Classification Society 2011 Annual Meetings

Carnegie Mellon University, Pittsburgh, PA: June 16th-18th, 2011
(registration and all meeting locations are in Baker Hall, Lower Level)

Wednesday, June 15th: Welcome Happy Hour, 6-8pm *Bridges Lounge, Holiday Inn*

Thursday, June 16th:

8:00-8:30am: Bagels, Pastries, Coffee *Baker Coffee Lounge*

8:30-8:45am: Conference Welcome/Opening Remarks

8:45-10:15am: **Classification Applications in Statistical Forensics**
Chair: Beth Ayers, Graduate School of Education, UC Berkeley

David Friedenberg, Battelle Memorial Institute

Use of GCxGC-TOFMS and Pattern Recognition for Detection and Attribution of Organophosphate Pesticide Signatures

Jared Schuetter, Battelle Memorial Institute

What Made This Hole? The Challenge of Image Analysis in Munition Forensics

Jennifer Wightman, Battelle Memorial Institute

A Classification Algorithm for the Detection of Chemicals Using Raman Spectroscopic Data

10:15-10:30am: Coffee Break *Baker Coffee Lounge*

10:30-12:00pm: **Classification Applications in Relational Data**
Chair: Rebecca Nugent, Dept of Statistics, Carnegie Mellon

David Krackhardt, Heinz School, Carnegie Mellon University

Efficient Extraction of Quality Subgraphs from Large Populations

Pavel Krivitsky, Heinz College/iLab, Dept of Statistics, Carnegie Mellon University

Latent Space Cluster Models for Social Networks

Andrew Thomas, Dept of Statistics, Carnegie Mellon University

Friends and Enemies: The Classification of Social Network Ties by Antagonism

12:00-1:30pm: Lunch Break

1:30-2:15pm: **Keynote Address**
Chair: Rebecca Nugent, Dept of Statistics, Carnegie Mellon

Stephen E. Fienberg, Maurice Falk University Professor of Statistics and Social Science, Carnegie Mellon University

Mixed-Membership Models for Disability, Text, and Network Analysis

2:15-2:30pm: Coffee Break *Baker Coffee Lounge*

2:30-4:00pm: **Recent Work in Classification in Education**
Chair: Elizabeth Hohman, Naval Surface Warfare Center

Elizabeth Ayers, Graduate School of Education, UC Berkeley
Analyzing the Learning Progressions in the Assessing Data Modeling and Statistical Reasoning Curriculum

Tracy Sweet, Dept of Statistics/PIER, Carnegie Mellon University
Latent Space Social Network Models for Interventions in Education Policy

Turadg Aleahmad, Human Computer Interaction/PIER, Carnegie Mellon University
Automatic Rating of User-Generated Math Solutions

4:00-4:15pm: Coffee Break *Baker Coffee Lounge*

4:15-5:30pm: **Some Recent Work in Classification Methodology**
Chair: Beth Ayers, School of Education, UC Berkeley

Gabrielle Flynt, Dept of Statistics, Carnegie Mellon University
Clustering Trajectories in the Presence of Informative Patterns of Monotone Missingness

Stephen France, Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee
Boosting Unsupervised Additive Clustering

Daniel McDonald, Dept of Statistics, Carnegie Mellon University
Spectral approximations in machine learning

5:30-7:30pm: Reception (drinks and light snacks) *Phipps Conservatory*

Friday, June 17th:

8:00-9:00am: Bagels, Pastries, Coffee *Baker Coffee Lounge*

9:00-9:45am: **Presidential Address**
Chair: Stan Sclove, Dept of Information and Decision Sciences, UIC

William Shannon, Divisions of General Medical Sciences and Biostatistics,
Washington University of St Louis School of Medicine

Ranking Physicians, Hospitals, Teachers, Auto-Mechanics,...
(or anything else that takes something, does something to it, and has a measurable outcome)

9:45-10:00am: Coffee Break *Baker Coffee Lounge*

10:00am-12:00pm: **Classification Society Dissertation Award Finalists**
Chair: Samantha B. Prins, Dept of Mathematics & Statistics, James Madison University

Frank Busing, Leiden University (WINNER)
Advances in Multidimensional Scaling

Hongxia Yang, Dept of Statistical Science, Duke University
Nonparametric Bayes Models for High-Dimensional and Sparse Data

Jorge Tendeiro, Faculty of Behavioural and Social Sciences, Heijmans Institute
Some Mathematical Results on Three-Way Component Analysis

David Casado, Dept of Statistics, Universidad Carlos III de Madrid
Classification Techniques for Time Series and Functional Data

12:00-1:30pm: Lunch Break (CS Board Meeting in Baker 154R)

1:30-3:15pm: **Graphs, Risks, and Clusters: The Work of Bernie Harris**
A memorial session celebrating the life and work of Bernard Harris (1926-2011)
Chair: Willem Heiser, Faculty of Social and Behavioural Sciences, Leiden University

Willem Heiser, Faculty of Social and Behavioural Sciences, Leiden University
Remembering Joseph Kruskal and Doug Carroll

David Banks, Dept of Statistical Science, Duke University
Adversarial Risk Analysis

Mel Janowitz, DIMACS, Rutgers
Continuity! To be or not to be! Is that even a question?

