2008 Spring Research Conference on

Statistics in Industry and Technology

May 19, 20, and 21, 2008

Georgia Institute of Technology
Atlanta, GA

Sponsors

Table of Contents

Program Committee

Paul Kvam	Georgia Institute of Technology
Jye-Chyi Lu	Georgia Institute of Technology
Kwok Tsui	Georgia Institute of Technology
Roshan Joseph Vengazhiyil	Georgia Institute of Technology
Jeff Wu	Georgia Institute of Technology

Spring Research Conference Management Committee Chair

Andrew J. Booker	The Boeing Company

Contributed Sessions Chairs

Winson Taam	The Boeing Company
William Brennemen	The Proctor and Gamble Company

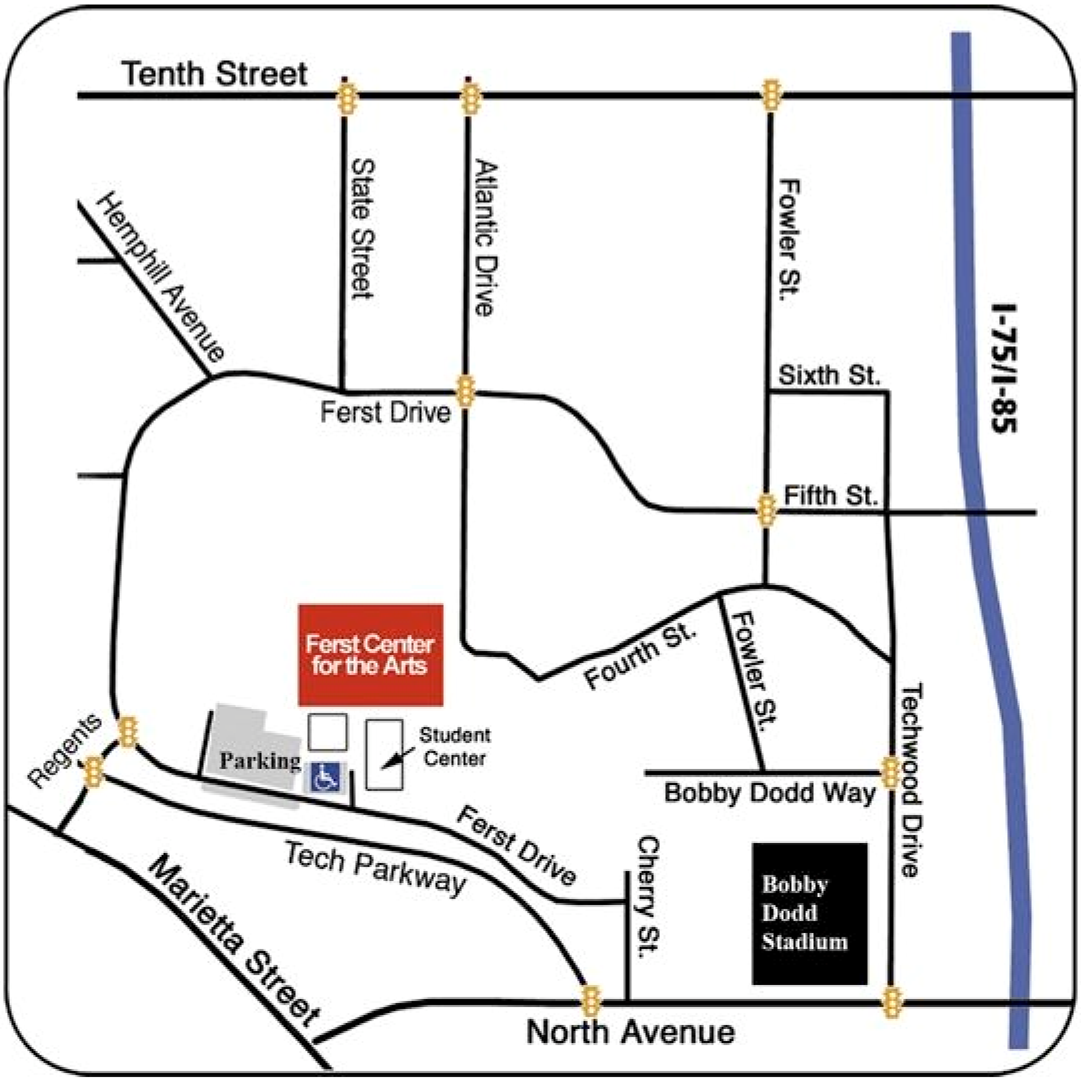
Session Organizers

Derek Bingham	Simon Fraser University
Tirthankar Dasgupta	Harvard
Brad Jones	JMP
Sue Lewis	University of Southampton
William Meeker	Iowa State University
Peter Qian	Wisconsin
Henry Rolka	CDC
Nicoleta Serban	Georgia Institute of Technology
David Steinberg	Tel Aviv University
Agus Sudjianto	Bank of America
Hongquan Xu	UCLA
Ming Yuan	Georgia Institute of Technology

Conference Sponsors

SAS
JMP
American Statistical Association
Institute of Statistical Mathematics
The Stewart School of Industrial and Systems Engineering, Georgia Tech
Boeing

Map of Ferst Center on the Georgia Tech Campus



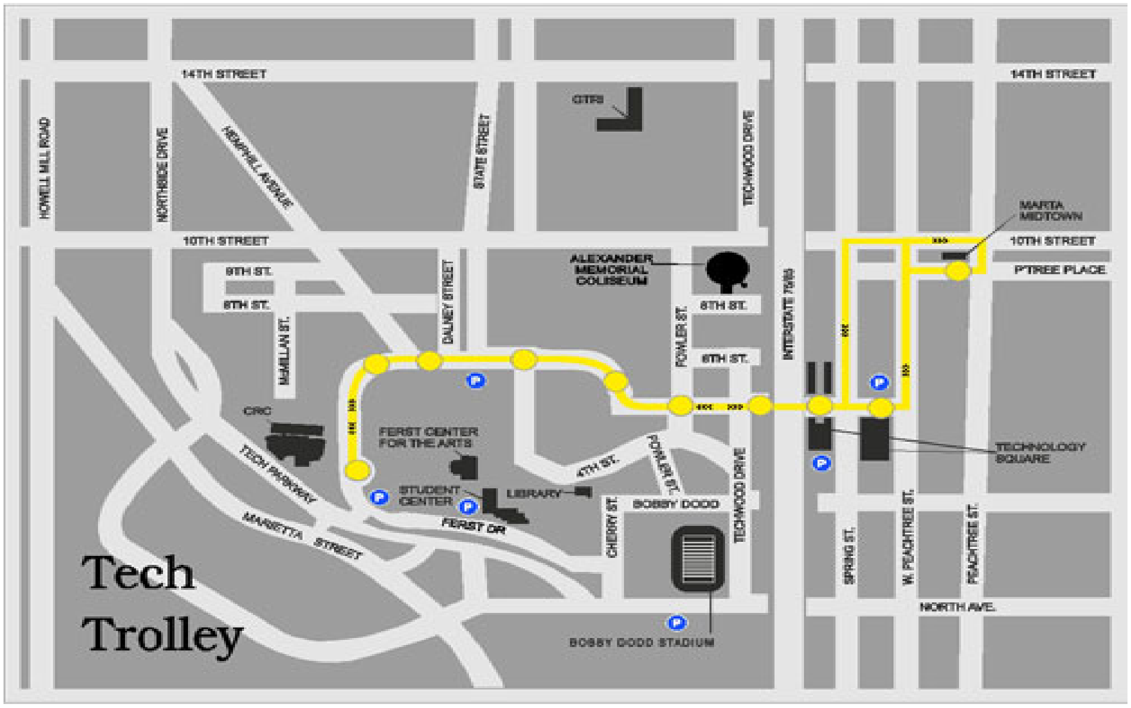
Directions:

