Interface 2012 ABSTRACT BOOK

43rd Symposium on the Interface of
Computing Science and Statistics

Theme:

Future of Statistical Computing:
Internet Scale Data, Flexible Modeling, and Visualization

May 16-18, 2012

Wednesday 8 am - Friday noon

Program Co-Chairs:

David W Scott, Rice University
Hadley Wickham, Rice University
Jeffrey Morris, MD Anderson

(Typeset May 18 at 1:00 pm)

Wednesday, May 16, 2012

8:00 am - 9:45 am Welcome and Keynote Address Auditorium

- **Keynote Address** Auditorium

Matrix Completion and Large-Scale SVD Computations
Trevor Hastie, Stanford University

The Singular Value Decomposition (SVD) is a fundamental tool in all branches of data analysis - arguably one of the most widely used numerical tools. Over the last few years, partly inspired by the "Netflix problem", the SVD has again come into focus as a solution to the "matrix completion" problem. One partially observes a very large matrix, and would like to impute the values not observed. By assuming a low-rank structure, the SVD is one approach to the problem - a SVD with large amounts of missing data. In this talk we discuss an approach for building a path of solutions of increasing rank via nuclear-norm regularization. An integral part of this algorithm involves repeatedly computing low-rank SVDs of imputed matrices. We show how these tasks can be efficiently handled by parallel computational algorithms, allowing the method to scale to very high-dimensional problems.

10:15 am - 12:00 noon Technical Sessions

- **Software Development in R** Auditorium
  Organizer: Duncan Murdoch, University of Western Ontario

RStudio - Integrated Development Environment for R
JJ Allaire, Founder RStudio Project

RStudio is a free and open source integrated development environment (IDE) for R. It includes a console, R syntax-aware code editor, and integrated plotting, history, workspace, and help components. RStudio also includes facilities for easily switching between multiple R projects and integrates with other tools such as TeX and version control. It is designed to both ease the learning curve for new R users as well as provide high productivity coding tools for more advanced users. RStudio works on Windows, Mac, and Linux systems and can also be deployed as a server to enable web access to R sessions running on remote systems.

Efficient R Parallel Loops on Long-Latency Platforms
Norm Matloff, University of California Davis

Many statistical computations take the form of loops with independent iterations. Much study has been conducted on the parallel processing of such loops, concerning tradeoffs between overhead and load balance, but most has been in the shared-memory setting. The present research concerns message-passing settings, not only in traditional cluster environments, but in newer paradigms such as the cloud and GPU. Focus will be on parallel vehicles for R, particularly Snow. A new type of random assignment of iterations will also be developed.

Debugging Support in R
Duncan Murdoch, University of Western Ontario

R maintains "source references" as it parses code. These are visible to the main R evaluator, so that source-level debugging is possible. For example, the setBreakpoint() function can specify a line in a source file, and R will break execution when it reaches code that originated from that line. In this talk I'll describe the internals and some of the built-in functions that make use of them, as well as plans for future improvements. If time permits I will also mention how the source references are used in Sweave to help in debugging the LaTeX documents that it creates.

- **Using Statistical Modeling for Pricing Applications** Room 1064
Organizer: John Salch, PROS Revenue Management, Inc

The Complexity of Petroleum Pricing at PROS
Daniel Covarrubias, PROS Revenue Management, Inc

PROS has implemented pricing solutions across multiple channels within the petroleum industry. In this talk we give an overview of the challenges and solutions related to two of those channels, Branded and Unbranded Petroleum Rack Pricing. We briefly describe the underlying complex nature of petroleum pricing in a competitive environment, and the process of developing appropriate forecasting models that capture the interaction amongst those competitors. We present the results of model evaluation for both channels and the significance of proper outlier identification in a real-time pricing environment. We conclude with an extension of the Branded and Unbranded Rack Pricing methodology to a third petroleum channel, Dealer Tank Wagon Pricing.

Segmentation for Extremely Large Datasets
Evan Brott, PROS Revenue Management, Inc

Segmentation the process of grouping similar data elements together can be of great value for forecasting and decision support. Although a wide variety of segmentation

algorithms are commonly used, most run into considerable difficulty as the amount of data they must sift through increases often becoming extremely inefficient or outright unusable for problem sizes encountered in common business applications. In this talk, we will present an algorithm developed for our airline customers who have thousands of potential segmentation variables and tens of millions of passenger records that has been shown to greatly increase forecast efficacy for multiple carriers.

A New Approach for Forecasting Demand via Willingness to Pay
Ed Kambour, PROS Revenue Management, Inc

Historically demand forecasts have been based on historical bookings and availability. We will be presenting a new model to estimate customer willingness to pay and demand elasticity, based solely on historical bookings and purchase prices. We will examine the underlying statistical model, along with showing some real world examples.

- **High-Dimensional Graphical Models** Room 1070
Organizer: Genevera Allen, Baylor College of Medicine and Rice University

A Log-Linear Graphical Model
Genevera Allen, Baylor College of Medicine and Rice University

Sparse Gaussian graphical models have become a popular way to visualize, model, and understand relationships in high-dimensional data. For high-dimensional count data, arising for example from text mining, call-logs, site visits, and RNA sequencing, these Gaussian graphical models are not appropriate. In this paper, we propose a Log-Linear Graphical Model to estimate sparse Poisson graphical models from high-dimensional count data. The model assumes that conditional on all other nodes, each node is Poisson distributed. We fit our model via neighborhood selection using $L1$ penalized log-linear models and develop a fast parallel algorithm to infer high-dimensional networks. Through simulations and a novel application to microRNA sequencing networks, we demonstrate the effectiveness of our methods.

Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification
Tyler McCormick, University of Washington

We propose an online binary classification procedure for cases when there is uncertainty about the model to use and parameters within a model change over time. We account for model uncertainty through Dynamic Model Averaging (DMA), a dynamic extension of Bayesian Model Averaging (BMA) in which posterior model probabilities may also change with time. We apply a state-space model to the parameters of each model and we allow the data-generating model to change over time according to

a Markov chain. Calibrating a "forgetting" factor accommodates different levels of change in the data-generating mechanism. We propose an algorithm which adjusts the level of forgetting in an online fashion using the posterior predictive distribution, and so accommodates various levels of change at different times.

We apply our method to data from children with appendicitis who receive either a traditional (open) appendectomy or a laparoscopic procedure. Factors associated with which children receive a particular type of procedure changed substantially over the seven years of data collection, a feature that is not captured using standard regression modeling. Because our procedure can be implemented completely online, future data collection for similar studies would require storing sensitive patient information only temporarily, reducing the risk of a breach of confidentiality. This is joint work with David Madigan, Adrian Raftery and Randall Burd.

Greedy Algorithms for Learning Discrete and Gaussian Graphical Models

Pradeep Ravikumar, University of Texas Austin

Undirected graphical models, also known as Markov random fields, are widely used in a variety of domains, including biostatistics, natural language processing and image analysis among others. They compactly represent distributions over a large number of variables using undirected graphs which encodes conditional independence assumptions among the variables. Recovering this underlying graph structure is thus important for many of these applications of MRFs, especially under constrained settings where the number of variables is large, and the samples are limited.

In this talk, we address the problem of learning the structure of a pairwise graphical model from samples in a high-dimensional setting. Classical approaches to this problem have ranged over various heuristic approaches, greedy methods being one class of them, but a line of recent research have proposed principled convex regularization based M-estimators, that have been shown to be not only computationally practical, but to also enjoy strong statistical guarantees. In this talk, we revisit a classical and simple greedy algorithm, that just iteratively adds and deletes edges. Surprisingly, we show that when these forward and backward steps are performed appropriately, we obtain a state of the art method for recovering graphical model structure. Indeed, not only does it enjoy obvious computational advantages over regularized convex optimization based approaches, but we also show that it is sparsistent, or consistent in sparsity pattern recovery, under weaker conditions, and with a smaller sample complexity.

Joint work with Ali Jalali, Christopher Johnson.

1:30 pm - 3:15 pm Technical Sessions

- **Man AND Machine: the Conversation, using the Language of Interactive Graphics** Auditorium

  Organizer: Heike Hofmann, Iowa State University

  Interactive Graphics in R: The Plumbing and the Painting
  Michael Lawrence, Genentech

  Adding support for interaction to a graphic requires an extra level of attention paid to both the design of the graphic itself and the design of the software underlying the graphic. This talk is concerned with the latter, in particular software infrastructure for efficient, scalable interactive graphics. The implementation of an interactive visualization consists of both dynamic drawing and dynamic computation (filtering, summarizing, transforming, etc), and a graphics system needs to scale in both respects. For drawing, we have developed a hardware-accelerated renderer, with multi-layered buffers for incremental updating and efficient spatial algorithms for mapping user actions to the data. It is low-level and applicable to a variety of graphics problems. Towards scalable computation, we have applied the model-view-controller pattern to the design of computational pipelines that operate dynamically on subsets of the data and make feasible the flexible and rapid development of algorithms implemented in the R language.

  cranvas: Building from Plumbing and Painting
  Yihui Xie, Iowa State University

  A long missing feature in R graphics systems is the support of interactivity, and the root reason is the lack of an underlying data pipeline. That is, only the painting model was implemented. Once a plot is drawn, it loses connection with the original data object. We introduce the concept of mutable data objects, on which events can be attached so that whenever there are changes in the data, these events can be executed. Interaction on plots often involves with modifying attributes of underlying data, and we can use a series of events to update plot layers. The data object serves as a central commander which may send signals to multiple plots at the same time, so that plots can communicate through a pipeline with several steps plumbed together. The other problem of R graphics systems is efficiency. We often have to redraw the whole plot if we want to update a part of it. Qt is a comprehensive toolkit which has a powerful graphics system. It supports event callbacks on plots, and plots can be constructed with layers. Most importantly, layers can be cached so they are not updated if not desired when we interact with plots. Qt also has a significant speed advantage over base R graphics systems.

  In this talk we introduce a new R package "cranvas", which is built upon the plumbing ideas and new painting engines. The infrastructure is based on other R packages plumbr, objectSignals (for plumbing), qtbase, qtpaint (for painting) and a series of

ggplot2-related packages like scales (for aesthetics). We will show examples of usage and our future plans towards a grammar of interactive graphics.