Stan Sclove, Dept of Information and Decision Sciences, UIC
Bernie Harris' Contributions to Cluster Analysis

3:15-3:30pm: Coffee Break *Baker Coffee Lounge*

3:30-5:00pm: **Combining Classification Trees and Regression Models**
Chair: TBA

Yuning He, Dept of Applied Mathematics & Statistics, UC Santa Cruz
Predicting Variable-Length Functional Outputs for Emulation of a NASA Flight Simulator

Herbie Lee, Dept of Applied Mathematics & Statistics, UC Santa Cruz
Optimization Under Unknown Constraints

Bobby Gramacy, Dept of Econometrics & Statistics, U of Chicago Booth School of Business
Dynamic Trees for Dynamic Trees for Response Surface Learning and Optimization

5:00-5:30pm: Classification Society General Meeting

7:00-10:00pm: Dinner Cruise on the Three Rivers *Station Square*
(Boarding starts at 6pm; more details in registration packet)

Saturday, June 18th:

8:00-8:30am: Bagels, Pastries, Coffee *Baker Coffee Lounge*

8:30-10:00am: **A Survey of Some Classification Applications**
Chair: TBA

Elizabeth Hohman, Naval Surface Warfare Center
Hypergraphs from Twitter Data

Michael Kurtz, Harvard-Smithsonian Center for Astrophysics
Using Multipartite Graphs for Recommendation and Discovery

Tim Brennan, Northpoint Institute
An Exploratory Taxonomic Study of Womens Pathways to Crime:
From Qualitative to Quantitative Patterns

10:00-10:15am: Coffee Break *Baker Coffee Lounge*

10:15-11:45am: **How Can We “Better Use” Classification/Clustering?**
Chair: TBA

Ahmed Albatineh, Dept of Epidemiology and Biostatistics, Florida International University
Effects of Some Design Factors on the Shape of Similarity Indices in Cluster Analysis

Shuai Sun, University of Massachusetts, Boston
Classification Society Automated Search Service

David Dubin, Graduate School of Library and Information Science, UIUC
Data Theory and Scientific Data Management

11:45-12:30pm: Coffee/Lunch Break (sandwiches provided) *Baker Coffee Lounge*

12:30-2:00pm: **Classification Applications in the Biological Sciences**
Chair: TBA

Li Liu, Dept of Statistics, Carnegie Mellon University
Investigating New Typologies for Classifying Aphasia Patients

Patricio La Rosa, Divisions of General Medical Sciences and Biostatistics,
Washington University of St Louis School of Medicine
Object Data Analysis of Taxonomic Trees from Human Microbiome Data

William Shannon, Divisions of General Medical Sciences and Biostatistics,
Washington University of St Louis School of Medicine
Dirichlet-Multinomial Power Calculations And Statistical Tests For Microbiome Data

2:00pm: Conference Closing Remarks

Classification Applications in Statistical Forensics

Chair: Beth Ayers, Graduate School of Education, UC Berkeley

Use of GCxGC-TOFMS and Pattern Recognition for Detection and Attribution of Organophosphate Pesticide Signatures

David A. Friedenberg; Erich D. Strozier; Douglas D. Mooney; Theodore P. Klupinski; Cheryl A. Dingus
Battelle Memorial Institute, Columbus, OH USA

Organophosphorus pesticides (OPPs) are a group of highly toxic compounds that are widely available in many countries. Thus, OPPs may be attractive to terrorist and criminal elements for use as chemical threat agents. In the event of a criminal action involving OPP, identifying the source of the chemical used could lead investigators to the responsible party. Components other than the parent OPP such as synthetic precursors or byproducts are often present in commercial preparations of OPPs and may offer information on the source or synthesis methods used, thus providing a fingerprint for different sources and/or manufacturing processes. A study was conducted to identify chemical attribution signatures (CAS) for commercially available OPPs by applying statistical pattern recognition techniques to analytical data acquired by GC?GC-TOFMS. Because the number of predictor variables is far greater than the number of observations, the data could not be classified using classical statistical methods such as logistic regression or discriminant analysis. Instead, penalized regression and random forest classification techniques were evaluated for their ability to separate the samples into groups corresponding to the different materials for a given compound. Our research demonstrates that applying statistical pattern recognition techniques to GC?GC-TOFMS analytical data can be useful in the attribution of OPP chemicals from specific sources. Furthermore, the results suggest that this combination of analytical chemistry and statistical approaches can be applied to chemical forensic analysis for source attribution or the discovery of attribution signatures.