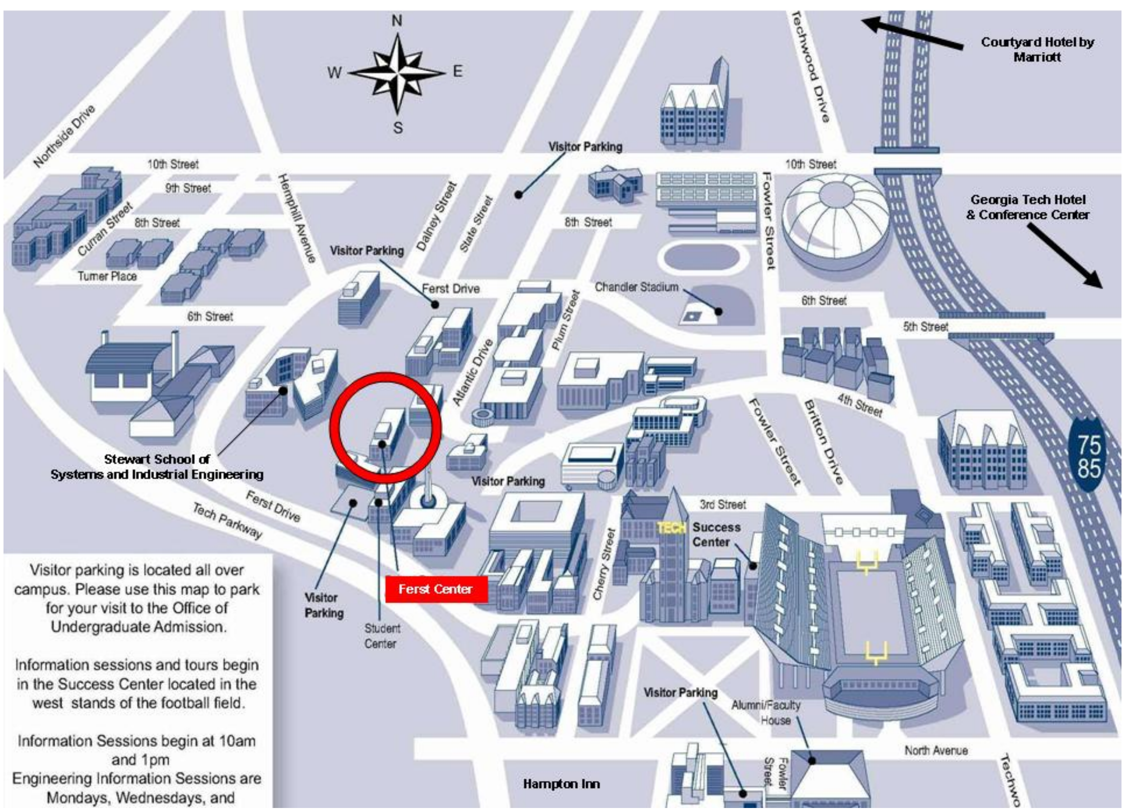
From Midtown Hotels (Georgia Tech Hotel, Regency Suites, and Wyndham): Look for the Tech Trolley – a yellow and black bus designed like an old-time trolley – and that will take you to the Georgia Tech campus.

There is no charge to ride the Tech Trolley. It runs about every 15 minutes. Exit the Tech Trolley across the street from the Campus Recreation Center (CRC). Take the sidewalk that heads away from the street to the Ferst Center, which is on the left about 100 yards. Look for a white sculpture in front of the building.

For those staying at the Marriott, we will provide a shuttle to the conference at 8:00 AM. From the Hampton Inn, the Ferst Center is within walking distance.

Georgia Tech and the Tech Trolley





Program at a Glance

Monday, May 19	**Theater**	**Richards Gallery**	**Westbrook Gallery**
8:30 – 8:45	Announcements		
8:45 – 9:45	Plenary Session 1		
9:45 – 10:30	**Break**		
10:30 – 12:00	Invited Session 1	Invited Session 2	Contributed Session 1
12:00 -1:00	**Lunch**		
1:00 – 2:30	Invited Session 3	Contributed Session 2	Contributed Session 3
2:30 – 3:00	**Break**		
3:00 – 4:30	Invited Session 4	Invited Session 5	Contributed Session 4
4:30 – 6:00	Invited Session 6	Contributed Session 5	
Tuesday, May 20	**Theater**	**Richards Gallery**	**Westbrook Gallery**
8:30 – 10:00	Invited Session 7	Contributed Session 6	
10:00 – 10:30	**Break**		
10:30 – 12:00	Invited Session 8	Invited Session 9	Contributed Session 7
12:00 -1:00	**Lunch**		
1:00 – 2:30	Invited Session 10	Contributed Session 8	
2:30 – 3:00	**Break**		
3:00 – 4:30	Invited Session 11	Invited Session 12	Contributed Session 9
4:30 – 5:30	Plenary Session 2		
7:00 – 9:30	**Conference Dinner** Guest Speaker: Max Morris, Iowa State University		
Wednesday, May 21	**Theater**	**Richards Gallery**	**Westbrook Gallery**
8:30 – 10:00	Invited Session 13	Contributed Session 10	
10:30 – 11:30	Plenary Session 3		
End of Conference			

