40th Symposium on the Interface: Computing Science and Statistics

INTERFACE 2008

RISK : REALITY

Durham Marriott, Durham, NC
21 - 24 May 2008

Sponsored by:
Interface Foundation of North America, Inc.

Hosted by:
National Institute of Statistical Sciences

Cooperating Organizations:
ASA, CSNA, ENAR, IASC, IMS, INFORMS, SIAM, WNAR

This symposium is a long-standing forum focusing on the interface between computing science and statistics.

Welcome

The Interface Foundation of North America welcomes you to the 40th Interface Symposium, the premier annual conference on the interface of computing science and statistics. The Foundation is a non-profit educational corporation founded in 1987 to sponsor the symposium and to publish the proceedings. For further information about IFNA, visit our website at: http://www.galaxy.gmu.edu/stats/IFNA.html

The theme of Interface 2008 is RISK : REALITY from *Information* to *Inference* and from *Technical Developments* to *Specific Contexts*.

Through the dual focus on the problems of information extraction, risk modeling, analysis and decision-making and on the computational technology and advances in tools to make characterization, quantization and evaluation of risk possible, this symposium will address issues central to understanding real risks and to conceptualizing potential risks and potential risks and risk management scenarios.

The Keynote Speaker is **Siddhartha Dalal**, Senior Advisor to the President for Technology at Rand Corporation. He brings an extraordinary background of theory, application and vision for statistics, statistical computing and information technology to hard problems modeling reality, developing technology for prediction and setting policy.

Contact Information

Conference Chairs:

Dr. Alan Karr, Director, NISS	karr@niss.org
Dr. Nell Sedransk, Associate Director, NISS	sedransk@niss.org

Phone: (919) 685-9300 Fax: (919) 685-9310

Mail: Interface Foundation of North America, Inc.
P.O. Box 7460
Fairfax Station, VA 22039-7460

Conference Program

Invited sessions have been organized by the Program Committee and were designed to cover broadly areas of research related to the theme. Contributed paper sessions address some of the highlighted topics for this year and also report advances in computational statistics / computation sciences / statistical computation / statistical graphics. See the full program schedule for details and the abstracts.

Registration and Financial Support

Interface Registration onsite (including the Interface banquet) is $350 for regular Interface members; $465 for non-members; and $150 for students. Members of cooperating societies can register at a discount: $400. Single-day registration (not including the banquet) is $175. Guest banquet tickets may be purchased for $65 each. Short course registration is $165 for members; the student rate is $75. In order to enroll in the short course on "OMICS," full Interface registration is required.

Registration for either related conference on 21 May 2008 can be done onsite and must be made separately.

Conference Location – North Carolina Triangle

The Interface 2008 conference hotel is the **Durham Marriott at the Civic Center** located at 201 Foster Street in Durham, North Carolina 27701. The Raleigh-Durham Airport (RDU) serves the Triangle Area with domestic and international flights. Ground service to Durham (16 miles) is available via taxi or direct shuttle service to the Durham Marriott.

The Triangle area includes the old communities of Durham, Chapel Hill and Raleigh with their many universities and colleges, particularly Duke University, the University of North Carolina and North Carolina State University respectively. Situated in their midst is Research Triangle Park, established as a joint venture and home to the National Institute of Statistical Sciences and the Statistical and Applied Mathematical Sciences Institute.

Interface 2008   Symposium Theme

RISK : REALITY

Everywhere today RISK is assessed, projected, analyzed, managed. Decisions based on perceived RISK are made daily in every economic sector and, indeed each aspect of life. The pharmaceutical industry continually deals with risk of rare and unexpected side effects of marketed drugs. Homeowners and federal agencies like NOAA deal with the risk of Categories IV and V hurricanes. Security agencies deal with risk of terrorists crossing borders into the US. Federal agencies and companies risk unauthorized penetration of classified and/or proprietary databases. Nuclear power plants must manage risk of radiation escape; while other energy resource managers evaluate strategies to manage risk to the power grid. The focus of Interface 2008 will encompass all these aspects.

Understanding RISK depends crucially on *Information* –

- First - *Acquiring Information*:
  - data acquisition
  - data bases
  - expert opinion
- Second - *Extracting Knowledge* using computational/statistical tools for:
  - massive data / data bases
  - high-dimensional data
  - streaming data
  - text
- Third - *Drawing Inference*:
  - modeling complex systems
  - dynamic analytic processes
  - verification and validation of models and algorithms
  - decision theory

Managing RISK depends on understanding specific *Contexts* –

- extreme events
- natural disasters
- infrastructure (e.g., energy)
- communications & internet
- portfolios
- adversaries & human threats

Program Committee

David Banks (Duke), **Dipak Dey** (U Connecticut), **David Dunson** (NIEHS**),** **Lutz Edler** (German Cancer Research Center, Heidelberg), **Alan Gelfand** (Duke), **Karen Kafadar** (U Indiana), **David Marchette** (NSWC), **Steve Marron** (UNC), **Wendy Martinez** (ONR), **George Michailidis** (Michigan), **Amy Nail** (NCSU), **Michael Porter** (NCSU), **David Rios-Insua** (U Rey Juan Carlos), **Yasmin Said** (George Mason), **Richard Smith** (UNC), **Michael G. Schimek** (U Graz), **Haipeng Shen** (UNC), **Jeffrey Solka** (NSWC), **David van Dyk** (UC Irvine), **Ed Wegman** (George Mason), **Stanley Young** (NISS), **Helen Zhang** (NCSU)

Local Host

National Institute of Statistical Sciences
Alan F. Karr, Director and Nell Sedransk, Associate Director
(919) 685-9300

Short Course

Wednesday 21 May 2008

Short Course : "OMICS"

Stan Young, NISS
Kejun (Jack) Liu, OmicSoft
Time 8:00 am – 12 noon

High-dimensional data such as microarray, proteomics and metabolomics data typically have many more variables than samples. Massive genomic data sets, for example, are now coming on line. Such a data set might include hundreds of thousands of patients and on the order of 1 million predictors.   Technical problems involve computer science, visualization and statistics issues. This tutorial will cover new matrix factorization methods. We will also explore examples from the microarray area using data drawn from the publicly available NCBI GEO database. There will be hands on use of free demonstration-version software.

S. Stanley Young and Kejun (Jack) Liu

Dr. Young's interests include analysis of complex data sets, algorithms, statistical strategies, and confidence that claims are valid. He has experience in all phases of drug discovery. He is a Fellow of the ASA and an adjunct professor at three research universities.

Dr. Liu graduated from NCSU in statistics and bioinformatics and completed a postdoctoral fellowship at NISS. He has 16 years experience in writing complex analysis software. He is the CTO of Omicsoft, a spin out company from GlaxoSmithKline.

INTERFACE 2008   KEYNOTE ADDRESS

Risk Analysis and Detection of Nuclear Material in Containers Entering US: It's Not Computing Prowess Alone

Siddhartha R. Dalal

Senior Technology Adviser to the President
RAND Corporation

Given the mandate by US Congress that all goods entering US should be inspected for illicit nuclear material, DHS, Department of Homeland Safety, is moving towards one hundred percent inspection of all containers entering US at various ports of entry for nuclear material. Around ninety-five percent of the containers entering at these ports of entry are currently being inspected. This has enabled collection of terabytes of data on millions of containers and their contents. The data include text, radiation portal and other data. Different sets of methods have emerged to analyze each set of data. But, up to now no effort has been made to analyze these sets of data in a unified manner for detecting illicit nuclear material. In this talk we discuss a number of challenges in creating a real time unified decision system from a risk analysis perspectives and the corresponding challenges in data analysis. The methodologies proposed here are based on a number of machine learning and statistical approaches and are generalizable to all situations where screening is involved.

Siddhartha Dalal

Siddhartha Dalal is the Senior Technology Adviser to the President at RAND. Sid's industrial research career began at Math Research Center at Bell Labs followed by Bellcore/Telcordia Technologies. Most recently he was a vice president of research at Xerox. He has co-authored over seventy publications, several patents and two NRC reports covering the areas of risk analysis, econometrics modeling, image processing, stochastic optimization, data/document mining, software engineering and Bayesian methods. He is recipient of several best paper awards from IEEE, ASA and ASQC.

CONFERENCE SCHEDULE

DATE	EVENTS
Wednesday 21 May	**Related Conferences** **ASA-QMDNS – Quantitative Methods in Defense and National Security** (8:30 am – 5:30 pm) **SAMSI – RISK Revisited: Progress and Challenges** (8:30 am – 5:30 pm) **Interface 2008 RISK : Reality** **Short Course: "OMICS"** Stanley Young (NISS) & Kejun (Jack) Liu (OmicSoft) (8:00 am – 12 noon) **Interface 2008 – ASA – SAMSI Evening Mixer** (8:00 pm – 10:00 pm)
Thursday 22 May	**Interface 2008** Registration (7:30 am - 5:30 pm) **Keynote Address** Siddhartha Dalal (Rand Corporation) *Risk and Public Policy* (8:30 am - 10:00 am) **Technical Sessions** (10:30 am - 5:30 pm) **Interface 2008 Banquet** (7:00 pm - 10:00 pm)
Friday 23 May	**Interface 2008** **Technical Sessions** (8:30 am - 5:30 pm)
Saturday 24 May	**Interface 2008** **Technical Sessions** (8:30 am – 12 noon)

INTERFACE 2008 Program

Technical Talks

THURSDAY 22 May

Thursday 10:30 am – 12:15 pm

Modeling of Extreme Events and Analysis of Risk

Organizers and Chairs: Dipak Dey & David Rios-Insua

Room 106

Elijah Gaioni	U Conn	*Semiparametric Functional Estimation using Quantile-based Prior Elicitation*
Sourish Das	U Conn	*Hurricane Activity in Context of Changing Environment*
Jesus Rios	SAMSI&Aalborg U	*Risk Analysis for Auctions*

Enhancing Knowledge and Assessing Risk through Analysis of Massive Data

Organizer: Karen Kafadar
Chair: Ginger Davis

Room 107

Ginger Davis	UVA	*Statistical Methods for Detecting Computer Attacks from Streaming Internet Data*
Amy Braverman	JPL/Cal Tech	*Massive Data Set Analysis for NASA's Atmospheric Infrared Sounder*
Michael Trosset	Indiana U	*What Kind of Knowledge Does Locally Linear Embedding Extract?*

