The SAS System

day=1) Thursday June 11th time=8:30-10:00

Presenter_Name	Title	Abstract
David Banks	Statistical Issues in Agent Based Models	Agent Based Models (ABMs) have become extremely popular tools for simulating certain kinds of complex phenomena. They have strong advantages in terms of both programming and validation. However, the statistical properties of the outputs of such models have not received significant theoretical attention. This talk will review the field, and then describe opportunities for research contributions.

day=2) Thursday June 11th time=10:00-12:0

Presenter_Name	Title	Abstract
Hadi Rezazad	Enhancing Computer Network Robustness and Efficiency	It is important to design network configurations with special consideration for their various aspects, such as security, integrity, scalability and cost. It is especially important for a network to be built as robust as possible to protect against failures, attacks and intrusions. In this paper, we develop an innovative method to assess and improve the robustness and efficiency of computer networks. This method uses computer network analysis, social network analysis, evolutionary computing, statistical methods, and graph theory. Specifically, the aim is to achieve optimized network robustness and efficiency with a primary focus on architecture of networks. We develop metrics for measuring the robustness and efficiency elements of networks and construct an evolutionary method for the enhancement of these elements and then apply the method to various networks, including random networks, biased networks and real-life networks.

The SAS System

Presenter_Name	Title	Abstract
Walid K Sharabati	Multi-Mode Social Networks	Social network analysis has become a staple in the analysis of organizations and societies. Most work has focused on single-mode networks and the analysis of their structures and substructures including dyads, triads and cliques. A modest amount of work has been done on two-mode networks and methods for reducing these to one-mode networks. In this paper we develop the mathematical foundations for multi-mode networks. In the two-mode network case, the two-mode network can be decomposed into two one-mode networks. Because there is only one unique way of forming the transpose with a two-dimensional matrix, only two one-mode networks result. Consider a three-mode network. Then the three-dimensional cuboid adjacency matrix is a tensor of rank 3. The transpose of the rank 3 tensor can be done in many different ways so that much richer substructures can be formed that can tease out a wide variety of social behaviors. This paper will discuss mathematics necessary for forming simpler one- and two-mode networks from higher order multi-mode networks and illustrate the theory with some applications.
Yasmin H Said	On Some Styles of Author-Coauthor Networks with Implications on Peer Review	It is desirable that the peer review have three important traits: independence of the referees, unbiased referees, and referees knowledgeable in the field. As any hard-working editor or associate editor knows, finding independent, unbiased, and knowledgeable referees for a paper or proposal is a difficult chore. This is especially true in a rather narrow field where there are not many experts so that issues of independence arise quickly. Clearly as a field becomes increasingly specialized, there are not as many independent experts. Thus finding someone who is both independent and knowledgeable is difficult. In the past, when many more authors adopted a solo style of authorship, finding someone who was not a co-author was relatively easy. Nonetheless, the issue of unbiasedness still was an issue. The introduction of double-blind refereeing focused on the unbiasedness issue. It is only natural for a referee to act in a favorable way towards someone he or she admires scientifically or with whom he or she may have a friendship. Double-blind refereeing, however imperfect it might be, at least removes the perception of referee bias by removing the name of the authors. In an era of Google Scholar, it is usually not very hard to learn the names of the authors of a paper simply by googling the title of the paper. Even if the authors of a paper are unknown, the topic of a paper and its similarity to the work of the referee can bias the referee to look upon the paper in question favorably. In this paper, I examine some implications for peer review by classifying styles of author-coauthor networks.

day=4) Thursday June 11th time=2:30-4:30

Presenter_Name	Title	Abstract
Hernando Ombao	Functional Connectivity As a Potential Biomarker for Classification	In this talk, we will discuss models that use functional connectivity as a potential biomarker for classification. This work is motivated by an experiment where the goal was to study brain network that mediate motion. Participants responded to a visual cue by slightly displacing the joystick from the starting position to either left or right. Here, we build a time-frequency network in the multi-channel EEG signals using the SLEX library and choose the features that best discriminate between rightward and leftward movements. The SLEX library consists of time- localized Fourier waveforms and thus can capture transient oscillatory features in the data. Moreover, our approach estimates the partial coherence between each pair of channels and thus uses the entire brain connectivity as measured by auto-spectra, cross-spectra, coherence and partial coherence for classification.