Conference Schedule

	Monday, May 19, 2008	Tuesday, May 20, 2008	Wednesday, May 21, 2008
8:30 – 10:00	**Plenary** 8:45 – 9:45 Peter Bickel	**Invited 7 (Health Monitoring)** Henry Rolka Howard Burkom Kwok Tsui **Contributed 6 (DoE 3)** Roger Longbotham Donghong Wu Mariano Amo-Salas Minghung Kao	**Invited 13 (DoE)** Ching-Shui Cheng John Stufken Boxin Tang **Contributed 10 (Inference)** Robert Parody Chia-Jung Chang Heeyoung Kim Mihai Giurcanu
10:30 – 12:00	**Invited 1 (DoE Comp Expt)** Dave Higdon Devon Lin Shane Reese **Invited 2 (Reliability)** Mike Hamada Huaiqing Wu Michael Luvalle **Contributed 1 (Stat Comp 1)** Heping Liu Kee-Hoon Kang Ying Hung Shu-Chuan Lin	**Invited 8 (Multi-scale Model 2)** Martha Grover Lance Waller **Invited 9 (DoE Comp Expt)** Max Morris David Steinberg Jason Loeppky **Contributed 7 (Medical Application)** Thomas Burr Moonsu Kang Jun Ye Ana Moura Bargo	**Plenary** James Glimm
13:00 – 14:30	**Invited 3 (Credit Risk)** Agus Sudjianto Aijun Zhang Shuchun Wang **Contributed 2 (DoE 1)** Peter Qian Eric Schoen Ryan Lekivetz Hyejung Moon **Contributed 3 (Spatial Data Analysis)** Yichuan Zhano Robertas Gabrys Russ Lenth Xuyuan Liu	**Invited 10 (Technometrics)** Agus Sudjianto Galit Shmueli **Contributed 8 (Reliability)** Yili Hong Broderick Oluyede Bo Henry Lindqvist Peter Hovey	
15:00 – 16:30	**Invited 4 (Machine Learning)** David Mease Ingo Steinwart John Langford **Invited 5 (Multi-Scale Model 1)** Marco Ferreira Carol Gotway Crawford **Contributed 4 (DoE 2)** Peng Zeng Chien-Yu Peng Lulu Kang Abhyuday Mandal	**Invited 11 (DoE)** William Li Robert Mee Brad Jones **Invited 12 (Nano-technology)** Xinwei Deng Tirthankar Dasgupta **Contributed 9 (Stat Comp)** Ben Haaland Minghua Zhu Jian Zhang Wheyming Song	
16:30 – 18:00	**Invited 6 (DoE)** Martina Vandebroek Dave Woods Jesus Lopez-Fidalgo **Contributed 5 (SPC)** Asokan Mulayath Variyath Scott Grimshaw Charles Champ Sung Won Han	**Plenary** Zhong Lin Wang	

Plenary Session Speakers

Peter Bickel **Monday, May 19, 8:45 – 9:45**
University of California, Berkeley

Peter Bickel is past President of the Bernoulli Society and of the Institute of Mathematical Statistics, a MacArthur Fellow, a COPSS prize winner, and a member of the American Academy of Arts and Sciences and of the National Academy of Sciences. Dr. Bickel uses asymptotic theory to guide the development and assessment of semiparametric models. His studies of hidden Markov models are directed toward understanding how well the method of maximum likelihood performs. He is also interested in the bootstrap, in particular in constructing diagnostic measures to detect malfunction of this technique. Recently he has become involved in developing empirical statistical models for genomic sequences.

James Glimm **Wednesday, May 21, 10:30 – 11:30**
State University of New York at Stony Brook

James Glimm has been noted for contributions to C*-algebras, quantum field theory, partial differential equations, fluid dynamics, scientific computing,the modeling of petroleum reservoirs, and geometric models for structural biology. Dr. Glimm received a 2002 National Medal of Science. The honor - one of the most prestigious in academia for researchers who make major impacts in fields of science and engineering through career-long, ground-breaking achievements - is given for work that spawned many advances in scientific theory and developments leading to new technologies. Glimm was honored for his work in shock wave theory and other cross-disciplinary fields in mathematical physics.

Zhong Lin Wang **Tuesday, May 20, 4:30 – 5:30**
Georgia Institute of Technology

Dr. Wang received his Ph.D in Physics from Arizona State University in 1987. After a year of post-doctoral in the State University of New York at Stony Brook in 1988, Dr. Wang was awarded a Research Fellowship by the Cavendish Laboratory, University of Cambridge, England. He received a U.S. Department of Energy Research Fellowship at Oak Ridge National Laboratory in 1989, and one year later he was appointed as a Research Associate Professor by the University of Tennessee. In 1993, he moved to the National Institute of Standards and Technology (NIST) to set up the microscopy facility. He joined Georgia Tech in 1995.

Banquet Speaker

Max Morris
Iowa State University

Tuesday, May 20, 7:00 – 9:30

Max Morris has a joint appointment at Iowa State University as Professor of Statistics, and Professor of Industrial Engineering. He held previous positions at Oak Ridge National Laboratory, Mississippi State University, and the University of Texas Health Sciences Center at San Antonio. He became a Fellow of the American Statistical Association in 1994, and served as Editor for Technometrics from 1996 to 1998. Professor Morris has earned numerous awards, including the ASQ Jack Youden Prize (2001), NISS Jerome Sacks Award for Cross-Disciplinary Research (2002) and the Iowa STAT-ers Teacher of the Year Award (2004).

Session Schedule and Abstracts

May 19, 2008 :: Monday :: 8:30 to 8:45

Welcome Session
Paul Kvam, Program Committee
Andrew Booker, SRC Program Chair
Mark G. Allen, Georgia Tech Senior Vice Provost for Research and Innovation

May 19, 2008 :: Monday :: 8:45 to 9:45

Plenary Session 1
Chair: Paul Kvam

Low effective dimension in models or data: a key to high dimensional inference?
Peter Bickel

Theoretical analysis seems to suggest that standard problems such as estimating a function of high dimensional variables with noisy data (regression or classification) should be impossible without detailed knowledge or absurdly large amounts of data. Nevertheless, algorithms to perform classification of images or other high dimensional objects are remarkably successful. The generally held explanation is the presence of sparsity/low dimensional structure. I'll discuss analytically and with examples why this may be right.

May 19, 2008 :: Monday :: 10:30 to 12:00

Invited Session 1 – DOE Computer Experiments
Chair: Derek Bingham

Inference from combining detailed computer simulations and experimental data
Dave Higdon

Inference regarding complex physical systems (e.g. subsurface aquifers, charged particle accelerators, shock physics) is typically plagued by a lack of information available from relevant, experimental data. What data is available is usually limited and informs indirectly about the phenomena of interest. However, when the physical system is amenable to computer simulation, these simulations can be combined with experimental observations to give useful information regarding calibration parameters, prediction uncertainty, and model inadequacy. This talk describes a framework for carrying out such simulation-based predictive investigations which involves experimental design, sensitivity analysis, response surface modeling, parameter estimation and accounting for systematic discrepancies between the simulation output and experimental data. An application from cosmology will be used as the context for this talk.

A Flexible Method for Constructing Designs for Computer Experiments
C. Devon Lin

It is becoming increasingly popular to perform scientific experiments on computer simulators since rapid growth in computer power has made it possible to study complex physical phenomena that might otherwise be too time-consuming, expensive, or impossible to observe. In many situations, the dimensionality of the inputs to the computer simulators can be very large. In others, a large simulation of the complex phenomena may be conducted, which requires a new approach to design of experiments. In this talk, I will introduce methods for constructing a rich class of Latin hypercubes of flexible run size with appealing projection and space-filling properties. The class includes many orthogonal Latin hypercubes that are not available in the literature, as well as nearly-orthogonal Latin hypercubes, and cascading Latin hypercubes. This is joint work with Derek Bingham, Randy Sitter and Boxin Tang at Simon Fraser University.