Streaming Data Analysis

Organizer and Chair: Edward Wegman

Room 102

Bill Szewczyk	NSA	*Data Analysis on Streams*
Werner Stuetzle	Washington	*Using Labeled Data to Evaluate Change Detectors in a Multivariate Streaming Environment*
Shen-Shyang Ho	JPL/ Cal Tech	*Change Detection in Data Streams by Testing Exchangeability*

Contributed Paper Session 1

Room 105

Zhiliang Ma	Johns Hopkins	*Combining Dissimilarity Representations in Embedding Product Space*
Adam Cardinal-Stakenas	Johns Hopkins	*Comparing Dissimilarity Representations of Disparate Information*
Joel Bernanke	Boston U	*Network Mapping of large data sets*

LUNCH BREAK: 12:15 pm – 1:45 pm

Thursday 1:45 pm – 3:30 pm

Probabilistic Models in Risk Assessment

Organizer and Chair: David Banks

Room 102

Mehmet Sahinoglu	Troy U	*Security Risk for Computer Systems*
Alyson Wilson	LANL	*Bayesian Reliability Analysis*
David Banks	Duke U	*Adversarial Risk Analysis*

Air Pollution Risk Assessment: from Research to Regulation

Organizer and Chair: Amy Nail

Room 106

Allen Lefohn	ASL & Associates	*Realistic Biological and Exposure/Dose Relationships: How They Modify Perceived Human Health & Ecological Risk*
Roger Peng	Johns Hopkins U	*Statistical Methods for Assessing the Health Risks of Particulate Matter Components*
Yongku Kim	SAMSI	*How Changing the Ozone Standard Might Affect Respiratory Mortality*

New Developments in Machine Learning and Statistical Modeling for Massive Data

Organizer and Chair: Helen Zhang

Room 107

Jerry Zhu	U Wisconsin	*Online Semi-Supervised Learning*
Yufeng Liu	UNC	*Robust Large-Margin Classifiers*
Howard Bondell	NCSU	*Simultaneous Feature Selection and Structure Identification for ANOVA Models*

Contributed Paper Session 2

Room 105

Roy E. Welsch	MIT	*Robust Risk: Using Robust Methods to Improve Investment Performance*
Bonnie K. Ray	IBM	*Challenges in Integrated Risk Management for the Enterprise*
Leming Qu	Boise State U	*Copula Density Eestimation by Total Variation Penalized with Constraints*

Thursday 3:45 pm – 5:30 pm

Multivariate Extremes

Organizer and Chair: Richard Smith

Room 106

Richard Smith	UNC	*Multivariate Extremes and Risk*
Jan Heffernan	Lancaster U	*A Conditional Approach to Modeling Multivariate Extremes*
Dan Cooley	Colorado State	*Prediction for Max-stable Processes via an Approximated Conditional*

Model-based Risk Assessment in Life Science

Organizer and Chair: Lutz Edler

Room 102

C. Portier	NIEHS	*Finding the Right Path: Using Structurally-Enhanced Pathway Enrichment Analysis to Identify Targets for High-Throughput Screening*
Lutz Edler	German Cancer Research Center	*Data Gaps and Needs in Model-based Risk Assessment*
Matthew Wheeler	UNC	*Dose Response Uncertainty and Model Averaging*

Recent Developments in Machine Learning and Classification - to appear in the *Journal of Computational and Graphical Statistics*

Organizer: David van Dyk
Chair: Michael Trosset

Room 107

George Michailidis	U Michigan	*An Iterative Algorithm for Extending Learners to a Semi-supervised Setting*
Tong Tong Wu	U Maryland	*An MM Algorithm for Multicategory Vertex Discriminant Analysis*
Han-Ming (Hank) Wu	Tamkang U	*Kernel Sliced Inverse Regression with Applications to Classification*

FRIDAY 23 May

Friday 8:30 am – 10:15 am

Statistics and Modern Image Analysis, I

Organizer: Steve Marron
Chair: Haipeng Shen

Room 106

S.M. Pizer	UNC	*M-reps, Curved Feature Space, Bayesian Segmentation*
R. E. Broadhurst	UNC	*Quantile Functions for Texture Analysis and M-rep Segmentation*
Suman Sen	UNC	*Manifold SVM for M-rep Data*

SNP Analysis Methods and Software

Organizer and Chair: Stan Young

Room 107

Danyu Lin	UNC	*HapStat*
Kejun (Jack) Liu	OmicSoft	*Analysis and Visualization of SNP Data*
Dmitri Zaykin	SAS	*Whole-genome SNP Analysis*

Text Mining Applications

Organizers and Chairs: Edward Wegman & Yasmin Said

Room 102

Andris Abakuks	U London-Birbeck	*The Synoptic Gospels Problem and the Trips-Link*
Walid Sharabati	American U	*The Relationship between Prophets and Chapters in the Quran: A Two-Mode Social Network Model*

Contributed Paper Session 3

Room 105

Dusan Maletic	Rutgers U	*Bayesian Methodology for Precision Astrometry of Highly Undersampled Images*
Amy Nail	N.C. State U	*Quantifying Local Creation and Regional Transport Using a Hierarchical Space-time Model of Ozone as a Function of Observed NOx, a Latent Space-time VOC Process, Emissions, and Meteorology*
Mariana Toma-Drane	USC	*Post-Chernobyl Psychological Effects on Individuals in Belarus*

Friday 10:30 am – 12:15 pm

Statistics and Modern Image Analysis, II

Organizer: Steve Marron
Chair: Steve Pizer

Room 106

Brad Davis	Kitware & UNC	*Smoothing over Diffeomorphisms*
Hongtu Zhu	UNC	*Intrinsic Regression Model for Positive Definite Matrices*
Haipeng Shen	UNC	*Supervised Singular Value Decomposition for Independent Component Analysis of fMRI*

Statistical and Computational Issues in Analyzing Sensor Networks

Organizer and Chair: Alan Gelfand

Room 102

George Michailidis	U Michigan	*Robust Target Detection & Localization in Wireless Sensor Networks*
Carol Y. Lin	CDC	*Statistical Issues in Designing an Optimal Detection System with Multiple Heterogeneous Sensors*
Soumendra Lahiri	Texas A&M	*Analysis of Microsensor Networks from a Statistical Perspective*

Text Data Analysis

Organizer and Chair: Jeffrey Solka

Room 107

Elizabeth Hohman	NSWC	*Generalization of the Vector Space Model for a Streaming Corpus of Text Documents*
Kendall Giles	VCU	*Interactive Text Mining with Iterative Denoising*
Avory Bryant	NSWC	*Cross Corpus Discovery via Nearest Neighbor Change-point Analysis*

Contributed Paper Session 4

Room 105

Zhenyu Liu	GWU	*A Triangle Test for Equality of Distribution Functions in High Dimensions*
Ori Rosen	UTEP	*A Bayesian Model for Multivariate Functional Data*
Shih-Chuan Cheng	Creighton U	*Confidence Estimation of the Parameter Involving in the Distribution of the Total Time on Test for Censored Data*
E. James Harner	WVU	*LifeStats: An Interactive Environment for Teaching Statistics*

LUNCH BREAK: 12:15 pm – 1:45 -pm

Friday 1:45 pm – 3:30 pm

Statistics and Evolutionary Biology, I

Organizer and Chair: Haipeng Shen

Room 107

Joel Kingsolver	UNC	*Evolutionary Analyses of Function-valued Traits*
Travis Gaydos	UNC	*Quantification of Curves' Variation and Simplicity to Find Genetic Constraints*
Brian O'Meara	National Evolutionary Synthesis Center	*Extending Models of Character Coevolution*

Sensor Networks and Statistics - New Researchers Session

Organizer and Chair: George Michailidis

Room 102

Sheela Nair	UCLA	*Fault Detection for Embedded Networked Sensing*
Natallia Katenka	U Michigan	*A Cost-efficient Approach to Wireless Sensor Network Design*
Gavino Puggioni	Duke U	*Analyzing Space-time Sensor Network Data under Suppression and Failure in Transmission*

Contributed Paper Session 5

Room 106

Vincent A. Cicirello	R. Stockton	*Statistically Modeling the Performance of a Multistart Randomized Heuristic Algorithm*
Eric Tassone	Google	*Keeping a Search Engine Index Fresh: Risk Versus Optimality Trade-offs in Estimating Frequency of Change in Web Pages*

Friday 3:45 pm – 5:30 pm

Statistics and Evolutionary Biology, II

Organizer and Chair: Haipeng Shen

Room 107

Christina Burch	UNC	*Distribution of Mutation Effects and Adaptation in an RNA Virus*
Mihee Lee	UNC	*Deconvolution and Sieve Estimation of Mutation Effect Distribution*
Paul Magwene	Duke U	*Modularity in Biological Systems: Statistical Challenges and Evolutionary Insights*

Assessing Health Risk from Complex Data

Organizer and Chair: David Dunson

Room 102

Joseph Ibrahim	UNC	*A Bayesian Hidden Markov Model for Motif Discovery through Joint Modeling of Genomic Sequence and ChIP-chip Data*
Jason Fine	UNC	*Analysis of Left-truncated Semi-competing Risks Data with Application to Disease Registries*
Lianming Wang	NIEHS	*Semiparametric Bayes Modeling of Onset and Progression from Current Status Data*

Integration of Disparate Types of Information

Organizer: Wendy Martinez

Chair: Jeffrey Solka

Room 106

Carey Priebe	Johns Hopkins U	*Disparate Information Fusion: On the Exploitation of Multiple Disparate Dissimilarities*
Brent Castle	Indiana U	*Combining Disparate Information by Nonmetric Multidimensional Scaling*
Jeffrey Solka	NSWC	*Disparate Information Fusion on Images and Text*

