model Parameters* Edward I. George *Bayesian Treed Modeling*
6:00-7:30 P.M.	Poster Session/Reception (Keene Faculty Center in Dauer Hall)

Saturday, January 12

8:30-9:00 A.M. Breakfast (Room 2205, New Physics Building)

9:00-10:20 A.M.	Session 5 Neural Networks (Room 2205, New Physics Building)	
	Chair: Dave Wilson	
	Terrence L. Fine	*A Tutorial on Large Margin Classifiers*
	Robert L. Paige	*Bayesian Inference in Neural Networks*
10:20-10:40 A.M.	Break	
10:40-12:00 Noon	Session 6 MICRO ARRAY ANALYSIS (Room 2205, New Physics Building)	
	Chair: Sam Wu	
	Hongyu Zhao	*Making Sense of Clusters from Microarray Data*
	Mark van der Laan	*Statistical Inference with Microarray Data*
12:00-2:00 P.M.	Lunch	
2:00-3:20 P.M.	Session 7 IMAGE ANALYSIS (Room 2205, New Physics Building)	
	Chair: Alan Hutson	
	Anand Rangarajan	*A unified non-rigid feature registration method for brain mapping*
	Christopher Genovese	*Controlling the False Discovery Rate: Understanding and Extending the Benjamini-Hochberg Method*
4:30-8:30 P.M.	Barbecue	(at the home of George Casella, Chair, Department. of Statistics, 2245 NW 24th Avenue)

Abstracts

Nonparametric Estimation Of Component Distributions In A Multivariate Mixture

Peter Hall, Australian National University

In some problems involving classification and discrimination, the main task is estimating the proportions of data that come from the respective distributions. For example, if patients present with symptoms expressible as fc-vectors, can the proportions of patients that are respectively "diseased" and "not diseased" be estimated from the mixture data? More generally, can the two fc-variate distributions, as well as the mixture proportions, be estimated? The answer is clearly "yes" if there is an appropriate parametric model for those distributions. However, the answer from a nonparametric viewpoint is in general unclear. Nevertheless, if each population has independent components then some progress can be made. Clearly nothing is identifiable, in a nonparametric sense, when $\kappa = 1$. And neither the mixture proportions nor the component distributions are identifiable if $\kappa = 2$. But for $\kappa \geq 3$, and in the case of independent components, everything is identifiable, even from a nonparametric viewpoint.

The Statistical Analysis Of Dynamic Curves And Sections

Christopher G. Small

By a curve, we shall understand a one-dimensional smooth path lying in R^2 or R^s which can be naturally parametrized by a real coordinate. The coordinate could represent physical time or any other variable which can be interpreted dynamically. In some cases, the curve will arise as the linear section of a higher-dimensional structure such as the planar section of a surface in R^s. In this talk, we develop a model for the shape of planar curves, based on their curvatures, that is reasonably robust to the location of landmarks or knots used to approximate the contours of the curve. The measurement for the shape difference between two curves that we propose is also based on the curvatures of the curves and directly inherits the simple Euclidean property for averages.

Probability Models for Statistical Image Understanding: Are We There Yet?

Anuj S rivast ava, Florida State University

Statistical inferences are basic to image understanding but they require probability models that capture real data and are tractable. A large number of models have been proposed over the years. There are low level models that study the texture directly: Markov random fields and component-based (PCA, ICA, etc) are examples. Then there are high level models that capture object shapes: Grenander's deformable template model is an example. Seeking models that analyze texture and relate to shapes, we present a parametric family that models well the observed histograms of the "filtered" images. These probabilities are universal in that they apply to all modalities: video, infrared, range, MRI, PET etc, and to both natural and artificai scenes. In this talk we analyze this family, present a few applications in clutter classification and object recognition, and suggest some extensions.