Enhancing Web Pages with R in the Browser
Gabriel Becker, University of California-Davis

The Web browser is likely to play an increasingly important and dominant role in disseminating content (text, graphics, video, etc.) in the future. Many analysts, content authors and data curators (e.g. NY Times, Oakland Crime `http://oakland.crimespotting.org/`, Gapminder [http://www.gapminder.org]) are already using it to deliver rich, interactive Web pages which engage readers and allow them to explore the data, analyses and results being shown. In this talk we present the WebR `http://www.omegahat.org/WebR` framework which integrates the R interpreter into a Web browser. WebR implements a bi- directional interface between the R and JavaScript languages which offers three main capabilities. First, R functions and packages can be invoked via JavaScript at the viewing time of a document, allowing for a much richer set of computations to be incorporated into a dynamic, interactive Web page. Secondly, developers can embed one or more active R graphics devices within a Web page, update them asynchronously and interact with the elements of each plot. Finally, R code can be included within an HTML page as a replacement for JavaScript and can be used to dynamically manipulate the page and respond to user interaction. These features not only facilitate the creation of the advanced displays we see today, but also provide the framework for qualitatively new ways of disseminating research results that allow the reader to explore a data analysis in its entirety including actual computations, results and research process.

- **Unifying Statistical Sciences** Room 1064
  Organizer: Arnie Goodman, Collaborative Data Solutions

Modeling, Dependence, Classification, United Statistical Science, Many Cultures
Manny Parzen, Texas A&M University

Two fundamental problems of Modern Applied Statistics are the dependence problem and the classification-dependence problem. A basic dependence problem observes many variables (features); we seek to identify (select) which pairs of variables are most dependent, and on the scatter plot of each pair display non-parametrically computed conditional mean and conditional quantile. A basic classification dependence problem observes (Y, X) where Y is binary and X can have p variables; we seek rules to non-parametrically predict (classify) Y from X. Our approach will be outlined by discussing following topics: Four aspects for research to have impact; many cultures of united statistical science; modeling (X, Y), unification of discrete and continuous variables; comparison probability, copula, Bayes theorem; Classification, hepatitis data example; Dependence, detect novel association; Comparison

density, copula density of (X, Y), score functions; Conditional mean $E[g(Y)|X]$ estimation; LP comoments $\mathrm{LP}(j, k; X, Y)$ to identify most dependent variables; Logistic regression, maximum entropy density estimation.

Discussion
Don Ylvisaker, UCLA

Comments on paper by Manny Parzen.

Discussion
Joe Newton, Texas A&M University

Comments on paper by Manny Parzen.

- **Tensors: Decompositions and Applications** Room 1070
  Organizer: Eric Chi, UCLA

Nonnegative Tensor Factorizations
Eric Chi, University of California, Los Angeles

Tensors have found application in a variety of fields from signal processing to bioinformatics and neuroimaging. In the latter two examples, data is nonnegative and estimating nonnegative multilinear models can yield more interpretable physical model by representing the data as a sum of nonnegative components. Algorithms for estimating such models abound, and counted among the throng are tensor extensions of the popular multiplicative Lee-Seung (LS) nonnegative matrix factorization algorithm. LS algorithms remain popular due to their simplicity. Yet is is well known that they can converge to non-stationary points. Instead of resolving the convergence issue, however, efforts have focused on developing faster algorithms with alternative strategies for which convergence can be proven. Nonetheless in practice LS algorithms are competitive with respect to wall-clock speed. Because of their simplicity and performance, the convergence issues of the LS approach warrant further investigation. We introduce a block-coordinate descent nonnegative tensor factorization algorithm for which the LS algorithm is a special case. Our algorithm converges to KKT points under mild conditions. We demonstrate the method on both real and synthetic data.

Tensor Regression with Applications in Neuroimaging Data Analysis
Hua Zhou, North Carolina State University

Classical regression methods treat covariates as a vector and estimate a corresponding vector of regression coefficients. Modern applications in medical imaging generate covariates of more complex form such as multidimensional arrays (tensors). Traditional statistical and computational methods are proving insufficient for analysis of

these high-throughput data due to their ultrahigh dimensionality as well as complex structure. We propose a new family of tensor regression models that efficiently exploit the special structure of tensor covariates. Under this framework, ultrahigh dimensionality is reduced to a manageable level, resulting in efficient estimation and prediction. A fast and highly scalable estimation algorithm is proposed for maximum likelihood estimation and the asymptotics are studied. Regularized tensor regression will also be discussed. Effectiveness of the new methods is demonstrated on both synthetic and real MRI imaging data.

GSVD Comparison of Cancer Patient-Matched Genomic Profiles Predicts Survival and Novel Drug Targets

Preethi Sankaranarayanan* and Orly Alter, University of Utah

Despite recent large-scale profiling efforts, the best prognostic predictor of glioblastoma multiforme (GBM) remains the patient's age at diagnosis. We describe a global pattern of tumor-exclusive co-occurring DNA copy-number alterations (CNAs) that is correlated, possibly coordinated with GBM patients survival and response to chemotherapy (Lee, Alpert, Sankaranarayanan and Alter, PLoS One 2012). The pattern is revealed by GSVD comparison of patient-matched but probe-independent GBM and normal copy-number profiles from The Cancer Genome Atlas. We find that, first, the GSVD, formulated as a framework for comparatively modeling two composite datasets, removes from the pattern, copy-number variations that occur in the normal human genome (e.g., female-specific X chromosome amplification) and experimental variations, without a-priori knowledge of these variations. Second, the pattern includes most known GBM-associated changes in chromosome numbers and focal CNAs, as well as several previously unreported CNAs including the biochemically putative drug target, cell cycle-regulated kinase-encoding TLK2. Third, the pattern provides a better prognostic predictor than the chromosome numbers or any one focal CNA that it identifies, suggesting that the GBM survival phenotype is an outcome of its global genotype. The pattern is independent of age, and combined with age, makes a better predictor than age alone. Recent experimental results verify a computationally predicted genome-wide mode of regulation, and demonstrate that GSVD modeling of DNA microarray data can be used to correctly predict previously unknown cellular mechanisms. This GSVD comparative modeling, therefore, draws a mathematical analogy between the prediction of cellular modes of regulation and the prognosis of cancers.

- **Late Breaking Session** Room 1075

  Organizer: David Scott, Rice University

From Single-SNP to Wide-Locus: Increasing Resolution and Power of GWAS

Knut M. Wittkowski*, Rockefeller University; Vikas Sonakya, Rockefeller University; Tingting Song, Rockefeller University; Martin P. Seybold, Stuttgart

University; Mehdi Keddache, Cincinnati Children's Hospital Medical Center; and Kartina Durner, Mount Sinai School of Medicine

Genome Wide Association Studies (GWAS) have had limited success when applied to complex diseases. Analyzing SNPs individually requires several large studies to integrate the often divergent results. In the presence of epistasis between SNPs, intragenic regions, or genes, multi- variate approaches based on the linear model (including stepwise logistic regression) often have low sensitivity and generate an abundance of artifacts. Recent advances in distributed and parallel processing spurred methodological advances in non-parametric statistics. U-statistics for multivariate data (GWAS) are not confounded by unrealistic assumptions (linearity, independence) and were recently extended to incorporate information about hierarchical data structures. For GWAS, a par- ticular hierarchical structure reflects the sequence of neighboring SNPs and recombination hotspots. This computational biostatistics approach increases power and guards against artifacts, paving the way to comparative effectiveness research and personalized diagnostics. In particular, it can identify clusters of genes around biologically relevant pathways and pinpoint functionally relevant intragenic regions. A study of only 185 children with childhood absence epilepsy and publicly available controls sufficed to integrate previous findings into biologically plausible hypotheses about the interplay of genetic risk factors. While most anti-epileptic drugs target regulatory processes at the level of the nucleus or cell membrane, By moving from single-SNP to wide-locus GWAS, GWAS was able to identified a cluster of genes controlling functional processes in the cytoplasm, suggesting a gene, cur- rently investigated in the treatment of inflammatory diseases, as a potential additional drug target in epilepsy.

Multivariate Spatial Process Modeling: An Overview

William Kleiber, UCAR

Multivariate stochastic processes play an increasingly important role in the geophysical sciences including meteorology, hydrology, earth science, environmental modeling and economics. The key difficulty is in specifying the cross-covariance function, which describes the covariance between different processes across the domain. In this talk, we discuss recent advances in multivariate process modeling, including nonstationary constructions and space-time matrix-valued covariance functions.

Uncertainty in Regional Climate Experiments

Steve Sain, UCAR

Climate models are subject to a number of different sources of uncertainty. Regional climate modeling introduces additional uncertainties associated with the boundary conditions and resolution of the models. In this talk, based on the ensemble being generated as part of the North American Regional Climate Assessment Program

(NARCCAP), we will present statistical methodology for the analysis of the spatial-temporal output in the ensemble and quantifying the uncertainties associated with the NARCCAP experiment.

3:45 pm - 5:30 pm Technical Sessions

- **Statistical Models for Complex Functional Data** Auditorium
  Organizer: Veera Baladandayuthapani, UT MD Anderson Cancer Center

Regression Modeling with Images as Predictors
Todd Ogden, Columbia University

In many biomedical applications it is of interest to use imaging data or other very high dimensional data as predictors in regression models, e.g., to predict a patient's treatment outcome based on brain imaging data obtained at baseline. Obtaining meaningful fits in such problems requires some form of dimension reduction while taking into account the particular structure of the data. This talk will describe some of the tools that have proven effective in this context.

Efficient Spatial Smoothing Over Irregular Domains Using Functional PCA
Lan Zhou, Texas A&M University

In this talk, we consider functional data defined on irregular spatial domains and observations are sparsely sampled. Penalized B-splines on triangulations are used to smooth such data. We develop a functional principal analysis (PCA) method for finding interesting structures in a collection of 2-d functional objects obtained by spatial smoothing. The method is casted into a mixed-effects model framework for parameter estimation and efficient computation. The method is applied to analyze the temperature variation over space and time in Texas using 10 years of temperature data recorded by Texas weather stations.