SATURDAY 24 May

Saturday 8:30 am – 10:15 am

Spatial Risk Mapping: Prediction and Change Detection

Organizer and Chair: Michael Porter

Room 106

Jason Dalton	SPADAC	*Space-time Forecasting of Extreme Events in Complex Environments*
Ronald D. Fricker, Jr.	Naval Postgraduate School	*Using the Repeated Two-sample Rank Procedure for Detecting Anomalies in Space and Time*
Michael Porter	NCSU	*A Martingale Methodology for the Quick Identification of Point Process Anomalies*

Contributed Paper Session 6

Room 107

Andrejus Parfionovas	Utah State U	*Classification Trees with Oblique Splits for Multidimensional Datasets*
Rebecca Nugent	CMU	*Clustering with Confidence: A Binning Approach*
Joran Elias	U Montana	*Making Tree Ensembles More Robust to Noisy Data*

Saturday 10:30 am – 12:15 pm

Change Detection in Random Graphs

Organizer and Chair: David Marchette

Room 106

David Marchette	NSWC	*Detecting Activity Changes in Graphs*
Youngser Park	Johns Hopkins U	*Scan Statistics in Hypergraphs*
Elizabeth Beer	Johns Hopkins U	*Torus Graph Inference for Detection of Localized Activity*

Risk of Reaching False Conclusions

Organizer and Chair: Stan Young

Room 107

Stan Young	NISS	*The Problem of Observational Studies*
Robert Obenchain	SoftRx	*A Complete Illustration of Local Control for Observational Studies*
Patrick Ryan	GlaxoSmithKline	*Exploring the Effects of Medicines: Managing Risk across Multiple Outcomes*
Alice White	GlaxoSmithKline	*Discussant*

40th Symposium on the Interface:
Computing Science and Statistics

INTERFACE 2008

RISK : REALITY

ABSTRACTS

Durham Marriott, Durham, NC
21 - 24 May 2008

INDEX OF AUTHORS

INTERFACE 2008 ABSTRACTS

KEYNOTE ADDRESS

Risk Analysis and Detection of Nuclear Material in Containers Entering US: It's Not Computing Prowess Alone

Siddhartha R. Dalal
Senior Technology Adviser to the President
RAND Corporation

Given the mandate by US Congress that all goods entering US should be inspected for illicit nuclear material, DHS, Department of Homeland Safety, is moving towards one hundred percent inspection of all containers entering US at various ports of entry for nuclear material. Around ninety-five percent of the containers entering at these ports of entry are currently being inspected. This has enabled collection of terabytes of data on millions of containers and their contents. The data include text, radiation portal and other data. Different sets of methods have emerged to analyze each set of data. But, up to now no effort has been made to analyze these sets of data in a unified manner for detecting illicit nuclear material. In this talk we discuss a number of challenges in creating a real time unified decision system from a risk analysis perspectives and the corresponding challenges in data analysis. The methodologies proposed here are based on a number of machine learning and statistical approaches and are generalizable to all situations where screening is involved.

Siddhartha Dalal
Siddhartha Dalal is the Senior Technology Adviser to the President at RAND. Sid's industrial research career began at Math Research Center at Bell Labs followed by Bellcore/Telcordia Technologies. Most recently he was a vice president of research at Xerox. He has co-authored over seventy publications, several patents and two NRC reports covering the areas of risk analysis, econometrics modeling, image processing, stochastic optimization, data/document mining, software engineering and Bayesian methods. He is recipient of several best paper awards from IEEE, ASA and ASQC.

10:30 a.m.-12:15 p.m., Thursday

MODELING OF EXTREME EVENTS AND ANALYSIS OF RISK

Organizers: Dipak Dey, University of Connecticuct
David Rios-Insua, Universidad Rey Juan Carlos

Semiparametric Functional Estimation using Quantile-based Prior Elicitation

Elijah Gaioni, University of Connecticut

A methodology by which inconsistent prior information can be used to perform functional estimation is presented. Sharp qualitative information consisting of the functional form of the likelihood is assumed known. It is also assumed that vague quantitative information in the form of multiple possible quantiles is available. An optimality criterion is then applied for the purpose of determining a predictive distribution consistent with the expert provided information. The prior distribution is estimated semiparametrically, where an adaptive method for selecting basis elements is used to limit the computational difficulties associated with the solution of this problem.

* * *

Hurricane Activity in Context of Changing Environment

Sourish Das, University of Connecticut

A major obstacle to researchers for hurricane research is due to extreme high variability of the data. In view of that, we will present some simple techniques to develop a dynamic statistical model that can identify small changes in hurricane counts in the face of highly variable hurricane frequency counts. We considered only satellite era data for our analysis. We will also discuss some issues about prior elicitation and mention how one could use the data from pre-satellite era, considering those as historical data, by using the idea of power prior (Chen 2001) and implementing Das and Dey's method of analysis (Das and Dey, 2006).

* * *

Risk Analysis for Auctions

Jesus Rios, SAMSI and Aalborg University
David Rios-Insua, Universidad Rey Juan Carlos
David Banks, Duke University

Applications in counterterrorism and corporate competition have led to the development of new methods for the analysis of decision-making when there are intelligent opponents and uncertain outcomes. This field represents a combination of statistical risk analysis and classical game theory, and is sometimes called adversarial risk analysis. We describe several approaches to adversarial risk problems, providing a unified framework for analysis aimed at prescribing advice to one of the participants. The key issue in our framework is the assessment of the probabilities of adversaries' actions. We assume that adversaries are expected utility maximizers and, therefore, uncertainty in their actions stem from our uncertainty about their utilities and probabilities when used to analyzed the adversaries' decision problems.

ENHANCING KNOWLEDGE AND ASSESSING RISK THROUGH ANALYSIS OF MASSIVE DATA

Organizer: Karen Kafadar, Indiana University

Statistical Methods for Detecting Computer Attacks from Streaming Internet Data

Ginger Davis, University of Virginia
Karen Kafadar, Indiana University
David J. Marchette, Naval Surface Warfare Center

Successful strategies for network security depend upon collating massive data features into usable variables, high-performance and computationally efficient methods for anomaly detection, and ability to predict typical behavior so atypical and abnormal events can be better identified. We define data variables (packets, sessions, activities) from network data and how they can be used in models for user profiles, workload management, application verification, etc. New techniques to process these data will be discussed, including visualization of data sets, leading to methods for clustering images. We provide examples of these methods on motivating internet packet data streams.

* * *

Massive Data Set Analysis for NASA's Atmospheric Infrared Sounder

Amy Braverman, Jet Propulsion Laboratory, California Institute of Technology
Eric Fetzer, Jet Propulsion Laboratory, California Institute of Technology
Brian Kahn, Joint Institute for Regional Earth System Science and Engineering, UCLA
Joao Teixeira, Jet Propulsion Laboratory, California Institute of Technology

NASA's Atmospheric Infrared Sounder (AIRS) has been collecting large quantities of remote sensing data about the vertical structure of Earth's atmosphere since AIRS was launched aboard the Aqua spacecraft in mid-2002. These data pose a classic problem in the analysis of massive data sets: how do we understand the relationships among fine-scale phenomena within their global context? We answer that question here by partitioning the data on a coarse spatio-temporal grid, and estimating the multivariate distribution of the data within each grid cell. Then, we look for patterns in the evolution of those distributions as functions of space and time, and ultimately tie them back to physical phenomena generating the data sets. Quantifying this evolution is challenging because the data are high-dimensional and the distributions are complex. We attack the problem using the Wasserstein distance between distributions as a measure of similarity among grid cells' data, and therefore as a measure of similarity between the underlying physical processes. We close with some thoughts on how this strategy might be applied in other problems where massive data sets arise.

What Kind of Knowledge Does Locally Linear Embedding Extract?

Michael Trosset, Indiana University
Brent Castle, Indiana University

Locally Linear Embedding (LLE) is a seminal technique for manifold learning, a collection of techniques for nonlinear dimension reduction that suppose high-dimensional data lie on (or near) low-dimensional manifolds. Manifolds are locally linear. LLE posits that local structure is preserved by preserving "reconstruction weights", the coefficients used to represent each point as a linear combination of its neighbors. We present simple examples that challenge that premise. We also explore the role of regularization in LLE and describe a connection to Laplacian eigenmaps.

STREAMING DATA ANALYSIS

Organizer: Edward Wegman, George Mason University

Data Analysis on Streams

Bill Szewczyk, National Security Agency

* * *

Using Labeled Data to Evaluate Change Detectors in a Multivariate Streaming Environment

Werner Stuetzle, University of Washington

* * *

Change Detection in Data Streams by Testing Exchangeability

Shen-Shyang Ho, Jet Propulsion Laboratory / California Institute of Technology

In a data streaming setting, data points are observed sequentially. The data generating model for the data points may change as the data is streaming. In this talk, we introduce a martingale methodology for change detection in high dimensional data streams by testing data exchangeability. Empirical results are presented to show its feasibility and effectiveness. The martingale change detection method is used to implement (i) an online adaptive support vector machine for labeled data streams, and (ii) a single-pass video-shot change detector for unlabeled video streams.

CONTRIBUTED PAPER SESSION 1

Combining Dissimilarity Representations in Embedding Product Space

Zhiliang Ma, Johns Hopkins University
Adam Cardinal-Stakenas, Johns Hopkins University
Youngser Park, Johns Hopkins University
Carey E. Priebe, Johns Hopkins University

Dissimilarity representation provides an alternative, sometimes beneficial, way to represent observations compare to feature based representation in statistical pattern recognition. As a consequence of a variety of possible dissimilarity functions that can be applied on a given set of objects, there can be many dissimilarity representations.

Some work \cite{Pekalska2001,Miller2008} has been done to combine dissimilarity representations and shows classification accuracy improvement after combining. In this work, we explore three possible ways on combining dissimilarity matrices and investigate in detail one of them - performing combination after embedding. We propose a dimension reduction / model selection approach to help achieve efficient combining in the embedding space. Examples of simulated and real data are presented indicating that our combining approach is useful.