The SAS System

Presenter_Name	Title	Abstract
Robert Krafty, PhD	Classification of families of locally stationary time series	Existing methods in non-stationary time series classification assume time series from different units within a population are generated by the same underlying stochastic process characterized by a time-varying second order spectrum, and both the between-time series variability and the within-time series variability are results of the same underlying stochastic process. This is usually not true in real applications and can lead to misclassification. In this talk, we propose a model for a family of time series by imposing a hierarchical structure on their log-spectra. This model assumes that while a family of time series share some similarity characterized by the population-average spectrum, each time series has its own characteristics modeled by the unit-specific deviation in terms of its log-spectrum. We then propose nonparametric methods to estimate the population-average log-spectrum and the between-unit variance function. We develop a quadratic rule for discriminating between different populations based on the estimated mean log-spectra and the variance functions. A simulation study is presented to empirically demonstrate the benefits of accounting for the between-time-series variability and the proposed procedure is used to discriminate pre-seizure EEG time series from non-seizure baseline data.
Wesley Thompson, PhD	First biologically-relevant markers for mood disorder diagnosis	Psychiatric diagnosis is hampered by the lack of biological markers reflecting disease mechanisms. Bipolar disorder (BD) , a psychiatric illness characterized by severe mood dysregulation, is one of the most debilitating of all illnesses. The absence of biologically-relevant diagnostic markers of BD leads to frequent misdiagnosis as unipolar depression, inadequate treatment and high suicide rate for BD sufferers, and huge societal cost. We show that anatomic and functional connectivity in three brain regions supporting mood regulation and sensory processing accurately discriminate BD from mentally-healthy controls and BD from unipolar depression, and provide the first evidence that biological markers reflecting pathophysiological disease mechanisms can accurately differentiate these smood disorders.

day=5) Friday June 12th time=8:30-10:00

Presenter_Name	Title	Abstract
Ed Wegman	Document Clustering and Social Networks	Text Mining has become a specialized offshoot of Data Mining, Information Retrieval, and Natural Language Processing. One of the major tools of this area is the vector space representation of documents. On the other hand, social network analysis has found its mathematical underpinnings primarily in mathematical graph theory. A graph has a dual representation as an adjacency matrix. So-called two-mode social networks have actors of two different types, frequently individuals and organizations. The adjacency matrix for these two-mode social networks has the same structure as the so-called term-document matrices used in text mining. In the talk we discuss these connections and show how these ideas can be exploited in both fields. In particular, methods for block modeling in social network analysis can be used for document clustering.

The SAS System

day=6) Friday June 12th time=10:00-12:0

Presenter_Name	Title	Abstract
Lyn Hunt	The Multimix Class of Mixture Models: Model Selection	The finite mixture model approach to clustering is a model based approach that requires the specification of both the number of components to be fitted to the model and the form of the component distributions. The Multimix class of mixtures was proposed by Hunt and Jorgensen (1996, 1999) and allows the clustering of data containing both categorical and continuous attributes. In this talk, we illustrate the performance of some commonly used model selection in selecting both the number of components in the model and the form of the component distributions when using the Multimix model to cluster data containing mixed categorical and continuous attributes. Hunt, L. A., and Jorgensen, M. A., (1999). Mixture Model Clustering Using the Multimix Program. Austral. & New Zealand J. Statist. 41, 153-171. Jorgensen, M. A., and Hunt, L. A., (1996). Mixture Model Clustering of Data Sets with Categorical andContinuous Variables. In Proceedings of the Conference on Information, Statistics and Induction in Science, Melbourne, 1996, 375-384.
Seung Hyun Baek	Hybridized Support Vector Machine Recursive Feature Elimination with Information Complexity	In statistical data mining research, datasets are nonlinear and at the same time high dimensional. It has become difficult to analyze such datasets in a comprehensive manner using traditional statistical methodologies. In this paper, we develop and introduce a novel wrapper method based on a hybridized recursive feature elimination technique (HSVM-RFE) in the kernel-based support vector machines (SVMs) to classify high dimensional data sets and to carry out subset selection of the variables in the original data space to determine the best subset of variables which are discriminating between the groups. Recursive feature elimination (RFE) ranks variables based on information complexity (ICOMP) criterion of Bozdogan. ICOMP plays an important role in not only choosing the optimal kernel function from a portfolio of many other kernel functions, but also, in selecting important subset(s) of variables as an effective measure of fitness function. We compare two different support vector machines based on the RFE technique with different measures (i.e., weight and gradient) for variable rankings. We illustrate the potential and the flexibility of our approach on two real benchmark data sets, one on Aorta data in early detection of atheroma (heart attack), and the other one is on Ionosphere data. We compare our approach with other recursive feature elimination techniques.