Bayesian Nonparametric Functional Models for High-Dimensional Genomics Data
Veera Baladandayuthapani, UT MD Anderson Cancer Center

Due to rapid technological advances, various types of genomic, epigenomic, transcriptomic and proteomic data with different sizes, formats, and structures have become available. These experiments typically yield data consisting of high-resolution genetic changes of hundreds/thousands of markers across the whole chromosomal map. Modeling and inference in such studies is challenging, not only due to high dimensionality, but also due to presence of structured dependencies (e.g. serial and spatial correlations). Using genome continuum models as a general principle we present

a class of Bayesian methods to model these genomic profiles using functional data analysis approaches. Our methods allow for simultaneous characterization of these high-dimensional functions using non-parametric basis functions, joint modeling of spatially correlated functional data and detection of local features in spatially heterogeneous functional data to answer several important biological questions. We illustrate our methodology by using several real and simulated datasets and propose methods to integrate various types of genomics data as well. (Joint work with Jeff Morris)

- **Some Data Based Analyses in Real World Finance** Room 1064
  Organizer: James Thompson, Rice University

Estimating the Term Structure With a Semi-Parametric Bayesian Population Model: An Application to Corporate Bonds and Ratings
Katherine B. Ensor*, Rice University; Alejandro Cruz-Marcello, Capital One Bank; and Gary Rosner, Johns Hopkins University

The term structure of interest rates is used to price defaultable bonds and credit derivatives, as well as to infer the quality of bonds for risk management purposes. We introduce a model that jointly estimates term structures by means of a Bayesian hierarchical model with a prior probability model based on Dirichlet process mixtures. The modeling methodology borrows strength across term structures for purposes of estimation. The main advantage of our framework is its ability to produce reliable estimators at the company level even when there are only a few bonds per company. After describing the proposed model, we discuss an empirical application in which the term structure of 197 individual companies is estimated. The sample of 197 consists of 143 companies with only one or two bonds. In-sample and out-of-sample tests are used to quantify the significant improvement in accuracy that results from approximating the term structure of corporate bonds with estimators by company rather than by credit rating, the latter being a popular choice in the financial literature.

Model Specification Error with High-Speed Computational Propagation as Seen in the Subprime Meltdown
John A. Dobelman, Rice University

The "subprime meltdown" is a general term used in the USA to refer to events purporting to be causative in the global financial crisis of 2008-2010. We show the exponential accumulation of model mis-specification errors, and their subsequent near light-speed propagation in computational-based trading strategies, as seen in the global subprime mortgage collateralized debt obligation and credit default swap (CDO/CDS) markets. We trace the development of a new financial species, the CDS on CDO for subprime mortgage loans, and outline how model errors propagated

throughout the system, culminating in the market turmoils of summer 2007. Our analysis also validates and further explains the observations and "unwind hypothesis" in Khandani and Lo regarding the equity hedge fund crisis of August 6-10, 2007.

Empirical Data Based Alternatives to Classical Techniques in Portfolio Formation
James R. Thompson, Rice University

Empirical research indicates that, contrary to the Nobel Prize winning work of William Sharpe, it is possible to build data based rules for buying portfolios which lie above the Capital Market Line. Three of these are discussed. Two of these, the equal weight S&P 100 rule and the MaxMedian S&P 500 rule backtested for 40 years best the annual growth rate of the Vanguard S&P 500 by 50%. The third, the patented SIMUGRAM algorithm used on the S&P 100 population best the Vanguard S&P 500 annual growth rate by 100%.

- **Bayesian Multiple Comparison Procedures** Room 1070
Organizer: Peter Mueller, University of Texas at Austin

A Bayesian Discovery Procedure
Michele Guindani, M.D. Anderson Cancer Center

We discuss a Bayesian discovery procedure for multiple-comparison problems. We show that, under a coherent decision theoretic framework, a loss function combining true positive and false positive counts leads to a decision rule that is based on a threshold of the posterior probability of the alternative. Under a semiparametric model for the data, we show that the Bayes rule can be approximated by the optimal discovery procedure, which was recently introduced by Storey (2007). Improving the approximation leads us to a Bayesian discovery procedure, which exploits the multiple shrinkage in clusters that are implied by the assumed non-parametric model. We compare the Bayesian discovery procedure and the optimal discovery procedure estimates in a simple simulation study and in an assessment of differential gene expression based on microarray data from tumor samples. We extend the setting of the optimal discovery procedure by discussing modifications of the loss function that lead to different single-thresholding statistics. Finally, I will discuss the extension of the previous arguments to develop a class of data-driven procedures for False Discovery Control in Large-Scale Spatial Multiple Testing. Most of this presentation stems from a joint work with Peter Mueller and Song Zhang

Bayesian Multiplicity Control for RNA-Seq Data on Differential Expression Using Gene Ontology Information
David B. Dahl, Texas A&M University

We adapt the Bayesian Discovery Procedure of Guindani, Mueller, and Zhang (2009) for Bayesian control of multiplicity in the context of a negative binomial sampling

model for RNA-seq data to identifying differentially expressed genes. Further, we extend their methodology by replacing the random partition prior from the Dirichlet process with a random partition prior indexed by distances from Gene Ontology (GO) annotations. This methodological innovation can be applied broadly, irrespective of the sampling model for the data. We show that the use of GO annotations in the clustering prior improves statistical power over the original Bayesian Discovery Procedure. We also compare against leading methods for differential expression of RNA-seq data. Finally, we explore the choice of the loss function and discuss the computational aspects of our approach.

Bayesian Decision Theoretic Multiple Comparison Procedures: An Application to Phage Display Data
Luis Leon, University of Florida

We discuss inference for a human phage display experiment with three stages. The data are tripeptide counts by tissue and stage. The primary aim of the experiment is to identify ligands that bind with high affinity to a given tissue. We formalize the research question as inference about the monotonicity of mean counts over stages. The inference goal is then to identify a list of peptide-tissue pairs with significant increase over stages. We use a semi-parametric Dirichlet process mixture of Poisson model. The posterior distribution under this model allows the desired inference about the monotonicity of mean counts. However, the desired inference summary as a list of peptide-tissue pairs with significant increase involves a massive multiplicity problem. We consider two alternative approaches to address this multiplicity issue. First we propose an approach based on the control of the posterior expected false discovery rate. We notice that the implied solution ignores the relative size of the increase. This motivates a second approach based on a utility function that includes explicit weights for the size of the increase.

Thursday, May 17, 2012

8:15 am - 10:00 am Technical Sessions

- **JCGS Highlights at the Interface** Auditorium
  Organizer: Richard Levine, JCGS Editor, San Diego State University

Symbolic-Covariance Principal Component Analysis and Visualization for Interval-Valued Data

Jennifer Le-Rademacher, Medical College of Wisconsin; and Lynne Billard, University of Georgia

This paper proposes a new approach to principal component analysis (PCA) for interval-valued data. Unlike classical observations which are represented by single points in p- dimensional space R^p, interval-valued observations are represented by hyper-rectangles in R^p and as such have an internal structure which does not exist in classical observations. As a consequence, statistical methods for classical data must be modified to account for the structure of the hyper-rectangles before they can be applied to interval-valued data. This paper extends the classical PCA method to interval-valued data by using the so-called symbolic covariance to determine the principal component (PC) space to reflect the total variation of interval-valued data. The paper also provides a new approach to constructing the observations in a PC space for better visualization. This new representation of the observations reflects their true structure in the PC space.

Local Derivative-Free Approximation of Computationally Expensive Posterior Densities

Nikolay Bliznyuk, University of Florida

Bayesian inference using MCMC is computationally prohibitive when the posterior density of interest, π, is computationally expensive to evaluate. We develop a derivative-free algorithm GRIMA to accurately approximate π; by interpolation over its highest probability density (HPD) region that is initially unknown. Our local approach reduces waste of computational budget on approximation of π; in the low-probability region inherent to global experimental designs. However, estimation of the HPD region is nontrivial when derivatives of π; are not available or are not informative about the shape of the HPD region. Without relying on derivatives, GRIMA iterates (i) sequential knot selection over the estimated HPD region of π; to refine the surrogate posterior and (ii) re-estimation of the HPD region using an MCMC sample from the updated surrogate density, which is inexpensive to obtain. GRIMA is applicable to approximation of general unnormalized posterior densities. To determine the range of tractable problem dimensions, we conduct simulation experiments on test densities with linear and nonlinear component-wise dependence, skewness, kurtosis and multimodality. Subsequently, we use GRIMA in a case study to calibrate a computationally intensive nonlinear regression model to real data from the Town Brook watershed.

Fitting Social Network Models Using Varying Truncation Stochastic Approximation MCMC Algorithm

Faming Liang, Texas A&M University

The exponential random graph model (ERGM) plays a major role for social network analysis. However, parameter estimation for the ERGM is a hard problem due to the intractability of its normalizing constant and the model degeneracy. The existing algorithms, such as Monte Carlo MLE (Geyer and Thompson, 1992) and stochastic approximation (Snijders, 2002), often fail for this problem in the presence of model degeneracy. In this paper, we introduce the varying truncation stochastic approximation Markov chain Monte Carlo (SAMCMC) algorithm to tackle this problem. The varying truncation mechanism enables the algorithm to choose an appropriate starting point and an appropriate gain factor sequence and thus to produce a reasonable parameter estimate for the ERGM even in the presence of degeneracy. The numerical results indicate that the varying truncation SAMCMC algorithm can significantly outperform the Monte Carlo MLE and stochastic approximation algorithms: For degenerate ERGMs, Monte Carlo MLE and stochastic approximation often fail to produce any reasonable parameter estimates, while SAMCMC can do; for non-degenerate ERGMs, SAMCMC can work as well as or better than Monte Carlo MLE and stochastic approximation.

- **Automatic, Flexible Computational Methods with Applications in Biostatistics** Room 1064

  Organizer: David Scott, Rice University

  Statistical Analysis of Computer Algorithms for Assigning Cause-of-Death Codes

  Diba Khan*, Centers for Disease Control & Prevention; National Center for Health Statistics; Myron Katzoff, Centers for Disease Control & Prevention; National Center for Health Statistics; Charles Sirc, Centers for Disease Control & Prevention; National Center for Health Statistics; Donna L. Hoyert, Centers for Disease Control & Prevention; National Center for Health Statistics; Alaina Elliott, Centers for Disease Control & Prevention; National Center for Health Statistics

  International Classification of Disease (ICD) codes are employed to represent the cause-of-death (CoD) terminology that physicians, medical examiners, and coroners report for more than two million death events in the United States. The code assignments are based upon information contained in death certificates filed in state vital statistics offices. Due to the numbers of death records and the needs for (1) a high degree of consistency in code assignments and (2) ease in adapting to periodic changes in ICD code structure, CDC's National Center for Health Statistics has developed computer algorithms for assigning CoD codes. In 85 percent of cases, the coding is done entirely by a computer, but there are some routine types of cases that are coded manually. Coding done by a computer depends upon the information on the death certificates. For special circumstances that require manual coding, the determinations of code assignments vary with the skill and experience of those assigning the codes and the information supplied on the death certificates. This presentation

describes the statistical techniques and procedures to be applied in order to verify the conformance of CoD code assignments with the rules established for that purpose and to quantify the concordance and coherence among professional medical code assignment experts and algorithms created by computer scientist's applications of the coding rules.