* * *

Comparing Dissimilarity Representations of Disparate Information

Adam Cardinal-Stakenas, Johns Hopkins University
Zhiliang Ma, Johns Hopkins University
Youngser Park, Johns Hopkins University
Carey Priebe, Johns Hopkins University

We present a methodology for applying pattern classification techniques to disparate data sources using the dissimilarity representation. This approach is of interest because it naturally leads to a principled method of fusing disparate information sources to improve classifier performance. Furthermore, within the dissimilarity framework we can apply machinery from the multidimensional scaling literature to evaluate the suitability of a dissimilarity metric in a classifier independent way.

In solving this modern statistical inference problem, the statistical analyst is faced with the issue of having to choose from a wide range of techniques, starting with the basic question of how to extract features from the observed data. Given a battery of feature extraction and comparison techniques, we propose a method for identifying the data representation that best separates the classes. If we have data observed as feature vectors, X, or a dissimilarity matrix Δ, we can calculate the congruence between Δ (or $d(X),$ the distance between the feature vectors) and Δ_Y, the discrete (0,1) dissimilarity between the class labels. In this way we measure how closely a data set approximates the ``perfect information'' of the class labels.

We will discuss our methodology in detail and provide an example of inference using these techniques on a data set consisting of image and caption pairs.

Network Mapping of Large Data Sets

Joel Bernanke, Boston University School of Public Health
Al Ozonoff , Boston University School of Public Health

Past studies of massive linked data sets, such as the physical connections between routers on the Internet, have shown that there is valuable information encoded in the network topology of such data. This technique has been extended to data sets without rigidly defined linkages, such as protein interaction networks. When linkages are not predefined, suitable criteria for identifying linkages must be developed. In this paper, we propose a natural mapping of a data set onto a network: variables map to nodes and the associations among variables map to edges. An edge exists between two variables when the strength of the association between the variables exceeds an arbitrary cutoff. Defined in this way, data sets map to undirected networks.

To illustrate the approach, we consider the topological properties of the 2002 National Health and Nutrition Examination Survey (NHANES). The NHANES data contain several hundred biological outcomes and other health-related variables, collected on a nationally representative sample of several thousand U.S. adults. We map a subset of 305 continuous variables from this survey to a collection of nodes and edges, and present an exploratory analysis of the resulting network.

1:45-3:30 p.m., Thursday

PROBABILISTIC MODELS IN RISK ASSESSMENT

Organizer: David Banks, Duke University

Security Risk for Computer Systems

Mehmet Sahinoglu, Troy University

The idea of the Game Theory is incorporated to the quantitative Security-Meter [1] design technique. The offensive side in a two-player zero-sum game is a set of involuntary (uncontrollable) vulnerability-threat probability (risk) combinations from the hackers' side and the bad-fellows on the malware front. The defensive side is a set of countermeasure probabilities on the mitigation front in the endeavor of determining a strategic Cost-Optimal-Countermeasure-Action (COCA) plan, or a roadmap, to optimize a minimum cost allocation for a voluntary controllable) list of countermeasure actions such as firewalls or anti-virus or anti-malware [3]. An interactive user-friendly software will be introduced to facilitate the three pillars of this ongoing applied research before a final product cycle [4]. Namely: i) the end-user solicitation for data-bank entry into the security-meter design [2] ii) the game theoretical application for countermeasure allocation, or the COCA interface, and iii) the improved security-meter solution with cost parameters and what-to-do list as a feedback. Privacy aspects for a probabilistic risk management will also be studied [5].

* * *

Bayesian Reliability Analysis: Statistical Challenges from Science-Based Stockpile Stewardship

Alyson Wilson, Los Alamos National Laboratory

In this talk, I examine two problems that have arisen from my work in Science-Based Stockpile Stewardship (SBSS) at Los Alamos National Laboratory. The goal of SBSS is the assessment of safety and reliability in aging warheads in the absence of nuclear testing. As a statistician, this leads to complex problems in system reliability and test planning. As a first example, I present Bayesian networks and their use in system reliability assessment and develop a complete Bayesian solution for inference with multilevel data. As a second example, I present a develop a Weibull reliability assurance test plan using hierarchical data.

* * *

Adversarial Risk Analysis

David Banks, Duke University

Traditional risk analysis assumes that the probabilities of costly outcomes are not affected by an intelligent opponent. Traditional game theory assumes that the consequences for each set of choices by intelligent opponents are known exactly. In practical applications (e.g., in counterterrorism, corporate competition, etc.) one needs an analysis that combines features of both. This talk describes two such analyses, one pertaining to counterbioterrorism and the other to competitive bidding in an auction.

AIR POLLUTION RISK ASSESSMENT: FROM RESEARCH TO REGULATION

Organizer: Amy Nail

Realistic Biological and Exposure/Dose Relationships: How They Modify Perceived Human Health & Ecological Risk

Allen Lefohn, ASL & Associates

* * *

Statistical Methods for Assessing the Health Risks of Particulate Matter Components

Roger Peng, Johns Hopkins University

* * *

How Changing the Ozone Standard Might Affect Respiratory Mortality

Yongku Kim, Statistical and Applied Mathematical Sciences Institute
Bahjat Qaqish, University of North Carolina at Chapel Hill
Rosalba Ignaccolo, Statistical and Applied Mathematical Sciences Institute
Michela Cameletti, Statistical and Applied Mathematical Sciences Institute
Richard L. Smith, University of North Carolina at Chapel Hill

We present a risk assessment analysis of the potential effect that various regulatory standards for ozone may have on the incidence of respiratory-related mortality. The analysis uses roll-back functions as models for the potential effect of regulatory standards. The statistical methods are based on hierarchical Bayesian models. The objective is to obtain estimates of the effects of various regulatory standards, estimates of their variability, and the effects of various modeling assumptions on those estimates. We also introduced a parametric rollback approach.

NEW DEVELOPMENTS IN MACHINE LEARNING AND STATISTICAL MODELING FOR MASSIVE DATA

Organizer: Helen Zhang, North Carolina State University

Online Semi-Supervised Learning

Jerry Zhu, University of Wisconsin

* * *

Robust Large-Margin Classifiers

Yufeng Liu, University of North Carolina at Chapel Hill

* * *

Simultaneous Feature Selection and Structure Identification for ANOVA Models

Howard Bondell, North Carolina State University

CONTRIBUTED PAPER SESSION 2

Robust Risk: Using Robust Methods to Improve Investment Performance

Roy E. Welsch, Massachusetts Institute of Technology

The Markowitz minimum variance (risk) portfolio (MVP) selection procedure involves the estimation of the sample covariance matrix of asset returns and can perform poorly due to estimation error. We focus on two ways to address this problem--robust optimization and robust estimation. We also consider shortfall risk.

Robust optimization (Goldfarb and Iyengar, 2003) addresses estimation errors by introducing uncertainty sets for market parameters like the individual covariances or matrix. Robust optimization provides worst-case solutions to the MVP problem as the market parameters vary within their uncertainty sets.

Robust estimation (Maronna, et al., 2006) focuses on estimating the covariance matrix and other market parameters robustly, that is, making them less sensitive to small fractions of outlying or unusual returns. There are many approaches to the robust estimation of the covariance matrix including both affine and non-affine equivariant methods. Another approach (Welsch and Zhou, 2007) avoids computation of the covariance matrix directly and allows the use of penalized robust regression methods.

We compare these approaches on several data sets and discuss their relative merits. We also consider ways to combine robust estimation with robust optimization and address the determination of the level of estimation robustness (efficiency loss and breakdown) and optimization robustness (size of uncertainty sets).

This is joint work with Rajiv Menjoge and Dung Tri Nguyen (Ph.D, students at MIT).

Challenges in Integrated Risk Management for the Enterprise

Bonnie K. Ray, IBM Watson Research Lab

I will present an overview of recent and on-going risk analysis projects at IBM Research, spanning all aspects of risk analysis, including acquiring information, extracting knowledge, and drawing inference. Current work on development of an integrated framework for enterprise risk management will be used as an organizational structure. A discussion of areas identified as requiring new methodological development will also be included.

Copula Density Estimation by Total Variation Penalized with Constraints

Leming Qu, Boise State University

Copulas are full measures of dependence among random variables. It is increasingly popular among academics and practitioners in statistics, finance and economics for modeling comovements between markets, risk factors and other relevant variables. A copula's hidden dependence structure that couples a joint distribution with its marginals makes a parametric copular non-trivial. For bivariate copula density estimation problem, we introduce a penalized likelihood approach with a total variation penalty. The constraint of the uniform marginal distributions for the copula is imposed. Adaptive choice of the amount of regularization is based on approximate Bayesian Information Criteria (BIC) type scores. Performance is evaluated through the Monte Carlo simulation.

3:45-5:30 p.m., Thursday

MULTIVARIATE EXTREMES

Organizer: Richard L. Smith, University of North Carolina at Chapel Hill

Multivariate Extremes and Risk

Richard L. Smith, University of North Carolina at Chapel Hill

Extreme value theory is the branch of statistics that is concerned with characterizing and estimating the distributions of extreme events. It is of great importance to risk analysis because in many contexts, the most extreme events are the ones that dominate risk calculations. Univariate extreme value theory is used when there is a single main variable of interest, and is now widely applied in numerous fields. Multivariate extreme value theory is used when the focus of interest is the joint distribution of extreme values in several variables, or in a single variable measured at different points in space and time. Although the origins of the theory stretch back several decades, it is much less widely known and used in wider statistical practice. In this overview talk, I describe the development of the subject from the characterization of multivariate extreme value distributions, multivariate threshold methods, and alternative characterizations of extremal dependence due to Ledford and Tawn. This will be followed by a review of recent developments.

* * *

A Conditional Approach to Modeling Multivariate Extremes

Jan Heffernan, Lancaster University and J. Heffernan Consulting

We review the Conditional Approach to modeling multivariate extreme values of Heffernan and Tawn (2004). This approach is based on an assumption about the asymptotic form of the joint distribution of a d-dimensional random variable conditional on it having an extreme component. The method is semi-parametric and overcomes practical restrictions which limited both the dimension and the shape of tail regions of multivariate distributions which could be explored using previously available methods. We illustrate the method using air pollution data which reveals complex extremal dependence behaviour. We also mention recent advances in the field building on this work.

