Abstracts

Wednesday, June 13

Plenary Session I

Visual Computing in Connectomics
Keynote Speaker: Hanspeter Pfister, Harvard University

Our modern ability to acquire and generate huge amounts of data can potentially enable rapid progress in science and engineering, but we may not live up that promise if our ability to create data outstrips our ability to make sense of that data. Visual computing tools are essential to gain insights into data by combining computational and statistical analysis with the power of the human perceptual and cognitive system and enabling data exploration through interactive visualizations. In this talk I will present our work on visual computing in Connectomics, a new field in neuroscience that aims to apply biology and computer science to the grand challenge of determining the detailed neural circuitry of the brain. I will give an overview of the computational challenges and describe visual computing approaches that we developed to discover and analyze the brain's neural network. The key to our methods is to keep the user in the loop, either for providing input to our fully-automatic reconstruction methods, or for validation and corrections of the reconstructed neural structures. The main challenges we face are how to analyze petabytes of image data in an efficient and scalable way, how to automatically reconstruct very large and dense neural circuits from nanoscale-resolution electron micrographs, and how to analyze the brain's neural network once we have discovered it.

Invited Session 1: Computer Experiments I

Computationally Efficient Use of Derivatives in Emulation of Complex Computational Models
Brian J. Williams, Los Alamos National Laboratory

We will investigate the use of derivative information in complex computer model emulation when the correlation function is of the compactly supported Bohman class. To this end, a Gaussian process model similar to that used by Kaufman et al. (2011) is extended to a situation where first partial derivatives in each dimension are calculated at each input site (i.e. using gradients). Simulation studies are conducted to assess the utility of the Bohman correlation function against strictly positive correlation functions when a high degree of sparsity is induced. (Joint work with Peter Marcy.)

Designs for Computer Experiments with Gradient Information
Fred J. Hickernell, Illinois Institute of Technology

Sometimes it is possible to obtain gradient values along with function values when performing computer experiments. These gradient values are obtained via adjoint methods. For large numbers of factors and a moderate number of runs, low degree polynomial regression models may be the best choice for constructing a surrogate. Spreading the design points uniformly over the domain of possible factor values maintains reasonable estimation efficiency over a large class of regression models when the model is unknown. If the form of regression model is known, then the optimal design can be computed using semi-definite programming. These optimal designs differ from those where only function values are used.

Cases for the Nugget in Modeling Computer Experiments
Robert B. Gramacy, University of Chicago

Most surrogate models for computer experiments are interpolators, and the most common interpolator is a Gaussian process (GP) that deliberately omits a small-scale (measurement) error term called the nugget. The explanation is that computer experiments are, by definition, "deterministic", and so there is no measurement error. We think this is too narrow a focus for a computer experiment and a statistically inefficient way to model them. We show that estimating a (non-zero) nugget can lead to surrogate models with better statistical properties, such as predictive accuracy and coverage, in a variety of common situations.

Invited Session 2: Innovation

Innovation, Quality Engineering, and Statistics
William H. Woodall, Virginia Tech

Some general characteristics and principles of innovation and the innovation process will be described. Examples will be given from industry and from statistical science. General advice on how individuals and companies can become more innovative will be summarized. The relationship between quality engineering, including Six Sigma, and innovation will be briefly discussed. How statistics can be used in the innovation process will be outlined. Ideas will be shared and solicited on how our profession can become more actively involved in innovation and more adequately recognized for its contributions.

The Role of a Statistician in Innovation
Willis A. Jensen, W. L. Gore & Associates

Innovation is a very hot topic of discussion in many circles, including those in the statistics field. Our focus in this presentation is the use of existing statistical methods to spur innovation, which we define as some improvement in a product or process that results in increased profits or a more efficient use of existing resources. We provide some specific examples in our work where statistical methods have led to incremental innovations. We share some important principles related to innovation that are

common in these examples. In addition, we discuss the role a statistician can play in facilitating innovation based on these principles.

Invited Session 3: Industrial Statistics in Europe

Industrial Experiments Using Supersaturated Designs
Dave Woods, University of Southampton, UK

Supersaturated designs are useful tools for industrial screening experiments that include many factors and limited resource. Such designs are often generated and assessed using criteria such as E(s2)- or D-optimality, which encapsulate aspects of design performance based on pairwise column correlations. Such measures do not necessarily relate directly to the aim of the experiment, which is typically to identify a high proportion of the active, or important, factors whilst declaring few unimportant factors as active. Usually, it will also be necessary to consider linear dependencies between more than two factor columns. This talk will discuss the selection, assessment and application of supersaturated designs, motivated and demonstrated by examples from manufacturing in the pharmaceutical and chemical industries. Assessment is via simulation studies that vary many standard features of an experiment: the number of factors, the design, and the data generating process. From these results, some guidance on the effectiveness of supersaturated designs is established. Some new design selection criteria are motivated from the simulations, and demonstrated with respect to the industrial examples. (Joint research with Chris Marley and Sue Lewis (University of Southampton), and scientists from Glaxo-SmithKline and Lubrizol.)

Statistical Process Control for Time Varying Processes – Cases, Issues, Ideas
Bart De Ketelaere, Katholieke Universiteit Leuven, Belgium

Statistical Process Control (SPC) is a powerful framework that is used in many industries to decrease process variability and to pinpoint special cause variation. Although a broad range of techniques are developed many years ago, often the real-life situation does not fully comply with the basic assumptions that are made in SPC resulting in poor results. One of the main violations against the assumptions is the fact that processes rarely behave stationary – this is evidently the case for biological processes (monitoring humans, crops or animals) but is also an important issue when monitoring industrial processes. Besides, the ever increasing amount of data, with a clear shift towards multivariate and even multiway quality control, makes that the classical univariate SPC approach is not feasible anymore. These two observations pose important challenges to statisticians to develop novel SPC algorithms that are broadly applicable in modern industries. In this talk we discuss both issues and use two very different case studies to show recent directions and developments in the SPC landscape. (Joint work with Kristof Mertens, Tjebbe Huybrechts and Josse De Baerdemaeker.)

Analysis of Categorical Data from a Polypropylene Experiment
Steven Gilmour, University of Southampton, UK

An increasing number of industrial experiments are run using multi-stratum designs, the simplest examples of which are split-plot and strip-plot designs. Often, these experiments span more than one processing stage. The challenge is to identify an appropriate multi-stratum design, along with an appropriate statistical model. In this talk, we introduce Hasse diagrams in the response surface context as a tool to visualize the unit structure of the experimental design, the randomization and sampling approaches used, the stratum in which each of the experimental factors is applied, and the degrees of freedom available in each stratum to estimate main effects, interaction effects and variance components. We illustrate the usefulness of the Hasse diagrams by focusing on several responses measured in the context of a large study conducted for investigating the adhesion properties of coatings to polypropylene for the automobile industry in Belgium. We discuss various types of quantitative responses, binary responses and ordered categorical responses, for designs ranging from a simple split-plot design to a strip-plot type of design involving repeated measurements of the responses.

Invited Session 4: Active Learning and Sequential Design

The Label Complexity of Active Learning
Steve Hanneke, Carnegie Mellon University

This talk describes recent progress on generally characterizing the number of label requests sufficient for active learning to achieve a given accuracy (called the label complexity), in both noisy and noise-free pattern recognition settings. In particular, we will focus on sufficient conditions for this label complexity to be significantly smaller than that of learning from random labeled samples (passive learning).

Exploiting Saliency in Compressive and Adaptive Sensing
Jarvis Haupt, University of Minnesota

Recent developments in compressive and adaptive sensing have demonstrated the tremendous improvements in sensing resource efficiency that can be achieved by exploiting sparsity in high-dimensional inference tasks. In this talk we discuss a framework for interpreting saliency as a natural generalization of sparsity, and we describe how compressive and adaptive sensing techniques can be extended and applied to this more general model. We will discuss our recent work quantifying the performance of these “compressive saliency sensing” procedures, and we demonstrate the approach in a two-stage active compressive imaging approach to automated surveillance.

Variational Approximations of Bayesian Inference for Large Scale Automated Decision Making
Matthias W. Seeger, EPFL, Switzerland

Bayesian decision making is driven by queries to the posterior distribution, obtained by conditioning a probabilistic model and prior knowledge on acquired data. This process, which has to be repeated many times in a sequential pipeline, is computationally challenging in general and remains out of reach for large scale coupled models of images or image sequences. We give an overview of recent variational techniques, with which the Bayesian computational challenge can be met approximately. Intractable posterior integrations are relaxed to variational optimization problems. Using novel decoupling techniques and double loop algorithms, these problems can be mapped to standard large scale optimization primitives such as penalized least squares estimation, as well as numerical mathematics and randomized techniques in order to estimate posterior covariance information over signals with several hundred thousand coefficients. We present results of our methodology for the problem of optimizing magnetic resonance imaging acquisitions by Bayesian sequential optimal design.

Invited Session 5: Technometrics Session

Bayesian Computation Using Design of Experiments-Based Interpolation Technique
V. Roshan Joseph, Georgia Institute of Technology

A new deterministic approximation method for Bayesian computation, known as Design of Experiments-based Interpolation Technique (DoIt), is proposed. The method works by sampling points from the parameter space using an experimental design and then fitting a kriging model to interpolate the unnormalized posterior. The approximated posterior density is a weighted average of normal densities and therefore, most of the posterior quantities can be easily computed. DoIt is a general computing technique which is easy-to-implement and can be applied to many complex Bayesian problems. Moreover, it does not suffer from the curse of dimensionality as much as some quadrature methods. It can work using fewer posterior evaluations, which is a great advantage over the Monte Carlo and Markov chain Monte Carlo methods especially when dealing with computationally expensive posteriors.