A Computational Approach for the Identification of Pollution Source Directions
Shane Reese

Pollution source apportionment (PSA) is the practice of identifying and describing pollution sources and their contributions. PSA frequently requires the identification of source directions, often as a post-analysis check to ensure that the contribution estimates are reasonable. In this talk, we develop a method of identifying source directions which solves the inverse problem for a particle dispersion model. MCMC is used to evaluate the complex relationship among observed pollutant concentrations, available meteorological information, and unknown source direction parameters. The method is flexible enough to identify multiple source directions for cases in which a species or source type of interest is emitted at more than one location, and Reversible Jump MCMC is used to evaluate the appropriate number of sources. The approach is demonstrated on the St. Louis EPA supersite airshed, demonstrating promising results in identifying known emitters of a variety of pollutants.

Invited Session 2 – Reliability
Chair: William Meeker

Bayesian Assessment of Repairable System Reliability and Availability
Mike Hamada

This talk presents the reliability assessment of repairable systems using failure count and failure time data using a Bayesian framework. With this framework, we can analyze these data with standard repairable system models and naturally handle situations that require hierarchical models. Using the analysis results, we can then evaluate current reliability and other performance criteria of repairable systems. We also consider availability, which accounts for the time to make repairs, and show how simulation simplifies this evaluation.

Analysis of Window-Observation Recurrence Data
Huaiqing Wu

Many systems experience recurrent events. Recurrence data are collected to analyze quantities of interest, such as the mean cumulative number of events. Methods of analysis are available for recurrence data with left and/or right censoring. Due to practical constraints, however, recurrence data are sometimes recorded only in windows. Between the windows, there are gaps over which the process cannot be observed. We extend existing statistical methods, both nonparametric and parametric, to window-observation recurrence data. The nonparametric estimator requires minimum assumptions, but will be inconsistent if the size of the risk set is not positive over the entire period of interest. There is no such difficulty when using a parametric model for the recurrence data. For cases in which the size of the risk set is zero for some periods of time, we propose and compare two alternative hybrid estimators. The methods are illustrated with two example applications.(This is joint work with Jianying Zuo and William Q. Meeker).

The application of statistical kinetics to reliability: status and current research problems
Michael Luvalle

Statistical kinetics is the use of statistical methods to extract kinetic information from degradation and failure data. While there are plainly advantages to being able to model the internal process causing degradation and failure, few statisticians have applied the ideas since their introduction in the mid 1980's. Part of the reason is the need to be familiar with some of the mathematics of chemical kinetics and physics, and part is that perhaps only 10-20% of new material system environment combinations require the effort. This talk will be in two parts, in the first part, I will address the 1st problem by describing "statistical kinetics light" particularly for analysis of failure time data consisting of: (1) Five structural models based on a 1step expansion of the kinetics around the accelerated life model. (2) A way to generate families of failure time models, amenable to standard maximum likelihood analysis, based on extending an approach by Meeker to include link functions similar to those used in GLM. (3) A simplified computational approach to identifying parts of the parameter space in alternative kinetic models that are unidentifiable from current experiments. In the second part I will discuss current research problems in applying statistical kinetic models to degradation data.

Contributed Session 1 - Statistical Computing I
Chair: Russ Lenth

SiZer for Time Series Using Estimated Autocovariance Function
Kee-Hoon Kang, CheolWoo Park, Jan Hannig

SiZer (Chaudhuri and Marron, 1999) is a visualization method based on nonparametric curve estimates. It is based on scale-space ideas from computer vision. Scale-space is a family of kernel smooths indexed by the scale, which is the smoothing parameter or bandwidth h. The statistical inference of SiZer makes heavy use of the assumption of independent errors. This assumption is inappropriate in time series contexts. The focus of this paper is on SiZer for time series. For SiZer to fulfill its potential to flag significant trends in time series, its underlying confidence intervals must be adjusted to properly account for the correlation structure of the data. This adjustment is not straightforward when the correlation structure is unknown. This is because of the identifiability problem between trend and dependence artifacts. Rondonotti, Marron, and Park (2007) addressed this issue and proposed an approach to this dilemma via a visualization which displays the range of trade-offs. While the original SiZer for time series is useful, there is still room for improvement. The estimation of the quantile in a confidence interval relies on a heuristic idea rather than on theory, and the estimation of an autocovariance function is not accurate in some situations. Moreover, any theoretical properties of the proposed method are not provided. This paper aims to remedy these problems in a moderately correlated time series. We propose to estimate the quantile by extreme value theory and the autocovariance function based on differenced time series. In addition, we study asymptotic properties of SiZer for time series in scale-space. Weak convergence of the empirical scale-space surface to its theoretical counterpart has been established under appropriate regularity conditions. A numerical study is conducted to demonstrate the sample performance of the proposed tool.

Binary Time Series Modeling with Application to Adhesion Frequency Experiments
Ying Hung

Repeated adhesion frequency assay is the only published method for measuring the kinetic rates of cell adhesion. Cell adhesion plays an important role in many physiological and pathological processes. Traditional analysis of adhesion frequency experiments assumes that the adhesion test cycles are independent Bernoulli trials. This assumption can often be violated in practice. Motivated by the analysis of repeated adhesion tests, a binary time series model incorporating random effects is developed in this paper. A goodness-of-fit statistic is introduced to assess the adequacy of distribution assumptions on the dependent binary data with random effects. The asymptotic distribution of the goodness-of-fit statistic is derived and its finite-sample performance is examined via a simulation study. Application of the proposed methodology to real data from a T-cell experiment reveals some interesting information, including the dependency between repeated adhesion tests.

Robust Estimation for Spatial Markov Random Field Model
Shu-Chuan Lin

Markov Random Field (MRF) models are useful in analyzing spatial lattice data collected from semiconductor device fabrication and printed circuit board manufacturing processes or agricultural field trials. When outliers are present in the data, classical parameter estimation techniques (e.g., least squares) can be inefficient and potentially mislead the analyst. This presentation extends the MRF model to accommodate outliers and proposes robust parameter estimation methods such as robust M- and RA-estimates. Asymptotic distributions of the estimates with a non-differentiable robustifying function are derived. Extensive simulation studies explore robustness properties of the proposed methods in situations with various amounts of outliers in different patterns. Also provided are studies of analysis of grid data with and without the edge information. Three data sets taken from the literature illustrate advantages of the methods. Joint work with Prof. J.-C. Lu.