* * *

Prediction for Max-stable Processes via an Approximated Conditional

Dan Cooley, Colorado State University

The dependence structure of a max-stable random vector is characterized by its spectral measure. Given only the spectral measure, we present a method for approximating the conditional density of an unobserved component of a max-stable random vector given the other components of the vector. The approximated conditional density can be used for prediction. We also present a new parametric model for the spectral measure of a multivariate max-stable distribution. This model is used to perform prediction for both a time series and spatial process.

MODEL-BASED RISK ASSESSMENT IN LIFE SCIENCE

Organizer: Lutz Edler, German Cancer Research Center

Finding the Right Path: Using Structurally-Enhanced Pathway Enrichment Analysis to Identify Targets for High-Throughput Screening

Christopher J. Portier, National Institute of Environmental Health Sciences

Biomedical research has resulted in vast amounts of results and data usually stored on different databases. Using such data, we have developed a framework for generating hypotheses relevant to the understanding of complex diseases (like cancer), viewed as an interplay between genetic and environmental factors. The links between the different complex diseases and environmental factors were derived using a novel algorithm (Structurally Enhanced Pathway Enrichment Analysis, SEPEA) that finds significantly enriched networks of genes or proteins. In the first part of the talk, I will elaborate on the description and evaluation of SEPEA. I will then describe the network of diseases and environmental factors that resulted from the application of SEPEA to the integrated data obtained from a gene/disease polymorphism database, a gene/environmental factor database and a biochemical pathway database. Here I will specifically talk about the overall validation of the network, validation of the metabolic diseases sub-cluster of this network and insights and hypothesis that we obtain by an evaluation of the cancer sub-cluster.

* * *

Data Gaps and Needs in Model-based Risk Assessment

Lutz Edler, German Cancer Research Center

* * *

Dose Response Uncertainty and Model Averaging

Matthew Wheeler, University of North Carolina at Chapel Hill

RECENT DEVELOPMENTS IN MACHINE LEARNING AND CLASSIFICATION - TO APPEAR IN THE *JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS*

Organizer: David van Dyk, University of California-Irvine

An Iterative Algorithm for Extending Learners to a Semi-supervised Setting

George Michailidis, University of Michigan
Mark Culp, West Virginia University

In this talk, we discuss an iterative algorithm, whose objective is to extend learners from a supervised setting into a semi-supervised setting. The algorithm is based on using the predicted response values for observations where missing (unlabeled data) and then incorporates the predictions appropriately at subsequent stages. Convergence properties of the algorithm are investigated for particular learners, such as linear/logistic regression, kernel smoothers, generalized additive models, tree partitioning methods, partial least squares, etc. The algorithm is illustrated on a number of real data sets using a varying degree of labeled responses.

* * *

An MM Algorithm for Multicategory Vertex Discriminant Analysis

Tong Tong Wu, University of Maryland
Kenneth Lange, UCLA

This talk introduces a new method of supervised learning based on linear discrimination among the vertices of a regular simplex in Euclidean space. Each vertex represents a different category. Discrimination is phrased as a regression problem involving ϵ-insensitive residuals and a quadratic penalty on the coefficients of the linear predictors. The objective function can be minimized by a primal MM (majorization-minimization) algorithm. Limited comparisons on real and simulated data suggest that the MM algorithm is competitive in statistical accuracy and computational speed with the best currently available algorithms for discriminant analysis.

* * *

Kernel Sliced Inverse Regression with Applications to Classification

Han-Ming (Hank) Wu, Tamkang University

Sliced inverse regression (SIR) was introduced by Li (1991) to find the effective dimension reduction directions for exploring the intrinsic structure of high-dimensional data. In this study, we propose a hybrid SIR method using a kernel machine which we call kernel SIR. The kernel mixtures result in the transformed data distribution being more Gaussian like and symmetric; providing more suitable conditions for performing SIR analysis. The proposed method can be regarded as a nonlinear extension of the SIR algorithm. We provide a theoretical description of the kernel SIR algorithm within the framework of reproducing kernel Hilbert space (RKHS). We also illustrate that kernel SIR performs better than several standard methods for discriminative, visualization, and regression purposes. We show how the features found with kernel SIR can be used for classification of microarray data and several other classification problems and compare the results with those obtained with several existing dimension reduction techniques. The results show that kernel SIR is a powerful nonlinear feature extractor for classification problems.

8:30-10:15 a.m., Friday

STATISTICS AND MODERN IMAGE ANALYSIS, I

Organizer: J. S. Marron, University of North Carolina at Chapel Hill

Bayesian Segmentation via Probabilities of Shape in Curved Feature Spaces

Stephen M. Pizer, University of North Carolina at Chapel Hill
Ja-Yeong Jeong, University of North Carolina at Chapel Hill
J. S. Marron, University of North Carolina at Chapel Hill

Probability distributions on the geometry of one or more objects benefit from including orientation information, which makes the feature space curved. The estimated value of such probability distributions can be combined with an estimated likelihood distribution on image intensity patterns in object-relative coordinates to allow the segmentation of objects from images so challenging as to require the method to understand shape. The segmentation of multiple objects is best done at multiple scales, leading to the need to estimate conditional probability distributions, of objects given others and of refinements at one scale given the segmentation at a larger scale. I discuss the PCA-like methods for estimation of these distributions in High-Dimension Low-Sample-Size situations, and I demonstrate such segmentations of the prostate, bladder, and rectum from CT images and of the caudate nucleus from MR images.

Quantile Functions for Texture Analysis and M-rep Segmentation

R.E. Broadhurst, University of North Carolina at Chapel Hill

Discriminative models often represent an object using the probability distribution of a set of local features. Understanding the within-class variation of these objects requires understanding the variability of probability distributions. This talk explores the usefulness of quantile functions for describing distribution variability in two tasks: texture classification and image segmentation. In both tasks the objects of interest are represented by the distribution of pixel features across an image region. Both tasks model distributions that do not fit existing parameter families, yet also undergo nonlinear within-class variation when represented nonparametrically by histograms. This talk considers the nonparametric option of the quantile function, which is shown to linearly model a useful class of variation. Principal component analysis is shown to learn effective class-specific quantile function subspaces, which in effect describe learned, class-specific parametric distribution families.

Manifold SVM for M-rep Data

Suman Sen, University of North Carolina at Chapel Hill
Mark Foskey, University of North Carolina at Chapel Hill
J.S. Marron, University of North Carolina at Chapel Hill
Martin Styner, University of North Carolina at Chapel Hill

Support Vector Machine (SVM) is a powerful tool for classification. We generalize SVM to work with data objects that are naturally understood to be lying on curved manifolds, and not in the usual d-dimensional Euclidean space. Such data arise from medial representation (m-rep) of medical images, Diffusion Tensor-MRI (DT-MRI), diffeomorphisms, etc. Considering such data objects to be embedded in higher dimensional Euclidean space results in invalid projections (on the separating direction) while conventional kernel embedding does not provide a natural separating direction. We use geodesic distances, defined on the manifold to formulate our methodology. This approach addresses the important issue of analyzing the change that accompanies the difference between groups by implicitly defining the notions of separating surface and separating direction on the manifold. The methods are applied in shape analysis with target data being m-reps of 3 dimensional images of human hippocampi and simulated distorted ellipsoids.

SNP ANALYSIS METHODS AND SOFTWARE

Organizer: S. Stanley Young, National Institute of Statistical Sciences

HapStat

Danyu Lin, University of North Carolina at Chapel Hill

* * *

Analysis and Visualization of SNP Data

Kejun Jack Liu, OmicSoft

* * *

Whole-genome SNP Analysis

Dmitri Zaykin, SAS Institute Inc.

TEXT MINING APPLICATIONS

Organizers: Edward Wegman & Yasmin Said

The Synoptic Gospels Problem and the Trips-Link

Andrus Abakuks, University of London, Birbeck College

In New Testament studies, the synoptic problem is concerned with the relationships between the gospels of Matthew, Mark and Luke. Specifications in probabilistic terms are set up of versions of Honore's triple-link model. Using data of Tyson and Longstaff, counts of the numbers of verbal agreements between the gospels are examined to investigate which of the possible triple-link models appears to give the best fit to the data.

To carry out an exploratory investigation of synoptic relationships at a more detailed level, individual sections of the text (pericopes) are compared to measure the pairwise similarities in the text between the gospels, and a simple principal component analysis is performed.

* * *

The Relationship between Prophets and Chapters in the Quran: A Two-Mode Social Network Model

Walid Sharabati, American University

CONTRIBUTED PAPER SESSION 3

Bayesian Methodology for Precision Astrometry of Highly Undersampled Images

Dusan Maletic, Rutgers University
Carlton Pryor, Rutgers University
Slawomir Piatek, New Jersey Institute of Technology

This article discusses a Bayesian statistics based method for precision astrometry when the analyzed data is severely undersampled. Precision astrometry poses a particular requirements on the image analysis where the crucial information is not a shape or content but the exact position and motion of the object. Particular application to the problem of determination of the proper motions of Dwarf Spheroidal Sattelite Galaxies of the Milky Way will be presented. An accuracy sufficient to measure the proper motion of these objects over period of just few years when observed by the Hubble Space Telescope is of the order of 0.005 pixel (0.25 milliarcseconds) as the order of magnitude of the expected motion over one year is 0.009 pixel. For this purpose, although it is the best instrument available at the moment, the Hubble Space Telescope provides severely undersampled image with the full width at half maximum of a stellar point spread function of approximately 1 pixel. Results will be compared with the traditional (Anderson&King (2000)) method of fitting the effective point spread function.