Microscopic Brain Tissue Imaging To Identify And Map Neuronal Distributions In Molecular Pharmacology And Functional Neuroanatomy

Nicholas Lange, Harvard University Schools of Medicine and Public Health

Abstract: A microscopic approach to brain imaging is required when one seeks to investigate the neuronal and neurochemical basis of brain activity in greater detail than is available through functional magnetic resonance imaging. Functional magnetic resonance imaging data are, in fact, quite removed from neuronal activity in both space and time since image elements each contain roughly two million neurons and time series at each of these elements are blurred temporally by local hemodynamics. The work to be presented involves digital micrographs of post-mortem brain tissue from two experiments conducted recently in my laboratory. Imaging data in the first experiment consist of a collection of two-dimensional (2-D) images of brain tissue from a designed experiment with two factors, animal and antipsychotic drug, that depict protein production encoded by a particular gene in cellular nuclei. Nuclear protein activity is separated from other image features through use of 2-D mathematical morphology and a multivariate discriminant function that employs size and shape measurements of segmented image objects. Prediction of spatial surface intensities derived from the resultant heterogeneous replicated spatial point patterns is performed though use of a weighted Poisson likelihood with tessellation-based weights in a generalized linear mixed-model framework. Imaging data in the second experiment consist of stacks of high-resolution through focus 2-D images, separated by less than one micron, containing profiles of various human brain cell types. Three-dimensional (3-D) objects are reconstructed from these profiles using 3-D mathematical morphology. Volume, shape, surface features and grayscale intensities are measured and 3-D objects are classified as being either a neuron or not and, if a neuron, as a pyramidal neuron or not.

Cluster identification in Time series and its applications to prediction

Deng-Shan Shiau and Mark C. K. Yang

Traditional time series analysis, regardless stationarity, usually tries to model the whole time series. This is sometime unrealistic because for many time series the pattern becomes regular (predictable) only at certain episodes. Morover, these episodes may have more meaning and importance than other occasions. A statistical algorithm is found to identify the most predictable patterns in time series data. This algorithm is shown to be very effective from our simulation studies and to be able to make much better predictions over traditional time series models from EEG and sunspots data.

Plaid Models and DNA Microarrays

Art Owen, Stanford University

This talk describes the plaid model, a tool for exploratory analysis of multivariate data. The motivating application is the search for interpretable biological structure in gene expression microarray data. Interpretable structure can mean that a set of genes has a similar expression pattern, in the samples under study, or in just a subset of them (such as the cancerous samples).

A set of genes behaving similarly in a set of samples, defines what we call a “layer”. These are very much like clusters, except that: genes can belong to more than one layer or to none of them, the layer may be defined with respect to only a subset of the samples, and the role of genes and samples is symmetric in our formulation.

The plaid model is a superposition of two way anova models, each defined over subsets of genes and samples. We will present the plaid model, an interior point style algorithm for fitting it, and some examples from yeast DNA arrays and other problems.

This is joint work with Laura Lazzeroni of Stanford University.

Simultaneous Estimation of a Phylogenetic Tree and its Evolutionary Model Parameters

Dennis Pearl, Ohio State University

New Markov chain Monte Carlo techniques for finding tree distributions and stochastic optimization methods for finding maximum likelihood trees can be readily adapted to include the estimation of the parameters of the assumed model of molecular evolution. This talk will discuss simultaneous confidence and credible regions for the tree and parameters underlying a bifurcating evolutionary process operating on nucleotide sequences. The ideas will be illustrated with examples from human viral evolution (HIV and HPV).