Invited Session 6: Classical Experimental Design

Templates for Design Key Construction
C. S. Cheng, University of California, Berkeley

We present and justify some useful templates for implementing design key construction of factorial designs with simple block structures, in particular those for the construction of unblocked and blocked split-plot and strip-plot factorial designs. The traditional method of constructing such designs is to use some independent treatment factorial effects to partition the treatment combinations into blocks, rows, columns, etc. One advantage of the design key construction is that a set of independent generators and the constraints imposed by the structures of the

experimental units are built in the template, which facilitates a systematic and simple construction of the design layout and eliminates the need to check some conditions for design eligibility when the traditional method is used.

Covariate-Adaptive Designs for Personalized Medicine
Feifang Hu, University of Virginia

In decades, scientists have identified genes (biomarkers) that seem to be linked with diseases. To translate these great scientific findings into real-world products for those who need them (Personalized Medicine), clinical trials play an essential and important role. New approaches to the drug-development paradigm are needed, especially new designs for clinical trials so that genetics and other biomarkers can be incorporated to assist in patient and treatment selection. In this talk, I will be focusing on clinical trial designs that use genetics or other biomarkers. New designs are proposed and their properties are discussed. Some further research problems will also be discussed.

On an Extension for Identifying Locally Optimal Designs for Nonlinear Models
John Stufken, University of Georgia

We describe an extension to an approach in Yang (*Annals of Statistics* **38**, 2010, 2499-2524) for identifying locally optimal designs for nonlinear models. We demonstrate the consequences of this extension through examples. We will see that, for many of the cases considered, the extension enables the identification of locally optimal designs with the minimal number of support points. This presentation is based on joint work with Min Yang.

Invited Session 7: Statistical Process Control

Uncertainty Quantification in Change Detection
Snigdhansu Chatterjee, University of Minnesota

Change detection, stability analysis, and process control and monitoring have a wide range of applications. Different methodologies have evolved to handle these topics in different disciplines, and it is not always clear what level of uncertainty is associated with a declared change, or whether there is any uncertainty quantification at all. We propose resampling-based measurement of uncertainty for change detection and process control problems. A p-value approach seems the most suitable for many cases, and details of this approach are studied. (This research is joint with Zhonghua Li, Ying Lu, Jaya Kawale, Karsten Steinhauser, Stefan Leiss, Peihua Qiu and Zhaojun Wang.)

Monitoring the Covariance Matrix with Fewer Observations than Variables
Edgard Maboudou, University of Central Florida

In real applications, when a change occurs in a multivariate process, it occurs in either location or scale. Several methods have been proposed recently to monitor the

covariance matrix. Most of these methods deal with a full rank covariance matrix, i.e., a situation where the number of rational subgroups is larger than the number of variables. In high dimensional problems, where the number of features is nearly as large as, or larger than the number of observations, existing methods do not provide a satisfactory solution because the estimated covariance matrix is singular. In this talk, we present a new method for sequentially detecting changes in the covariance matrix of a multivariate Gaussian process when the number of observations available is less than the number of variables.

Profile Control Charts Based on Nonparametric L-1 Regression Methods
Ying Wei, Columbia University

Classical statistical process control often relies on univariate characteristics. In many contemporary applications, however, the quality of products must be characterized by some functional relation between a response variable and its explanatory variables. Monitoring such functional profiles has been a rapidly growing field due to increasing demands. This paper develops a novel nonparametric L-1 location-scale model to screen the shapes of profiles. The model is built on three basic elements: location shifts, local shape distortions, and overall shape deviations, which are quantified by three individual metrics. The proposed approach is applied to the previously analyzed vertical density profile data, leading to some interesting insights. (Joint work with Zhibiao Zhao and Dennis K. J. Lin.)

Invited Session 8: Change-Point Detection

PELT: Optimal Detection of Changepoints with a Linear Computational Cost
Rebecca Killick, Lancaster University, UK

We consider the problem of detecting multiple changepoints in large oceanographic data sets. In this setting the amount of data being collected is increasing and consequently the number of changepoints will also increase with time. An efficient and accurate analysis of such data is of considerable interest to those working in the energy sector as understanding the characteristics of the ocean environment is central to reliable design and operation of marine and coastal structures. Detecting the presence of changepoints in oceanographic time-series is of particular importance, since statistical and engineering modelling of the ocean environment, structural loading and response typically assumes stationarity of the environment (in time). Traditional methods for identifying multiple changepoints within this type of data are either computationally efficient and statistically approximate, or computationally slow and statistically exact. For large data sets computational efficiency is important but statistical approximations can have a large impact on any potential inferences. Thus we introduce a new, statistically exact, method for identifying the optimal number and location of changepoints which has a computationally cost, under mild conditions, that is linear in the number of observations. We demonstrate this method, called PELT, on an oceanographic data set. (Joint work with Paul Fearnhead and Idris Eckley.)

Approximate Simulation-Free Bayesian Inference for Multiple Changepoint Models with Dependence within Segments
Nial Friel, University College Dublin

This paper proposes approaches for the analysis of multiple changepoint models when dependency in the data is modelled through a hierarchical Gaussian Markov random field. Integrated nested Laplace approximations are used to approximate data quantities, and an approximate filtering recursions approach is used which results in a computational saving when detecting changepoints. Analysis of real data demonstrates the usefulness of this approach. The new models which allow for data dependence are compared with conventional models where data within segments is assumed independent. This work is in collaboration with Jason Wyse (Trinity College, Dublin) and Havard Rue (NTNU, Trondheim).

Group Fused Lasso for Multiple Change-Point Detection
Kevin Bleakley, INRIA, France

We present the group fused Lasso for detection of change-points shared by a set of co-occurring one-dimensional signals. Change- points are detected by approximating the original signals via a constraint on the multidimensional total variation, leading to piecewise-constant approximations. Fast algorithms are proposed to solve the resulting optimization problems, either exactly or approximately. Conditions are given for consistency of these algorithms as the number of signals increases, and empirical evidence is provided to support the results on simulated and biological (aCGH) data.

Invited Session 9: Special Session on Experimental Design

The Recondite Ability of Potential Outcomes in Experimental Design
Donald B. Rubin, Harvard University

The talk will point out that, despite the critically important role that Neyman's (1923) introduction of the POTENTIAL OUTCOME notation played in the formal understanding of the physical act of randomization for formal unbiased estimation and valid hypothesis testing, and Fisher's implicit use of this sort of thinking as far back as 1918, mistakes were made by both (1935) and apparently never corrected (e.g., in Latin squares). Moreover, because of computational limitations until the late 20th century, the "mindless" application of OLS normal-theory super-population models took over the field of experimental design, including its standard textbooks. As a consequence, there appear, even currently, a variety of not fully correct statements that are widely accepted, for example, about the importance of additive treatment effects (e.g., correlations between potential outcomes), even in 2^k designs. Another consequence was the use of generally inappropriate language to define confidence intervals and hypothesis tests as "valid" and "conservative" – which CAN contradict the literal meanings of those English words. Furthermore, the focus on OLS estimation still leads to the widespread thoughtless use of models for the implicit imputation of missing potential outcomes that assumes additivity and linearity, despite the availability of Bayesian computational tools with far greater flexibility to conduct

realistic modeling. In some sense, this talk will be a "Back to the future" one that highlights the enlightening role that potential outcomes and randomization, keystones of experimental design in the field's infancy, but replaced in subsequent decades with the automatic application of OLS with its often inapposite assumptions, should still play in the future development of experimental design. (Based on joint projects involving Arman Sabbaghi, Cassandra Wolos Pattanayek, Valeria Espinosa, Natesh Pillai and Tirthankar Dasgupta.)

Three-Phase Sequential Design for Sensitivity Experiments
C. F. Jeff Wu, Georgia Institute of Technology

In sensitivity testing the test specimens are subjected to a variety of stress levels to generate response or non-response. These data are used to estimate the critical stimulus (or threshold) level of the experimental object. Because of its versatile applications, several sensitivity testing procedures have been proposed and used in practice. There remains the outstanding question of finding an efficient procedure, especially when the sample size is small and the interest lies in the extreme percentiles. In the paper we propose a novel three-phase procedure, dubbed 3pod, which can be described as a trilogy of "search-estimate-approximate". A core novel idea is to choose the design points to quickly achieve an overlapping data pattern which ensures the estimability of the underlying parameters. Simulation comparisons show that 3pod outperforms existing procedures over a range of scenarios in terms of efficiency and robustness. (Joint work with Yubin Tian.)

Invited Session 10: Reliability

Reliability and Risk Issues and the Nuclear Deterrent
Aparna Huzurbazar, Los Alamos National Laboratory

Deterrence, as a component of military strategy, is certainly not a concept confined to nuclear deterrence. At its simplest, as defined by the Department of Defense (DoD), deterrence is "the prevention from action by fear of the consequences. Deterrence is a state of mind brought about by the existence of a credible threat of unacceptable counteraction." (DoD Dictionary, available at: http://www.dtic.mil/doctrine/dod_dictionary/). At Los Alamos National Laboratory (LANL), our mission is to develop and apply science and technology to ensure the safety, security, and reliability of the U.S. nuclear deterrent. To do this, we must use the cutting edge methods from reliability and risk assessment that ultimately inform decision makers. We will illustrate with methods and applications from our work at LANL.