Application of Statistical Estimators to Expanding Pareto Fronts
Heping Liu, Alice Smith, Haluk Yapicoglu, Gerry Dozier

With the emergence of computational intelligence, metaheuristics have been used to solve multi-objective optimization problems (MOPs). Metaheuristics are less susceptible to the shape or continuity of the Pareto front, which gives them some advantages in comparison with traditional programming methods. However, the nondominated solutions are identified one at a time and these metaheuristics provide less information on the contiguity of the Pareto front. Also, metaheuristics require frequent function evaluations, which can exhaust computational sources. This paper explores using statistical interpolation methods to expand optimal solutions obtained by metaheuristic search. The goal is to significantly increase the number of Pareto optimal solutions while limiting computational effort. The interpolation approaches studied are Kriging and general regression neural networks (GRNN). Kriging is a best linear unbiased estimator and GRNN is a strongly statistically based neural network paradigm. Both of them are nonlinear statistical interpolators. Several test functions are used to examine their performance. The results show both Kriging and GRNN successfully expand and enrich the Pareto fronts of MOPs from metaheuristic search.

May 19, 2008 :: Monday :: 13:00 to 14:30

Invited Session 3 – Modeling Credit Risk Portfolios in Retail Banking
Chair: Agus Sudjianto

Credit Portfolio Modeling in Retail Banking and Its Statistical Challenge
Agus Sudjianto

Credit risk modeling is a very active research topic in computational finance; in particular, for the purpose of modeling corporate default. In retail banking, however, the subject has received less attention. The purpose of this talk is to introduce the challenge of modeling the credit risk of very large and heterogeneous credit portfolios in retail banking from a statistical perspective. Various existing approaches will be presented accompanied by examples. Problems encountered in practice will be presented to stimulate further research.

Recent Advances in Default Risk Modeling Based on Vintage Data
Aijun Zhang

Vintage data commonly arise in financial risk management, especially in today's retail banking. They have two special features: triangular structure and dual-time coordinates. Existing credit risk models cannot be applied in this situation. In this paper, we propose a dual-time intensity model for analyzing the default risk, which is decomposable into maturation, macroeconomic and vintage effects. [Joint work with Dr. Agus Sudjianto, Ph.D. and Dr. Vijay Nair, Ph.D.]

Semiparametric Modeling of Macroeconomic Drivers to Credit Performance
Shuchun Wang

A central issue in credit risk management is the identification of the driving forces of credit performance. Empirical approaches often shed little light on the causal relationship among state variables whereas theoretical approaches often fail to capture the typical characteristics observed in practice. In this work, we present a semiparametric structural model to bridge this gap. The model explains how macro-economy changes may affect credit performance. With rigorous econometric analysis, we show how this model helps us understand the complex relationship between state economy and credit performance using a real data example. [Joint work with Dr. Ming Yuan, Ph.D.]

Contributed Session 2 – Design of Experiments I
Chair: Roger Longbotham

Sliced Space-Filling Designs
Peter Qian

Design construction for computer experiments with qualitative and quantitative factors is an important but unsolved issue. In this work a general approach is proposed for constructing a new type of design called sliced space-filling design to accommodate these two types of factors. It starts with constructing a Latin hypercube design based on a special orthogonal array for the quantitative factors and then partition the design into groups corresponding to different level combinations of the qualitative factors. The points in each of these groups are guaranteed to have good space-filling properties in low dimensions. This is joint work with Jeff Wu at Georgia Tech.

A Simple Orthogonal Blocking Strategy for Mixed Irregular Designs of Strength Two
Eric Schoen

Orthogonal arrays (OAs) of strength 2 permit independent estimation of main effects. An orthogonally blocked OA could be considered as an OA with one additional factor. I propose to search for suitable blocking arrangements by studying projections of arrays into optimum arrays with one factor less. I illustrate with a complete catalogue of blocked pure or mixed 18-run arrays.

Partially Clear Two Factor Interactions for Fractional Factorial Designs
Ryan Lekivetz

In a $2^{\{m-p\}}$ design, a two-factor interaction (2fi) is said to be clear if it is not aliased with any main effect or any other 2fi. A clear 2fi can be estimated under the weak assumption that all interactions of three or more factors are negligible. However, the existence of a clear 2fi can be restrictive on the number of factors that can be used in a design. We examine the usefulness of "partially-clear" 2fi's - those that can be estimated when certain 2fi's are also negligible, and report our findings for designs of 32 and 64 runs.

Two-stage Group Screening for Computer Experiments
Hyejung Moon, Thomas Santner, Angela Dean

Computer codes are used to simulate physical processes in many fields of research. It is not uncommon for runs of the computer code to be extremely time consuming. Since a large fraction of the inputs may have relatively little effect on the output, rapid identification of the influential inputs can be crucial. This research proposes a method for identifying active inputs using a two-stage group screening procedure. The proposed procedure first partitions individual inputs into groups and fits a model to the groups. Active groups are identified at the first stage and only inputs in those groups that appear active proceed to the second stage. After collecting additional data, the final active inputs are determined. First-order and total sensitivity indices are used to determine whether an input is active. The sensitivity index of each input is compared with a yardstick value obtained from the distribution of the sensitivity index of a dummy input. The dummy input is assumed to be partially related to the output so that the dummy input dominates non-active real inputs, that is, those with small effects. The success of two-stage group screening is very dependent on the design used in each stage. Orthogonal array-based Latin hypercube designs under a minimax correlation criterion are investigated for both group variables and individual inputs at the first stage and for only the chosen individual inputs at the second stage. The performance of the two-stage group screening is compared with one-stage variable selection in terms of error rates, the expected number of experiment runs, and computation times. Empirical performance of the procedure is studied by simulation. The two-stage group screening is shown to be fast and accurate when the proportion of non-active inputs is large.

Contributed Session 3 – Special Data Analysis
Chair: Abhyuday Mandal

Omnibus Tests for Comparison of Competing Risks with Covariate Effects via Additive Risk Model
Yichuan Zhao

It is of interest that researchers study competing risks in which subjects may fail from any one of K causes. Comparing any two competing risks with covariate effects is very important in medical studies. We develop omnibus tests for comparing cause-specific hazard rates and cumulative incidence functions at specified covariate levels. In the paper, the omnibus tests are derived under the additive risk model by a weighted difference of estimates of cumulative cause-specific hazard rates. Simultaneous confidence bands for the difference of two conditional cumulative incidence functions are also constructed. A simulation procedure is used to sample from the null distribution of the test process in which the graphical and numerical techniques are used to detect the significant difference in the risks. In addition, we conduct a simulation study, and it shows that the proposed procedure has a good finite sample performance. A melanoma data set in clinical trial is used for the purpose of illustration.