Bayesian Treed Modelling

Edward I. George, University of Pennsylvania

When simple parametric models such as linear regression fail to adequately approximate a relationship across an entire set of data, an alternative may be to consider a partition of the data, and then use a separate simple model within each subset of the partition. Such an alternative is provided by a treed model, which uses a binary tree to identify such a partition. However, treed models go further than conventional trees (eg CART, C4.5) by fitting models rather than a simple mean or proportion within each subset. In this talk, I will discuss a Bayesian approach for finding and fitting parametric treed models, in particular focusing on both linear and logistic regression. The potential of this approach is illustrated by a cross-validation comparison of predictive performance with neural nets, MARS, and conventional trees on simulated and real data sets. This is joint work with Hugh Chipman, University of Waterloo and Rob McCulloch, University of Chicago. A paper is available at:

http://www-stat.wharton.upenn.edu/~edgeorge/Research_papers/treed-models.pdf

A Tutorial on Large Margin Classifiers

Terrence L. Fine, Cornell University

Our discussion focusses on binary classification and is extended at the close to M-ary classifiers. We assume the existence of a training set. T_n = {(æ^U) : i = 1 : n} of n correctly labelled examples, with the ṙ-t.h such example having a d-dimensional numerical feature vector x_i and a label ti G {—f, f}. We make no assumptions concerning the probability model for this data beyond the examples being independent and identically distributed.

Given a {—1, l}-valued classification function f : IR^ i—► {—1,1}, its margin on (x_i, t_i) is $t'if(Xi)$· The ṙ-th example is correctly classified if and only if γ_i > 0 and confidently correctly classified only when γ_i is "substantially" greater than zero. Large margin classifiers (LMCs) are classification functions for which most (many times, all) examples are classified with large margins. Functions capable of large margin classification on an arbitrary training set T_n can only be selected by an appropriate training algorithm A operating on a highly flexible family T of potential classifiers. For support vector machines (SVMs) this family T is the relatively simple one of classification by a hyperplane but the necessary complexity is gained by embedding the feature vector x inR^d into a new feature vector y G IR^D with D » d. We will not discuss SVMs. Instead we will consider JF that arise as the convex hull of a set B of relatively simple base functions (e.g., small single hidden layer neural networks).

The notion of such classifiers has been primarily developed and pursued by those concerned with so-called machine learning. An LMC is generally a quite complex function and one so complex that at first one would believe that it must overfit. the training data T_n, achieving near zero training set errors, but performing poorly on independent testing data C_m. However, the training algorithms that have been developed (we will discuss AdaBoost) come with Vapnik-Chervonenkis-like theorems guaranteeing that, with high probability, the true error probabilities are not much larger than the large margin empirical error probabilities observed on the training data. There is an encouraging confluence of supportive empirical studies for the good performance of LMCs and theorems providing accessible upper bounds on the resulting true error probabilities.

Bayesian Inference in Neural Networks

Rob Paige, Texas Tech University

Approximate marginal Bayesian computation and inference are developed for neural network models. The marginal considerations include determination of approximate Bayes factors for model choice about, the number of nonlinear sigmoid terms, approximate predictive density computation for a future observable, and determination of approximate Bayes estimates for the nonlinear regression function. The choice of prior and the use of an alternative sigmoid lead to posterior invariance in the nonlinear parameter which is discussed in connection with the lack of sigmoid identifiability. A principal finding is that parsimonious model choice is best determined from the list of modal estimates used in the Laplace approximation of the Bayes factors for various numbers of sigmoids. The proposed methods are illustrated in the context of two nonlinear datasets that involve respectively univariate and multivariate nonlinear regression models.

Making Sense of Clusters from Microarray Data

Hongyu Zhao, Yale University School of Medicine

Recent advances in large-scale RNA expression measurements, e.g. cDNA microarrays, and proteomics technologies have opened the opportunity for massively parallel biological data acquisition and thus have shifted our attention towards an integrated understanding of the genetic networks underlying complex biological phenotypes. Many existing statistical procedures, most notably a variety of clustering algorithms, have been applied to analyze microarray data to identify genes or samples with correlated expression patterns. However, compared to the large number of clustering algorithms, there is a lack of statistical and computational methods to interpret the observed gene expression patterns. In this talk, we will describe our attempts to integrate various genomics information to identify biological mechanisms underlying gene clusters. The developed methods will be illustrated through their applications to yeast cell cycle data.