Bayesian Multistate Models for Nuclear Power Plant Reliability
David Collins, Los Alamos National Laboratory

We present several multistate stochastic process models for predicting the reliability of nuclear power plant (NPP) piping subsystems; this is an important area, since typical NPPs may have up to 40 miles of piping, much of it safety-critical (e.g., the main reactor coolant loop). We represent subsystems as statistical flowgraphs, with vertices

representing states of partial or complete failure and edges representing probability distributions for transitions between states. Failure transitions are driven by processes based on material properties of the pipes, and the physical and chemical dynamics of the fluid being carried. Repair transitions are based on detection of leakage by visual inspection, or non-visible flaws by radiography or ultrasound. Since failures in these subsystems are rare, Bayesian methods are used to incorporate scientific and engineering judgment into probability distributions for state waiting times.

We first present a simple Markov model, which can be quickly iterated thousands of times with parameter samples generated using Markov Chain Monte Carlo (MCMC); a solution in each iteration can be computed using either flowgraph methods or differential equations. A more realistic semi-Markov model is then presented, which allows arbitrary waiting-time distributions. Iterating the numerical solution of this model, using integral equations or standard flowgraph techniques, is computationally intensive; we discuss simulation and the fast Fourier transform as alternatives. (Joint work with Aparna V. Huzurbazar, Brian J. Williams, Richard L. Warr.)

Application of Reliability Modeling to Non-Linear Repeated Measure Degradation Stability Data
Fangyi Luo* and William Brenneman, Procter & Gamble

In product development we conduct accelerated product stability testing to establish the relationship between stability at use conditions and at accelerated conditions. Stability data are typically collected as degradation data over time. A two step approximate degradation analysis is currently used to model this type of data. First, a non-linear mixed effects model is fit to the repeated measure degradation data. Second, a reliability model is developed for the pseudo time to stability failure data estimated from the first step. By designing an experiment with accelerated and non-accelerated factors, the resulting reliability model can be used to make predictions of product stability for new products. The statistical challenges we are facing include designing accelerated non-linear degradation tests with multiple accelerating and non-accelerating factors, developing a non-linear degradation model to predict degradation at use conditions, and incorporating the prediction errors estimated from the non-linear modeling to the reliability modeling in the approximate degradation analysis.

Contributed Session 1: Machine Learning

Undirected Graphical Model Selection Using Tree Decompositions: Tractable Algorithms and Active Learning
Divyanshu Vats, University of Minnesota

An undirected graphical model is a joint probability distribution defined on an undirected graph G = (V,E), where the nodes in the graph index a collection of random variables and the edges encode conditional independence relationships amongst random variables. The undirected graphical model selection (UGMS) problem is to estimate the graph G given observations drawn from the undirected graphical model. We show how tree decompositions can be used to decompose the UGMS problem into

multiple subproblems over a small number of nodes. Tree decompositions have been used for inference over graphical models (the junction tree algorithm), however, to our knowledge, there has not been any work that explores the use of tree decompositions for UGMS. We show that under certain conditions on the graphical model, using tree decompositions for UGMS leads to weaker sufficient conditions for high-dimensional consistent graph estimation. Further, when the graph is sparse, the tree decomposition approach leads to tractable algorithms for UGMS since the computational complexity depends on the largest subproblem. Finally, we propose an algorithm for active learning for UGMS using tree decompositions that sequentially draws observations from the graphical model based on prior observations. In the high-dimensional setting, we show that the sufficient conditions on the number of scalar observations needed for the active algorithm is less than that needed for the non-active algorithm. Intuitively, the active learning algorithm draws more measurements from the parts of the graph that are difficult to learn and less measurements from the parts of the graph that are easy to learn.

Multi-Relational Learning via Hierarchical Nonparametric Bayesian Collective Matrix Factorization
Hongxia Yang, IBM Watson Research Center

Relational learning addresses problems where the data come from multiple sources and are linked together through complex relational networks. Two important goals are pattern discovery (e.g. by (co)-clustering) and predicting unknown values of a relation, given a set of entities and observed relations among entities. In the presence of multiple relations, combining information from different but related relations can lead to better insights and improved prediction. For this purpose we propose a nonparametric hierarchical Bayesian model that improves on existing collaborative factorization models and frames a large number of relational learning problems. In contrast to exiting methods, the proposed model naturally incorporates (co)clustering and prediction analysis in a single unified framework, and allows for the estimation of entire missing row or column vectors. We develop an efficient Gibbs algorithm and a hybrid Gibbs using Newton's method to enable fast computation in high dimensions. We demonstrate the value of our framework on simulated experiments as well as two real world problems: discovering kinship systems and predicting the authors of certain articles based on article-word co-occurrence features.

Probabilistic Hashing Methods for Fitting Massive Logistic Regressions and SVM with Billions of Variables
Ping Li, Cornell University

In modern applications, many statistics tasks such as classification using logistic regression or SVM often encounter extremely high-dimensional massive datasets. In the context of search, certain industry applications have used datasets in 2^{64} dimensions, which are larger than the square of billion. This talk will introduce a recent probabilistic hashing technique called b-bit minwise hashing (Research Highlights in Comm. of ACM 2011), which has been used for efficiently computing set similarities in massive data. Most recently (NIPS 2011), we realized that b-bit minwise hashing can be seamlessly integrated with statistical learning algorithms such as

logistic regression or SVM to solve extremely large-scale prediction problems. Interestingly, for binary data, b-bit minwise hashing is substantially much more accurate than other popular methods such as random projections. Experimental results on 200GB data (in billion dimensions) will also be presented.

Wisely Using a Budget for Crowd Sourcing
Seyda Ertekin, Massachusetts Institute of Technology

The problem of "approximating the crowd" is that of estimating the crowd's majority opinion by querying only a subset of it. Algorithms that approximate the crowd can intelligently stretch a limited budget for a crowdsourcing task. We present an algorithm, "CrowdSense," that works in an online fashion where examples come one at a time. CrowdSense dynamically samples subsets of labelers based on an exploration/exploitation criterion. The algorithm produces a weighted combination of a subset of the labelers' votes that approximates the crowd's opinion. We then introduce two variations of CrowdSense that make various distributional assumptions to handle distinct crowd characteristics. In particular, the first algorithm makes a statistical independence assumption of the probabilities for large crowds, whereas the second algorithm finds a lower bound on how often the current sub-crowd agrees with the crowd majority vote. Our experiments on CrowdSense and several baselines demonstrate that we can reliably approximate the entire crowd's vote by collecting opinions from a representative subset of the crowd.

Plenary Session II

Nonparametric Bayesian Models for Sparsity, Networks, Time Series and Covariances
Keynote Speaker: Zoubin Ghahramani, University of Cambridge, UK

Probability theory offers a powerful framework for modelling which can be applied to nearly all inference and prediction problems. I will outline the basics of the probabilistic framework, and motivate the use of Bayesian nonparametrics as a natural approach to flexible modelling of complex data sources. I will then illustrate four examples of our recent work in this area: nonparametric hidden Markov models for time series, the Indian Buffet Process as a general approach to modelling sparse matrices, probabilistic models of social and biological networks, and models for covariance and volatility based on copulas and generalised Wishart processes.

Invited Session 12: New Challenges and Directions in Industrial Statistics

Predictive Consumer Modeling with Product Variables vs. Technical Measurements
Pradipta Sarkar, Procter & Gamble

Typical consumer surveys are observational studies. Analyses of these studies help us to understand the correlation between consumer propensity to buy and other attributes. Models of these studies are usually unable to reliably predict changes in consumer ratings resulting from product composition changes. In this presentation we will discuss statistical methodologies to predict consumer response from changes in product composition. Some of the topics to be included are designing consumer studies in “product space” vs. “technical measure space”, sample size determination for large consumer design of experiments, logistic difficulties associated with such tests and how to handle them, selection of consumer design variables, importance of validated technical test methods for predictability of technical-to-consumer models, validation and re-usability of models, choice of variable selection methods. We’ll also talk about some modeling challenges and look forward to ideas from the audience.

Uncertainty Quantification and Optimization Under Uncertainty for a Hypersonic Vehicle
Andrew Booker, The Boeing Corporation

This talk will discuss experiences and challenges at Boeing with Uncertainty Quantification (UQ) and Optimization Under Uncertainty (OUU) in conceptual design problems that use complex computer simulations. The talk will describe tools and methods that have been developed and used by the Applied Math group at Boeing and their perceived strengths and limitations. Application of the tools and methods will be illustrated with an example in conceptual design of a hypersonic vehicle. Finally I will discuss future development plans and needs in UQ and OUU.

Thermal Zone Mapping in Data Center Management Using Prototype-Based Spatio-Temporal Clustering
Angela Schoergendorfer* and Huijing Jiang, IBM Watson Research Center

To operate a large data center efficiently and economically, a number of air conditioning units (ACU's) need to be monitored. For this purpose, the concept of creating practical thermal zones has been proposed, where each zone represents a region influenced by a particular ACU. Such thermal zones can provide a practical representation of the cooling system for ACU operation management. This talk introduces a Bayesian hierarchical model-based clustering approach for thermal zone representation when a network of thermal sensors is deployed to monitor real-time temperature in a data center. In our approach, thermal zones are delineated by associating sensor temperature time-series with ACU discharge temperature profiles – the set of fixed cluster prototypes. A functional data analysis approach incorporating spatial dependence is developed to model patterns over time and space.

Contributed Session 2: Computer Experiments

New Research Directions in Computer Experiments: Clustered Designs
Selden Crary, NewallStreet, Inc.

I describe four fertile research areas in optimal design and analysis of computer experiments that follow from the discovery of proximal points (clusters) in such designs, under the minimum-Gaussian-process-*IMSE* objective and assuming fixed covariance parameters, as follows: (1) *Design generation*: I discuss numerical precautions and procedures useful in such searches. (2) *Phases and phase transitions*: Each design falls into one of the proper subgroups of the symmetry group of the design domain. In the domain of the covariance parameters, there are contiguous regions of optimal designs with identical subgroup membership, as well as phase transitions corresponding with boundaries between these regions. I demonstrate one, especially rich, phase diagram. (3) *Theory of clustered designs*: The IMSE objective function of clustered designs is, almost always, an essential discontinuity. In addition, IMSE is a continuous function of the coordinates of the design points, with all derivatives continuous, unless points actually merge. (4) *Applications*: I discuss the application of clustered designs to sequential generation of metamodels, as well as the application of clustered-design theory to the numerical inversion of highly ill-conditioned covariance matrices.