Bayesian Survival Trees for Clustered Observations with Application to Tooth Prognosis

Richard Levine, San Diego State University

Tooth loss from periodontal disease or dental caries (decay) afflicts most adults over the course of their lives. Survival tree methods for correlated observations have shown potential for developing objective tooth prognosis systems, however the current technology suffers either from prohibitive computational expense or unrealistic simplifying assumptions to overcome computational demands. In this talk Bayesian tree methods are developed for correlated survival data, relying on a computationally feasible, yet flexible, frailty model with piecewise constant hazard function. Bayesian stochastic search methods, using a Laplace approximated marginal likelihood, are detailed for tree construction and posterior ensemble averaged variable importance ranking and amalgamation procedures are developed to identify indicators of tooth prognostic groups from a forest of trees. The proposed methods are used to assign each tooth from the VA Dental Longitudinal Study to one of five prognosis categories and evaluate the effects of clinical factors and genetic polymorphisms in predicting tooth loss. The prognostic rules established may be used in clinical practice to optimize tooth retention and devise periodontal treatment plans.

Looking Beyond the Lamppost: Flexible Methods Bringing Light into the Dark Alleys of Complex Data

Jeffrey S. Morris, The University of Texas M.D. Anderson Cancer Center

Modern science is characterized by new measurement instruments producing an explosion of data, ever-growing in their quantity and complexity. These data raise numerous quantitative challenges, among them the challenge of efficiently and reliably extracting the valuable scientific information they contain while managing the practical challenges raised by their size and subtleties. The absence of sufficiently flexible methods and frameworks often forces scientists to first simplify their data using simple summaries to eliminate some of their vexing complexities. This can work well if these summaries retain all relevant information in the data, but many times that is not the case. This convenience-driven oversimplification of complex data is like "looking for the lost keys under the lamppost," where we hope them to be, while ignoring the "dark alleys," which may in fact be where the keys reside. One primary objective of modern statistics is to develop efficient, flexible methods and

modeling frameworks that can model the complex data as they are, while avoiding oversimplifications, and thereby better modeling the systems that generate the data and potentially uncovering more of the treasure trove of information they contain. We could say that these methods are intended to "bring light to the dark alleys of complex data," emboldening researchers to venture to places they fear to go and perhaps uncover new insights as a result. I will discuss this principle in the context of various areas of application, including genomics, proteomics, activity studies, and functional brain imaging. I will briefly summarize and discuss a flexible and efficient framework for modeling complex object data such as functions and images developed in the past several years. Our framework is based on the functional mixed model framework, a generalization of linear mixed models that can model simultaneous effects of multiple factors on the objects and handle correlation between objects induced by the experimental design. The multi-domain modeling approach and software developed for this framework is appropriate for high-dimensional objects with complex features, can accommodate missing data and outliers, and yields Bayesian inference for all model components. This work serves as an example of methodological development motivated by the premise of developing flexible, efficient frameworks for modeling complex object data. There are similar ongoing efforts currently underway by many other researchers, and I expect these joint efforts will collectively produce a rich, flexible set of modeling tools that we as the statistical community can offer to the broader scientific community to help them uncover insights they could not find without our contributions.

- **Information Mining** Room 1070
  Organizer: William Szewczyk, National Security Agency

Choosing a Dissimilarity for Classification
Adam Cardinal-Stakenas, Department of Defense

The dissimilarity representation is a vital component of modern statistical analysis. By examining all pairs of dissimilarities between the elements of an observed data set, one can leverage the more than 50-year history of statistical pattern recognition on high- and infinite-dimensional spaces, spaces that exhibit non-Euclidean geometry, and spaces that defy analysis by traditional means. However, for most data sets observed in spaces like these, there are many dissimilarities that could be successfully applied. A critical question facing the researcher is: how should one choose which dissimilarity to use in these circumstances, and, if there are many that perform well, can they be combined to optimize inferential performance? We will begin to address these questions by applying a variety of methods from matrix analysis, factor analysis, optimization, combinatorics, and statistics.

Strategies for Streaming Exploratory Data Analysis

William Szewczyk, National Security Agency

One of the often overlooked lessons of data analysis is the importance of the default model during explorations, since choice of this model guides the discovery task. For years, if not centuries, the default model has been the Gaussian one. While this has served well for many situations, as the method of data processing changes from storage and batch processing to processing as the data is in motion, the assumptions needed for the Gaussian model begin to break down. In this talk I will present an alternative model that can be applied to streaming data, as well as an example of how its use simplifies the analysis of an open-ended data set.

Solving a Story with Multiple Unknowns

Andy Frenkiel, IBM/T.J. Watson Research

In this talk we present an informal exploration of the types of unknowns that emerge as select information is removed or changed in a story. Starting with a news wire article, we elide story elements such as names, dates, and quantities and conjecture different classes of unknowns that we observe as the information content of the story becomes more sparse. This exercise is motivated by the practical and significant challenge of providing automation to assist in filling in the gaps in the stories and knowledge bases central to real-world activities such as news reporting, financial analysis, medical diagnosis, cyber defense, among others. We conclude by discussing how some types of unknowns cannot be repopulated solely using text analysis and question-answering methods and we suggest research directions for filling in those unknowns.

- **Contributed Paper Session I** Room 1075

Organizer: David Scott, Rice University

Optimal Reduced Isotonic Regression

Janis Hardwick and Quentin Stout*, University of Michigan

Isotonic regression is a shape-constrained nonparametric regression in which one assumes that the ordinate is a nondecreasing function of the abscissa. An isotonic regression forms a sequence of steps. For a set of N data points it may contain as many as N steps and thus has been criticized as overfitting the data or making the representation too complicated. So-called "reduced" isotonic regression constrains the number of steps permitted. Previously, the fastest algorithms for exact solutions for the $L2$ metric take $O(N + K^2L)$ time, where K is the number of steps in the unconstrained isotonic regression and L is the maximum number of steps allowed in the reduced isotonic regression. Some researchers found this to be too slow and used approximation approaches. Here, we decrease the time for the exact solution

to $O(N + KL \log K)$. Our approach is based on an algorithm for finding an optimal histogram for increasing values which is more efficient than applying Bellman's dynamic programming algorithm for arbitrary values.

Exponential-Family Random Network Models

Ian Fellows* and Mark S. Handcock, University of California, Los Angeles

Random graphs, where the connections between nodes are considered random variables, have wide applicability in the social sciences. Exponential-family Random Graph Models (ERGM) have shown themselves to be a useful class of models for representing complex social phenomena. We generalize ERGM by also modeling nodal attributes as random variates, thus creating a random model of the full network, which we call Exponential-family Random Network Models (ERNM). We demonstrate how this framework allows a new formulation for logistic regression in network data. We develop likelihood-based inference for the model and an MCMC algorithm to implement it.

We use this new model formulation to analyze a peer social network from the National Longitudinal Study of Adolescent Health. We model the relationship between substance use and friendship relations, and show how the results differ from the standard use of logistic regression on network data.

A Projection Pursuit Index Based on Kernel PCA with Gaussian Kernels

Victor Muniz*, Johan Van Horebeek, and Rogelio Ramos, Research Center in Mathematics. Monterrey, Mexico

Kernel based methods have become very popular to extract nonlinear structures in high dimensional data; they extend linear methods without increasing significantly the complexity by making use of implicit transformations, and this is particularly useful for complex data and for large p small n data. For applications with a large number of observations, the kernel approach becomes computationally intractable. To this end we propose an approximation using random projections as was done for SVM's by Rahimi and Recht (2007). Moreover this leads to an intuitive interpretation of Kernel PCA in the original data space. Based on this, we introduce a Projection Pursuit index sensitive to contrasts in the density of the observations. We discuss the underlying optimization problem and present an extensive set of experiments, including comparisons with otherIndexes.

Dependent Pólya Urn Schemes

Bernardo Nipoti, University of Texas MD Anderson Cancer Center

The proposal and study of dependent nonparametric priors has been a major research focus in the recent Bayesian nonparametric literature. We introduce a flexible class of dependent nonparametric priors and derive a suitable sampling scheme which

allows their concrete implementation. The proposed class is obtained by normalizing dependent completely random measures, where the dependence among the completely random measures arises by virtue of a suitable construction of the underlying Poisson random measures. We obtain an expression for the partially exchangeable partition probability function that forms the basis for the determination of a Markov Chain Monte Carlo algorithm for drawing posterior inferences. This may be thought of as a generalization to the two-dimensional setting of the well-known Blackwell MacQueen Pólya urn scheme and it involves the update of three independent urns.
First we provide general distributional results for the whole class of dependent completely random measures, then we specialize to two specific priors, which represent the natural candidates for concrete implementation due to their analytic tractability: the bivariate Dirichlet and normalized σstable processes. In these settings we provide the details for the actual implementation of the algorithm.
Such an algorithm allows to perform a full Bayesian analysis both in terms of density estimation and of clustering of the data. The analysis is completed by some illustrations concerning synthetic and real twosample datasets.

10:30 am - 12:15 pm Technical Sessions

- **Big, Fast, and Interactive Data** Auditorium
  Organizer: Michael Kane, Yale University

Elastic Computing with R and Redis
Bryan Lewis, Paradigm4

Redis is a networked key value store with many interesting capabilities. We illustrate the use of Redis as a framework for parallel computing with R suited to elastic computing environments like EC2.

How Google Estimates Traffic for Millions of Queries
Tim Hesterberg, Google

Google estimates traffic for millions of queries continuously, and looks for spikes. Among other uses, this feeds Google Hot Trends. We describe how Google efficiently processes a wide variety of queries with frequencies ranging from tens to millions of hits per day, with widely varying activity across time of day and week, and updates these models continuously as new data is observed.

EDA, Visualization and Collaboration on the Web
Carlos Scheidegger, AT&T Labs

This talk will present ongoing work on an R-based EDA environment the web. As organizations and business become more data-driven, EDA has become increasingly important. In large organizations, sharing derived data, experiments, graphics and visualizations is still a hard problem, and is addressed mainly by stand-alone version-control systems. In contrast, we envision a system where users collaborate with each other, fluidly moving from a full-fledged R command prompt to a simplified "pick your analysis-visualization-data". Sharing a new dataset, analysis or visualization should be as easy as creating it. We will show the current state of the tool in a live demo.