Statistical Inference with Microarray Data

Mark J. van der Laan

Large-scale gene expression studies are becoming increasingly common as new microarray technology makes it possible to capture the gene expression profiles for thousands of genes at once. Statistical inference with such high dimensional data structures (and, all too often, relatively small samples) is a challenging analytical problem. Firstly, we address multiple testing and provide optimal multiple testing procedures. In the current microbiology literature, (hierarchical) cluster analysis methods have been used to find groups of genes with similar patterns of expression. Such methods are purely exploratory and, thus, do not provide any type of significance level for features in the data or any opportunities for purposeful experimental design. We propose the use of a deterministic rule, applied to the parameters of the gene expression distribution, to select a target subset/clustering parameter of biological interest. We focus on parameters that are functions of the mean vector and covariance (i.e. correlation) matrix; we also employ the output of general clustering algorithms (i.e. "partitioning around medoids" or PAM) to define the clustering parameter. An estimate of the target parameter is obtained by applying the procedure to the sample statistics (e.g. sample mean and covariance). The bootstrap is used to estimate the distribution of these estimated subsets/clusters; relevant summary measures of this distribution are also proposed. We prove consistency of the subset estimates and asymptotic validity of the bootstrap under the assumption that the sample size converges faster to infinity than the logarithm of the number of genes. The method has also been used to analyze cancer-patient, data.

A unified non-rigid feature registration method for brain mapping

Anand Rangarajan, University of Florida

We describe the design, implementation and results of a unified non-rigid feature registration method for the purposes of anatomical MRI brain registration. The method draws on our previous work on robust non-rigid point matching but with a crucial difference. In contrast to previous work, we pose the feature registration problem in a many-to-many as opposed to one-to-one matching framework. Our new non-rigid registration method implements an iterative joint clustering and matching (JCM) strategy which effectively reduces the computational complexity, and ensures many-to-many matching. We demonstrate the application of the method using two different types of cortical anatomical features: the outer cortical surface and major sulcal ribbons. Points sub-sampled from each type of feature are fused into a common 3D point-set. representation. We have conducted carefully designed synthetic experiments to study the effect of using different types of features either separately or together. A validation study examining the accuracy of non-rigid alignment of many brain structures is also presented. Finally, we extend the approach to the construction of anatomical atlases.

Controlling the False Discovery Rate: Understanding and Extending the Benjamini-Hochberg Method

Christopher Genovese, Carnegie Mellon University

A common analysis of functional neuroimaging data involves performing a hypothesis test at every volume element in the brain to locate brain activity. Radio astronomers perform tests at every location in a map to distinguish sources from background. In the analysis of DNA microarrays, comparisons across experimental groups involve simultaneous tests at thousands of genes. These are all examples of multiple testing problems, which arise frequently in applications. The challenge of multiple testing is to define a decision rule that provides good power while controlling some overall measure of error. Traditional methods seek strong control of the familywise error rate. Benjamini & Hochberg (1995) introduced a. new criterion - the False Discovery Rate - and put forward a procedure to control it. Their procedure is widely applicable, computationally simple, and often provides better power than competing methods. The reasons underlying the method's success are, however, somewhat mysterious at first. In an effort to de-myst.ify, I will present two alternate perspectives on the procedure - one asymptotic and one structural - and will examine its operating characteristics under several risk measures. In particular, I will introduce the dual quantity, called the False Nondiscovery Rate, and use it to assess performance. These ideas lead to methods for constructing confidence intervals on the realized false discovery rate and then to new procedures for controlling both rates, including a bootstrap method for bounding tail probabilities on the realized FDR.

This is joint work with Larry Wasserman at Carnegie Mellon Statistics.