Designs for Computer Experiments that Minimize the Expected Integrated Mean Square Prediction Error
Erin Leatherman, Ohio State University

A computer experiment uses a computer simulator based on a mathematical model of a physical process as an experimental tool to determine "responses" or "outputs" at a set of user-specified "input" sites. When it is of interest to predict simulator output over the entire input space, classical design criteria for computer experiments select designs that are space-filling. This research investigates an alternative design criterion which minimizes the Expected

Integrated Mean Square Error (EIMSE). The EIMSE is calculated assuming a hierarchical Bayesian model for the unknown output function. In addition to applying the minimum EIMSE criterion, this research also searches for a surrogate for EIMSE, and explores subclasses of designs that produce small criterion values.

Integrating Analytical Models with Finite Element Models: An Application in Micromachining
Shan Ba, Georgia Institute of Technology

We consider the problem of integrating analytical models with finite element simulations. We show that computationally cheap analytical models can be used to perform a sensitivity analysis which can reveal critical information about the underlying system prior to conducting the computationally intensive simulation study. We propose a two-stage sequential strategy, which can efficiently absorb the prior information from the sensitivity analysis and assign a customized number of levels for each input variable in the finite element simulations. The method is also applicable for integrating other types of models having different levels of accuracy and speed. A case study for developing force metamodels in micromachining is presented to illustrate the effectiveness of the proposed method. (This is joint work with Nikhil Jain, V. Roshan Joseph and Ramesh Singh.)

Adaptive Designs for Modeling and Optimization of Computer Experiments
Xu Xu, University of Wisconsin-Madison

Adaptive designs are appealing for experiments with expensive computer codes. Instead of using a space-filling design like a Latin hypercube design, a criterion is needed for such an adaptive scheme. Existing criteria include maximum predictive variance for the prediction purpose and maximum expected improvement for the optimization purpose. We propose a robust alternative to these criteria. The new criterion is a hybrid function of the cross-validation error estimate and a distance-based weighting function. Numerical examples are provided to illustrate the effectiveness of the proposed method in approximation and global optimization of computer experiments. This is joint work with Peter Qian at the University of Wisconsin-Madison.

Invited Session 13: Machine Learning

Removing Confounding Factors via Constraint-Based Clustering: An Application to Finding Homogeneous Groups of MS Patients
Carla Brodley, Tufts University

Confounding factors in unsupervised data can lead to undesirable clustering results. For example in medical datasets, age is often a confounding factor in tests designed to judge the severity of a patient's disease through measures of mobility, eyesight and hearing. In such cases, removing age from each instance will not remove its affect from the data as other features will be correlated with age. We present a method based

on constraint-based clustering to remove the impact of such confounding factors and compare it to the standard approach of detrending. Motivated by the need to find homogenous groups of MS patients, we apply our approach to remove physician subjectivity from patient data. The result is a promising novel grouping of patients that can help uncover the factors that impact disease progression in MS.

Modeling Time Series Dependence for Scoring Sleep in Mice
Abraham Wyner, University of Pennsylvania

Current methods for scoring sleep behavior in mice are expensive, invasive, and labor intensive, thus leading to considerable interest in high-throughput automated systems which would allow many mice to be scored cheaply and quickly. Previous efforts have been able to differentiate sleep from wakefulness, but cannot differentiate the rare and important state of REM sleep from non-REM sleep. Key difficulties in detecting REM are that (i) REM is much rarer than non-REM and wakefulness, (ii) REM looks similar to non-REM in terms of the observed covariates, (iii) the data is noisy, and (iv) the data contains strong time dependence structures crucial for differentiating REM from non-REM. We develop a novel statistical approach which embeds statistical learning methods into generalized Markov models to account for this time dependence. Our methodology can accommodate very general and very long-term dependence structures in an easily estimable and computationally tractable fashion. We show improved differentiation of REM from non-REM sleep in our application to sleep scoring in mice.

Learning Models of Human Action from Image and Videos on the Web
Stan Sclaroff, Boston University

To date, most research in human action recognition in video has focused on videos taken in controlled environments working with limited action vocabularies; however, real world videos rarely exhibit such consistent and relatively simple settings. Instead, there is a broad range of environments where the actions can possibly take place, together with a large variety of possible actions that can be observed. We are developing methods that can work better with uncontrolled videos, for instance, home movies and YouTube videos. The main idea is to use images and videos collected from the Web to learn representations of actions and use this knowledge to automatically annotate actions in videos. We exploit features of the detected humans in the images and videos, features of the detected moving objects, and features of the overall scene. Our approach is unsupervised in the sense that it requires no human intervention other than the action keywords to be used to form text queries to Web image and video search engines. Thus, we can easily extend the vocabulary of actions, by simply making additional search engine queries. Experiments show the benefits of this approach in two areas: improving the retrieval precision of human action images, and automatically tagging human actions in YouTube videos. This is collaborative work with Nazli Ikizler-Cinbis at Hacettepe University and Shugao Ma at Boston University.

Growing a List
Cynthia Rudin, Massachusetts Institute of Technology

We would like to intelligently grow a long list, starting from a small seed of examples. Our algorithm for solving this problem takes advantage of the wisdom of the crowd, in the sense that there are many experts who post lists of things on the Internet. We want both to find these experts, and aggregate their lists in an intelligent way in order to produce a single concise, complete, and meaningful list. We give examples of finding a list of events in and around Boston, and finding a list of Jewish foods. (This is work with Ben Letham and Katherine Heller.)

Invited Session 14: Computer Experiments II

Multiobjective Optimization of Expensive Black-Box Functions via Expected Maximin Improvement
Thomas Santner, Ohio State University

Many engineering design optimization problems contain multiple objective functions all of which must be minimized, say. This talk proposes a method to solve such problems by identifying the Pareto Set of inputs; a given input (vector) belongs to the Pareto Set if and only if there is no competing vector of inputs that simultaneously decreases the value of all the objective functions. The proposed methodology is particularly intended to handle cases when the objective functions are expensive to compute. In brief, the method replaces the objective function evaluations by a rapidly computable approximator based on an interpolating Gaussian process (GP) model. It sequentially selects new input sites guided by an improvement function, i.e., the next input at which each output is to be evaluated is that input vector that maximizes the conditional expected value of this improvement function given the current data. The proposed method provides two advances within this framework. First, it introduces an improvement function based on the modified maximin fitness function. Second, it implements a family of GP models that allow for dependent output functions but which permits zero covariance should the data be consistent with a model of no association. Examples from the literature are presented to show that the proposed procedure can improve substantially previously proposed statistical improvement criteria for the computationally intensive multiobjective optimization setting. (Joint work with Josh Svenson.)

A New Class of Alpha-Stable Processes: Modeling and Bayesian Computation
Rui Tuo, Georgia Institute of Technology

In this talk we consider the conditional inference for alpha-stable processes. We introduce a new class of alpha-stable processes. The finite dimensional distributions of these stochastic processes can be represented using independent stable random variables. This representation allows for Bayesian inference for the proposed statistical model. We can obtain the posterior distributions for the untried points as well as the model parameters through an MCMC algorithm. The computation for the representation requires some geometrical information given by the design points. We

propose an efficient algorithm to solve this computational geometry problem. We use two examples to illustrate the proposed method and its potential advantage.

PBART: Parallel Bayesian Additive Regression Trees
Matt Pratola, Los Alamos National Laboratory

The Bayesian Additive Regression Tree (BART) is a statistical model that represents observed data as a sum of weak learners. Each tree represents a weak learner, and the overall model is a sum of such trees. In this work, we extend the BART model to handle massive datasets using a parallelized MCMC algorithm. Our approach handles datasets too massive to fit on any single data repository, and scales linearly in the number of processor cores.

Invited Session 15: Journal of Quality Technology Session

Algorithms and Model Spaces for Model-Robust Experiment Design
Byran J. Smucker, Miami University, Oxford, OH

Optimal experiment design, utilizing criteria such as D-optimality, requires that the form of the model be specified before the experiment is conducted. The resulting design is quite dependent upon this specification. To reduce or eliminate this dependence, model-robust designs can be considered. One standard approach in the literature is to design with a set of model forms in mind, instead of just one. In this talk, we describe procedures used in constructing these model-robust designs and give empirical results demonstrating their advantages. We also discuss potentially large model sets and the computational problems associated with such sets. We suggest a solution to this problem, and give some initial, promising results.

I-optimal Versus D-optimal Split-plot Response Surface Designs
Peter Goos, University of Antwerp, Belgium

Response surface experiments often involve only quantitative factors, and the response is fit using a full quadratic model in these factors. The term response surface implies that interest in these studies is more on prediction than parameter estimation because the points on the fitted surface are predicted responses. When computing optimal designs for response surface experiments, it therefore makes sense to focus attention on the predictive capability of the designs. However, the most popular criterion for creating optimal experimental designs is the D-optimality criterion, which aims to minimize the variance of the factor effect estimates in an omnibus sense. Because I-optimal designs minimize the average variance of prediction over the region of experimentation, their focus is clearly on prediction. Therefore, the I-optimality criterion seems to be a more appropriate one than the D-optimality criterion for generating response surface designs. Here we introduce I-optimal design of split-plot response surface experiments. We show through several examples that I-optimal split-plot designs provide substantial benefits in terms of improved prediction compared with D-optimal split-plot designs, while also performing very well in terms of the precision of the factor effect estimates.