- **Bayesian Hierarchical Models for High Dimensional Spatial Data** Room 1064
  Organizer: Sudipto Banerjee, University of Minnesota

Space-Time Data Fusion Under Error in Computer Model Output: An Application to Modeling Air Quality
Veronica J. Berrocal, University of Michigan

In order to effectively investigate the linkage between ambient exposure and health outcomes, accurate estimates of exposure are needed. The US EPA monitors pollutant levels using information from monitoring networks as well as estimates generated by deterministic numerical models. The former measure pollutant concentrations using instruments at a sparse set of locations; the latter yield estimates of the average concentration in a large number of grid cells of pre-specified dimensions by numerically solving complex systems of differential equations capturing various diffusion, chemical and atmospheric processes. Combining these information sources can improve exposure assessment at high, in fact, point level resolution.
In this talk, we present two methods to fuse the two sources of data. Both models are extensions of an earlier downscaler model and address two potential concerns with the model output. One recognizes that there may be useful information in the computer model output for grid cells that are neighbors of the one in which a monitoring site lies. The second acknowledges potential spatial misalignment between a station and its putatively associated grid cell. The first model is a Gaussian Markov random field smoothed downscaler that relates monitoring station data and computer model output via the introduction of a latent Gaussian Markov random field linked to both sources of data. The second is a smoothed downscaled with spatially varying random weights defined through a latent Gaussian process and an exponential kernel function. Both models allows to address the large dimensionality of the computer model output efficiently.

An Adaptive Spatial Model for Precipitation Data From Multiple Satellites Over Large Regions

Avishek Chakraborty, Texas A&M University

Satellite measurements have of late become an important source of information for climate features such as precipitation due to their near-global coverage. In this talk, we look at a precipitation dataset during a 3-hour window over tropical South America that has information from two satellites. We develop a flexible hierarchical model to combine instantaneous rainrate measurements from those satellites while accounting for their potential heterogeneity. Conceptually, we envision an underlying precipitation surface that influences the observed rain as well as absence of it. The surface has been specified using a mean function centered around a set of knot locations, to capture the local patterns in the rainrate, combined with a residual Gaussian process to account for global correlation across sites. To improve over the commonly used prefixed knot choices, an efficient reversible jump scheme has been used to allow the number of such knots as well as the order and support of associated polynomial terms to be chosen adaptively. To facilitate computation over a large region, a reduced rank approximation for the parent Gaussian process has been employed.

Flexible Predictive Process Spatial Factor Models for Misaligned Data Sets
Qian Ren and Sudipto Banerjee, University of Minnesota

We present joint modeling for a large number of geographically referenced outcomes observed over a very large number of locations. We seek to capture associations among the variables as well as the strength of spatial association for each variable. In addition, we reckon with the common setting where not all the variables have been observed over all locations, which leads to *spatial misalignment*. Dimension reduction is needed in two aspects: (i) the length of the vector of outcomes, and (ii) the very large number of spatial locations. Latent variable (factor) models are usually used to address the former, while low-rank spatial processes offer a rich and flexible modeling option for dealing with a large number of locations. We merge these two ideas to propose a class of hierarchical low-rank spatial factor models. Our framework pursues stochastic selection of the latent factors without resorting to complex computational strategies (such as reversible jump algorithms) by utilizing certain identifiability characterizations for the spatial factor model. A Markov chain Monte Carlo (MCMC) algorithm is developed for estimation that also deals with the spatial misalignment problem. We recover the full posterior distribution of the missing values (along with model parameters) in a Bayesian predictive framework. Various additional modeling and implementation issues, including a special class of priors for the spatial range, are presented as well. We illustrate our methodology with simulation experiments and an environmental data set.

- **Computing in Statistics Education** Room 1070
  Organizer: Webster West, Texas A&M University

Teaching Formulas in Statistics Classes: When Is It Beneficial?
David Lane, Rice University

With the advent of powerful and easy-to-use statistical software, statistics instructors can focus on concepts rather than long and complicated computational formulas. This is a very positive development and has greatly improved statistical education. However there are some concepts for which the formula is so tightly tied to the concept that teaching the formula facilitates student understanding. This paper presents examples of concepts that are better taught conceptually and without formulas as well as examples of concepts that are better taught together with formulas. It is concluded that most formulas do not need to be taught but that the exceptions are important.

Using Simulations to Teach Statistical Inference
Beth Chance, California Polytechnic

A current "hot" trend in statistics education is use computer simulations of randomization distributions to take advantage of current computer power to introduce students to concepts in statistics inference in a more intuitive, visual manner. These tools have the potential to replace traditional parametric methods in statistics education and practice. We will discuss several options for implementation along with some research results on their effectiveness. We will also discuss methods for blending technologies to enrich the student learning experience.

The Impact of Technology on the Teaching of Statistics
Webster West, Texas A&M University

Over the past two decades, we have seen rapid technological advancements that have had a tremendous effect on statistical education both in terms of its content and its delivery. In this talk, we will take a nostalgic look back at this technological journey, and we will also look into the crystal ball to see where new technology may take statistical education in the future.

- **Contributed Paper Session II** Room 1075
Organizer: David Scott, Rice University

A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data
Xueying Chen* and Minge Xie, Rutgers University

If there are extraordinarily large data, too large to fit into a single computer or too expensive to perform a computationally intensive data analysis, what should we do? To deal with this problem, we propose in this paper a split-and-conquer approach and illustrate it using a computationally intensive penalized regression method, along

with a theoretical support. Consider a regression setting of generalized linear models with n observations and p covariates, in which n is extraordinarily large and p is either bounded or goes to infinity at a certain rate of n. We propose to randomly split the data of size n into K subsets of size $O(n/K)$. For each subset of data, we perform a penalized regression analysis and the results from each of the K subsets are then combined to obtain an overall result. We show that the combined overall result still retains all the desired properties of penalized estimators such as the model selection consistency and asymptotic normality under mild conditions. When K is less than $O(n^{1/5})$, we also show that the combined result is asymptotically equivalent to the corresponding analysis result of using the entire data all together, assuming that there were a super computer that could carry out such an analysis. In addition, the split-and-conquer approach involves a random splitting and a systemic combining. We demonstrate that there were a super computer that could carry out such an analysis. In addition, the split-and-conquer approach involves a random splitting and a systemic combining. We demonstrate that the approach has an inherent advantage of being more resistant to false model selections caused by spurious correlations, and we further establish an upper bound for the expected number of falsely selected variables and a lower bound for the expected number for truly selected variables. Furthermore, when a computational intensive algorithm is used in the sense that its computing expense is at the order of $O(n^a)$, $a > 1$, we show that the split-and-conquer approach can substantially reduce computing time and computer memory requirement. The proposed methodology is demonstrated numerically using both simulation and real data examples.

Manipulating Dates and Times in R With the Lubridate Package

Garrett Grolemund* and Hadley Wickham, Rice University

This talk presents the lubridate package for R, which facilitates working with dates and times. Date and times create various technical problems for the data analyst. Parsing date-times into a computer program is difcult because date-times may be represented in many ways. Formatting choices and conventions such as time zones and military times will affect how a moment of time is described and saved. Modifying date-times is difficult because time spans have inconsistent lengths depending on when and where they occur due to conventions such as daylight savings time, leap years, and leap seconds. lubridate gives an analyst the power to use or ignore these conventions with three new time span object classes for R. This talk will offer practical advice on how to solve date-time related problems in R with lubridate. The talk also introduces a conceptual framework for arithmetic with date-times in R.

mpoly*: Multivariate Polynomials in R*

David Kahle, Baylor University

The **mpoly** package is a general purpose collection of tools for symbolic computing with multivariate polynomials in R. In addition to basic arithmetic, **mpoly** can take derivatives of polynomials, compute Gröbner bases of collections of polynomials, and convert polynomials into a functional form to be evaluated.

2:00 pm - 3:45 pm Technical Sessions

- **Inference on Graphs** Auditorium
  Organizer: Organizers: Carey Priebe and Dave Marchette, Johns Hopkins University and Naval Surface Warfare Center

  Graph Inference with Imperfect Edge Classifiers
  Michael Trosset, Indiana University

  We test simple hypotheses about a random graph with a fixed set of vertices and random edges, each of which possesses one of K mutually exclusive attributes. The edge attributes are inferred by means of a fallible classifier. Suppose that E and F are the confusion matrices of two such classifiers. Using results from statistical decision theory, we demonstrate that, if there exists a K x K stochastic matrix R such that ER = F, then most powerful (MP) tests based on E are necessarily more powerful than MP tests based on F. By means of an example, we also demonstrate that entry-wise superiority of E to F does not guarantee that an MP test based on E is more powerful than an MP test based on F.

  Consistent Embedding of Stochastic Blockmodels
  Minh Tang, Johns Hopkins University

  A stochastic block model consists of a random partition of n vertices into K blocks for which, conditioned on the partition, every pair of vertices has probability of connection determined entirely by their block memberships. Suppose a realization of the n-by-n vertex adjacency matrix is observed but the underlying partition of the vertices into blocks is not observed. The main inferential task is thus to partition the vertices into blocks. This talk describes a spectral partitioning algorithm for adjacency matrices that is consistent for the above inferential task. The algorithm is particularly simple and requires only an upper bound on the rank of the K-by-K probability matrix that underlies the stochastic block model. We illustrate the methodology by presenting examples related to the detection of community structure in networks.

  Title Vertex Nomination: Improved Fusion of Content and Context
  Glen Coppersmith, Johns Hopkins University

We expand our previous investigations of vertex nomination to include a more principled fusion of content and context that exhibits superior performance. Our data are a collection of communications encoded as an attributed graph. Vertices represent the actors and edges connect pairs of actors that have communicated. Specifically, the edges are attributed with the human language content of these communications and the vertices are attributed with class membership. One class of vertices is of interest to us, and exhibit different behavior (both in terms of the content they are exposed to and the other actors they communicate with – both content and context). We observe the class label for only a small number of the vertices from the class of interest, and we wish to find the remainder of the class. This is a specific instance of a general 'more like this' problem, and thus has a number of applications. We demonstrate that a principled fusion of information derived from the content and the context provides superior inference over either alone, and that tuning this fusion further improves performance.

- **Statistical and Computational Methods for Large Spatial Data Sets** Room 1064
  Organizer: Jianhua Huang, Texas A&M University

Covariance Decomposition with Low Rank and Sparse Representation for Large Spatial Datasets
Huiyan Sang, Texas A&M University

Bayesian hierarchical models have been widely used in spatial statistics but face tremendous computational challenges for very large data sets. With regard to this challenge, we propose what we call full scale approximation models for spatial data. In this talk, we will show the utility of our method in various spatial model settings, including non-stationary,non-Gaussian and multivariate spatial processes models in the context of large data sets. We illustrate the approach with simulated and real data sets.