What Made This Hole? The Challenge of Image Analysis in Munition Forensics

Jared M. Schuetter; David A. Friedenberg; Douglas D. Mooney
Battelle Memorial Institute, Columbus, OH USA

When an unknown munition is fired at a target, the damage to the target at the point of impact can be used to identify the munition. This signature is affected by many sources of variability. These include, but are not limited to, the type of munition used (e.g. small arms, grenades, explosives), the angle and distance from the target, and the target material. Using images of the blast damage on such targets, scientists at the Battelle Memorial Institute have been developing feature extraction and pattern recognition techniques to predict the type of munition used. There are a number of difficulties involved in this process, primarily due to the large amount of variability from image to image, even for the same type of weapon. This variability can be attributed to the causes mentioned above, but also include environmental nuisance effects (e.g., lighting, camera angle, obstructions) that appear in the images. As a result of this variation, designing feature extraction algorithms that work well in all images can be tricky. A choice of a good classifier is crucial as well. This presentation will describe the challenges present in this classification context, some possible solutions, and open questions to resolve in the future.

A Classification Algorithm for the Detection of Chemicals Using Raman Spectroscopic Data

Jennifer L. Wightman, Jared M. Schuetter, Douglas D. Mooney, Theodore J. Ronningen *Battelle Memorial Institute, Columbus, OH USA*

We describe a classification algorithm for the detection of chemicals based on Raman spectroscopic data. Automated chemical detection and classification has applications in national security, forensics, and manufacturing and industrial quality control. The algorithm under consideration is based on the model that

a given test spectrum can be represented as a linear combination of the individual spectra in the training set. A constrained optimization problem is solved to determine the combination of training spectra that optimally represents the test spectrum. The method of backward selection, based on an F-test criterion, is applied to determine which of the contributing spectra are statistically significant in the classification. The application of backward selection is useful in removing weights that numerically improve the residual error of the optimization problem but have no chemical significance. One particular strength of this classification algorithm is the ability to identify mixtures of chemicals.

Classification Applications in Relational Data

Chair: Rebecca Nugent, Dept of Statistics, Carnegie Mellon

Efficient Extraction of Quality Subgraphs from Large Populations

David Krackhardt, *Heinz School, Carnegie Mellon University*

TBA

Latent Space Cluster Models for Social Networks

Pavel Krivitsky, *Heinz College/iLab, Dept of Statistics, Carnegie Mellon University*

Latent space models for social networks postulate the existence of a latent "social space", where the probability of a relation between entities depends on their relative positions within this space. The latent cluster model for social networks models groups of entities as clusters on the latent space, and Bayesian fitting of this model via MCMC allows the latent space position estimation to borrow strength from the cluster process. This talk describes some extensions to this model: representing inhomogeneity through covariates and generalize the model to non-binary data. We use this model to detect and identify demographics of magazine subscribers based on the number of subscribers each pair of magazines share.

Friends and Enemies: The Classification of Social Network Ties by Antagonism

Andrew Thomas, *Dept of Statistics, Carnegie Mellon University*

The clustering and classification of groups in standard social networks has been actively studied under the name ”community detection” for some time, but these methods rely largely on binary relations, as if two people in the network could only be friends or strangers. The introduction of antagonistic relations, or ”enemy ties”, complicates this matter by changing many of the assumptions behind this methodology. I review a number of the complications when it comes to empirically determining the clustering relationships in social networks once this information is introduced, and illustrate their meaning in determining whether a single tie can best be classified as friendly or antagonistic in the context of the entire network.

Keynote Address

Chair: Rebecca Nugent, Dept of Statistics, Carnegie Mellon

Mixed-Membership Models for Disability, Text, and Network Analysis

Stephen E. Fienberg
Maurice Falk University Professor of Statistics and Social Science, Carnegie Mellon University

In traditional clustering problems there are two tasks: determine the number of clusters or groups, and allocate units to one and only one cluster. Mixed membership models are formal statistical models that relax this clustering approach by allowing each unit to belong more than one group, a cording to a probability-like membership vector. I will illustrate the mixed-membership approach to statistical analysis through three different applications involving the classification of research articles using text and references, groupings in social networks, and longitudinal profiles based on survey measurement of disability.

Recent Work in Classification in Education

Chair: Elizabeth Hohman, Naval Surface Warfare Center

Analyzing the Learning Progressions in the Assessing Data Modeling and Statistical Reasoning Curriculum

Elizabeth Ayers, *Graduate School of Education, UC Berkeley*

The Assessing Data Modeling and Statistical Reasoning (ADMSR) project has developed a framework of seven basic constructs that describe the elements of statistical learning in middle school students. The seven constructs, or progress variables, considered in this framework were developed through a series of design experiments to explore the typical patterns of change as students learned to construct and revise models of data as a part of the Model Measure curriculum. This presentation explores initial clustering work on pre-post test data to model changes in groups of students between the exams. In addition, post-test data from a large scale test is examined.