Contributed Session 3: Design and Analysis of Experiments: Some Cutting-Edge Applications

Simulation of a Nanoscale Experiment Satisfying a Generalized Langevin Equation
Martin Lysy, Harvard University

A problem of contemporary interest in nanoscale biophysics is the reconstruction of the force profile exerted on an Atomic Force Microscope (AFM) by the water molecules in a solvent, as a function of distance between the AFM and a given solute. In principle, this profile can be trivially obtained from the AFM's stationary distribution under the assumption of thermal equilibrium. In practice, experimental complications require that the stochastic dynamics of the process be statistically modeled. A natural framework for such modeling is given by the Generalized Langevin Equation (GLE), which begins with a physically intuitive decomposition of the forces acting on the AFM. These forces are then separated into internal and external components, the former being balanced by the GLE according to fundamental physical laws.

Here, a rigorous definition of the stochastic process underlying a GLE is presented, along with a simple simulation method for realizing the process with continuous sample paths. The method is used to assess and calibrate a force profile reconstruction experiment with real AFM data.

Optimal Design of Experiments with Linear Network Effects
Ben Parker, University of Southampton, UK

We investigate how the connections between experimental units affect the design of experiments on those experimental units. Specifically, where we have unstructured treatments, whose effect propagates according to a novel linear network effects model which we introduce, we show that optimal designs are no longer balanced; we further demonstrate how experiments which do not take a network effect into account can lead to a much higher variance than necessary and/or a large bias.

In complex engineered systems, for example data networks, changing one component can bring about changes in other connected components. We show how this methodology can be used in a wide range of experiments from different areas of industry and in agricultural trials, crossover trials, as well as experiments on connected individuals in a social network.

An Application of Fractional Factorial Designs to Study Drug Combinations
Jessica Jaynes, University of California, Los Angeles

Herpes simplex virus type 1 (HSV-1) is known to cause diseases of various severities. There is increasing interest to find drug combinations to treat HSV-1 by reducing drug resistance and cytotoxicity. Drug combinations offer potentially higher efficacy and lower individual drug dosage. In this paper, we report a new application of fractional factorial designs to investigate a biological system with HSV-1 and six antiviral drugs, namely, Interferon-alpha, Interferon-beta, Interferon-gamma, Ribavirin, Acyclovir, and TNF-alpha. We show how the sequential use of two- and three-level fractional factorial designs can screen for important drugs and drug interactions, as well as determine

potential optimal drug dosages through the use of contour plots. Our initial experiment using a two-level fractional factorial design suggests that there is model inadequacy and drug dosages should be reduced. A follow-up experiment using a blocked three-level fractional factorial design indicates that TNF-alpha has little effect and HSV-1 infection can be suppressed effectively by using a right combination of the other five antiviral drugs. These observations have practical implications in the understanding of antiviral drug mechanism that can result in better design of antiviral drug therapy.

Analysis of Cell Adhesion Experiments Based on Hidden Markov Models
Yijie Wang, Georgia Institute of Technology

Cell adhesion plays an important role in physiological and pathological processes. It is mediated by specific interactions between cell adhesion proteins (called receptors) and the molecules to which they bind (called ligands). This study is motivated by cell adhesion experiment conducted at Georgia Tech, which uses decrease/resumption of thermal fluctuations of a biomembrane probe to pinpoint association/dissociation of receptor-ligand bonds. More than one type of bond is commonly observed and they correspond to different fluctuation decrease due to their string strength difference. Existing approach is not robust in estimating the association/dissociation points and can only detect one type of bond. A hidden Markov model is developed to tackle the problems. It works by assuming that the probe fluctuates differently according to the underlying binding states of the cells, i.e., no bond or distinct types of bonds. These states are unobservable but their changes can be captured by a Markov chain. Applications of the proposed approach to real data demonstrate robustness and accuracy of estimating bond lifetimes and waiting times, which form basis for estimation of kinetic parameters.

Friday, June 15

Invited Session 17: New Directions in Optimal Experimental Design

Optimal and Sequential Design for Bridge Regression
Sarah Carnaby, University of Southampton, UK

Motivated by an experiment from organic chemistry, general methods are developed for the selection of an optimal design for a linear model fitted using bridge regression, where accurate estimates of the model coefficients are required. Bridge regression is a family of coefficient shrinkage methods, including ridge regression and the lasso as special cases, that perform continuous subset selection. They can provide lower prediction error than ordinary least squares through trading variance for bias, can alleviate problems of multicollinearity and allow more predictors than runs to be considered.

The relationship between bridge regression and Bayesian inference is exploited to develop a class of D-optimal designs. A necessary approximation to the variance-covariance matrix of coefficient estimators is derived and designs are then found using algorithmic search. In addition, a sequential design criterion is developed based on maximising prediction variance, which is estimated using a bootstrapping procedure.

This criterion allows additional points to be selected that may enhance or repair an existing design.

The methods are demonstrated for a variety of screening experiments under ridge and lasso regression, including experiments to understand the predictors that influence the melting point of organic compounds. This application is fundamental in the development of chemicals with desired thermophysical behaviour. (Joint research with Dave Woods.)

Multi-Objective Optimal Experimental Designs for Event-Related fMRI Studies
Abhyuday Mandal, University of Georgia

Functional magnetic resonance imaging (fMRI) is an advanced technology for studying brain functions. Due to the complexity and high cost of fMRI experiments, high quality multi-objective fMRI designs are in great demand. We propose an efficient approach to find optimal experimental designs for event-related functional magnetic resonance imaging (ER-fMRI). We consider multiple objectives, including estimating the hemodynamic response function (HRF), detecting activation, circumventing psychological confounds and fulfilling customized requirements. Taking into account these goals, we formulate a family of multi-objective design criteria and develop a genetic-algorithm-based technique to search for optimal designs. Our proposed technique incorporates existing knowledge about the performance of fMRI designs, and its usefulness is shown through simulations. We also consider a nonlinear model to accommodate a wide spectrum of feasible HRF shapes, and propose an approach for obtaining maximin efficient designs. Our approach involves a reduction in the parameter space and an efficient search algorithm. The designs that we obtain are much more robust against mis-specified HRF shapes than designs widely used by researchers. (Joint research with Ming-Hung (Jason) Kao, Dibyen Majumdar and John Stufken.)

Optimal Design for Partial Likelihood in Survival Analysis
Jesús López Fidalgo, Universidad de Castilla-La Mancha, Spain

In a follow-up study, the time–to–event may be censored either by withdrawal or by end of the study. This time variable is modeled through a Cox-proportional hazards model including covariates, which are under the control of the experimenter. At the time the model is being fitted, it is known whether the time observed is censored or not. This is not the case when the experiment is to be designed and some additional prior probability distribution has to be assumed for the withdrawal. This adds an important degree of complexity to the problem, and requires dealing with two sources of imprecision when the experiment is to be scheduled. On the one hand, the censored times when the event has not happened yet; on the other hand, the probability distribution for censoring. Moreover, the Cox partial likelihood for these models is usually considered instead of the full likelihood. A partial information matrix is built in this case and optimal designs are computed and compared with the traditional optimal designs for the full likelihood information. The use of partial information means some of the tools for computing optimal designs with full likelihood are not valid anymore. Some general results are provided in order to deal with this approach and then an application to a simple case with two possible treatments is solved. The

partial information matrix depends on the parameters and therefore a sensitivity analysis is made in order to check the robustness of the designs for the choice of the nominal values of the parameters.

Invited Session 18: Statistics for Engineering Design

An Information-Theoretic Metric of System Complexity with Application to Engineering System Design
Douglas Allaire, Massachusetts Institute of Technology

System complexity is considered a key driver of the inability of current system design practices to at times not recognize performance, cost, and schedule risks as they emerge. We present here a definition of system complexity and a quantitative metric for measuring that complexity based on information theory. We also derive sensitivity indices that indicate the fraction of complexity that can be reduced if more about certain factors of a system can become known. This information can be used as part of a resource allocation procedure aimed at reducing system complexity. Our methods incorporate Gaussian process emulators of expensive computer simulation models and account for both model inadequacy and code uncertainty. We demonstrate our methodology on a candidate design of an infantry fighting vehicle.

Support Vector Autoregression in the Structural Health Monitoring Paradigm
Luke Bornn, University of British Columbia, Canada

Although the application of statistical techniques to structural health monitoring has been investigated in the past, these techniques have predominantly been limited to identifying damage-sensitive features derived from linear models fit to the output from individual vibration sensors. As such, they are typically limited to identifying only that damage has occurred. In general, these methods are not able to identify which sensors are associated with the damage in an effort to locate the damage within the resolution of the sensor array, nor are the linear models particularly accurate at modeling complex systems.

To improve upon this approach to damage detection, we use autoregressive support vector machines (SVMs) to model sensor output time histories and show that such nonlinear regression models more accurately predict the time series than linear autoregressive (AR) models. Here the feature for this comparison is the residual errors between the measured response data and predictions of the time series model. Although SVMs have been used for structural health monitoring, these approaches predominantly focus on one and two class SVMs, which are used for outlier detection and group classification, respectively. Our approach is unique in its combination of support vector regression, autoregressive techniques, and residual error analysis. Thus while earlier approaches look at classifying sections of the time-series response as damaged or undamaged directly (the dependent variable being a binary indicator), our methodology works by using support vector regression to model the raw time-series data, then subsequently predicting damage by monitoring the residuals of the model.