Nonstationary Cross-Covariance Models for Multivariate Processes on a Globe
Mikyoung Jun, Texas A&M University

In geophysical and environmental problems, it is common to have multiple variables of interest measured at the same location and time. These multiple variables typically have dependence over space (and/or time). As a consequence, there is a growing interest in developing models for multivariate spatial processes, in particular, the cross-covariance models. On the other hand, many data sets these days cover a large portion of the Earth such as satellite data, which require valid covariance models on a globe. We present a class of parametric covariance models for multivariate processes on a globe. The covariance models are flexible in capturing nonstationarity in the

data yet computationally feasible and require moderate numbers of parameters. We apply our covariance model to surface temperature and precipitation data from an NCAR climate model output. We compare our model to the multivariate version of the Matërn cross-covariance function and models based on coregionalization and demonstrate the superior performance of our model in terms of AIC (and/or maximum loglikelihood values) and predictive skill. We also present some challenges in modeling the cross-covariance structure of the temperature and precipitation data. Based on the fitted results using full data, we give the estimated cross-correlation structure between the two variables.

Bayesian Estimation for Large Spatial Datasets Observed on a Lattice
Jonathan R. Stroud, George Washington University

This talk proposes a new Bayesian MCMC method for parameter estimation for Gaussian processes observed on a lattice. The main computational trick is to use circulant embedding of the covariance matrix which allows us to use Fast Fourier Transforms to evaluate the likelihood function. Unlike existing approaches for large spatial datasets, the method provides exact inference and does not rely on approximations. Missing data are easily handled using imputation, and the method is scalable to very large datasets. We also provide a related algorithm for maximum likelihood estimation via the stochastic EM algorithm. We apply the methods using simulated data and real satellite images, and show that they outperform existing approaches including the Whittle approximation and Covariance Tapering.

- **Woman VS Machine: The Inference Battle** Room 1070
Organizer: Di Cook, Iowa State University

Statistical Inference after Model Selection
Andreas Buja, Wharton School, University of Pennsylvania

It is common practice in statistical data analysis to perform data-driven variable selection and derive statistical inference from the resulting model. Such inference enjoys none of the guarantees that classical statistical theory provides for tests and confidence intervals when the model has been chosen a priori. We propose to produce valid "post-selection inference" by reducing the problem to one of simultaneous inference. Simultaneity is required for all linear functions that arise as coefficient estimates in all submodels. By purchasing "simultaneity insurance" for all possible submodels, the resulting post-selection inference is rendered universally valid under all possible model selection procedures. Importantly the inference does not depend on the truth of the selected submodel, and hence it produces valid inference even in wrong models.
Joint with: Richard Berk, Larry Brown, Kai Zhang, Linda Zhao

Facing Off: Power of Visual and Classical Tests
Heike Hofmann, Iowa State University

Lineups (Buja et al, 2009; Wickham et al, 2010) have been established as tools for visual testing similar to standard statistical inference tests, allowing us to evaluate the validity of graphical findings in an objective manner. In simulation studies (Majumder et al, 2011) lineups have been shown as being efficient: the power of visual tests is comparable to classical tests while being much less stringent in terms of distributional assumptions made. This makes lineups versatile, yet powerful, tools in situations where conditions for regular statistical tests are not or cannot be met. Here, we want to introduce lineups as a tool for evaluating the power of competing graphical designs. We highlight some of the theoretical properties and then show results from two studies evaluating competing designs: both studies are designed to go to the limits of our perceptual abilities to highlight differences between designs. We use both accuracy and speed of evaluation as measures of a successful design.

Turk Experiments for Visual Inference
Mahbub Majumder, Iowa State University

It has been found that the visual test, which does not have distributional assumptions, has the power comparable to the classical tests (Majmder et al, 2011). To examine the power of visual statistical inference we recruited human subjects from Amazon Mechanical Turk (`http://aws.amazon.com/mturk/`) for evaluating lineups. The Turk website is designed to recruit workers for simple and easy tasks. Even though the task of evaluating lineups is easy, the technical design of the experiment is complex and the tools available to design this from Turk website is just too simple. In this paper we present how we deal with this trouble and to conduct our experiment.

- **Contributed Paper Session III** Room 1075
Organizer: David Scott, Rice University

Bayesian Multiplicity Control for Graphs
Riten Mitra*, University of Texas MD Anderson Cancer Center; Peter Mueller, University of Texas at Austin; and Yuan Ji, University of Texas MD Anderson Cancer Center

We consider a fully Bayesian framework for joint inference on multiple graphical models. Imposing a suitable prior on latent indicators has been shown to control for multiplicity in mean effects and variable selection models. This is usually achieved by imposing a prior distribution on a hyper parameter representing the probability of inclusion. We extend this idea first to the analysis of dependence structure implied by a single graphical model, and then to the inference on multiple graphical models. The joint prior distribution of the vector of edge inclusion probabilities is extended

from an Uniform to a Dirichlet distribution. We show formally that this choice of the prior distribution improves posterior inference greatly compared to an independent analysis of the multiple models. Mathematically, the KL divergence between two posterior diverges as the number of edges go to infinity. We recommend that this fully Bayesian model based on the Dirichlet prior for the hyperparameters be used for joint estimation of graphical models. We apply this model to the expression data of protein markers obtained from a novel Mass Cytometry technology called CyTOF.

A Nonparametric Bayesian Model for Local Clustering

Juhee Lee, University of Texas MD Anderson Cancer Center

We propose a nonparametric Bayesian local clustering (NoB-LoC) approach for heterogeneous data. The NoB-LoC model defines local clusters as blocks of a two-dimensional data matrix and produces inference about these clusters as a nested bidirectional clustering. Using protein expression data as an example, the NoB-LoC model clusters proteins (columns) into protein sets and simultaneously creates multiple partitions of samples (rows), one for each protein set. In other words, the sample partitions are nested within the protein sets. Any pair of samples might belong to the same cluster for one protein set but not for another. These local features are different from features obtained by global clustering approaches such as hierarchical clustering, which create only one partition of samples that applies for all proteins in the data set. As an added and important feature, the NoB-LoC method probabilistically excludes sets of irrelevant proteins and samples that do not meaningfully co-cluster with other proteins and samples, thus improving the inference on the clustering of the remaining proteins and samples. Inference is guided by a joint probability model for all random elements. We provide extensive examples to demonstrate the unique features of the NoB-LoC model.

Joint with Peter Mueller (Department of Mathematics, University of Texas at Austin) and Yuan Ji (NorthShore University HealthSystem, Evanston, IL)

Testing Goodness of Fit of Protein Conformational Sampling Models

Mehdi Maadooliat, IAMCS, Texas A&M University

Regardless of considerable progress in the past years, protein structure prediction remains as one of the major unsolved problems in computational biology. To predict protein structure, there has been much work on both template-based and template-free modeling methods, where each has its own advantages and disadvantages. Fragment assembly methods combine the advantages of the template-based and the template-free modeling to achieve more successful results in demonstrating the protein backbone structure.

The literature has focused on using variety of parametric models on sequential dependencies between the angle pairs along the protein chains. Despite the presence

of different statistical methods for modeling the protein backbone structure, there is no statistical tool to evaluate the goodness of fit of such models. We propose to consider the joint distribution of the dihedral angles over different lags based on a nonparametric approach and monitor the behavior of the sequence of angles using the principal components analysis. As a result, we can develop graphical tools and numerical measurements to compare and evaluate the performance of different models in fitting the data.

Improving Compton Scatter Camera Image Resolution Using Classification and the Metropolis-Hastings Algorithm

Dennis Mackin, University of Texas MD Anderson Cancer Center

Relative to conventional external beam x-ray radiation, proton therapy allows for the precise targeting of tumors and the sparing of healthy tissue. Unfortunately, the full benefits of proton therapys precision cannot be realized because treatment margins must account for day-to-day changes in patient anatomy and positioning. Research is ongoing to develop methods for verifying the radiation dose delivery because dose verification will enable a reduction in the treatment margins. One promising verification method is to produce images of the proton beam from the prompt gamma rays emitted by nuclear de-excitations in the irradiated tissue. The gamma rays can be detected with Compton scatter cameras, multi-stage detectors which measure a gamma rays incident energy and restrict its propagation direction to the surface of a cone. However, the resolution of the Compton camera images is limited by (1) the characteristics of the detectors and by (2) the algorithm used to reconstruct the image. In our talk we discuss the ways we are using logistic regression and support vector machines to remove poorly detected gamma rays, thereby addressing problem (1). We also will discuss how a novel application of Markov chains and the Metropolis-Hastings algorithm to the Compton camera image reconstruction problem addresses (2).

4:15 pm - 5:30 pm Banquet Keynote Auditorium

- **Banquet Address** Auditorium

Repetition and Surprise, Rehearsal and Reinvention

Mark Hansen, UCLA

Far from virtual, inert quantities, data exert real forces in the physical world. They are incendiaries wielding the power of once-secret diplomatic cables; mores initiated

with the invention of a privacy setting; and physical laws shaping the built environment with the quiet, persistent action of zoning regulations. Data rarely act in isolation, gaining power through combination, "join"ing forces and moving into new terrain. Their presence is thought to guarantee transparency, their absence is seen as suspicious, and restrictions on their movement appear to be temporary, at best.
In this talk, I will take a broad view of data (and companion ideas like "algorithm," "model" and "visualization") and explore their use in creative practices. I will present a selection of work from my artistic collaborations over the last decade – From a permanent display in the lobby of the New York Times building and a new work for the 9/11 Memorial Museum in New York City; to a performance designed as part of the New York Public Librarys centennial celebration last June.
I will tie these artworks to a larger movement in which data and data processing are seeping into almost every academic discipline on campus. I'll end this talk with a proposal to aggregate the data practices from science, the humanities, and even art and design under a single umbrella – Data science.

Friday, May 18, 2012

8:15 am - 10:00 am Technical Sessions

- **Modeling, Analyzing, and Visualizing High-Dimensional Data in Genomics**
  Auditorium
  Organizer: Karen Kafadar, Indiana University

  Fast Graphical Model Estimation and Its Applications
  Daniela Witten, University of Washington

  The graphical lasso, recently proposed for Gaussian graphical modeling in high dimensions, involves estimating an inverse covariance matrix under a multivariate normal model by maximizing the L1-penalized log likelihood. I will begin by presenting a very simple but previously unknown necessary and sufficient condition that can be used to identify the connected components in the graphical lasso solution. This condition can be used to achieve massive computational gains: computing the graphical lasso solution with 20,000 features now takes minutes on a standard desktop machine, whereas previously the computations were prohibitive. This opens up new doors for rigorous network analysis of high-dimensional biological data. As a specific example, I will discuss estimation of graphical models under distinct biological conditions, in which we expect some, but not all, aspects of the networks to differ between

conditions. An extension of the necessary and sufficient condition developed for the graphical lasso allows for extremely fast network estimation in this setting. Parts of this work are joint with J Friedman, N Simon, P Wang, and P Danaher.