Detecting Changes in the Mean of Functional Observation
Robertas Gabrys

Principal component analysis (PCA) has become a fundamental tool of functional data analysis. It represents the functional data as $X_i(t) = \mu(t) + \sum_{1<l<\infty} \eta_{i,l}\varphi_l(t)$; where μ is the common mean, $\varphi_l(t)$ are the eigenfunctions of the covariance operator, and the $\eta_{i,l}$ are the scores. Inferential procedures assume that the mean function $\mu(t)$ is the same for all values of i. If, in fact, the observations do not come from one population, but rather their mean changes at some point(s), the results of PCA are confounded by the change(s). It is therefore important to develop a methodology to test the assumption of a common functional mean. We develop such a test using quantities which can be readily computed in the R package fda. The null distribution of the test statistic is asymptotically pivotal with a well-known asymptotic distribution. The asymptotic test has excellent finite sample performance. Its application is illustrated on temperature data from Prague, England and Greenland. The comprehensive asymptotic theory for the estimation of a change{point in the mean function of functional observations is developed. Both the case of a constant change size, and the case of a change whose size approaches zero, as the sample size tends to infinity are considered. We show how the limit distribution of a suitably defined change-point estimator depends on the size and location of the change. The theoretical insights are confirmed by a simulation study which illustrates the behavior of the estimator in finite samples.

Regression Modeling for Computer Model Validation with Functional Responses
Xuyuan Liu

Statistical analysis of functional responses based on functional data from both computer and physical experiments has gained increasing attention due to the dynamic nature of many engineering systems. However, the complexity and huge amount of functional data bring many difficulties to apply traditional or existing methodologies. The objective of the present study is twofold: (1) prediction of functional responses based on functional data and (2) prediction of bias function for validation of a computer model that predicts functional responses. In this paper, we first develop a functional regression model with linear basis functions to analyze functional data. Then combining data from both computer and physical experiments, we use the functional analysis modeling to predict the bias function which is crucial for validating a computer model. The proposed method, following the classical nonparametric regression framework, uses a single step procedure which is easily implemented and computationally efficient. Through an application example of motor engine analysis to predict acceleration performance and gear shift events, we demonstrate our approach and compare it to using the Gaussian process modeling approach.

StatWeave: Flexible Software for Literate Programming in Statistics
Russell V. Lenth

The R function Sweave() and related programs such as SASweave make it easy to incorporate R (or SAS) code and output in a LaTeX document. StatWeave unifies and extends these. It supports several statistical languages (possibly within the same document), and both LaTeX and OpenOffice file formats. It is written in Java (making it portable across platforms), and it is possible to add new file formats and new languages simply by writing and configuring additional Java classes. The talk will include a discussion of StatWeave's design, and some examples.

May 19, 2008 :: Monday :: 15:00 to 16:30

Invited Session 4 – Machine Learning
Chair: Agus Sudjianto

Evidence Contrary to the Statistical View of Boosting
David Mease

The statistical perspective on boosting algorithms focuses on optimization, drawing parallels with maximum likelihood estimation for logistic regression. In this talk we present empirical evidence that raises questions about this view. Although the statistical perspective provides a theoretical framework within which it is possible to derive theorems and create new algorithms for general contexts, we show that there remain many unanswered important questions. Furthermore, we provide examples that reveal crucial flaws in the many practical suggestions and new algorithms that are derived from the statistical view. We examine experiments using simple simulation models to illustrate some of these flaws and their practical consequences. This is joint work with Abraham Wyner at the University of Pennsylvania.

Estimating Conditional Quantiles with Support Vector Machines
Ingo Steinwart

We first recall a recently proposed support vector machine (SVM) formulation for the problem of estimating conditional quantiles. Since this SVM is based on the so-called pinball loss we then investigate this loss and its quantitative relation to conditional quantiles in detail. Finally, we apply these findings to describe the learning performance of the corresponding SVM.

Learning without the Loss
John Langford

In many natural situations, you can probe the loss (or reward) for one action, but you do not know the loss of other actions. This problem is simpler and more tractable than reinforcement learning, but still substantially harder than supervised learning because it has an inherent exploration component. I will discuss two algorithms for this setting:
(1) Epoch-greedy, which is a very simple method for trading off between exploration and

exploitation. (2) Offset Tree, which is a method for reducing this problem to binary classification.

Invited Session 5 – Multi-scale Modeling
Chair: Lance Waller
Gaussian multiscale spatio-temporal models
Marco Ferreira

We develop a new class of multiscale spatio-temporal models for Gaussian areal data. Our framework decomposes the spatio-temporal observations and underlying process into several scales of resolution. Under this decomposition the model evolves the multiscale coefficients through time with structural state-space equations. The multiscale decomposition considered here, which includes wavelet decompositions as particular case, is able to accommodate irregular grids and heteroscedastic errors. The multiscale spatio-temporal framework we develop has several salient attributes. First, the multiscale decomposition leads to an extremely efficient divide-and-conquer estimation algorithm. Second, the multiscale coefficients have an interpretation of their own; thus, the multiscale spatio-temporal framework may offer new insight on understudied multiscale aspects of spatio-temporal observations. Finally, deterministic relationships between different resolution levels are automatically respected for both observations, the latent process, and the estimated latent process. We illustrate the use of our multiscale framework with an analysis of a spatio-temporal dataset on agriculture production in the state of Espirito Santo, Brazil (Joint work with Scott Holan and Adelmo Bertolde)

Linking Environmental and Health Data from Different Spatial Scales: A Case Study from Florida's Environmental Public Health Tracking Initiative
Carol Gotway Crawford

The Centers for Disease Control and Prevention (CDC) created the Environmental Public Health Tracking (EPHT) Program to integrate hazard monitoring, exposure, and health effects surveillance into a cohesive tracking network. Few new data are being collected in the EPHT effort. The emphasis has been, and likely will continue to be, on the synthesis of existing environmental and health data systems. Part of Florida's effort to move toward implementation of EPHT is to develop models of the spatial and temporal association between myocardial infarctions and ambient ozone levels in Florida. Existing data were obtained from Florida's Agency for Health Care Administration, Florida's Department of Environmental Protection, the U.S. Census Bureau, and CDC's Behavioral Risk Factor Surveillance System. In this presentation, we highlight the opportunities and challenges associated with combining disparate spatial data for EPHT analyses. We compare the results from two different approaches to data linkage, focusing on the need to account for spatial scale and the support of spatial data in the analysis.

Contributed Session 4 – Design of Experiments II
Chair: Peter Hovey

Testing the Significance of Predictors via Link Free Method
Peng Zeng

One important step in regression analysis is to identify significant predictors from a pool of candidates so that a parsimonious model can be obtained using these significant predictors only. However, most of the available methods assume linear relationships between response and predictors, which may be inappropriate in some applications. In this talk, we discuss a link-free method that avoids specifying how the response depends on the predictors. Therefore, this method has no problem of model misspecification, and it is suitable for selecting significant predictors at the preliminary stage of data analysis. A test statistic is suggested and its asymptotic distribution is derived. Examples are used to demonstrate the proposed method.