Furthermore, we also show how the residuals from the SVM prediction of each sensor may be combined in a statistically rigorous manner to provide probabilistic

statements regarding the presence of damage as assessed from the amalgamation of all available sensors. In addition, this methodology allows us to pinpoint the sensors that are contributing most to the anomalous readings and therefore locate the damage within the sensor network's spatial resolution. The process is demonstrated through a simulation study as well as on a test structure where damage is simulated by introducing an impact type of nonlinearity between the measured degrees of freedom.

The "New" Matrix Statistics of Random Matrix Theory and the Julia Programming Language
Alan Edelman, Massachusetts Institute of Technology

Traditional statistics may be described as scalar and vector statistics. The scalar and multivariate normal distributions are familiar examples. Random matrix statistics investigations began in the early 20th century with major innovations from the physics community half a century later. Today, with improvements in computational methods and underlying theory, random matrix theory is now being applied to many diverse areas including finance, an Aids study, computational biology, traffic flows, and seeing through apparently opaque objects. New applications are discovered every day.

As a "Part 2" of this talk, we will introduce the Julia Programming Language, one that is as easy and familiar as R, Python, or MATLAB, is open source, and aspires to solve the "two language problem," which requires libraries and production code be written outside the language.

Invited Session 19: Military Applications of Statistics

Space Filling and Optimal Experimental Designs for Use in Studying Computer Simulation Models with an ISR Application
Rachel Silvestrini, Naval Postgraduate School

Computer simulation models play an increasingly important role in the military acquisition and Test and Evaluation (T&E) process, where some physical experiments on the real system or even a prototype are prohibitively expensive. Typically, researchers employ traditional factorial or fractional factorial designs, or even more recently, optimal designs to study stochastic computer simulation models. Alternatively, space-filling designs are usually employed to study deterministic computer models (such as a finite element analysis model). This talk covers a brief survey of experimental design choices for computer simulation models and the modeling strategy for analyzing results of the experiments. Then a hybrid optimal and space-filling design approach is considered as an alternative. We illustrate the construction of these designs with examples, and demonstrate their performance in response prediction with respect to an Intelligence, Surveillance, and Reconnaissance (ISR) simulation model involving Unmanned Aerial Vehicles (UAVs).

Methodological Considerations and Statistical Modeling of U.S. Army Mental Health
Shayne Gallaway[1]*, Amy M. Millikan[1], David S. Fink[1] and Michael R. Bell[2],
[1]U.S. Army Institute of Public Health; [2]Uniformed Services University of the Health Sciences

Sustained combat operations in Iraq and Afghanistan have coincided with an increased prevalence of behavioral health outcomes (e.g., suicide, depression, posttraumatic stress) among U.S. Army Soldiers. Population assessment of risk factors for and development of strategies to mitigate negative behavioral outcomes requires examination of constructs important to leadership that are not always easily estimated. Data is collected and ascertained from multiple modes and sources. In this paper we describe the methodological considerations and calculation of constructs to approximate combat intensity and individual behavioral health risk profiles, as well as statistical modeling of factors associated with these constructs. This paper will also describe the individual linkage of population data sources (not collected for the purposes of research) to retrospectively construct synthetic longitudinal cohorts; and the analysis of these data to identify significant exposures and risk factors among populations of concern.

Development of Non-linear Mixed Effects Models for Assessing Effectiveness of Spending in Iraq
Maj. Nicholas J. Clark* and LTC John Jackson, United States Military Academy

This paper describes an approach for assessing the effectiveness of Commander Emergency Relief Program (CERP) funds during contingency operations. Data of this nature is often messy and has variations both within cities and across cities. The data is also temporal in nature and affected by many latent factors. Traditional analysis fails to fully account for these sources of variation and subsequently do not allow the user to generalize their results. Furthermore, cities that tend to have high levels of violence are generally treated in the same manner as cities that tend to have low levels of violence.

To handle the variation we embed correlated random effects similar to longitudinal models commonly used in Biostatistics to measure the effectiveness of drugs. We differentiate between cities with a high propensity towards violence and cities with a low propensity towards violence enabling analysis of several cities simultaneously. This provides a more accurate description of the effectiveness of Information Operations weapons similar to CERP funds. We demonstrate the model with a case study of CERP funds spent in Iraq between the years 2005 and 2007 in 80 cities across the country.

Invited Session 20: Modern Statistical Computing

A Bayesian View of Sliced Inverse Regression with Interaction Detection
Jun Liu, Harvard University

Previously we have proposed a Bayesian partition model for detecting interactive variables in a classification setting with discrete covariates. This framework takes advantage of the structure of the naïve Bayes classifier and introduces latent indicator

variables for selecting variables and interactions. In our effort to extend the methods to continuous covariates, we found interesting connections with semi-parametric index models and the Sliced Inverse Regression method. In index models, the response is influenced by the covariates through an unknown function of several linear combinations of the predictors. Our finding of the Bayesian formulation of such models enabled us to propose a set of new models and methods that can effectively discover second-order effects and interactions among the covariates. A two-stage stepwise procedure based on likelihood ratio test is developed to select relevant predictors and a Bayesian model with dynamic slicing scheme is derived. The performance of the proposed procedure in comparison with some existing method is demonstrated through simulation studies. (Joint work with Bo Jiang.)

Practices and Perils in Multiphysics Simulations
Efthimios Kaxiras, Harvard University

Multiphysics and multiscale simulations are becoming an essential tool in modeling complex physical systems with the accuracy and resolution to provide quantitative description. In this presentation I will give some examples of complex phenomena in which multiple scales are involved and describe some methods for modeling these phenomena. I will also attempt to highlight limitations and challenges that arise in such simulations from the algorithmic and numerical perspective.

Statistical Computation and Computational Statistics: An Interweaving Perspective
Xiao-Li Meng, Harvard University

The two phrases “Statistical Computation” and “Computational Statistics” are often used interchangeably. But the former should emphasize more the use of statistical theory and methods in constructing and evaluating algorithms, and the latter more on using computational methods for statistical purposes. One could even push further that the latter should include those statistical theory and methods that are inspired by algorithmic considerations. This talk reports on a recent example of how interweaving computational considerations and statistical thinking can enhance each other. The simple statistical fact that regressing Y on X is not the same as regression X on Y has led to a very promising strategy of improving data augmentation algorithms by interweaving two kinds of residual augmentations. In return, the constructions of residual augmentations in a joint posterior space have led to the realization that there are two kinds of Bayesian ancillarity, in contrast to the notion of sufficiency, which is the same (almost surely) from either the Bayesian perspective or the frequentist perspective.

Contributed Session 4: Quality and Reliability

EWMA p Charts Under Sampling by Variables — Ideas, Numerics and Properties
Sven Knoth, Helmut Schmidt University, Hamburg

The data is sampled in batches of size n in order to monitor stability in terms of yield. For given lower and upper specification limits the probability of nonconforming quality is estimated via the sample mean (and variance). This proportion estimate is plugged into an EWMA chart. Then, the control chart with only an upper limit signals for decreased quality. Thus, this monitoring scheme alarms only if the quality level deteriorated considerably. It allows, for example, that the pre-run (in-control) sample mean does not match the target mean value such as the center of the specification interval. This mismatch and other imperfect in-control situations are quite common in control charting practice. In order to calculate properties such as the ARL (zero-state, worst-case) one has to face similar issues as for variance schemes, because the support of the chart statistic is bounded. This is solved with appropriate numerical methods. Eventually, the resulting and reasonably calibrated EWMA p variables charts are compared to classical EWMA and CUSUM charts for the mean and also to classical EWMA p charts based on simply the sample proportions.

Robust Leak Tests for Transmission Systems Using Nonlinear Mixed-Effect Models
Kamran Paynabar, Georgia Institute of Technology

Leakage of the transmission fluid or oil in powertrain systems (i.e. transmissions, cylinder heads, engine blocks, etc.) can cause engine overheating and/or permanent damages. Therefore, it is crucial to run a leak test to inspect for any possible porosity in the casting parts. However, the inspection results of the amount of leakage at given testing times are sensitive to the tested part's temperature, which varies from part to part but was not previously incorporated in the leak testing systems. The objective of this paper is to develop a robust leak testing system that is insensitive to the part temperature variations in real-world production processes. To achieve a robust leak test, we propose a temperature compensation algorithm to adjust the measured leak flow, which takes both tested part and calibration temperatures into account. For this purpose, a nonlinear mixed-effect model is first developed for modeling the leak flow profile as a function of both the leak testing time and the part temperature. Then, the fitted mixed-effect model is used to adjust the leak flow based on the calibration temperature. We evaluated the performance of the proposed method based on an independent test dataset. The results show that the average percentage of error reduction for the test samples is about 92%, which indicates a significant improvement in the leak testing system.

A Class of Tests for Exponentiality Against NBUE Alternatives
M.Z. Anis, Indian Statistical Institute

In this paper we develop a class of test statistics for testing exponentiality against NBUE alternatives. The test statistics are shown to be asymptotically normal and consistent. The exact distribution under the null hypothesis of exponentiality has been

found. This class of test statistics includes the test proposed by Hollander and Proschan (1975) as a special case. Efficiency studies have also been done.

Process Monitoring and Feedforward Control for Proactive Quality Improvement
Lihui Shi, University of Washington-Seattle

Process adjustment strategy is an important part of the process improvement methods, which is also called engineering process control (EPC), and it is often integrated with statistical process control (SPC) to improve the process control performance. While feedback control is used to compensate for the output deviation, feedforward control is a proactive control strategy based on a direct measurement of the disturbance, and it acts before the disturbance affects the system. Feedforward control is usually combined with feedback control for variation reduction. In this talk, we analyze and solve the process adjustment problem from a statistical perspective, the rationales for feedforward control are explained, and a new philosophy on its application is given. The feasibility condition for feedforward control application is illustrated from a new disturbance decomposition viewpoint, and the validity of some disturbance models which work well for feedforward control is investigated. Some relevant issues on process monitoring, feedback control and feedforward control are discussed and addressed.