Latent Space Social Network Models for Interventions in Education Policy

Tracy Sweet, *Dept of Statistics/PIER, Carnegie Mellon University*

Recent intervention studies in education the possibilities of both multiple replicates of similar networks as well as effects of interventions on the networks themselves. Within the network literature, subgroup identification is of particular interest, and there is reason to believe an intervention on a sample of schools can alter the subgroup structure of each school network. Since current social network statistical models focus on modeling one network at a time, I propose models to accommodate both multiple partially-exchangeable networks as well as network-level experiments. As an application, I present a simulated intervention that affects the number of subgroups within each network.

Automatic Rating of User-Generated Math Solutions

Turadg Aleahmad, *Human Computer Interaction/PIER, Carnegie Mellon University*

Intelligent tutoring systems adapt to users' cognitive, but not affective or conative factors. Crowd-sourcing may be a way to create materials that engage users along these differences. We build on earlier work in crowd- sourcing worked example solutions and offer a data mining method for automatically rating the contributions to determine which are worthy of presenting to students. We find that with 64 examples available, trained model on average exceeded the agreement of human experts. We also discuss the generalizability of this result and how examples might be generated.

Some Recent Work in Classification Methodology

Beth Ayers, Graduate School of Education, UC Berkeley

Clustering Trajectories in the Presence of Informative Patterns of Monotone Missingness

Gabrielle Flynt, *Dept of Statistics, Carnegie Mellon University*

Growth mixture models are a method for analyzing longitudinal data and have been recognized for their usefulness in identifying homogeneous subpopulations, potential meaningful classes of subject growth trajectories, within the larger heterogeneous population. Missing data is an inevitable obstacle present in longitudinal studies. It is common for subjects to miss intermittent measurements throughout the study, or even to drop-out of the study prematurely. Missingness is most often dependent on some unobserved variables and may cause biased estimates in statistical procedures that do not account for the informative missing values. Pattern mixture models are an approach to missing data analysis that model the drop-out mechanism present in longitudinal data with informative missingness. A common assumption used in pattern mixture models is known as the complete case missing variable restriction, which uses fully observed trajectories to make parameter estimates for trajectories with missing values. Attempting to cluster subject growth trajectories without accounting for informative missingness can easily lead to misclassification. It is probable that different patterns of missingness have different causes and will have different effects on subject outcomes. The goal of this work is to combine pattern mixture models and trajectory classification while investigating the validity of the commonly used complete case missing variable restriction. Results will be shown for several simulated data sets as well as a data set that measures clinical depression in subjects.

Boosting Unsupervised Additive Clustering

Stephen France, Ahmed Abbasi, *Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee*

We describe a methodology for large scale overlapping cluster analysis. In overlapping cluster analysis, each data item can be a member of multiple clusters or classes. This is opposed to partitioning clustering, where each item is assigned to exactly one cluster, and fuzzy clustering, where each item has a fuzzy membership probability for each cluster. The overlapping clustering techniques described in this paper utilize an additive decomposition of similarity. Thus, in the psychology literature, overlapping clustering is often called additive clustering. We describe some overlapping clustering models and describe optimization methods for fitting these models. We discuss scalability problems that occur when fitting overlapping clustering models and describe a methodology for scaling up overlapping clustering using supervised multi-label classification techniques. We perform experiments to test the quality of the scaled up solutions and then we give an illustrative example to show how overlapping clustering can be used as an exploratory data analysis tool to help learn categories from data.

Spectral approximations in machine learning

Daniel McDonald, Darren Homrighausen, *Dept of Statistics, Carnegie Mellon University*

In many areas of machine learning, it becomes necessary to find the eigenvector decompositions of large matrices. We discuss two methods for reducing the computational burden of spectral decompositions: the more venerable Nystr?m extension and a newly introduced algorithm based on random projections. Previous work has centered on the ability to reconstruct the original matrix. We argue that a more interesting and relevant comparison is their relative performance in clustering and classification tasks using the approximate eigenvectors as features. We demonstrate that performance is task specific and depends on the rank of the approximation.

Presidential Address

Chair: Stan Sclove, Dept of Information and Decision Sciences, UIC

Ranking Physicians, Hospitals, Teachers, Auto-Mechanics,...(or anything else that takes something, does something to it, and has a measurable outcome)

William Shannon,
Divisions of General Medical Sciences and Biostatistics, Washington University of St Louis School of Medicine

Comparative effectiveness research (CER) is designed to identify healthcare interventions having the best patient outcomes to direct patients to receive the best treatment and to direct our healthcare dollars to where they will be most productive. When comparing observational data to determine the best intervention, CER requires that we apply risk or case-mix adjustment methods before examining outcomes of care. For example, to compare survival in treatment or hospital for inpatient acute myocardial infarction (AMI) patients using the proportion surviving may be misleading if the severity of disease is significantly different across interventions or hospital. To make comparisons valid, risk adjustment must balance patient factors, such as disease severity and co-morbidities, which result in different likelihood of death. A standard approach to risk adjustment is to use measures of ”observed-to-expected” rates, where expected outcome for patients are estimated by an existing, often unknown and proprietary, regression model previously fit to a standard or reference population of patient data said to be representative of all patients. The observed outcome is obtained from the patient's discharge data. The goal of the risk adjustment is to determine if an intervention (or provider) on average shows better, worse, or the same observed outcomes compared to expected outcomes.