The SAS System

Presenter_Name	Title	Abstract
John L. Eltinge	Classification and Clustering of Complex Survey Data	"This paper provides an overview of some open questions in the application of tree-based classification methods to data collected through a complex sample design. First, we review some standard issues in the general analysis of complex survey data, including distinctions among sources of variability related to the sample design and the underlying superpopulation model; specific features of a complex sample design, including stratification, multistage selection and unequal probabilities of selection; and inferential methods intended to account for the abovementioned features. Second, we consider some standard classification-tree methods and extensions of these methods that account for distributions induced by a complex sample design; and highlight open questions regarding properties of these extensions. Third, we discuss subsampling designs for construction of learning and testing samples when the original sample units are selected through a complex design. Fourth, we outline five classes of potential practical applications of classification-tree methods to survey data.
Joseph Schafer	Causal Modeling When the Treatment is a Latent Class	In the potential-outcomes approach to causal inference, causal effects are differences among outcomes that would be realized if different treatments were applied to the same individual. The treatment is usually assumed to be a binary variable measured without error. In many settings, however, the treatment is measured imperfectly by multiple questionnaire items, and failure to account for measurement error may bias the estimated effects. We present models in which the treatment is a latent class. Treatment propensities and potential outcomes are regressed on a rich set of confounders and prognostic variables. After estimating parameters by EM, we estimate average causal effects within each treatment group by solving expected estimating equations. We apply our model to estimate the effects of naturalistic weight-loss strategies on body-mass index among adolescent girls, taking into account the complex survey design.

The SAS System

Presenter_Name	Title	Abstract
Michael Larsen	Latent Class Analysis of Bioeconomy Survey Data	Economic and political factors have led the United States, and particularly the state of Iowa, to a renewed interest in biorenewable fuels. Subsequent analytic work and media reports, however, have countered the initial enthusiasm. A survey was conducted in 2008 to assess Iowans? views on energy policy alternatives and local biofuel initiatives. Although information exists on preferences of agricultural producers and industry professionals, little research has assessed consumer viewpoints. We hypothesized that issues of energy policy and the bioeconomy would be more salient in communities hosting biofuel plants. Our objectives were to assess (1) knowledge and policy opinions regarding the bioeconomy and (2) the impact of proximity (and other individual and community characteristics) on local support for the biofuels industry. A stratified random sample telephone survey of 378 adults living in four Iowa counties was conducted in early 2008. The four counties included one metro and one nonmetro county with a biofuel plant, and one metro and one nonmetro county without a biofuel plant. The current work presented in this paper uses latent class analysis to identify domains within the sampled population and compare responses across identified latent classes.
Marcus Berzofsky	Survey Classification Error Analysis: Critical Assumptions and Model Robustness	Survey statisticians have recognized the need to quantify the classification error on key survey items. Two historical approaches for estimating the classification error are the gold standard approach (Bross, 1954; Tennenbein, 1972) and reliability analysis (Hansen, Hurwitz, and Pritzker, 1964). While each of these methods can be useful, they each have limitations that make them difficult to use in some survey settings. The gold standard requires that true response value for an item be known for each survey respondent which is often impractical and reliability analysis is difficult to interpret when a categorical outcome is being measured. A third approach now being used by survey statisticians is latent class analysis (LCA) based on work done by Lazarsfeld & Henry (1968). This approach is designed for categorical outcomes and does not require that a gold standard measurement be known. LCA is a modeling technique that uses the EM algorithm to estimate the false negative rate, the false positive rate, and the true prevalence for the item of interest. However, LCA does require that four critical assumptions be met: univocality, local independence, group homogeneity, and a simple random sample. These assumptions are often difficult to achieve in a survey setting. This paper will describe each assumption and, through simulation, demonstrate the marginal impact that a violation has on the classification error rates and how robust the model is to the failure of a model assumption. This paper will focus on surveys designed to measure a topic sensitive in nature like marijuana use or sexual victimization. Outcomes such as these often have negligible or zero false positive rates. In other words, for sensitive items it is very