Quantifying Local Creation and Regional Transport Using a Hierarchical Space-time Model of Ozone as a Function of Observed NOx, a Latent Space-time VOC Process, Emissions, and Meteorology

Amy Nail, North Carolina State University
Jackie Hughes-Oliver, North Carolina State University
John Monahan, North Carolina State University

We explore the ability of a space-time model to decompose 8-hour ozone concentrations into parts attributable to local emissions and regional transport, to predict ozone at a given point in space and time, and to assess the efficacy of past and future emission control programs. We model ozone as created plus transported ozone plus an error that has a seasonally varying spatial covariance. The created component uses atmospheric chemistry results to express ozone created on a given day at a given site as a function of the observed NOx concentration, the latent VOC concentration, and temperature. The ozone transported to a given day at a given site we model as a weighted average of the ozone observed at all sites on the previous day, where the weights are a function of wind speed and direction. The latent VOC process model has a mean trend that includes emissions, temperature, a workday indicator, and an error with a seasonally varying spatial covariance. Using likelihood methods, we fit the original model and obtain space-time predictions for comparison with a withheld dataset and with predictions from CMAQ, the deterministic model used by EPA to assess emission control programs. Our model produces one set of predictions based on the mean trend and spatial correlations and another based on the mean trend alone. The predictions based on the mean trend and spatial correlations have a lower root mean squared error (RMSE) when compared to point observations than do than do the 36 km gridcell averages from CMAQ; predictions based on the mean trend alone have the same RMSE as CMAQ but systematically under predict high ozone values.

* * *

Post-Chernobyl Psychological Effects on Individuals in Belarus

Mariana Toma-Drane, University of South Carolina
A. E. Frongillo, University of South Carolina
J. Vena, University of South Carolina
W. Karmaus, University of South Carolina
D. Friedman, University of South Carolina
A. Michalek, University of Buffalo
K. Moysick, University of Buffalo
M. Mahoney, University of Buffalo

Objective To examine if a large-scale nuclear disaster such as the Chernobyl nuclear accident induced long term psychological effects (i.e., perceptions of risk and health, food perceptions) and food related behaviors on the Belarusian exposed individuals, and if they differ based on family roles (i.e., mother, father, or child) at the time of the accident.

Methods Data were collected on psychological effects and food-related behaviors in 2002-2003 using questionnaires administered through interviews. The study sample was composed of children (n=145) and their parents (n=153) recruited as controls in a case-control pilot study in two Belarusian oblasts that had low (Mogilev) and high exposure (Gomel). Mixed linear models were

estimated to assess psychological effects and food-related behaviors among mothers, fathers and children.

Results Fathers were less likely to perceive themselves as being at risk than mothers, but children were more likely to perceive themselves as being at risk. Children were less likely to perceive themselves as having health problems than their mothers and fathers, and to perceive their food as being radioactively contaminated than their mothers and fathers. Fathers were less likely to report improvements on food-related behaviors. Assessing parents' perceptions of risk for their children indicated the parents were more likely to perceive their children as being at risk than themselves.

Conclusion An individual's role in the family, in the context of the Chernobyl nuclear accident helped explain long-term psychological effects and food related behaviors among family members induced by this large-scale nuclear accident. These results are consistent with the caregiver role of a mother within the Belarusian family.

10:30-12:15, Friday

STATISTICS AND MODERN IMAGE ANALYSIS, II

Organizer: J. S. Marron, University of North Carolina at Chapel Hill

Smoothing Over Diffeomorphisms

Brad Davis, Kitware and University of North Carolina at Chapel Hill

* * *

Intrinsic Regression Model for Positive Definite Matrices

Hongtu Zhu, University of North Carolina at Chapel Hill

The aim of this talk is to develop an intrinsic regression model for the analysis of positive-definite matrices as responses in a Riemannian manifold and their association with a set of covariates, such as age and gender, in a Euclidean space. The primary motivation and application of the proposed methodology is in medical image. Because the positive-definite matrices do not form a vector space, applying classical multivariate regression to modeling positive-definite matrices may undermine their association with covariates of interest, such as age and gender, in real applications. Our intrinsic regression model, as a semiparametric model, uses a link function to map from the Euclidean space of covariates to the Riemannian manifold of positive-definite matrices. We develop an estimation procedure to calculate parameter estimates and establish their limiting distribution. We develop score statistics to test linear hypotheses of unknown parameters and develop a test procedure based on a resampling method to simultaneously assess the statistical significance of linear hypotheses across a large region of interest. Simulation studies are used to demonstrate the methodology and examine the finite performance of the test procedure for controlling the family-wise error rate. We apply our methods to the detection of statistical significance of diagnostic effects on the integrity of white matter in a diffusion tensor study of human immunodeficiency virus.

* * *

Supervised Singular Value Decomposition for Independent Component Analysis of fMRI

Haipeng Shen, University of North Carolina at Chapel Hill

STATISTICAL AND COMPUTATIONAL ISSUES IN ANALYZING SENSOR NETWORKS

Organizer: Alan Gelfand, Duke University

Robust Target Detection & Localization in Wireless Sensor Networks

George Michailidis, University of Michigan

Detecting and localizing a target and estimating its signal are among the fundamental tasks of wireless sensor networks. Performing them efficiently while conserving energy and communications costs often requires distributed processing. We propose an algorithm for local neighborhood voting and show that performing these operations before global data fusion can significantly improve target detection and reduce communications costs compared to standard global methods. We also show that the same algorithm can improve target localization, where methods based on collecting signal data from a small subset of sensors determined by local operations can achieve performance levels similar to methods based on the full data, but at a fraction of the communications cost. In particular, EM-type algorithms we have developed for localizing the target and estimating its signal from this limited information are shown to perform as well as, or, at low SNRs, even better than maximum likelihood methods based on the full data (the current "gold standard" in the field). Some extensions to online tracking are also briefly discussed.

Statistical Issues in Designing an Optimal Detection System with Multiple Heterogeneous Sensors

Carol Y. Lin, Center for Disease Control
Lance Waller, Emory University
Robert Lyles, Emory University
P. Barry Ryan, Emory University

Combining individual tests or sensors in a detection system often improves overall diagnostic performance, and many optimal decision-theoretic combinations under many different criteria exist. However, when designing a detection system, cost-effectiveness is important, particularly when combining a set of sensors with heterogeneous individual cost and performance. We consider the expected cost of correct and incorrect decisions of the system, constrained by the budget available to select an optimal system with various combinations of different types of conditionally independent and dependent sensors. We illustrate the approach using a hypothetical network of two different types of air pollution monitors.

Analysis of Microsensor Networks from a Statistical Perspective

Soumendra Lahiri, Texas A&M
P. Banerjee, Texas A&M

Wireless micro-sensors are resource constrained and must use their energy efficiently to ensure a longer life of the network. In this talk, we consider a two-hop self-organizing, distributed clustering protocol for wireless sensor networks, that builds upon the work of Heinzelman et al. (2002: IEEE Transactions on Wireless Communications). We consider performance of the proposed protocol in terms of energy usage, and the quality of information gathered from a statistical point of view. We develop a suitable spatial statistical framework to address the issue of joint optimization of energy consumption and information loss. We will present some of our numerical and theoretical findings and outline some statistical challenges for future work.

TEXT DATA ANALYSIS

Organizer: Jeffrey Solka, Naval Surface Warfare Center

A Generalization of the Vector Space Model for a Streaming Corpus of Text Documents

Elizabeth Hohman, Naval Surface Warfare Center

In statistical text processing, the vector space model is often used to represent documents as vectors in a high-dimensional space with each dimension corresponding to a different word in the lexicon. The vector entries for each document are found by weighting the word counts of the document by a set of weights assigned to each word in the lexicon. These weights are inversely proportional to the probability of observing the word in a document. Consequently, common and content-free words are down-weighted while the contribution of rare and informative words is increased. In the case of a streaming corpus, where new documents are being observed and the corpus is changing, an approximation must be made to these weights. Additionally, the lexicon cannot be pre-determined at the start of the task and so the dimensions of the vector space cannot be assigned to specific words. This work introduces a generalization to the vector space model that allows the model to be used in the environment of a growing corpus of documents. The document frequency of words is approximated and a changing lexicon of words is managed. Results are shown on an example corpus of news articles.

Interactive Text Mining with Iterative Denoising

Kendall Giles, Virginia Commonwealth University

The tasks of proper data analysis and knowledge extraction in text are beset by multiple difficulties when the datasets are large and in high dimension. From a performance perspective, it can be prohibitively expensive to search in a high dimensional space. Also, complex datasets often have local relationships of interest, findings that might be missed with global searches. While some progress has been made with addressing Curse of Dimensionality issues, traditional data mining algorithms largely take a static approach to the data mining process —- simply tabulating the outputs of a particular algorithm from a given input, leaving the user to start the process over again with new inputs if another run is desired. With this static approach, the user is prevented from interacting with the data mining algorithm as well as with the data. In an effort to allow the user to dynamically analyze their data, we present our methodology called Iterative Denoising, which is a statistical pattern recognition framework for analyzing complex datasets. An important realization of our methodology is that users may want to interact with visualized representations of their data. We not only provide to the users lower-dimensional-space representations to highlight (possibly) desired structures in the data, but we also allow the user to interact with the data through an explicit interaction step. For example, the user may wish to change the displayed geometry relationships between objects, say to reflect some metadata intelligence the user has received that is not reflected in the original data. We highlight these contributions with examples from analysis of text data.

Cross Corpus Discovery via Nearest Neighbor Change-point Analysis

Avory Bryant, Naval Surface Warfare Center
Jeffrey Solka, Naval Surface Warfare Center

This talk will explore our recent work on latent semantic indexing updating. This works examines the effect on nearest neighbor structure of corpus combinations. We will also discuss the exploitation of this information to formulate a strategy for literature-based discovery. The developed strategy will be illustrated on several small corpora collections.

CONTRIBUTED PAPER SESSION 4

A Triangle Test for Equality of Distribution Functions in High Dimensions

Zhenyu Liu, George Washington University
Reza Modarres, George Washington University

A new nonparametric test statistic is proposed for testing the equality of the two multivariate distributions by comparing their interpoint distances. Given two p-dimensional random samples X and Y**,** a triangle can be formed by randomly selecting one data point from the X sample and two data points from the Y sample or one data point from the Y sample and two data points from the X sample. Our test statistic estimates the probabilities that the distance of two data points from the same distribution is the smallest, the middle or the largest in the triangle. We show that the test statistic is asymptotically distribution-free under the null and consistent against general alternatives. The powers of the test for some alternatives are compared to some other nonparametric tests through a simulation study. Since the computational complexity of our statistic is independent of the dimension, it is suitable for high dimensional data, especially for cases when p>n.