Poster Abstracts

A Unified Approach to Echocardiographic Image Analysis

David C. Wilson, University of Florida

The purpose of this presentation is to describe the development and design of computer-based algorithms, which make measurements on the 4 views of the heart, used in a. st.ress-echo test. The 4 views are: I), the parasternal short-axis, 2). the apical 4 chamber, 3). the apical 2 chamber, and 4). the parasternal long-axis.

The first step in the development process is to use the ideas of Procrustes shape analysis, thin-plate spline, and cluster analysis to build a geometric model from a given dataset of expert estimates. The second step is to use the geometric model and the thin-plate spline transformation to register a collection of images. The third step is to compute the average image from the collection of registered images. The fourth step is to use the average image to create a collection of convolution filters to locate key features in the image.

Once convolution filters have been created, the first step in the design of an algorithm is to locate the features through standard matched filter techniques. Once the locations have been found, the second step is to use the thin-plate spline transformation to embed the geometric model into the image.

One advantage of this program is that it can be applied to any view of the heart. A second benefit is that it provides an orderly process for incorporating the knowledge and experience of the expert into a computer-based method. In particular, it allows for the fact, that an expert, may not. place his estimates of the epica.rdia.1 and endocardial borders in locations of high intensity or high gradient..

This is a. joint, work with Yunmei Chen, Feng Huang, Edward A. Geiser, MD, John barocca., Jenifer Buxe, all at the University of Florida.

Mult.icategory Support. Vector Machines, with application to cancer classification using gene expression data.

Yoonkyung Lee, University of Wisconsin-Madison

The Support. Vector Machine (SVM) has become a. popular choice of classification tool in practice. Despite its theoretical properties and its empirical success in solving binary problems, generalization of the SVM to more than two classes has not. been obvious. Oftentimes multicategory problems have been treated as a. series of binary problems in the SVM paradigm. However, solutions to a. series of binary problems may not. be optimal for the original mult.icategory problem. We propose mult.icategory SVMs, which extend the binary SVM to the mult.icategory case, and encompass the binary SVM as a. special case. The proposed method deals with the equal misclassification cost, and the unequal cost, case in a. unified way. It. is shown that, the mult.icategory SVM implements the optimal classification rule for appropriately chosen tuning parameters as the sample size gets large. The effectiveness of the method is demonstrated through simulation studies and real applications to cancer classification problems using gene expression data.

Restaurant Guide to Gainesville[1]

Orientation

Gainesville's quadrants (NW, NE, SE, SW) have Streets running N-S, Avenues and Places running E-W. University Ave. separates north and south, Main St. separates east and west, and Archer Road radiates out of the city toward the southwest. The university falls south of University Ave. and west of SW 13th Street, and "downtown" is about a mile east, near the intersection of University and Main.

Key:

* has outdoor tables
Φ within walking distance of conference
A A food especially recommended.

Luncheon places

The Reitz Union (has) the usual type of student food on the first floor - a Food Court including Wendy's, Subway, Allegro Pasta, The Wokery and a Java Hut. The ground floor facing the pond has a pizza take-out, yogurt shop, donut shop, and the Baja Tortilla Grill.

West University Ave., between about 18th St. and 10th St., has a variety of small restaurants. All the ones listed below (except Wine and Cheese) are within a ten or fifteen minute walk of the Reitz Union and the Holiday Inn.

Cafe Gardens (1632 W. University Ave.) - burgers, sandwiches and good salads of various types, usually good grilled fish sandwich specials. * # AA

Copper Monkey (1700 W. University Ave.) - burgers, sandwiches, salads, no fish. #

Kotobuki (1702 W. University Ave.) - Japanese, including sushi; luncheon specials. #

Farah's On The Avenue (1120 W University Av) - Middle eastern food, Pita sandwiches, bar, a block east of the Holiday Inn * *Φ*

Saigon Cafe (1222 W University Av) - Vietnamese, a block east of the Holiday Inn * *Φ*

Chaucer's (112 NW 16th St) - Chili, Veggie burgers, sandwiches, Gourmet coffee #

Bageland (1717 NW 1st Ave.) - bagels with various toppings. * #

Burrito Brothers (16 NW 13th St.) - take out Mexican, inexpensive. *Φ*

El Indio (407 NW 13th St.) - Mexican, inexpensive. * #

Book Lover's Cafe (505 NW 13th St.) - Good vegetarian selection #

[1]List originally prepared by Alan Agresti (alias AA), who takes credit for your good meals but no responsibility for the disappointing ones or for poor service.