The SAS System

day=8) Friday June 12th time=2:30-4:30

Presenter_Name	Title	Abstract
David Friendenberg	Bounding the Rate of False Clusters in Telescope Images	New technologies like next-generation telescopes and high resolution fMRI scanners produce massive images, in which a diverse collection of objects are embedded in a noisy background. To extract meaningful information from these images we must separate and catalog the clusters of activity (eg galaxies, regions of brain activity) from noise caused by the detector. We have developed adaptive multiple testing procedures that control the proportion of false clusters, by examining different level sets of the image. Furthermore we present multi-scale transformations that enhance the clusters in the image making them easier to detect. We demonstrate our techniques on data from the Chandra X-ray observatory satellite and show that we can simultaneously make a probabilistic guarantee about the rate of false clusters while detecting objects with power comparable to the standard techniques used by astronomers that do not have such rigorous error control.
Joseph Richards	Efficient and Accurate Inference for High-Dimensional Astrophysical Data	As the number of astronomical sources observed by astronomical surveys computationally efficient methods that can draw accurate statistical inferences about these objects. In Richards et al.(2009), we introduced the diffusion map method to the astrophysical community as a computationally efficient method to parametrize and draw inferences from high-dimensional astrophysical data. Diffusion map is a non-linear technique that has proved to be effective for analyzing data that are complicated, high-dimensional, and noisy. The novelty of the method is that it uncovers a simple, natural representation of the data based on local interactions of data points. Statistical inferences made in diffusion space tend to be more accurate than inferences made in the original (high-dimensional) data space. Recently, we have used diffusion map to pursue a suite of astrophysical problems involving large databases of complicated, current projects: galaxy population synthesis using Sloan Digital Sky Survey (SDSS) spectra and quasar classification using SDSS photometry data. In this first project, the goal is to estimate the star formation history (SFH) of each galaxy in a database by fitting its spectrum. We show that using a basis of SSP prototypes selected by diffusion course-graining yields better SFH estimates than the SSP bases used in the literature and bases derived from other methods. Second, we show how to exploit the diffusion map parametrization of over a million SDSS objects to classify quasars from their SDSS photometry. Here, the ultimate goal is to create a quasar catalog that is both complete and efficient. We compare our levels of completeness and efficiency to values in the literature.

The SAS System

day=9) Saturday June 13th time=8:30 - 10:

Presenter_Name	Title	Abstract
Avory Bryant	Performing Scientometric Analysis through Document Clustering and Dynamic Graph Visualization	Scientometrics is performed by the analysis of the open source scientific literature in an attempt to analyze science. Bibliographic databases provide access to this scientific literature in the form of millions of publications from journals and conference proceedings amongst other resources. Document clustering refers to clustering based on free text content-based features such as a publications title or abstract. These features can be used to represent publications in the vector space model by a term-document matrix which clustering methodologies can be applied to. This presentation focuses on the 2-D graph visualization of these clustering solutions using two techniques. The first technique being a specified graph layout obtained by multi-dimensional scaling and the other a force directed graph layout obtained using distances in the ambient space. Nodes represent clusters while edges represent some relationship between the documents in clusters li ke overlapping citations or overlapping institution affiliations. Node color or size can also be used to highlight cluster specific features such as the number of documents in a cluster or the average growth rate, by publication year, of the documents belonging to a cluster. Note the focus of this work is not on document clustering or 2-D visualization of high dimensional data but on performing scientometrics analysis at the document cluster level using graph visualization. Using this dynamic graph visualization (node positions being static) scheme we hope to create a system that can be used to take advantage of the feature rich environment provided by the open source scientific literature.

The SAS System

Presenter_Name	Title	Abstract
Elizabeth Hohman	Enron Hypergraphs, Communication and Text	We represent the Enron email dataset as a time series of hypergraphs where each employee is a vertex and each email is a hyperedge containing the employees who sent or received the email. This encodes the email communications but does not use the text of the email. We investigate several methods for including the email text in the representation. First, we multiply the hypergraph incidence matrix by a matrix of word count histograms for each email. This is equivalent to representing each employee as the average of the word count vectors for emails he or she has seen. Next we investigate a novel random hypergraph projection method which embeds both vertices and hyperedges into the same space. We look at the same projection method on the term-document matrix to project both emails and words to the same space. We measure the distance between employees in the projections over subsequent weeks and use this to detect changes in the hypergraph structure.
Kristin Yancey	A Dissimilarity Approach to Detecting Out-of-Language Documents	Given a language model consisting of terms (frequent words, n-grams, etc.), we use statistical methods to select the optimal dissimilarity measure from a set of standard measures such as the Kullback-Leibler and rank-order distances. The dissimilarity measure can then be used to compute how similar documents are to the given language model, i.e. whether they are in the language represented by the given language model. This capability will be used to find anomalous or out-of-language documents within a corpus purporting to be in a single language. To perform the analysis, we use a Zipf distribution to create canonical document models from the language model and compare the distribution of their dissimilarities with that of the actual document models. We use a selection of the smaller Wikipediae (such as Zulu and Swati) as our test corpora.

day=OPEN time=OPEN

Presenter_Name	Title	Abstract
Adam Kehoe	The Anatomy of Topical Bursts: Understanding Patterns in Biomedical Pubishing	This project investigates a potential model for scientific publications concerning highly popular and novel topics in Medline. This model employs a statistical model to decompose topical bursts into multiple sub-bursts, based on specific subject areas. Analyzing series of sub-bursts offers insight into the adoption of popular topics as they traverse disparate fields. The model is motivated by an over-arching interest in identifying structural patterns in scientific communication around high- impact discoveries. This project also explores the potential utility of such a model for literature based discovery in translational medical research.