* * *

A Bayesian Model for Multivariate Functional Data

Ori Rosen, University of Texas at El Paso
Wesley Thompson, University of Pittsburgh

We propose a method for analyzing multivariate functional data with unequally spaced observation times that may differ among subjects. Fitting multivariate observations simultaneously rather than fitting each variable separately may be advantageous if the error terms corresponding to each variable are correlated. Our method is formulated as a Bayesian mixed-effects model in which the fixed part corresponds to the mean functions, and the random part corresponds to individual deviations from these mean functions. Covariates can be incorporated into both the fixed and the random effects. The methodology is studied by simulation and illustrated with real data.

Confidence Estimation of the Parameter Involving in the Distribution of the Total Time on Test for Censored Data

Shih-Chuan Cheng, Creighton University

Data mining has captured the interest of researchers from many different as well as diverse fields of study such as data base systems, machine learning, statistics, cluster analysis, knowledge based systems. One of the major issues in data mining is the analysis of attribute relevance [8]. The basic idea is to come up with a measure that can be effectively used to quantify the relevance of an attribute in identifying a class or a concept. One possible application is the total failure time of censored data [5]. That is, the total time on test for censored data of a system with several components often plays an important role in the reliability theory. In some cases, all components of a system with several identical components may be put on test until an r-th smallest failure time occurs and the total time on test is subsequently calculated.

The total time on test for censored data from exponentially distributed censored data has been proved to be an adequate statistic by Nair and Cheng [11] in light of the works by Skibinsky [12], Cheng and Mordeson [3], and others [1, 9, 13-15]. The test may be repeated for many times. As a result, since the total time on test until the r-th ordered failure time is observed will be recorded for each test, several total time on test (for censored data) for the same system are available for use in analyzing the reliability of the system.

In this proposal, we are investigating the interval estimation (confidence interval) of the parameter of underlying probability distribution for total failure time of censored data based upon the famous Cramér-Rao lower bound.

* * *

LifeStats: An Interactive Environment for Teaching Statistics

E. James Harner, West Virginia University
Dajie Luo, West Virginia University
Jun Tan, West Virginia University

LifeStats, a Java application (http://javastat.stat.wvu.edu), has an extensive list of statistical tools and data analysis capabilities which make it an innovative system for learning statistics. These include simulators and various interactive tools for learning statistical concepts. The centerpiece is the Five-Step procedure for simulating theoretical and data-based probability modes. However, other interactive tools are available for manipulating probability distributions and for illustrating sampling, the central limit theorem, confidence intervals, power and other statistical concepts. The data analysis components are dynamic and offer extensive features for problem solving. The data table is dynamically linked to the various analyses windows so that students can hi-lite or mask observations by a single click or by brushing. All the standard analyses can be done in a highly interactive way, which is conducive to exploration and learning. LifeStats uses XML to save and retrieve data files, which can be located remotely on a server.

The components of LifeStats are also available as Java applets (using the same LifeStats code base). This allows the above statistical tools to be popped up from a pdf-based interactive textbook (written using LaTex). JavaStat, which is based on the interactive data analysis and graphical components of LifeStats, acts as a front-end to R. The objective is to bring certain high-level functions of R to JavaStat without excessive duplicative effort. Results returned from R are wrapped and then displayed using dynamic graphics in JavaStat.

LifeStats can access and display Web content from IDEAL (Intelligent Distributed Environment for Adaptive Learning), a Web-based learning environment. Although LifeStats and IDEAL (http://ideal.stat.wvu.edu) are standalone systems, their integration makes a more powerful whole. LifeStats acts within a full client-server environment, which allows students to access network resources, e.g., datasets, examples, exercises, quizzes, exams, and assessment information.

1:45-3:30, Friday

STATISTICS AND EVOLUTIONARY BIOLOGY, I

Organizer: Haipeng Shen

Evolutionary Analyses of Function-valued Traits

Joel Kingsolver, University of North Carolina at Chapel Hill

* * *

Quantification of Curves' Variation and Simplicity to Find Genetic Constraints

Travis Gaydos, University of North Carolina at Chapel Hill

* * *

Extending Models of Character Coevolution

Brian O'Meara, National Evolutionary Synthesis Center

SENSOR NETWORKS AND STATISTICS – NEW RESEARCHERS SESSION

Organizer: George Michailidis

Fault Detection for Embedded Networked Sensing

Sheela Nair, University of California – Los Angeles

Recent advances in sensor technology, computing, and low-power communications have facilitated the development of embedded networked sensing (ENS). However, a major issue that limits the widespread use of ENS is the quality of sensor data, which are compromised by various faults and anomalies. Fault detection systems for sensor networks face a unique combination of statistical and computing challenges. The nature of the data collection as well as the processes being studied lead to large amounts of highly complex, non-stationary data. In addition, algorithms have computational and communication constraints. During a deployment, it is often desirable to assess the quality of data being collected by the network in near-real time, and fault detection algorithms must be both computationally feasible and scalable. We propose a signature-based approach for fault detection, borrowing on similar methods used successfully in fraud detection. An algorithm that adaptively estimates and tracks sensor signatures under both typical and faulty sensor behavior will be described. We will also outline current work in spatial-temporal modeling and the corresponding challenges within the computational context of sensor networks. Examples of ongoing environmental ENS deployments will be used to illustrate the problems and methodology.

A Cost-efficient Approach to Wireless Sensor Network Design

Natallia Katenka, University of Michigan

This study presents a general flexible approach for the design of wireless sensor network under the random deployment mechanism. The cost of sensing and communications is incorporated into the design of the network, while in addition allowing for unreliable sensors. In the proposed approach, cost is treated generically and can correspond to either a fixed acquisition cost, or an operational cost or a combination of both. The main objective is to minimize the overall network cost, while enforcing the coverage and connectivity constraints. The proposed approach should be regarded as part of any feasibility study during the planning stages for the deployment of a wireless sensor network, when decisions about its capabilities and cost are considered. An additional technical contribution is the derivation of a new simple bound on the probability of a network being connected, which exhibits a very good performance in simulations and unlike existing ones is shown to be better suited for network design studies.

* * *

Analyzing Space-time Sensor Network Data under Suppression and Failure in Transmission

Gavino Puggioni, Duke University

In this paper we present a fully model-based analysis of the effects of suppression and failure in data transmission with sensor networks. Sensor networks are becoming an increasingly common data collection mechanism across a variety of fields. Sensors can be created to collect data at very high temporal resolution. However, during periods when the process is following a stable path, transmission of such high resolution data would carry little additional information with regard to the process model, i.e., all of the data that is collected need not be transmitted. In particular, when there is cost to transmission, we find ourselves moving to consideration of suppression in transmission. Additionally, for many sensor networks, in practice, we will experience failures in transmission - messages sent by a sensor but not received at the gateway, messages sent but arriving corrupted. Evidently, both suppression and failure lead to information loss which will be reflected in inference associated with our process model. Our effort here is to assess the impact of such information loss under varying extents of suppression and varying incidence of failure. We consider two illustrative process models, presenting fully model-based analyses of suppression and failure using hierarchical models. Such models naturally facilitate borrowing strength across nodes, leveraging all available data to learn about local process behavior.

Contributed Paper Session 5

Statistically Modeling the Performance of a Multistart Randomized Heuristic Algorithm

Vincent Cicirello, Richard Stockton College

The complexity of many combinatorial optimization problems often preclude the application of exact problem solving techniques as the problem is scaled to real-world size. For example, for a problem that is NP-Hard in the strong sense, any algorithm that guarantees that its solutions are optimal is going to be limited in the size of the instance that it can solve given computational limitations. Rather than requiring "optimal" solutions, an alternative is to favor sufficiently good solutions that can be found efficiently. One approach to this uses a randomized heuristic algorithm. This paper specifically considers an algorithm known as Value Biased Stochastic Sampling (VBSS). VBSS uses a heuristic function to bias the random generation of a proposed solution to the problem. This process is repeated some number of times and retains only the best solution found over several trials. The key to the effectiveness of VBSS is the choice of the heuristic function. The job of the heuristic function is to provide problem dependent evaluation of the choices that can be made. This evaluation is used to bias the random decisions during the problem solving process. At the time of the design of such an algorithm, it may not be obvious which of several heuristics are best. During the past few years, we have been developing techniques for allowing VBSS to self-select which heuristic to use. Our approach relies on modeling the distribution of the quality of solutions obtained by VBSS over its allocated set of restarts. This paper presents some of our analysis of potential models, including goodness of fit tests for several possible assumptions we can make about the distribution of the quality of solutions. Our analysis leads us to the selection of the Generalized Extreme Value Distribution for our models of randomized heuristic performance.

Keeping a Search Engine Index Fresh: Risk Versus Optimality Trade-offs in Estimating Frequency of Change in Web Pages

Eric Tassone, Google
Carrie Grimes, Google

Search engines strive to maintain a "current" repository of all web pages on the internet to index for user queries. However, refreshing all web pages all the time is costly and inefficient: many small websites don't support enough load, and while some pages update content frequently, others don't change at all. As a result, estimated frequency of change is often used to decide how frequently a web page needs to be refreshed in an offline corpus. Here we consider a Poisson process model for the state changes of a page, where a crawler samples the page at some known (but variable) time interval and observes whether the page has changed in during that interval. Under this model, we first estimate the rate of change for the observed intervals using a Maximum Likelihood estimator described in Cho and Garcia-Molina (2000), and test the model on a set of 100,000 web pages by examining the outcome of a refresh policy based on these estimates. Second, we consider a constantly evolving set of web pages, where new pages enter the set with no information and estimation must begin immediately, but where we control the ongoing sampling of the page. In this setting, the refresh efficiency gained by an accurate estimator trades off with the risk of a page not being fresh due to incomplete information. In addition, the size of a search engine corpus (potentially billions of pages) requires that any estimator be computationally inexpensive. We implement a computationally simple empirical Bayes estimate that improves initial estimation. We demonstrate that the initial page sampling strategy that minimizes the risk of a stale corpus directly precludes optimal strategies for acquiring information to improve the accuracy of estimated rate of change and consider alternate strategies for initial estimation.