Row-wise Complementary Designs
Chien-Yu Peng, Shao-Wei Chang

The technique of (columnwise) complementary designs, proposed independently by Chen and Hedayat (1996) and Tang and Wu (1996), is powerful for characterizing designs with a large number of factors. In this work, we extend the idea and propose row-wise complementary designs which are particularly useful in handling designs with large run sizes. A pair of designs is mutually row-wise complementary of order r if they are row partition of a full factorial design with r replicates. Based on a polynomial representation approach for factorial designs called indicator function, we establish a series of relationships between a design and its row-wise complementary design, which includes isomorphism, orthogonality, generalized word length pattern, minimum aberration, moment aberration, and uniformity. In addition, we apply the technique of row-wise complementary design to identify minimum aberration two-level designs with larger run sizes. The method can be generalized and applied to higher-level, mixed-level, blocked factorial designs, or column-wise complementary designs.

Bayesian Optimal Single Arrays for Robust Parameter Design
Lulu Kang

It is critical to estimate control-by-noise interactions in robust parameter design. This can be achieved by using a cross array, which is a cross product of a design for control factors and another design for noise factors. However, the total run size of such arrays can be prohibitively large. To reduce the run size, single arrays are proposed in the literature, where a modified effect hierarchy principle is used for the optimal selection of the arrays. In this article, we argue that effect hierarchy is a property of the system and cannot be altered depending on the objective of an experiment. We propose a Bayesian approach to develop single arrays which incorporate the importance of control-by-noise interactions without altering the effect hierarchy. The approach is very general and places no restrictions on the number of runs or levels or type of factors or type of designs. A modified exchange algorithm is proposed for finding the optimal single arrays. We also explain how to design experiments with internal noise factors. The advantages of the proposed approach are illustrated using several examples.

G-SELC: Optimization by Sequential Elimination of Level Combinations Using Genetic Algorithm and Gaussian Processes
Abhyuday Mandal

Identifying promising compounds from a vast collection of feasible compounds is an important and yet challenging problem in pharmaceutical industry. An efficient solution to this problem will help reduce the expenditure at the early stages of drug discovery. In an attempt to solve this problem, Mandal, Wu and Johnson (2006) proposed SELC algorithm which was motivated by SEL algorithm of Wu, Mao and Ma (1990). However, SELC fails to extract substantial information from the data to guide the search efficiently as this methodology is not based on any statistical modeling of the data. The current approach uses Gaussian Process (GP) modeling to improve upon SELC method, and hence named as *G*-SELC. The performance of the proposed methodology is illustrated using four and five dimensional test functions, and its higher success rates are demonstrated via simulations. Finally, we use the proposed approach on a real pharmaceutical data set for finding a group of chemical compounds with optimal properties.

May 19, 2008 :: Monday :: 17:00 to 18:30

Invited Session 6 – Industrial Experiments: Recent European Advances in Design
Chair: Sue Lewis

Design issues in Stated Preference Experiments
Martina Vandebroek

In a stated preference or conjoint experiment respondents evaluate a number of products that are defined by their underlying characteristics. The resulting data yields information on the importance that respondents attach to the different characteristics, also called the part-worths. With this information, one can forecast consumer demand for new products. Data can be collected in various ways. In a discrete choice experiment several choice sets consisting of a number of products are presented to the respondents, these are then asked to choose their preferred product from each choice set. Alternatively, respondents can be asked to rank the alternatives in the choice set in a decreasing or increasing order of utility which is then called a rank order experiment. When respondents are asked to rate the alternative products, the experiment is called a rating-based conjoint experiment. The optimal design of a stated preference experiment consists of choosing the appropriate alternatives and of grouping the alternatives in choice sets such that the information gathered about the part-worths is maximized. In this presentation an overview will be given of the statistical models that are used to analyze stated preference data and of the design issues involved. Special attention will be given to the problem of assessing accurately the marginal rate of substitution which measures the consumer's willingness to give up an attribute of a product in exchange for another attribute. The marginal rate of substitution can be obtained by taking the ratio of two part-worths which leads to specific design problems.

Experiments in blocks for a non-normal response via generalised estimating equations
Dave Woods

Many industrial experiments measure a response that cannot be adequately described by a linear model with normally distributed errors. An example is an experiment in aeronautics to investigate the cracking of bearing coatings where a binary response was observed, success (no cracking) or failure (cracked). A further complication which often occurs in practice is the need to run the experiment in blocks, for example, to account for different operators or batches of experimental units. To produce more efficient experiments, block effects are often included in the model for the response. When the block effects can be considered as nuisance variables, a marginal (or population averaged) model may be appropriate, where the effect of individual blocks are not explicitly modelled. We discuss block designs for experiments where the response is described by a marginal model fitted using Generalised Estimating Equations (GEEs). GEEs are an extension of Generalised Linear Models (GLMs) that incorporate a correlation structure between experiment units in the same block; the marginal response for each observation follows an appropriate GLM. This talk will describe some design strategies for such models in an industrial context.

Optimal designs for models with potential censoring
Jesus Lopez-Fidalgo

Designing an experiment for a real life problem may involve new and complex situations. As motivation a medical problem of finding an experimental design to predict cardiopulmonary morbidity after lung resection with standardized exercise oximetry is considered. A degree of complexity appears when an experimental unit cannot complete the assigned experimental condition, e.g. the prescribed exercise time in the medical example. Thus, the explanatory variable has to be considered as potentially censored. This presentation deals with optimal design theory for models with potential censoring either on independent or dependent variables. On the one hand optimal approximate designs when an independent variable might be censored are considered. The problem is which design should be applied to obtain at the end of the experimentation an optimal approximate design when the censored distribution function is assumed known in advance. On the other hand optimal experimental design theory is adapted to a particular Cox Regression problem. The failure time is modelled according to a probability distribution depending on some explanatory variables through a linear model. In both cases equivalent theorems and algorithms are provided in order to calculate optimal designs. Some examples illustrate these approaches for D-optimality.

Contributed Session 5 – Statistical Process Control
Chair: Jun Ye

A Multivariate Robust Control Chart for Individual Observations
Asokan Mulayath Variyath

To monitor a multivariate process, classical Hoteling's T^2 control chart is often used. However, it is well known that such control charts are very sensitive to the presence of outlying observations in the historical Phase I data used to set up its control limits. In this paper, we propose a robust Hotelling's T^2 type control chart for individual observations based on highly robust and efficient estimators of location and scatter known as re-weighted minimum covariance determinant (RMCD) estimators. We propose this control chart for monitoring the Phase II data set where control limits are set by using the phase I robust estimates based on RMCD. We illustrate how to set the control limit for the proposed new control chart, study its performance in comparison with existing methods using simulations. We illustrate the proposed method using a real case example.

Note on the CUSUM Charts for Uys and Lombard
Charles Champ, Allison Jones-Farmer

Simulation was used by Uys and Lombard (2007) to analyze the run length performance of the CUSUM chart they proposed. They were not aware of an analytical method fir determining the performance of their chart. We give the integral equations that are useful in evaluating the run length performance of their chart. Some corrections are given for their selection of the upper control limit. Versions of their chart are introduced that would be useful for the practitioner both when the in-control mean and standard deviation are known and when these parameters are estimated.