We propose to develop and release an open-source HealthCare Rankings (HCR) case-mix adjustment software package combining methods from observational data analysis, operations research, statistics, and mathematics that have not been applied in combination previously in CER and health services research. The HCR algorithm ranks two or more interventions or providers simultaneously based on direct comparison of patient-level data. This algorithm avoids the need to have a reference database for observed-to-expected comparisons. This proposal is a joint effort of investigators in the Washington University School of Medicine (WUSM) Dept. of Medicine's Biostatistical Consulting Center and the BJC HealthCare Center for Clinical Excellence (CCE). There are 11 hospitals in the BJC network with a comprehensive informatics system of patient level clinical and administrative data available for developing and validating the HCR algorithm.

Classification Society Dissertation Award Finalists

Samantha B. Prins, Dept of Mathematics & Statistics, James Madison University

Advances in Multidimensional Scaling

Frank Busing, *Leiden University* (WINNER)

No advances without a proper description of the path that leaded up to the current status quo. At the end of the last century, nonmetric, least-squares, unfolding was more dead than alive, a state that was reached soon after its conception in the early sixties. Developments in multidimensional scaling by Shepard (1962) and Kruskal (1964) accelerated the technical foundations of multidimensional unfolding, for which the conceptual foundations were already laid down by Coombs (1950, 1964), but the degeneracy problem soon proved an insurmountable hurdle. With varying success, researchers tried to overcome the stubborn problem. Their work paved the road for researchers working on the problem in the twentieth century. At last, the hurdle was taken by Busing et al. (2005). Their penalized Stress function is able to avoid degenerate solutions, but at the same time introduces two penalty parameters. Apart from the suggested defaults for these parameters, the optimal choice remains an open issue. Nevertheless, the non-degenerate unfolding algorithm allows for some overdue maintenance on nonmetric, least squares, unfolding. We therefore conclude with some examples of recent (and future) developments.

Nonparametric Bayes Models for High-Dimensional and Sparse Data

Hongxia Yang, *Dept of Statistical Science, Duke University*

Latent class models (LCMs) are used increasingly for addressing a broad variety of problems, in- cluding sparse modeling of multivariate and longitudinal data, model-based clustering, and flexible inferences on predictor effects. Typical frequentist LCMs require estimation of a single finite num- ber of classes, which does not increase with the sample size, and have a well-known sensitivity to parametric assumptions on the distributions within a class. Bayesian nonparametric methods have been developed to allow an infinite number of classes in the general population, with the number represented in a sample increasing with sample size. In this article, we propose a new non- parametric Bayes model that allows predictors to flexibly impact the allocation to latent classes, while limiting sensitivity to parametric assumptions by allowing class-specific distributions to be unknown subject to a stochastic ordering constraint. An efficient MCMC algorithm is developed for posterior computation. The methods are validated using simulation studies and applied to the problem of ranking medical procedures in terms of the distribution of patient morbidity.

Some Mathematical Results on Three-Way Component Analysis

Jorge Tendeiro, *Faculty of Behavioural and Social Sciences, Heijmans Institute*

Tucker three-way PCA and Candecomp/Parafac are two well-known methods of generalizing principal component analysis to three way data. Candecomp/Parafac yields component matrices that are typically unique up to jointly permuting and rescaling columns. Tucker-3 analysis, on the other hand, has full transformational freedom. That is, the fit does not change when the component matrices are postmultiplied by nonsingular transformation matrices, provided that the inverse transformations are applied to the so-called core array. This freedom of transformation can be used to create a simple structure in the component matrices, and/or in the core array. We address the question of how a core array, or, in fact, any three-way array can be transformed to have a maximum number of zero elements. Direct applications are in Tucker-3 analysis (to facilitate the interpretation of a Tucker-3 solution), in constrained Tucker-3 analysis, and as a mathematical tool to examine rank and generic or typical rank of three-way arrays. Only arrays with symmetric frontal slices are considered.

Classification Techniques for Time Series and Functional Data

David Casado, Andres-M. Alonso, Juan-J. Romo, *Dept of Statistics, Universidad Carlos III de Madrid*
Sara Lopez-Pintado, *Columbia University*

The main aim of this doctoral thesis is to develop classification techniques for dependent and functional data. Methods for classifying time series and functional data are proposed. Although this work involves several type of data, the functional data play a central role. An important point of both classification methodologies is that the original problems are not directly dealt with: the time series problem is rewritten as a functional data problem while the functional data problem is solved using a multivariate technique. Nevertheless, it is worthwhile noticing the different role of the functional data in the two forthcoming proposals: in the time series problem functional, estimators are constructed; while in the functional data problem, curves are the primary data.