3:45 pm – 5:30 pm, Friday

STATISTICS AND EVOLUTIONARY BIOLOGY, II

Organizer: Haipeng Shen

Distribution of Mutation Effects and Adaptation in an RNA Virus

Christina Burch, University of North Carolina at Chapel Hill

* * *

Deconvolution and Sieve Estimation of Mutation Effect Distribution

Mihee Lee, University of North Carolina at Chapel Hill

* * *

Modularity in Biological Systems: Statistical Challenges and Evolutionary Insights

Paul Magwene, Duke University

ASSESSING HEALTH RISK FROM COMPLEX DATA

Organizer: David Dunson, Duke University

A Bayesian Hidden Markov Model for Motif Discovery through Joint Modeling of Genomic Sequence and ChIP-chip Data

Joseph Ibrahim, University of North Carolina at Chapel Hill

We propose a unified framework for the analysis of Chromatin (Ch) Immunoprecipitation (IP) microarray (ChIP-chip) data for detecting transcription factor binding sites (TFBSs) or motifs. ChIP-chip assays are used to focus the genome-wide search for TFBSs by isolating a sample of DNA fragments with TFBSs and applying this sample to a microarray with probes corresponding to tiled segments across the genome. Present analytical methods use a two-step approach:(i) analyze array data to estimate IP enrichment peaks then (ii) analyze the corresponding sequences independently of intensity information.The proposed model integrates peak finding and motif discovery through a unified Bayesian hidden Markov model (HMM) framework that accommodates the inherent uncertainty in both measurements. A Markov Chain Monte Carlo algorithm is formulated for parameter estimation, adapting recursive techniques used for HMMs. In simulations and applications to a yeast RAP1 dataset, the proposed method has favorable TFBS discovery performance compared to currently available two-stage procedures in terms of both sensitivity and specificity.

Analysis of Left-truncated Semi-competing Risks Data with Application to Disease Registries

Jason Fine, University of North Carolina at Chapel Hill

Semi-competing risks data occurs when an event of interest may be dependently censored by another event, but not vice versa. Such data are commonly encountered with chronic diseases, where co-morbidities may be potentially dependently censored by death. In the Denmark diabetes registry, a further complication arises in the analysis of nephropathy, a key disease landmark. There is both dependent censoring and left truncation on mortality. That is, only those individuals living long enough to be referred to the Steno hospital in Copenhagen enter the registry. In this talk, we consider two analyses of such data: one based on a semiparametric model for the marginal distribution of nephropathy and the second based on a nonparametric estimator of the cumulative incidence of nephropathy. Novel methodology addressing key issues in the data will be presented for each analysis, along with some practical discussion of the relative merits of the two approaches.

* * *

Semiparametric Bayes Modeling of Onset and Progression from Current Status Data

Lianming Wang, National Institute of Environmental Health Sciences

This article proposes a semiparametric Bayes approach for inference on onset and progression of disease based on cross sectional screening data. In particular, data for an individual consist of the current disease status and a continuous measurement of disease progression that provides a surrogate for the waiting time in the disease state. We propose an approach that avoids parametric assumption on the baseline time to disease onset and the density of the surrogate as a function of time since onset. This is accomplished using a Dirichlet process mixture model for the onset time distribution, and a restricted dependent Dirichlet process mixture for the surrogate. The latter model allows the surrogate to have an unknown density that increases stochastically with time since onset. We allow covariates to impact rates of onset and progression, with the onset time model having an accelerated failure time form and the progression model log-linear. Efficient methods are developed for posterior computation using a blocked Gibbs sampling algorithm. The method is assessed through a simulation study, and applied to an epidemiologic study of uterine fibroid tumors.

INTEGRATION OF DISPARATE TYPES OF INFORMATION

Organizer: Wendy Martinez, Office of Naval Research

Disparate Information Fusion: On the Exploitation of Multiple Disparate Dissimilarities

Carey Priebe, Johns Hopkins University

* * *

Combining Disparate Information by Nonmetric Multidimensional Scaling

Brent Castle, Indiana University
Michael Trosset, Indiana University

Consider the problem of classifying objects for which two (or more) very different kinds of information are available, e.g., text and images. Using each type of information, one can measure dissimilarities between pairs of objects; however, the different measures of dissimilarity are not comparable. To combine information, we normalize each dissimilarity matrix by replacing numerical values with ranks. We then study three strategies for constructing the representation space in which classification will be performed: (1) Combine ranks, then embed the objects by 2-way nonmetric MDS; (2) Use each set of ranks separately to embed the objects by 2-way nonmetric MDS, then form the product of the separate representations; (3) Use each set of ranks simultaneously to embed the objects by 3-way nonmetric MDS.

* * *

Disparate Information Fusion on Images and Text

Jeffrey Solka, Naval Surface Warfare Center

8:30-10:15, Saturday

SPATIAL RISK MAPPING:
PREDICTION AND CHANGE DETECTION

Organizer: Michael Porter, North Carolina State University

Space-time Forecasting of Extreme Events in Complex Environments

Jason Dalton, SPADEC

* * *

Using the Repeated Two-sample Rank Procedure for Detecting Anomalies in Space and Time

Ronald D. Fricker, Jr., Naval Postgraduate School

The Repeated Two-Sample Rank (RTR) procedure is a nonparametric statistical process control methodology that applies to both univariate and multivariate data. The method transforms a sample of data into univariate statistics; changes in the distribution of the data are then detected using nonparametric rank tests. In this discussion we explore its use as a spatio-temporal event detection and monitoring methodology for use in applications such as biosurveillance, crime mapping, or IED incidence change detection. The methodology is designed to sequentially incorporate information from individual observations as they arrive into an automated systems and thus can operate on data in real-time. Upon a signal of a possible distributional change, the methodology suggests a way to graphically indicate the likely location of the distributional change.

* * *

A Martingale Methodology for the Quick Identification of Point Process Anomalies

Michael Porter, North Carolina State University

A method is introduced to detect anomalies in space-time and higher dimensional point processes. Point process residuals are used to construct martingale statistics that are incorporated into common change detection procedures. These martingales can detect unusual observations (or regions of unusual activity) that may indicate changes to the baseline process. This procedure is robust to the type of change and can operate in real-time. The problem of detecting such changes is applicable in areas such as disease surveillance, computer intrusion detection, terrorist networks, intelligent site selections, and crime and terrorism.

CONTRIBUTED PAPER SESSION 6

Classification Trees with Oblique Splits for Multidimensional Datasets

Andrejus Parfionovas, Utah State University
Adele Cutler, Utah State University

We are developing an enhanced modification of the tree classification algorithm for high-dimensional datasets. Differently from the classical tree-based methods which focus on one variable at a time to separate the observation, we propose to perform the search for the best split in two-dimensional space formed by a linear combination of variables. Our theoretical investigation and the numerical simulations has demonstrated a number of improvements, such as: better accuracy and precision of the classification; smaller size of the trees; higher flexibility and adjustments to the complicated structure of the data. In addition, the trees with oblique splits provide a closer look to the data structure and its relationship to the classification process. The research is currently in progress and more results might be available by the time of the presentation.

* * *

Clustering with Confidence: A Binning Approach

Rebecca Nugent, Carnegie Mellon University
Werner Stuetzle, University of Washington

The goal of clustering is to identify distinct groups in a dataset and assign a group label to each observation. To cast clustering as a statistical problem, we regard the data as an i.i.d. sample from some unknown underlying probability density p(x). In nonparametric clustering, we adopt the premise that groups correspond to modes of the density. Our goal then is to find the modes and assign each observation to the ``domain of attraction" of a mode. We do this by estimating the cluster tree of p(x), a representation of the hierarchical structure of the level sets of the density. Leaves of the cluster tree correspond to modes of the density. An obvious way of estimating the cluster tree of the underlying density p(x) from a sample is to first compute a density estimate and then use its cluster tree as an estimate for the cluster tree of p(x). However, due to noise in the density estimate, the cluster tree estimate likely will not generate a one-to-one mapping of clusters to groups in the population. Spurious clusters may be falsely identified as groups. We introduce a bootstrap-based simultaneous confidence band that can be used in conjunction with cluster tree estimation methods to assign a level of significance to each cluster as well as to the estimated cluster tree. Clustering with Confidence could also be used as an automatic pruning procedure given the user input of a desired confidence level. Results for a low-dimensional binning approach that exactly computes level sets for piece-wise constant density estimates and, time permitting, a graph-based estimation approach, generalized single linkage, for use in high dimensions will be shown.

Making Tree Ensembles More Robust to Noisy Data

Joran Elias, University of Montana

Tree ensembles have proven to be a powerful and commonly used classifier in a wide range of applications. This is due, in part, to their documented robustness to noise in the class labels. For certain two-class classification problems it is common to have reliable data for only one of the two classes. Examples include species presence/absence data and document retrieval. In these situations one class is known fairly reliably (species presences, user flagged documents) while the other class inevitably contains a certain number of mislabeled observations (species absences, documents whose interest to the user is unknown). We hypothesize that a simple resampling strategy can vastly improve the robustness of tree ensembles to large amounts of one-sided noise in the class labels. This hypothesis is verified with a simulation study on several real and simulated data sets that demonstrates the effectiveness of this resampling strategy.

10:30 am – 12:15 pm, Saturday

CHANGE DETECTION IN RANDOM GRAPHS

Organizer: David Marchette, Naval Surface Warfare Center

Detecting Activity Changes in Graphs

David Marchette, Naval Surface Warfare Center

* * *

Scan Statistics in Hypergraphs

Youngser Park, Johns Hopkins University

* * *

Torus Graph Inference for Detection of Localized Activity

Elizabeth Beer, Johns Hopkins University

RISK OF REACHING FALSE CONCLUSIONS

Organizer: S. Stanley Young, National Institute of Statistical Sciences

The Problem of Observational Studies

S. Stanley Young, National Institute of Statistical Sciences

* * *

A Complete Illustration of Local Control for Observational Studies

Robert Obenchain, SoftRx

* * *

Exploring the Effects of Medicines: Managing Risk across Multiple Outcomes

Patrick Ryan, GlaxoSmithKline

* * *

Alice White, GlaxoSmithKline - Discussant