Presenter_Name	Title	Abstract
Bernard Harris	Graph theoretical methods for cluster verification	To motivate this investigation, consider the following scenario. If n observations of random data in the unit square are plotted, frequently one may conclude that the data show several clusters of values. A natural question is whether these apparent clusters are "real" or simply a consequence of random variation. In some simulation experiments conducted by the author and his associates, one may easily conclude that clusters are plots of random numbers. The results discussed in this report provide some techniques for deciding the above question. Realizations of n independent, identically distributed k-dimensional random vecctors are given. A metric ? and a threshold parameter, ?>0 are selected. The realizations are to be interpreted as vertices of a graph and two vertices are adjacent if the distance (?) between them is less than ?. Subject to some regularity assumptions on the distribution, the distributions of complete subgraphs, degrees of vertices and isolated vertices are determined. The asymptotic behavior of these distributions is studied. The efficacy of these various criteria is investigated. These methods are also useful in the study of mixtures. Some preliminary ideas about applying such methods to data taking values in Hilbert spaces are also discussed.
Bill Shannon	Microarray Dimension Reduction Based on Maximizing Mantel Correlation Coefficients Using a Genetic Algorithm Search Strategy	We present the GA-Mantel algorithm to find in high dimensional microarray data a subset of genes that captures relevant spatial relationships among the samples, in order to reduce the data for further analysis and eventually to identify meaningful biological markers. GA-Mantel uses a genetic algorithm to search over possible probe subsets using the Mantel correlation as the scoring measure for assessing the quality of any given probe subset, and consensus methods for selecting the final list of important genes. GA-Mantel is evaluated on both artificial data sets and on experimental microarray data taken from leukemia patients. Current results indicate the GA-Mantel method exhibits promise as a way of efficiently identifying information-rich gene subsets in large data sets while avoiding the curse of dimensionality.
Brent Fegley	The Puzzle of English-language Poetry: Discerning Stylistic Variation by Computational Means	What differentiates one poem from another? What makes a poet's oeuvre cohere? Can a computational approach to poetic analysis mimic human- grade discriminatory capabilities? The current study explicates some of the challenges and successes encountered during exploration of a body of 19th- and 20th-century American poetry, with the poetry of Langston Hughes at its center. Poetry offers interesting ground for study. Its complexity is not simply syntactic, but auditory and visual as well; and its parsing demands macro and micro levels of analysis. The variability and expressiveness of the English language, its rhythm and rhyme, are among the greatest computational challenges; but some statistical measures can be illuminating. This research will show that under the right conditions and from various perspectives, it is possible to discern someone's poetic style computationally.

The SAS System

Presenter_Name	Title	Abstract
Chandler Armstrong	Evolution of the Python Programming Language: Impact of an Online Discussion Forum	This project sought to study the importance of mailing list discussions for implementations in the Python programming language. A dataset was constructed by characterizing each discussion thread by many generic features (such as measures of tree depth/breadth and response times) and independently classifying it as important vs. not important (to language development). This dataset was used to fit a statistical model that attempts to distinguish important threads (from unimportant ones) using these generic features.
Dan Steinberg	Interaction Detection With TreeNet Boosted Tree Ensembles	Recent advances in machine learning technology make it possible to determine definitively whether or not interactions of any degree need to be included in a predictive model. We can thus establish conclusively, for example, for a given set of predictors, that an additive model (one with no interactions) cannot be improved upon with interactions. Or alternatively, one might prove that a model with interactions will outperform a model without them. Further, we can now identify precisely which interactions are supported by the data, and also the degree of interaction, even in very high dimensional data. The tools we use to acheive these results are extensions of Stanford University Professor Jerome Friedman's TreeNet, developed by the authors and embedded in the Salford Systems TreeNet 2.0 Pro Ex product. We illustrate the concepts in the context of a real world regression model where we are quickly able to identify all the important interactions with a modest number of boosted tree ensemble models.
Daniel R. Lawrence	A Forced-Classification Analysis of Ordered-Choice Data Subject to a Polychotomous Criterion Item	Forced classification is a procedure of dual scaling that enables the investigator to emphasize or "focus on" a specific item or collection of items in the analysis. It has been applied to bothincidence and dominance categorical data. If the investigator wishes to emphasize an incidence-type item (e.g., a demographic variable of some sort) in a forced classification, this is a fairly simple matter assuming the categorical data being analyzed are also incidence. In the case of dominance data, though, the incidence "criterion item" must somehow be subject to both dichotomous and the more general polychotomous criterion item, but for ordered-choice data (also known as successive-categories or Likert data), forced classification analyses can accommodate only dichotomous criterion items at present.This talk will introduce a procedure for doinga forced classification subject to any polychotomous criterion item.