Contributed Session 5: Advanced Statistical Modeling

Parabolic SPDEs Driven by a Levy Noise and their Numerical Approximation
Silika Prohl, Princeton University

This paper provides extensions of the work on subsampling by Bertail et al. (2004) for strongly mixing case to weakly dependent case by application of the results of Doukhan and Louhichi (1999). We investigate properties of smooth and rough subsampling estimators for sampling distributions of converging and extreme statistics when the underlying time series is η or λ-weakly dependent.

A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data
David S. Matteson, Cornell University

Change point analysis has applications in a wide variety of fields. The general problem concerns the inference of a change in the distribution for a set of time-ordered observations. Sequential detection is an online version in which new data is continually arriving and is analyzed adaptively. We are concerned with the related, but distinct, offline version, in which retrospective analysis of an entire sequence is performed. For a set of multivariate observations, we consider nonparametric estimation of both the number of change points and the positions at which they occur. We do not make any assumptions regarding the nature of the change in distribution or any distribution assumptions beyond the existence of first absolute moments. Estimation is based on hierarchical clustering and we propose both divisive and agglomerative algorithms. We compare the proposed approach with competing

methods in a simulation study and conclude with applications in finance, genetics, and operations. (This talk is based on joint work with Nicholas A. James, PhD Candidate, School of Operations Research and Information Engineering, Cornell University (E-mail: nj89@cornell.edu).)

Tractable Functional Response Modelling with Processes with Non-Separable Covariance Functions
Matthew A. Plumlee, Georgia Institute of Technology

Growth in the technology of sensors, data acquisition, and data storage has made it increasingly common to receive samples in the form of profile or image data, which require the modelling of an underlying functional response. Gaussian Process models have become increasingly popular to model functional responses; these give the flexibility needed to deal with non-linear functions. However, fitting these models requires computationally expensive matrix inversion or decomposition that costs on the cubic order. For functional responses, this problem is exacerbated, as the amount of data grows very quickly with the number of profiles or images received. Many works in the spatiotemporal and functional response domain have included the separability and stationarity assumptions on the covariance function of the GP model to ease the computational burden. This assumption allows one to separately preform matrix operations in the spatial and temporal dimensions, reducing the computational burden. However, in many applications, separability is an invalid assumption. Toward the aim of relaxing the separability assumption in large data sets, this work proposes a method to model these functions without the separability assumption while maintaining the computational advantages.

Poster Session Abstracts

Method Selection for Computing A-basis and B-basis
Xiaomi Hu, Wichita State University

The terms A-basis and B-basis in the aviation composite material industry are used to refer to the lower limits of one-sided confidence intervals of the first and the 10th percentiles of the breaking strength of the material with confidence coefficient 95%. In this talk it is shown that the coverage probabilities for A-basis and B-basis do not depend on population parameters, nor sample statistics. Thus the coverage probabilities, as the characteristics of computation methods, could be utilized to evaluate the methods in the process of method selection when multiple methods are available.

Internet: Information Source or Computer Infection Hazard? Epidemiological Models for Browser-Based Malware
Natallia Katenka, Boston University

Internet has become a convenient tool for an effective information search, social interaction, banking, shopping, and, unfortunately, the source of many types of computer viruses. The implications of the computer infections are the losses of billions of dollars and an exposure of personal and highly classified information. Previous research in the Internet security relies on epidemic models that either do not distinguish between different Internet applications or do not consider the structure of the computer communication network, the models that prove to be unrealistic, and hence no longer applicable. A vast majority of existing computer viruses depend on the protocol of a given Internet application and spread passively as a part of regular communication among computers using the same Internet application. In this work, we develop simple epidemic models for browser-based computer threats existing today, anticipated tomorrow and after-tomorrow. We partially adapt probabilistic models developed for sexually transmitted infections (STI) for heterosexual population. Our models incorporate the infectiousness and susceptibility to infection of both servers and clients and are built on the user communication network inferred from the NetFlow traffic data collected at core networks from a large European Internet Service Provider. Using these models, we explore analytically and numerically vulnerability of web application to different transmission rates of infection and investigate simple proactive strategies that would allow to control and prevent an epidemic outbreak on web application.

Indicator Functions in the Linear-Quadratic Parametrization System
Arman Sabbaghi, Harvard University

A fractional factorial design is uniquely specified by its indicator function. As such, indicator functions can shed light on the structure and general properties of specific designs. Previous work on indicator functions has been focused on its applications under the orthogonal components parametrization system for contrasts. We consider

their construction and application under the linear-quadratic parametrization system. Methods for the construction of indicator functions, and their use for calculation of complex aliasing relations for three-level symmetrical fractional factorials under the Linear-quadratic system will be derived. We illustrate our results and the value of indicator functions for specific examples of orthogonal arrays. The indicator function will be seen to be particularly useful in studying the structure and general properties of fractional factorial designs under the linear-quadratic parametrization system.

Equivalence of Factorial Designs with Qualitative and Quantitative Factors
Tena Katsaounis, The Ohio State University at Mansfield

Two symmetric fractional factorial designs with qualitative and quantitative factors are equivalent if the design matrix of one can be obtained from the design matrix of the other by row and column permutations, relabeling of the levels of the qualitative factors and reversal of the levels of the quantitative factors. In this paper, necessary and sufficient methods of determining equivalence of any two symmetric designs with both types of factors are given. An algorithm used to check equivalence or non-equivalence is evaluated. If two designs are equivalent the algorithm gives a set of permutations which map one design to the other. Fast screening methods for non-equivalence are considered. Extensions of results to asymmetric fractional factorial designs with qualitative and quantitative factors are discussed.

A Posterior Predictive Check Approach to Analyze 2 level Factorial Designs
Valeria Espinosa, Harvard University

Factorial designs have been widely used in many scientific and industrial settings. The traditional ways of analyzing such experiments assume an underlying normal population and restrict the range of effects that can be tested to the means. We explore two methods based on the Rubin Causal Model framework which allow the relaxation of these assumptions: a randomization version of the Loughin and Noble (1997) proposal and a Posterior Predictive Check approach. Both are based on sequential testing with the same starting point: a Fisher randomization test for a sharp null. The use of two discrepancy measures is compared. There are some simulation results available. Further assessment of the Bayesian methodology is being performed due to the difference in the distribution of traditional p values (uniform under the null) and the posterior predictive ones.

Sensitivity Analysis of Expected Shortfall by Means of a Second-Order Approximation
Guven Gul Polat, Istanbul Technical University, Turkey

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-order derivative 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, the most efficient risk estimation method, utilizing high performance computing techniques. Totally, in addition to the increase in the accuracy of the risk estimation by a higher order approximation, we demonstrate the acceleration of the simulation by a parallel execution.

Resolving Groupings of Subsets
Kalyan Veeramachaneni, Massachusetts Institute of Technology

We introduce a new problem we call Resolving Groupings of Subsets which emerges from scenarios in which multiple computational agents (algorithms or people) group distinct or overlapping, relatively small, subsets of a very large dataset. The goal is to fuse this partial information into a globally coherent grouping of the entire dataset. We contribute fusion strategies and a means of determining whether there is sufficient information from the subsets' groupings to yield a stable consensus grouping. For the opportunity when subsets can be actively composed while the solution framework is executing, i.e. in an online setting with active learning, we devise multiple strategies for adaptive subset composition. The strategies rely on computationally inexpensive, non-linear transformations of local evidence, such as accumulated pairwise element co-occurrence in a group. The local evidence facilitates efficient subset composition which contributes to attaining better accuracy more quickly. Frequent adaptation rather than delayed adaptation provides the most accurate empirical results.

A New Adaptive Design for Environmental Sampling
Huijuan Li, Rutgers University

The spatial distribution of a natural resource is an important consideration for designing an efficient survey in environmental sampling. A new adaptive design is proposed which is flexible and takes into account the spatial distribution to improve estimation efficiency. A New algorithm is introduced for the search of optimal designs. The proposed design and its efficiency are demonstrated by a water quality study.

Multiscale Recurrence Analysis of Long-Term Nonlinear and Nonstationary Time Series
Yun Chen, University of South Florida

Recurrence analysis is an effective tool to characterize and quantify the dynamics of complex systems, e.g., laminar, divergent or nonlinear transition behaviors. However, recurrence computation is highly expensive as the size of time series increases. Few, if any, previous approaches have been capable of quantifying the recurrence properties from a long-term time series, while which is often collected in the real-time monitoring of complex systems. This paper presents a novel multiscale framework to explore recurrence dynamics in the complex systems and resolve the computational issues for the large-scale datasets. As opposed to the traditional recurrence analysis in a single

scale, we characterize and quantify the recurrence dynamics in multiple wavelet scales, which captures not only nonlinear but also nonstationary behaviors in a long-term time series. The proposed multiscale recurrence approach was utilized to: 1) identify heart failure subjects from the 24-hour time series of heart rate variability (HRV), and 2) analyze the 3-lead vectorcardiogram (VCG) signals for the detection of myocardial infarctions. The classification models were shown to identify the conditions of congestive heart failure with an average sensitivity of 92.1% and specificity of 94.7%. In addition, the multiscale recurrence analysis of 3-lead VCG leads to a superior classification model that detects the myocardial infarction with an average sensitivity of 96.8% and specificity of 92.8%, which is much better (i.e., 5.6% increase in terms of correct rates) than the single-scale recurrence analysis in previous investigations. The proposed multiscale recurrence framework can be potentially extended to other nonlinear dynamic methods that are computationally expensive for large-scale datasets.