For the classification of time series, their integrated periodograms are considered instead of the time series themselves. Subsequently, a new element is assigned to the group minimizing the distance from its integrated periodogram to the group mean of integrated periodograms. Although the periodogram is defined only for stationary time series, the application of the methodology to nonstationary series is still possible by calculating these periodograms locally. Finally, functional data depth is applied to make the classification robust.

The classification of functional data arises naturally in the previous framework. More- over, the problem of selecting the most appropriate form (crude functions, their integrals or their derivatives) to express the data is also suggested. Without loss of generality, this second problem is equivalently formulated in terms of functions and their derivatives of different order, without integrals. In this thesis, a single methodology is proposed to cope with these two tasks at the same time. Following the same criterion of classifying a curve by using the distances from the function or its derivatives to group representative (usually the mean) functions or their derivatives, the combination of those distances is proposed in our method. The proposal works with a multivariate variable defined in terms of the distances. Moreover, an automatic form of ranking the original functions and their derivatives by discriminant power is obtained.

Graphs, Risks, and Clusters: The Work of Bernie Harris

A memorial session celebrating the life and work of Bernard Harris (1926-2011)

Chair: F. James Rohlf, Dept. Ecology & Evolution, Stony Brook University

Adversarial Risk Analysis

David Banks, *Dept of Statistical Science, Duke University*

Bernie Harris was a pioneer in risk analysis. This talk attempts to combine traditional risk analysis with a game-theoretic perspective. Specifically, classical risk analysis has assumed that the opponent is non-adversarial (i.e., “Nature”) and thus is inapplicable to many situations. This work explores Bayesian approaches to adversarial risk analysis, in which each opponent must model the decision process of the other, but there is the opportunity to use human judgment and subjective distributions. The approach is illustrated in the analysis of two important applications: sealed bid auctions and a simple form of poker; some related work on counterbioterrorism is also covered. The results in these three applications are interestingly different from those found from a minimax perspective.

Continuity! To be or not to be! Is that even a question?

Mel Janowitz, *DIMACS, Rutgers*

The input to a clustering problem can sometimes only have ordinal significance. Yet ordinal properties of appropriate cluster algorithms will yield continuous results. If distances between dissimilarities are meaningless, why should a continuous cluster method result? Why even think about continuity? Some reasons for this are discussed. It is shown that single-linkage clustering is the unique cluster method whose image is the set of all ultrametrics, and which is isotone and idempotent: neither property would seem to have any connection with a metric. Yet single-linkage clustering is a continuous cluster method. The link between ordinal and metric conditions is tied to the fact that for a DC $d^{'}$ sufficiently close to a DC d, it follows that $d \preceq d^{'}$ in the sense that $d(a;b) < d(x;y)$ implies $d^{'}(a;b) < d^{'}(x;y)$. It follows that every level clustering of d appears as a level clustering of $d^{'}$. If a sequence (d_n) of dissimilarities converges to d, then for some positive integer N, $n \geq N$ implies that $d \preceq d_n$ for all n. It is argued that continuous cluster methods F should have the additional property that $d \preceq d^{'}$ implies $Fd \preceq Fd^{'}$, and that this rather than continuity is the useful property. Thus if $d^{'}$ is an estimate with small errors of a true dissimilarity d, then if $d \preceq d^{'}$, such an F will hopefully produce a useful estimate of the true clusters Fd from those of $Fd^{'}$. There remains a tantilizing extra condition still to be found, and indications are given as to its nature.

Bernie Harris’ Contributions to Cluster Analysis

Stan Sclove, *Dept of Information and Decision Sciences, UIC*

This talk will describe some of Bernie Harris’ contributions to cluster analysis. The problem of cluster analysis is, given observation vectors on a sample of objects or individuals, group them. Harris focused on the detection of clustering by viewing the situation via the number of edges and cliques in graphs, in particular, via the expected number by chance alone. The talk also includes personal reminiscences, collected over years of interaction at various conferences, esp. the Classification Society (Bernie was an active member for a number of years and organized and hosted one of our meetings) and sessions of the American Statistical Association’s Risk Analysis Section (Bernie was a founding member).

Combining Classification Trees and Regression Models

Chair: TBA

Predicting Variable-Length Functional Outputs for Emulation of a NASA Flight Simulator

Yuning He, *Dept of Applied Mathematics & Statistics, UC Santa Cruz*

The ability to understand and analyze computer simulation models can rely heavily on the ability to approximate the model with a good statistical surrogate. We develop methods for emulating a NASA flight simulator where both of the inputs and outputs are high-dimensional, and the outputs are variables that are functions of time. A challenging, novel aspect of this application is that the length of the output functions, which represent the length of simulated flights, vary with the input. This happens because the flight controller in the simulator sometimes fails to control the flight back to a stable state and fails prematurely. We propose a multi-stage scheme for predicting each output variable. We first classify the inputs according to the characteristic (length) of their corresponding output curves. Then we employ suitable basis representations for each of the output classes, followed by learning of the mappings between inputs and the basis coefficients for each class.