Leonardo's by the Slice (1245 W. Univ. Ave., across from Holiday Inn) - several types of pizza by the slice and daily pasta specials and espresso bar, inexpensive. * # AA

Bistro (1245 W. Univ. Ave., next to Leonardo's) - salads, sandwiches in the $7-8 range including recommended eggplant and grilled tuna sandwiches, and a couple of pasta dishes.* # AA

The Wine and Cheese Gallery (113 N Main St) - nice variety of American and European lunch dishes. Highly recommended. * JB

Dinner recommendations

The nicest restaurant strip is downtown along SE 1st Ave. and nearby, just east of Main St. The first eight restaurants listed below are in this location. This is about a mile east of Holiday Inn and two miles from the Reitz Union. Though one can walk there from Holiday Inn directly east on University Ave., it is recommended that you not walk at night in the area north of University between NW 13th St. and Main. All restaurants listed take credit cards. None are "dress-up," though jeans would probably not be acceptable for Sovereign, Mr. Han's, and possibly Amelia's.

Amelia's (235 S. Main, in Sun Center, 373-1919) - very good Italian (owner from Naples, takes reservations. * AA

Emiliano's (7 SE 1st Ave., 375-7381) - Puerto Rican cafe serves sandwiches with salad or dinner entrees; pleasant, informal ambiance, good desserts and cafe con leche. * AA

Harry's (SE 1st Ave. near Emiliano's, 372-1555) - New Orleans Style, inexpensive and informal, though food recommended not as highly as others in this area. *

Sovereign (12 SE 2nd Ave., 378-6307) - continental cuisine in converted coach-house, rather formal and somewhat more expensive (similar to Wolfgang's) AA

Pura Vida (12 SW 1st Av, 378-3398) - Inexpensive, eclectic variety of foods in an informal cafe setting. Dishes usually include jerk chicken, penne pasta, a spicy shrimp dish, sometimes an Indian lamb dish.

Market Street Pub (120 SW 1st Av) - Good beer selection, English pub food * JB

Mark's Downtown (Sun Center, 336-0077). Nouvelle cuisine, at least for Gainesville. *BP

The Brasserie (101 SE 2nd Place (Sun Center), 375-6612). Nouvelle cuisine, at least for Gainesville. BP

Leonardo's 706 (706 University, 378-2001) - good Italian, more casual than Amelia's, "California-style" Italian food includes pizzas, innovative entrees and excellent desserts. * # AA

Mildred's Big City Food (3445 W University Ave. (in a shopping center/strip mall), 371-1711). Nouvelle cuisine, at least for Gainesville. *BP

Bahn Thai (1902 SW 13th St., 335-1204) - Thai food, very good, not too expensive . BP

Szechuan Panda (3830 SW 13th St., 336-6464) - Chinese

Mr. Han's (6944 NW 10th Place, 331-6400) - probably the best Chinese food in G'ville, though about 6 miles west of university. AA

Saki Drops (NW 39th Av, 372-8686) - Japanese/Asian, "some people think this is the best restaurant in Gainesville" AA

Archer Road strip - Archer Road, between about the 3200 and blocks, contains a large selection of chain restaurants, such as Texas Roadhouse, Outback Steak House, TGIF, Alehouse, Steak and Ale, etc. Most of these have long lines of students on weekends. Keep driving out Archer Road another 55 miles and you come to Cedar Key, a small fishing village on the Gulf of Mexico with several restaurants facing the Gulf.