Modeling, Experimental Design, and Analysis of Stereolithography Process for Direct 3-D Printing
Jizhe Zhang, University of Southern California

Stereolithography (SLA) is one of the most widely employed techniques of 3-D printing. However, final product shrinkage due to material phase changing during printing is unavoidable. It often leads to dimensional deviations, which require post-machining step for correction. The shrinkage has traditionally been analyzed through finite element simulation or experimental trial-and-error methods. Systematic models for compensating shrinkage are hardly available, particularly for online control. To achieve direct printing, we develop a systematic approach to model and predict the shrinkage process, design the experimental plan to identify the process model, and analyze the experimental outcomes for optimized compensation. The experimental design takes advantage of the symmetry property of the product and a variant of Latin Square design is applied. The proposed shrinkage model and the corresponding design of experiments are also verified both theoretically and experimentally.

Sensitivity Analysis for Computer Code with Non-Rectangular Regions
Fangfang Sun, Ohio State University

Computer experiments study input/output relationships using computer code implementations of physics-, economics-, biology-, or engineering- based mathematical models. An input is said to be 'influential' if changes in this input will result in large variations in the output. Sensitivity analysis is widely used for identifying influential input variables. Two approaches to evaluating statistically sensitivity are (1) estimating global sensitivity indices based on Sobol's variance decomposition (Sobol (1990)), and (2) evaluating local sensitivity indices based on a gradient measure using a one-at-a-time sampling design (Morris(1991)). Although both approaches have been studied for (hyper-) rectangular input regions, they have not been considered carefully for the non-rectangular input region setting. In this poster, we propose gradient based method to evaluate sensitivity indices for non-rectangular regions. It is shown by examples that the proposed method works well in both the standard and the new setting.

Some Aspects of Modeling Dependence in Copula-based Markov Chains
Martial Longla, University of Cincinnati

Dependence coefficients have been widely studied for Markov processes defined by a set of transition probabilities and an initial distribution. This work clarifies some aspects of the theory of dependence structure of Markov chains generated by copulas that are useful in time series econometrics and other applied fields. The main aim of this paper is to clarify the relationship between the notions of geometric ergodicity and geometric ρ-mixing; namely, to point out that for a large number of well-known copulas, such as Clayton, Gumbel or Student, these notions are equivalent. Some of the results published in the last years appear to be redundant if one takes into account this fact. We apply this equivalence to show that any mixture of Clayton, Gumbel or Student copulas generate both geometrically ergodic and geometric $\rho-$mixing stationary Markov chains, answering in this way an open question in the literature. We shall also point out that a sufficient condition for $\rho-$mixing, used in the literature, actually implies Doeblin recurrence.

Efficient Cross-validation of Predictive Accuracy via Statistical Designs
Qiong Zhang , University of Wisconsin-Madison

Cross-validation is used routinely for assessing the prediction error of a regression model. Despite its popularity, this method is known to have high variability. We propose an experimental design approach to significantly reduce this variability. This approach borrows Latin hypercube designs, originally motivated in computer experiments, to construct a cross-validation sample such that the input values in each fold achieves uniformity. The proposed method applies to various regression models such as linear regression models and additive models. Theoretical results are derived to show that the proposed cross-validation method provides estimates with significantly smaller variability than its counterpart under the independent and identically distributed sampling. Numerical examples are provided to corroborate the derived theoretical results.

Better Latin Hypercube Designs: Controlling Correlations, Achieving Two-Dimensional Stratification, or Both?
Jiajie Chen, University of Wisconsin-Madison

The Latin hypercube design is the most popular choice for running a computer experiment. Various methods have been proposed to construct Latin hypercube designs with small column-wise correlations. Another direction of active research is to construct Latin hypercube designs with multi-dimensional stratification. To kill two birds with one stone, we propose a simple method for constructing new Latin hypercube designs with both controlled correlations and uniformity in two dimensions. This is a direct method and does not entail iterative updates. When used in numerical integration, the constructed designs can filter out not only one- and two-dimensional functional ANOVA components but also bilinear terms more efficiently. Examples are given to illustrate the construction and sampling properties of the proposed designs. This is joint work with Peter Qian.

The Orthogonalizing EM Algorithm
Tian Jin, University of Wisconsin-Madison

We propose an algorithm, called the Orthogonalizing EM (OEM) algorithm, for computing generalized inverses and fitting penalized regression models. The algorithm actively expands an arbitrary model matrix to an orthogonal matrix and treats the extra rows as missing data. Then in each iteration, the algorithm imputes the missing responses based on the current estimates of the parameters and obtains a closed-form solution for the complete data. The proposed algorithm has several desirable properties. First, the OEM solution to the penalized least squares problem with the SCAD or MCP penalty can achieve the oracle property. Second, under various penalties, the solution path of OEM has the same coefficient for any fully aliased columns. Third, for a singular regression matrix, the ordinary least squares solution of OEM converges to the Moore-Penrose generalized inverse. In terms of computational speed, the algorithm con-verges very fast for problems with large n but can be slow for the small n large p case. This is joint work with Shifeng Xiong at Chinese Academy of Sciences, and Bin Dai and Peter Qian at the University of Wisconsin-Madison.

Model Calibration through Minimal Adjustments
Chia-Jung Chang, Georgia Institute of Technology

Model calibration refers to estimating unknown parameters in a physics-based model from real data. When model assumption is violated, the estimates become inaccurate leading to poor model prediction. Usually, the Gaussian process model provides a powerful methodology for calibrating a physical model in the presence of model uncertainties. However, if the data contains systematic experimental errors, then the Gaussian process model can lead to an un-necessarily complex adjustment of the physical model. Besides, all existed works ignore the potentially important bias that can occur in the observations. In this work, we develop a methodology for calibrating the physical model in the presence of both model and experimental biases. Two real case studies are presented to demonstrate the prediction ability.

Designs for Diagnostics for GP Emulators
Yan Chen, University of Wisconsin-Madison

The Gaussian process emulator has been widely used in computer experiments. This method is known to have a mixed performance: working well for some examples and not suitable for others. Diagnosis of this emulator for a particular computer experiment becomes critical. We consider the design aspect of this problem. An experimental design strategy is proposed to construct efficient training and testing samples for validating a Gaussian process emulator. Examples are given to illustrate the effectiveness of the proposed method. This is joint work with Qiong Zhang and Peter Qian at the University of Wisconsin-Madison.

Interpretable Predictive Models
Cynthia Rudin, Massachusetts Institute of Technology

I am working on the design of predictive models that are both accurate and interpretable by a human. These models are built from association rules such as "dyspepsia & epigastric pain -> heartburn." I will present three algorithms for "decision lists," where classification is based on a list of rules:

1) A very simple rule-based algorithm, which is to order rules based on the "adjusted confidence." In this case, users can understand the whole algorithm as well as the reason for the prediction.

2) A Bayesian hierarchical model for sequentially predicting conditions of medical patients, using association rules.

3) A mixed-inter optimization (MIO) approach for learning decision lists. This algorithm has high accuracy and interpretability - both owing to the use of MIO.

This is joint work with David Madigan, Tyler McCormick, Ben Letham, Allison Chang, and Dimitris Bertsimas.

Learning Variant of the Secretary Problem
Luis Voloch, Massachusetts Institute of Technology

In the standard secretary problem, one has no access to the distribution of the qualities of the applicants. The goal is to come up with a strategy that maximizes the probability of getting the single best applicant. We explore a variant of the secretary problem for which one knows the form of the distribution of the quality of the applicants, but not the parameters. We present a strategy that is asymptotically optimal (in the sense of maximizing the probability of selecting the best applicant) for all smooth distributions, and propose a strategy for a yet smaller class of distributions, and observe via simulations how the asymptotic success rate (which is the same for all) changes with the number of applicants.

Sequential Event Prediction
Ben Letham, Massachusetts Institute of Technology

In sequential event prediction, we are given a "sequence database" of past sequences to learn from, and we aim to predict the next event within a current event sequence. We focus on applications where the set of past events has predictive power and not the specific order of those past events. Such applications arise in recommender systems, equipment maintenance, medical informatics, and in other domains. Our formalization of sequential event prediction draws on ideas from supervised ranking. We show how specific choices within this approach lead to different sequential event prediction problems and algorithms. We apply our approach to an online grocery store recommender system as well as a novel application in the health event prediction domain.

Application of a Bayesian Approach to Classification of Particle Motion in Live Cells
Syuan-Ming Guo, Massachusetts Institute of Technology

Quantitative analysis of particle motion, either based on particle tracking or fluorescence fluctuation datasets, is a powerful approach to understanding the mechanism of transport of biological particles. However, inferring motion models from single-particle trajectories (SPTs) or fluorescence intensity traces is non-trivial due to noise from both sampling limitations and heterogeneity present in biological samples. We present and apply a systematic Bayesian approach to multiple hypothesis testing of competing motion models for mean-square displacement (MSD) curves from SPT datasets, and temporal autocorrelation functions (TACFs) from fluorescence correlation spectroscopy (FCS) datasets. By explicitly including the noise covariance matrix into fitting and marginalizing out the model parameters, model complexity is appropriately penalized to avoid over-fitting. We test this procedure rigorously using simulated trajectories and intensity traces for which the underlying physical process is known, demonstrating that it properly accounts for noise by choosing the simplest physical model that describes the observed data. Further, we show that computed model probabilities provide a reliability test for the downstream interpretation of associated parameter values. We subsequently illustrate the broad utility of the approach by applying it to disparate biological systems including chromosomes, kinetochores, membrane receptors, and developing embryos undergoing a variety of complex biological motions. This automated and objective Bayesian framework naturally scales to large amounts of data, making it ideal for classifying motion of large numbers of SPTs from high-throughput screens or large numbers of TACFs from imaging FCS measurements.