Conditional Network Testing in High-dimensional Dependent Data

Gary Gadbury, Kansas State University

Multiple testing research has undergone renewed focus in recent years as advances in high throughput technologies have produced data on unprecedented scales. Much of the focus has been on false discovery rates and related quantities that are estimated (or controlled for) in large scale multiple testing situations. Some estimators may have high variance in the presence of correlation, and the effect of this variance on interpretations of estimators has received less attention in the literature. Recent papers by Efron have directly addressed this issue and incorporated measures to account for the correlation in false discovery rate estimates and density estimates. This work begins by demonstrating the effect of dependence structure on the variance of the number of discoveries and the false discovery proportion (FDP). A variance of the number of discoveries is shown and the density of a test statistic, conditioned on the status (reject or failure to reject) of a different correlated test, is derived. It is shown that the correlations among the test statistics affect the conditional density and alter the threshold for significance of a correlated test. The concept of performing tests within networks is introduced and called conditional network testing (CNT). These tests are based on the conditional density mentioned above. Initial results illustrate that this method stabilizes the variance of the number of discoveries under dependence and reduces the FDP. A method for simulating realistic data is also discussed and illustrated with CNT.

Robust Identification of Conditional Gene Expression in Development of Onthophagus Beetles

Guilherme V Rocha, Indiana University

Multi-cellular organisms develop different tissues through cellular differentiation regulated by gene regulatory networks. Onthophagus taurus stand out as a model organism in evolutionary developmental biology, due to the varied responsiveness of their phenotype to environmental factors, including the expression of horns: a novel complex trait with no homologous structure in other organisms. Identifying the genes involved in the differentiation of tissues according to gender and nutrition factors provides understanding of the molecular mechanisms involved in tissue development and provides insight into how novel traits evolve. A large microarray experiment was designed to assess the expression of genes in four tissue types of male and female beetles exposed to high and low levels of nutrition. We describe the analysis of the data from this study, which involves problems of multiple testing and estimating the relative sizes of differentially expressed genes under different conditions.

This is joint work with Karen Kafadar (IU Statistics), Armin Moczek (IU Biology), Emilie Snell-Rood (U Minnesota, Biology), Teiya Kijimoto (IU Biology), and Justen Andrews (IU Biology)

- **Visualization and Computational Methods for Actigraphy Data** Room 1064
  Organizer: Jürgen Symanzik, Utah State University

Powerful Actigraphy Data Through Functional Representation
Jimin Ding, Washington University of St. Louis

An actigraph is a watch-like device, usually attached to the wrist or leg, that contains accelerometers to measure movements in the form of activity counts every minute or every few seconds. As an emerging clinical technology, actigraphy data is often collected over several days for each participant to evaluate sleep, daytime activity, and circadian activity rhythms in people. In this talk, we view the measured activities of each day for each person as a function of time and analyze them using functional data analysis (FDA). The functional linear mixed effects model is applied to those clustered curves. Here, subject effects are captured through random effects while treatment effects are modeled through fixed effects. We employ principal components analysis for both within-subject and between-subject covariances to fit the model.

Reliability and Reproducibility Issues in Accelerometer-Based Estimates of Physical Activity
Julia Kozlitina, UT Southwestern Medical Center

Accelerometer-based monitors have become a widely used tool for the objective assessment of physical activity (PA) over the past few years. Although these small devices promise to provide accurate measures of free-living physical activity, the validity of accelerometer-derived estimates depends in large part on intra- and inter-monitor reliability as well as an adequate study design. In this talk we will discuss the different sources of variability in accelerometer data. Our results are based both on lab experimentation and preliminary analysis of a population-based study of physical activity. We will describe some approaches to reducing the known sources of variability and bias in order to improve estimates of PA and discuss the implications of our findings for the design of future studies.

Movelets: A Dictionary of Movement
Jeff Goldsmith, Johns Hopkins Bloomberg School of Public Health

Recent technological advances provide researchers with a way of gathering real-time information on an individual's movement through the use of wearable devices that record acceleration. In this paper, we propose a method for identifying activity types,

like walking, standing, and resting, from acceleration data. Our approach decomposes movements into short components called "movelets", and builds a reference for each activity type. Unknown activities are predicted by matching new movelets to the reference. We apply our method to data collected from a single, three-axis accelerometer and focus on activities of interest in studying physical function in elderly populations. An important technical advantage of our methods is that they allow identification of short activities, such as taking two or three steps and then stopping, as well as low frequency rare (compared with the whole time series) activities, such as sitting on a chair. Based on our results we provide simple and actionable recommendations for the design and implementation of large epidemiological studies that could collect accelerometry data for the purpose of predicting the time series of activities and connecting it to health outcomes.

- **Developing Intelligence in Unmanned Ground Systems** Room 1070
  Organizer: Barry Bodt, U.S. Army Research Laboratory

  Some Thoughts on Experimentation Philosophy in the Robotics CTA
  Barry Bodt, U.S. Army Research Laboratory

  The Robotics Collaborative Technology Alliance (RCTA) is an alliance of government, industrial, and academic institutions performing research in robotics to enable the development of unmanned ground systems for the military. The principal way the Robotics CTA shows progress is through the Integrated Research Assessment (IRA). The target is Integrated Research in which more than one RCTA technology working together is necessary to achieve component behaviors of an operationally relevant mission task. And the target is also Formal Assessment, where statistical rigor and sound experimentation practices are preferred over demonstrations that sometimes have neither. The question is not "Can it do it?" in a specific circumstance, but rather "How well does it do it?" over a relevant space of circumstances. A key purpose of the IRA is to objectively measure the current capability in light of some operational context and assist the developers in documenting their progress. An assessment is merely a data point in the development cycle, an opportunity to stress the technology in a system and to identify what it does well and what it could do better. It is not pass/fail. When advancements are made and a follow-on IRA occurs, the cycle continues with the bar a little higher. In this paper, I will discuss the role of the IRA, supporting task-based assessments, and give some thoughts on competing experimentation models that encourage advancement in robotics.

  Preliminary Performance Evaluation of Autonomous Mobility in Small UGVs
  Alberto Lacaze, Robotic Research, LLC

  A system for autonomous mobility and coordination of groups of unmanned ground vehicles (UGVs) that can execute a variety of military relevant missions in dynamic

urban environments has been developed. Historically, UGV operations have been primarily performed via tele-operation, requiring at least one dedicated operator per robot, and requiring substantial real-time bandwidth to accomplish those missions. The system provides long-term value to the war-fighter. To that end, we self-imposed a set of constraints that would force us to develop technology that could readily be used by the military in the near term: (1) Use a relevant (deployed) platform; (2) Use low-cost, reliable sensors; (3) Develop an expandable and modular control system with innovative software algorithms to minimize the computing footprint required; (4) Minimize required communications bandwidth and handle communication losses; and (5) Minimize additional power requirements to maximize battery life and mission duration.

Using Expectations to Drive Cognitive Behavior
Unmesh Kurup, Carnegie Mellon University

Generating future states of the world is an essential component of high-level cognitive tasks such as planning. We explore the notion that such future-state generation is more widespread and forms an integral part of cognition. We call these generated states expectations, and propose that cognitive systems constantly generate expectations, match them to observed behavior and react when a difference exists between the two. We describe an ACT-R model that performs expectation-driven cognition on two tasks pedestrian tracking and behavior classification. The model generates expectations of pedestrian movements to track them. The model also uses differences in expectations to identify distinctive features that differentiate these tracks. During learning, the model learns the association between these features and the various behaviors. During testing, it classifies pedestrian tracks by recalling the behavior associated with the features of each track. We compare the models performance to a simple K-nearest-neighbor classifier.

10:15 am - 12:00 noon Technical Sessions

- **Computational Tools and Statistical Methods with Medical Applications**
  Auditorium
  Organizer: Bradley Broom and Kim-Anh Do, University of Texas M.D. Anderson Cancer Center

Graph-Based Signal Integration for High-Throughput Phenotyping
Jorge Herskovic*, University of Texas M.D. Anderson Cancer Center and Elmer Bernstam, UT Health Sciences Center at Houston

Electronic Health Records aggregated in Clinical Data Warehouses (CDWs) promise to revolutionize Comparative Effectiveness Research and suggest new avenues of research. However, the effectiveness of CDWs is diminished by the lack of properly labeled data, and labeling techniques are cumbersome and expensive. I will present a novel approach that integrates clinical knowledge from the CDW generated during the course of care, the biomedical literature, and the Unified Medical Language System (UMLS) to perform high-throughput phenotyping. I will explain how we automatically construct a graphical knowledge model and then use it to phenotype breast cancer patients, and highlight the future of this approach.

Extending the Grammar of Graphics for Genomic Data: an R Implementation

Tengfei Yin, Iowa State University; Dianne Cook*, Iowa State University and Michael Lawrence, Genentech

This talk introduces new methodology to visualize and explore high-throughput genomic data, such as second generation sequencing data, in the context of genomic annotations. The methods leverage the statistical functionality available in R, build on the grammar of graphics (as implemented by ggplot) and the data handling capabilities of the Bioconductor project. The plots provide detailed views of genomic regions, edge-linked interval to data views, summary views of sequence alignments and splicing patterns, as well as genome-wide overviews with stacked, circular and grand linear layouts. Statistical summaries displayed in the overview guide the user to the interesting regions for closer inspection. Color schemes are carefully selected based on biological conventions, and cognitive perceptual principles. The methods are available in a new R package called ggbio. The package provides a high-level generic plot function to generate graphics with intelligent defaults, based on the flavor of the data.