The SAS System

Presenter_Name	Title	Abstract
Fionn Murtagh	Ultrametric Wavelet Regression of Multivariate Time Series: Application to Colombian Conflict Analysis	We first pursue the study of how hierarchy provides a well-adapted tool for the analysis of change. Then, using a time sequence-constrained hierarchical clustering, we develop the practical aspects of a new approach to wavelet regression. This provides a new way to link hierarchical relationships in a multivariate time series data set with external signals. Violence data from the Colombian conflict in the years 1990 to 2004 is used throughout. We conclude with some proposals for further study on the relationship between social violence and market forces, viz. between the Colombian conflict and the US narcotics market.
Geoff McLachlan	Mixture Models in Multiple Hypothesis Testing	There are many important problems these days where consideration has to be given to carrying out hundreds or even thousands of hypothesis testing problems at the same time. For example, in forming classifiers on the basis of high-dimensional data, the aim might be to select a small subset of useful variables for the prediction problem at hand. In the field of bioinformatics, there are many examples where a large number of hypotheses have to be tested simultaneously tested. For example, a common problem there is the detection of genes that are differentially expressed in a given number of classes. The problem of testing many hypotheses at the same time can be expressed in a two-component mixture framework, using an empirical Bayes approach; see, for example, Efron (2004). In this framework, we extend the results of McLachlan et al. (2006) on the adoption of normal mixture models to provide a parametric approach to the estimation of the so-called local false discovery rate. The latter can be viewed as the posterior probability that a given null hypothesis does hold. With this approach, not only can the global false discovery rate be controlled, but also the implied probability of a false negative can be assessed. The methodology is demonstrated on some problems in bioinformatics.
Karen Wickett	Examining Structural Variation in Journal Articles	This talk will demonstrate the use of a representation of textual documents as mathematical graphs to characterize the structural variability across a large body of scientific articles. Each graph represents the structure of an individual document and is generated from the XML supplied by the PubMed Central Open Access Subset. The analysis will be based on a pairwise similarity measure between graphs that captures the relevant kinds of structural variability, such as variations in the levels of declared sections and subsections, or the distribution of references throughout the text. The similarity measure may be used to create clusters of documents to complement topic-based groupings of documents and could be correlated to external measures of document use or impact.

The SAS System

Presenter_Name	Title	Abstract
Lynette A Hunt	Development of a Model to Predict Resolution of Diabetes Following Gastric Bypass Surgery	Obesity and Type 2 diabetes are being described as a worldwide epidemic. They are closely linked and are a particular problem in the developed countries. For the severely obese, life expectancy and quality of life are reduced as a result of the associated co-morbidities. These include conditions such as diabetes, hypertension, dyslipidaemia, heart disease, asthma, restrictive lung disease, obstructive sleep apnoea and weight bearing joint problems. For society, the loss of productivity and the enormous cost of health care have important implications for future economic wellbeing. A variety of surgical procedures have been devised for addressing severe obesity and these have become increasingly employed and accepted the World over. The gastric bypass operation is regarded as the gold standard of these procedures and has come to be recognised as having major metabolic benefits for patients. For example, it is now known that 75-85% of all severely obese Type 2 diabetics undergoing gastric bypass will have permanent resolution of their diabetes, and usually as early as 6 days after the surgery. The mechanism by which this occurs is a matter of intense international interest at the present time, as its discovery may herald a paradigm shift in the management of Type 2 diabetes and other metabolic diseases. This paper investigates the predictors of this resolution and explores the development of a model to predict the outcome (ie resolution or not) for any one individual prior to surgery.
Stephen L. France	Is the distance compression effect overstated? Some theory and experimentation	Previous experimental work in the document clustering literature has shown that the Minkowski-p distance metrics are unsuitable for clustering very high dimensional document data. This unsuitability is put down to the effect of "compression" of the distances created using the Minkowski-p metrics on high dimensional data. Previous work on distance compression has generally used the performance of clustering algorithms on distances created by the different distance metrics as a proxy for the quality of the distance representations created by those metrics. In order to separate out the effects of distances from the performance of the clustering algorithms we tested the homogeneity of the latent classes with respect to item neighborhoods rather than testing the homogeneity of clustering solutions with respect to latent classes. We show the theoretical relationships between the cosine, correlation, and Euclidean metrics. We propose that some of the performance difference between the cosine and correlation metrics and the Minkowski-p metrics is due to the inbuilt normalization of the cosine and correlation metrics. The normalization effect decreases with increasing dimensionality and the distance compression effect increases with increasing dimensionality. For document datasets with dimensionality up to 20,000, the normalization effect dominates the distance compression effect. We propose a methodology for measuring the relative normalization and distance compression effects.