Optimization Under Unknown Constraints

Herbie Lee, *Dept of Applied Mathematics & Statistics, UC Santa Cruz*

Optimization of complex functions, such as the output of computer simulators, is a difficult task that has received much attention in the literature. A less studied problem is that of optimization under unknown constraints, i.e., when the function must be evaluated not only to determine a value for optimization, but also to determine if a constraint has been violated, either for physical or policy reasons. We develop a statistical approach based on Gaussian processes and Bayesian learning to both approximate the unknown function and estimate the probability of meeting the constraints. A new integrated improvement criterion is proposed to recognize that responses from inputs that violate the constraint may still be informative about the function, and thus could potentially be useful in the optimization. The new criterion is illustrated on synthetic data and on a motivating problem from health care policy.

Dynamic Trees for Response Surface Learning and Optimization

Bobby Gramacy, *Dept of Econometrics & Statistics, U of Chicago Booth School of Business*

We introduce a new response surface methodology, dynamic trees, with an application to optimization of (noisy) black box functions under unknown constraints. The benefits of dynamic trees over more traditional models for such applications, like Gaussian processes, is that they natively accommodate non-stationarity and heteroskedasticity in the response, natively deal with categorical predictors and responses, can facilitate input sensitivity analyses. The talk will focus on the the dynamic tree process, and inference by sequential Monte Carlo which is ideal for application to optimization under constraints. It will also highlight an R package implementing the methods, called dynaTree, which is available on CRAN.

A Survey of Some Classification Applications

Chair: TBA

Hypergraphs from Twitter Data

Elizabeth Hohman, *Naval Surface Warfare Center*

We use data from the micro-blogging service Twitter. Public tweets from March, 2011 are used to form hypergraphs, where hyperedges result from words, hashtags, or topics. We explore the use of topic models on the hashtags and associated tweets in order to answer questions about the hashtags, such as how they change in time and whether they have multiple meanings. This is preliminary work and the majority of the presentation will focus on the problem definition and data processing.

Using Multipartite Graphs for Recommendation and Discovery

Michael Kurtz, *Harvard-Smithsonian Center for Astrophysics*

The Smithsonian/NASA Astrophysics Data System exists at the nexus of a dense system of interacting and interlinked information networks. The syntactic and the semantic content of this multipartite graph structure can be combined to provide very specific research recommendations to the scientist/user. We show the process we are currently using in our ADS Labs environment (adslabs.org/ui); and will discuss future changes and improvements.

An Exploratory Taxonomic Study of Womens Pathways to Crime: From Qualitative to Quantitative Patterns

Tim Brennan, *Northpoint Institute*

The classification of womens pathways to crime has been dominated by qualitative research. This has identified several typified female ”pathways” to crime. We report a quantitative taxonomic study of women prisoners (N = 718) assessed on psycho-social, personality and criminal history features and recently developed factors adressing gender specific issues linked to crime. Using several cluster analysis methods a hierarchical classification ranging from 4 to 8 clusters at successive K levels was developed. Cross method convergence was examined using bagged K-Means, Semi-supervised clustering and Wards method. Cluster stability was examined at several K levels using McIntyre-Blashfields validation approach. Class separation was examined using the Minimax Probability Machine. New case assignment was tested using the Support Vector Machine method. Eight female pathways are described and contrasted to the prior qualitative pathways research.

How Can We “Better Use” Classification/Clustering?

Chair: TBA

Effects of Some Design Factors on the Shape of Similarity Indices in Cluster Analysis

Ahmed Albatineh, *Dept of Epidemiology and Biostatistics, Florida International University*

This paper investigates the effects of number of clusters, cluster size, and correction for chance agreement on the shape of two similarity indices, namely, Jaccard (1912) and Rand (1971) indices. Skewness and kurtosis are calculated for the two indices and their corrected forms and compared with those of the normal distribution. Two clustering algorithms are implemented, namely, Ward and K-means algorithms. Data were randomly generated from bivariate normal distribution with a specified mean and variance covariance matrix which has no clustering structure. Two clusterings of the data were obtained using random partitioning and a clustering algorithm by requesting 2, 5, 10, 15,.., 50 clusters. The similarity between the two clusterings is calculated at the same number of clusters. Three way ANOVA is performed to assess significance of the design factors using skewness and kurtosis of the indices as responses. Test statistics for testing skewness and kurtosis are calculated. Effect sizes measured using partial eta square and observed power were obtained. Simulation results showed that independent of the clustering algorithms or the similarity indices used, the interaction effect due to density X number of clusters and the main effect due to density and number of clusters were found significant for skewness and kurtosis all the times. The significance of the correction for chance agreement factor for either skewness or kurtosis was somehow dependent on the index being used. Hence, such design factors must be taken into consideration when studying the shape or distribution of such indices.