Massive Parallelization of Serial Inference Algorithms for a Complex Generalized Linear Model

Marc Suchard, UCLA; Shawn Simpson, Columbia University; Ivan Zorych, Columbia University; Patrick Ryan, Johnson & Johnson Pharmaceutical Research and Development; and David Madigan*, Columbia University

Following a series of high-profile drug safety disasters in recent years, many countries are redoubling their efforts to ensure the safety of licensed medical products. Large-scale observational databases such as claims databases or electronic health record systems are attracting particular attention in this regard, but present significant methodological and computational concerns. In this paper we show how high-performance statistical computation, including graphics processing units, relatively inexpensive highly parallel computing devices, can enable complex methods in large databases. We focus on optimization and massive parallelization of cyclic coordinate descent approaches to fit a conditioned generalized linear model involving tens

of millions of observations and thousands of predictors in a Bayesian context. We find orders-of-magnitude improvement in overall run-time. Coordinate descent approaches are ubiquitous in high-dimensional statistics and the algorithms we propose open up exciting new methodological possibilities with the potential to significantly improve drug safety.

- **Applications of Interactive Graphics in R** Room 1064
  Organizer: David Scott, Rice University

Using R and Web Technologies to Create Analytics and Apps: Part of Making Statistics Relevant in a Large Organization
Chad Shaw, Baylor College of Medicine

Large organizations – whether companies or academic research institutions – present opportunities and challenges for statisticians. To be relevant in the organization its necessary for statisticians to know their customers and to deliver analytics that are both timely and domain specific. Web technologies can be central in accomplishing this work. We have developed a web-based analysis server we call Rho that uses front end web technologies to process requests that are then analyzed by back end R analysis engines. The software has a web client, a web-server layer, and a back-end layer for statistical analysis. The system is compatible with data storage on disk or in relational databases. The Rho system permits rapid deployment of customized analysis tools over the internet by web-enabling R. All components of the Rho system are built from locally created code using open source components. The web client for Rho can be either a Java applet, HTML generated http request. The web server in Rho is a custom servlet that extends the Tomcat servlet container. The back-end data analyses are performed in the R statistical computing environment. The R analysis engines are maintained in pools of network distributed, live sessions so that analysis is rapid. The servlet in the Rho system performs structured communication between analysis requests submitted over the internet and the available R analysis engines to achieve security and scalability.

Exploring Statistical Strategies for Use in Challenge-Response Experiments
Matthew S. Shotwell*, Kenneth J. Drake, Veniamin Y. Sidorov, and John P. Wikswo: Vanderbilt University School of Medicine

We consider an experimental approach where a system under study is subjected to controlled challenges, with the expectation that responses to these challenges will be informative about the underlying mechanism. For example, Vanderbilt researchers have examined the association of cardiac electrophysiology and cardiac metabolism by observing the time course of electrophysiological measures, such as action potential duration, under intermittent anoxic stress. Challenges are designed to have a

significant impact on the response, but often cause the response to change abruptly, to be irregular in shape, or become more variable. We consider parametric and non-parametric alternatives to accommodate these data complexities, and to address the relevant scientific goals. Lastly, we demonstrate a custom interactive graphic to aid in optimal design of challenge-response experiments.

The Anscombe Data Sets: Explained and Expanded
Jürgen Symanzik, Utah State University

In his 1973 paper, Anscombe introduced four data sets, each consisting of 11 (x, y) pairs. All four data sets had identical summary statistics, such as means, variances, and correlations and, therefore, they also yielded the same regression lines. However, the plots showed rather different patterns, and a meaningful regression line should have been calculated only for one of these (x, y) pairs. Anscombe (1973) did not report how he created these "fictitious data sets". Chatterjee and Firat (2007) who described a general method to create data sets with identical summary statistics but different graphical representations via a genetic-algorithm-based approach, indicated: "It is not known, however, how Anscombe came up with his datasets." In this presentation, we will describe how Anscombe most likely created his data sets and we will demonstrate how additional data sets with identical summary statistics can be created in a deterministic way.

- **Generalized Parallel Coordinates** Room 1070
  Organizer: Rida Moustafa, George Washington University & dMining Technology

Cluster Detection and Visualization with Generalized Parallel Coordinates
Rida Moustafa, George Washington University & dMining Technology

Visual pattern discovery in large multivariate datasets is a challenging problem in the fields of data mining and exploratory data analysis. This is due, in part, to the visual cluttering problem, which depends on screen resolutions and the number of points. The cluttering defies most information visualization techniques in general and parallel coordinates in particular. The cluttering effect increases with the number of data records, which makes the visual detection of hidden clusters, trends, correlations, periodicity, and anomalies even more difficult.
In this talk we discuss our hybrid plots called the quantized generalized parallel coordinate plot (QGPCP). The QGPCP detects the frequency of the profile lines (or curves), which represent the multivariate observations in parallel coordinate space, and maps this frequency into a gray (or HSV) scale color to highlight the profile lines (or curves) in a crowded GPCP. The approach has shown a great success in mitigating cluttering and detecting clusters in very large data not only in parallel coordinates but also the Andrews plot and the scatterplot matrix. We demonstrate the QGPCP

on cluster tracking and visualization on Remote sensing, Computer Network, and Housing data sets.

Visual Cluster and Outlier Detection with L-plot
Michael D. Larsen*, George Washington University & dMining Technology; Rida E. Moustafa, George Washington University & dMining Technology; and Ali S. Hadi, American University in Cairo & dMining Technology

The L-plot is a simple yet powerful $2D$ projection of multivariate data based base on the $L1$ and $L2$ measures. The visualization of these measures captures the linear and nonlinear structures from the data and reveals relationships of multivariate cases. The measures have high preserving rate of underlying data structures, and clusters and outliers can be easily identified. We consider various complex linear and nonlinear structures in theory and in simulated data as well as some well-known data sets to demonstrate the effectiveness of the plot.

Visual Analytics Approach for Social Network Interaction
Jie Cong, George Washington University & dMining Technology and Rida Moustafa, George Washington University & dMining Technology

Social network analysis has come into our sight since so many IT miracles regarding it has happened in the 21st century. Data visualization is one of the essential ingredients in analysis of this particular kind of network. In the first part of the paper, we use a large dataset (Slashdot social network from November 2008), to investigate the social network. We focus on data visualization methods and clustering methods, using different scaling and transformation methods to mitigate the cluttering and reveal the existing structures, especially power transfer function to solve this problem. In the second part, we turn to a Facebook dataset. Facebook was ranked the No. 1 picture sharing website in the US. This may give us a hint about the relationship of pictures and online social networks. Pictures posted tell about how people would want others to see them. The survey conducted in 2007 (the fastest-growing year in Facebook history) assists in conducting statistical research and answer the questions and provide comprehensive view about Facebook pictures. We mainly use non-parametric ANOVA to do the tests.

- **Contributed Paper Session IV** Room 1075
  Organizer: David Scott, Rice University

Parallel Monte Carlo Simulation for the Sensitivity Analysis of Expected Shortfall by Means of a Second-Order Approximation
Güven Gül Polat, Istanbul Technical University

The financial crisis of 2007-2009 has motivated academic research and supervisory policy agenda to better understand risk contribution to the market risk in order to

capture systemic risk. To this end, sensitivity analysis is performed via first derivatives of the market expected shortfall (ES) with respect to market allocation. The rate of return on the market is given by the weighted combination of the underlying equities returns in terms of arithmetic return. Since it is more adequate to work with logarithmic returns in risk assessment and weighted combination equation is only approximately achieved in this case, we consider a second-order approximation for the market logarithmic return. The estimation of ES and its sensitivity is based on Monte Carlo simulation utilizing embarrassingly parallel computing. Totally, in addition to the increase in the accuracy of the estimation by a higher order approximation, we demonstrate the acceleration of the simulation by a parallel execution on a distributed memory system.

Comparison of Binary Discrimination Methods for High Dimension Low Sample Size Data

Addy Bolivar-Cime, Rice University and J.S. Marron, University of North Carolina, Chapel Hill

A comparison of some binary discrimination methods is done in the high dimension low sample size context. In particular we obtain results about the asymptotic behavior of the methods Support Vector Machine, Mean Difference (i.e. Centroid Rule), Distance Weighted Discrimination, Maximal Data Piling and Naive Bayes when the dimension d of the data sets tends to infinity and the sample sizes of the classes are fixed. It is concluded that, under appropriate conditions, the first four methods are asymptotically equivalent, but the Naive Bayes method can have a different asymptotic behavior when d tends to infinity.

Factor Model for Forecasting with Multi-collinearity and Nonlinear Dependence

Joseph Egbulefu, Rice University

Empirical analysis of financial time series has identified non-linear dependence properties inherent in financial variables. Factors based on Principal Component Analysis and Partial Least Squares, empirical methods used for forecasting under multicollinearity, can be deficient in extracting certain non-linear properties. We construct a dynamic factor model for asset prices and returns using non-linear least squares to identify dependencies inherent in variables and constructing factor loadings from singular vectors of the data matrix under a suitable non-linear transformation. The method has been shown to outperform PCA and PLS forecasting when applied to high frequency exchange rates.

Relations Between Attentional Structure and Attentional Function: Utilization of Alternative Statistical Approaches

Paulina Kulesz, University of Houston

Objective: Structure-function relations in the domain of attention are not well understood. Limited research findings may stem from problems in estimating these relations in small samples combined with data distributions that do not conform to the assumptions of the statistics used to estimate the relations. We examined the utility of using alternative statistics to estimate structure-function relations in a small mixed sample.

Participants and Methods: Participants were 61 children (43 spina bifida, 18 normal controls) evaluated in a larger study examining cognitive and neurobiological variability in spina bifida and related disorders. We used the Pearsons Correlation and four robust correlations: the Percentage Bend Correlation, the Winsorized Correlation, the Skipped Correlation using the Donoho-Gasko Median, and the Skipped Correlation using Minimum Volume Ellipsoid Estimator (MVE) to examine structure-function relations in the domain of attention. A bootstrap sampling process was used to compare performance of the five estimators in this field context.

Results: In general, three patterns of relations were observed: (a) all estimators performed similarly, (b) the Pearson estimate differed from the four robust estimators, (c) the Skipped correlation using MVE differed from the other three robust estimators, which were comparable to the Pearson estimate. The three patterns of results were not readily associated with deviations from bivariate normality in specific structure-function relation being studied as reflected in measures of multivariate skew and kurtosis.

Conclusions: Using alternative approaches to estimate relations can assist investigators when confronted with small samples and multivariate non-normal data. Utilization of the Pearson correlation along with robust correlations can strengthen inferences about variable relations. Using the bootstrap to obtain empirical distributions for the estimates can further strengthen conclusions about variable relations. The similarity of estimates across methods suggested that the lack of structure-function relations found in the literature is not easily attributed to violations of distributional assumptions.

Joint work with D.J. Francis, T.S. Tian, and J.M. Fletcher.

Index