The SAS System

Presenter_Name	Title	Abstract
Vetle Torvik	Text and Document Modeling	This is a proposal for a session of four graduate student speakers, abstracts to follow in separate submissions: Examining Structural Variation in Journal Articles. Speaker: Karen Wickett (wickett2@illinois.edu) The Anatomy of Topical Bursts: Understanding Patterns in Biomedical Publishing Speaker: Adam Kehoe (kehoe2@illinois.edu) The Puzzle of English-language Poetry: Discerning Stylistic Variation by Computational Means Speaker: Brent Fegley (fegley1@illinois.edu) Evolution of the Python Programming Language: Impact of an Online Discussion Forum Speaker: Chandler Armstrong (carmstr3@illinois.edu)
Carolyn J. Anderson	Simultaneous Estimation of Multinomial Logistic Regression Models: Factor Analysis of Polytomous Item Response Data with Covariates	The approach described here starts with Bock's (1972) nominal response model (NRM). The NRM is a multinomial logistic regression model for responses to items where the ordering of response options is not known a priori and the predictor or explanatory variable is unobserved (i.e., the latent construct). The latent variable in the multinomial logistic regression is replaced with an estimate based on responses to all other items. Given a set of items, there is one multinomial logistic regression for each. The problem then becomes one of simultaneously estimating multinomial logistic regressions with restrictions across the models. This approach allows us to go beyond what is typical in standard item response theory modeling in that we can handle multiple correlated latent constructs, impose (linear and/or ordinal) restrictions on category scores, test the effect of additional variables (e.g., a treatment versus control condition), create hybrid IRT models, and obtain measures on the latent constructs. An extension of the approach will also be described that is akin to a structural equation model for observed discrete data.
Carolyn J. Anderson	Simultaneous Estimation of Multinomial Logistic Regression Models: Factor Analysis of Polytomous Item Response Data with Covariates	The approach described in this talk starts with Bock's (1972) nominal response model (NRM). The NRM is a multinomial logistic regression model for responses to items where the ordering of response options is not known a priori and the predictor or explanatory variable is unobserved (i.e., the latent construct). The latent variable in the multinomial logistic regression is replaced with an estimate based on responses to all other items. Given a set of items, there is one multinomial logistic regression for each. The problem then becomes one of simultaneously estimating multinomial logistic regressions with restrictions across the models. This approach allows us to go beyond what is typical in standard item response theory modeling in that we can handle multiple correlated latent constructs, impose (linear and/or ordinal) restrictions on category scores, test the effect of additional variables (e.g., a treatment versus control condition), create hybrid IRT models, and obtain measures on the latent constructs. An extension of the approach will also be described that is akin to a structural equation model for observed discrete data.

The SAS System

Presenter_Name	Title	Abstract
Chia-Yi Chiu	Algorithms for Finding the Number of Clusters and Labeling Clusters in Cognitive Diagnosis	An asymptotic theory for clustering examinees based on their cognition profiles using cluster analysis has been developed and examined (Chiu, Douglas, & Li, 2009). The theory verifies that given a particular cognitive diagnosis model and a long test, K-means and hierarchical agglomerative cluster analysis with a sum-scores statistic as input are efficient and consistent alternatives, and perform nearly as well as the model-based method. However, cluster analysis does not directly determine the number of clusters or label the clusters, which, from the cognitive diagnosis perspective, are substantial in order to finely understand examinees' cognitive strengths and weaknesses. This paper presents methods to handle the problems. The number of clusters is chosen based on a scree plot of the fusion coefficients, the distances between clusters that are joined in each step of a HACA analysis. Regarding the issue of labeling clusters, an objective consistency index of the clusters with respect to underlying attribute patterns is proposed as a criterion to exhaustively search for all possible labeling ways. Empirical results from simulation studies are also provided in the paper.
Elilzabeth Ayers	Conditional Subspace Clustering of Skill Mastery Information	In educational research, a fundamental goal is identifying which skills students have mastered, which skills they have not, and which skills they are in the process of mastering. As the number of examinees, items, and skills increases, inference about skills in even simple cognitive diagnosis models becomes difficult. We adopt a faster, simpler approach: cluster a capability matrix representing each student's knowledge for each individual skill to generate skill set profile clusters of students. We further extend this approach with the introduction of an automatic subspace clustering method that identifies skills on which students are well-separated and then use clustering results from (possibly much) smaller conditional subspaces. We demonstrate the feasibility and scalability of our method on several simulated datasets and illustrate the difficulties inherent in real data using a subset of online mathematics tutor data. In addition, we explore the effect of the Q-matrix design on the performance of the methods and discuss some alternative definitions of the capability matrix. We also comment on computational advantages of our approach and show that it gives faster, more reasonable results.