Classification Society Automated Search Service

Shuai Sun, *University of Massachusetts, Boston*

The aim of the project is to build a search engine to search for bibliographic Classification Literature, to find additional bibliographic data elements like DOI. Crossref free DOI lookup service has been used to find the DOI for the bibliographic record. An interface has been carefully designed according to usability principles and techniques. This search engine has been developed using the open source technologies like JAVA, SOLR and J2EE. Some of the features include faceted Search, hit highlighting etc., more functionality will be added over time. A demonstration of the work till date is provided.

Data Theory and Scientific Data Management

David Dubin, *Graduate School of Library and Information Science, UIUC*

Practitioners of data-driven methods in the natural sciences are now facing challenges that recall those which prompted social scientists to develop theories of data and measurement in the last century. There are increasing expectations for natural scientists to share their data, and to make data management plans and strategies a regular part of their enterprise. But the goal of expressing data for later discovery and reuse confronts us with puzzling relationships among observations, quantities, and data values. A new field of research data management specialists are already proposing solutions to these puzzles in the form of abstract models that govern descriptive standards for data and the architecture of scientific data repositories. These current efforts are largely uninformed by the theories contributed over the past seventy years from the fields of classification, data analysis and the quantitative social sciences.

Classification Applications in the Biological Sciences

Chair: TBA

Investigating New Typologies for Classifying Aphasia Patients

Li Liu, *Dept of Statistics, Carnegie Mellon University*

Aphasia is a disorder that results from damage to portions of the brain that are responsible for language. The AphasiaBank is an NIH funded project for creating a shared database to study communication in aphasia. Data are contributed from labs at 12 sites. There are two main methods of classifying an aphasic patient to the existing aphasia typology. Clinicians may diagnose a patient through an interview, or patients may be diagnosed by taking a WAB test. In this paper, we analyzed demographic and biometric data from interviews with 77 aphasic patients. Principle component analysis was used to reduce the dimensionality of the data. From the scree plot, we chose to use the first two principle components to classify the patients by different unsupervised learning methods. K-means could only separate the two types with the largest number of observations (Anomic and Broca) into different groups, and was not useful in finding group structure; Model-Based clustering separated data into three groups, one of those groups captures almost all of the Anomics and another group captures all of Globals. These three groups could be viewed as a good overall typology with regards to the datas existing structure.

Object Data Analysis of Taxonomic Trees from Human Microbiome Data

Patricio La Rosa, Elena Deych, Berkley Shands, William Shannon, *Divisions of General Medical Sciences and Biostatistics, Washington University of St Louis School of Medicine*
Yanjiao Zhou, George Weinstock, Erica Sodergren *Genomic Institute, Washington University in St. Louis Medical School*

Human microbiome research uses next generation sequencing to characterize the microbial content from human samples to begin to learn how interactions between bacteria and their human host might impact health. As an emerging medical research area there are few formal methods for designing and analyzing these experiments, with most approaches being ad hoc and applicable to the particular problem being faced. Since microbiome samples can be represented as taxonomic trees, it is natural to consider statistical methods which operate on graphical structures such as tree objects. A unimodal probability model for graph-valued random objects has been derived and applied to several types of graphs (cluster trees, digraphs, and classification trees). In this work we apply this model to HMP taxonomic trees which allows for a fully statistical data analysis. This model allows us to calculate core microbiomes using statistical maximum likelihood estimation, test hypotheses and calculate P values of whether the core microbiomes are the same or different across patient subgroups using likelihood ratio tests. As an example, we apply our methodology to the HMP data on 24 subjects.

Dirichlet-Multinomial Power Calculations And Statistical Tests For Microbiome Data

Patricio S. LaRosa, Elena Deych, Erica Sodergren, George Weinstock, William Shannon, *Divisions of General Medical Sciences and Biostatistics, Washington University of St Louis School of Medicine*
J. Paul Brooks, Edward L. Boone, David J. Edwards, Qin Wang *Virgina Commonwealth University*

We provide a statistical framework to perform formal hypothesis testing, and to calculate power and sample size requirements for human microbiome experiments using taxonomical classification of metagenomic sequences. The methods proposed allow for modeling and comparing statistically the taxa abundance distributions of microbiotas from one and several populations. In particular, we use the Dirichlet-Multinomial (DM) distribution to model taxa counts from a set of samples. The DM model takes into account the variability of the taxa probabilities or relative abundance distribution (RAD) across samples providing a better fit to the data than a Multinomial model. We study the power and size of the following hypothesis tests: Multinomial goodness of fit against a Dirichlet multinomial alternative; one-sample, two-sample, and multiple-sample comparison of RAD means; and two-sample comparison of RAD distributions. More specifically, for a given taxa number, we provide guidelines for computing sample size, namely, the numbers of subjects and number of reads per subject required to obtain a desired statistical power. As an example, we apply our methodology to the HMP data on 24 subjects.