The SAS System

Presenter_Name	Title	Abstract
Hua-Hua Chang	Building an effective cognitive diagnostic CAT for students' Classifications	It should be noted that a new trend in psychometric research is to classify students' mastery levels for a given set of attributes the test is designed to measure, where an attribute is a task, subtask, cognitive process, or skill involved in answering an item. The objective of the research is to develop a Computerized Adaptive Testing (CAT) system mainly to meet the needs of classification. Our goal is to build an effective item selection algorithm which incorporates the function of cognitive diagnoses. The utility of the CAT system will enhance assessment because it also provides students and their teachers with useful diagnostic information in addition to the single "overall" score. A field test will be administrated in China for about 120,000 Grade 4 and Grade 9 math and English in May 2009. The results of the filed test will be reported at the conference.
Jeff Douglas	Performance of Several Distance Measures for Clustering Data Arising from Some Common Models in Educational Testing	In this study, we explore the behavior of cluster analysis under different distance measures, using some of the most common models in educational testing for data generation. Theoretical results on clustering accuracy are given for distance measures used in hierarchical agglomerative cluster analysis for data from unidimensional item response models, multidimensional item response models and restricted latent class models. An aim is to identify distance measures that work well for a variety of models, and provide theoretical justifications for using them. Potential applications in educational testing are discussed.
Jonathan Templin	Classification-based Psychological Measurement via Confirmatory Mixture	In recent years, the field of psychometrics has seen an increased research focus on classification-based procedures for evaluating the breadth of abilities a person may possess. Predominantly used in the measurement of cognitive abilities, such models have many names, including diagnostic classification models or cognitive diagnosis models. As latent variable models that utilize categorical latent variables to classify individuals, diagnostic models link the diagnostic status of a person with item responses of a test. The use of diagnostic models can provide highly informative feedback as the current diagnostic status of a person, feedback which may streamline the process of remediation in educational contexts. In this talk, I link existing diagnostic classification models with more commonly used psychometric methods, highlighting the statistical properties of diagnostic models. I further discuss the types of situations where diagnostic models are most eff ectively applied.

The SAS System

Presenter_Name	Title	Abstract
Chia-Yi Chiu	Cluster Analysis for Cognitive Diagnosis: Theory and Applications	Latent class models for cognitive diagnosis often begin with specification of a matrix that indicates which attributes or skills are needed for each item. Then by imposing restrictions that take this into account, along with a theory governing how subjects interact with items, parametric formulations of item response functions are derived and fitted. Cluster analysis provides an alternative approach that does not require specifying an item response model, but does require an item-by-attribute matrix. After summarizing the data with a particular vector of sum-scores, K-means cluster analysis or hierarchical agglomerative cluster analysis can be applied with the purpose of clustering subjects who possess the same skills. Asymptotic classification accuracy results are presented, along with simulations comparing effects of test length and method of clustering. An application to a language examination is provided to illustrate how the methods can be implemented in practice.
Daniel M Rice	Reduced Error Logistic Regression	Reduced Error Logistic Regression (RELR) is really an answer to the often raised question as to why there are no error terms in Logistic Regression, as the RELR formulation utilizes error terms. RELR is fundamentally different from arbitrary regression coefficient smoothing methods like Lasso and Penalized Logistic Regression, as it models non-arbitrary estimates of logit error consistent with known Extreme Value error properties in logistic regression. A very surprising algebraic result that follows directly from this logit error formulation is that RELR offers a solution to the dual curses of dimensionality and complexity. Given high dimensional data, RELR can rapidly perform all traditional stages in regression model building, including parsimonious and reliable variable selection, as a single stage optimization process with no arbitrary parameters. Because it is a single stage optimization process without arbitrary analyst decisions, RELR represents significant savings in time and complexity to build a parsimonious regression model given high dimensional data. Based upon a high dimensional and multicollinear dataset, we provide evidence that RELR models can have substantially reduced error in validation sample outcome predictions compared to standard methods including Penalized Logistic Regression, Neural Networks, Decision Tree, Stepwise Select Logistic Regression, Full Standard Logistic Regression, Support Vector Machines, and Partial Least Squares. RELR does not use validation sample information in training, so its substantially reduced error compared to Penalized Logistic Regression in both logit coefficients and outcome predictions is especially significant.