Control Charts for Spatial Data
Scott Grimshaw

New technology to measure thickness of bottles is non-destructive and provides the opportunity to obtain measurements at any location. While there is a hope control charts using more measurements will be more sensitive to detecting flaws in local regions, the intuition is that measurements close together will be redundant. This talk treats the bottle measurements as spatial data and investigates the properties of Shewhart control charts.

A Comparison between SCAN and CUSUM for Detecting Increased Rates of Poisson Observations for On-line Monitoring
Sung Won Han

SCAN statistics are used in health surveillance to detect increased rates of diseases or symptoms in finite time domains. However, for on-line monitoring, scan methods have been modified to detect rate increases as soon as possible. In this paper, we investigate the properties of the scan statistics modified for on-line monitoring under independent Poisson observation. We also compare the SCAN method with the CUSUM method, known as an optimal method, under stationary baseline and sustained rate changes. We compare these methods based on the conditional expected delay (CED), given various increased rates. For the comparison, we adjust the scale by optimal envelops. We also show how much CED's differ among the methods. If we know, even roughly, what is the shift size, we highly recommend the CUSUM method over the CUSUM, as CUSUM is better with regard to overall criteria such as CED and robustness.

May 20, 2008 :: Tuesday :: 8:30 to 10:00

Invited Session 7 – Health Process Monitoring
Chair: Henry Rolka

Recent Background and Influences on the Direction of Research in Public Health Biosurveillance
Henry Rolka

Biosurveillance for Public Health has been a focus of recent legislative and policy initiatives such as the Pandemic and All Hazards Preparedness Act (PAHPA) and Homeland Security Presidential Directives (HSPDs). The consequences of these initiatives produce increased technical requirements for data analysis, information science and communicating biosurveillance results involving the characterization of uncertainty. In addition to traditional analytic epidemiological studies, data and information processing is used to establish a "common operating picture" and "situational awareness" for public health on an ongoing bases and especially to be used in responding to emergencies. These evolving public health operational requirements augment the complexity of practical biosurveillance. This talk will provide a brief background and history with some examples of challenging components and areas of potential research.

Bridging the Gap Between Statistical Research and Public Health Practice in Infectious Disease Surveillance
Howard Burkom

Recent years have seen considerable research in adaptation of methods in statistics, machine learning, data mining, and related fields for the evolving application of advanced disease surveillance. This presentation addresses a mismatch between some developed methodology and the needs of the health monitoring community. In preparation for a recent roundtable discussion convened by the U.S. Medicine Institute, practicing epidemiologists were polled regarding the utility of their automated systems. Responses indicated that users are getting benefits from these systems, but often not the benefits conceived by developers. Another finding was that "situational awareness" is meaningful and multifaceted, varying according to the objectives and purview of the monitoring institutions. Many of these facets bring statistical challenges. Meeting these challenges requires technology adaptations and combinations, but only after surveillance needs and the associated data environments are well understood. This talk will elaborate, enumerate some of these challenges, and provide illustration using one example drawn from a surveillance dataset.

Research in Public Health and Disease Surveillance
Kwok Tsui

Due to various outbreaks of influenza and continuing bioterrorism threat, research efforts on healthcare and public health surveillance have become very important worldwide. In this talk we will present overview and review of the general issues involved in healthcare, public health, and syndromic surveillance. In particular, we review the latest research in disease transmission models, surveillance systems, monitoring methods, and performance measures. We also discuss the research challenges and illustrate them with various problems and examples.

Contributed Session 6 – Design of Experiments III
Chair: Peter Qian

Random Design Space for Lung's Retention Model
Mariano Amo-Salas, J. Lopez-Fidalgo, J.M. Rodriguez-Diaz

A model of lung retention of radioactive particles is considered in this work. When a leak of radiation happens in facilities with workers the accident is detected only at the end of the workers' shift when the filters are checked. Thus, the actual time when the leak happens in not known and therefore the design space of times to perform bioassays on the workers should be considered as a random set. Optimal designs for different possible times of the accident are computed and they are compared with the worst case, when the accident happens at the beginning of the shift. Moreover, different distributions of probability are assumed as distribution of the moment of the accident and designs computed for those situations. Another research line is based on the study of the covariance matrix. The observations taken on the same worker are correlated. We can consider the covariance parameter as a parameter of interest (to be estimated). The "Virtual Noise" method is applied to compute the optimal designs for this model. Different covariance structures are used in this investigation.

Challenges to Online Experimentation
Roger Longbotham

Designed experiments with online properties (websites) have a number of unique aspects that present challenges and opportunity for further research. Some of the differences from "offline" experimentation are the very large sample sizes, highly skewed distributions, very dynamic environment (i.e. non-stationarity) and ability to measure many aspects of the user experience. In addition, you often have inconsistent and limited information about users. I will present the challenges we have faced with online experiments and solutions we have found for some of them. Some of the areas where there are unique challenges are experimental design, degraded power for certain metrics, issues regarding user identification, robot and fraud detection, and measuring trends in treatment effects.

A Computational Algorithm for Searching for Optimal Orthogonal Arrays for Estimating Main Effects and Some Specified Two-factor Interactions
Dong Hong Wu, Boxin Tang

We consider the problem of finding optimal orthogonal arrays for estimating main effects and some specified two-factor interactions. Based on a theoretical result from Tang and Zhou (2008), we develop a computational algorithm for this purpose. The performance of the algorithm is evaluated by comparing the results obtained by the algorithm with those from complete search. Finally, we present a useful collection of optimal orthogonal arrays with small run sizes.

Multi-objective Optimal Experimental Designs for Event-Related fMRI Studies
Ming Hung Kao

Well planned experimental designs are crucial to successfully achieving statistical goals under psychological restrictions in functional magnetic resonance imaging (fMRI) studies. With sophisticated allocations of stimuli, researchers can gather valuable fMRI time series and acquire precise information about human brain activities. However, due to the nature of fMRI experiments, the underlying design space is very large and irregular. This makes it difficult to find an optimal design that simultaneously accomplishes various goals of a study and fulfills the scientific restrictions. In this work, we carefully define design measurements to evaluate the “goodness” of a design with respect to statistical aims and psychological constraints. Achievable bounds for these measurements are also derived to provide global normalization. We then create a multi-objective design criterion by combining globally normalized measurements and use it to evaluate fMRI designs. Based on this criterion, a modified genetic algorithm, which efficiently searches over the design space, is purposed for generating multi-objective optimal designs. We show via simulation that our approach outperforms those in the literature.

May 20, 2008 :: Tuesday :: 10:30 to 12:00

Invited Session 8 – Statistical Analysis and Design
Chair: Abhyuday Mandal

An Experimental Design Approach to Process Design
Martha Grover Gallivan

When designing a new process, one rarely has a perfect model, but in the case of nanoscale systems, there may be several candidate models with unknown coefficients, and none will be the "correct" one. Also, technology and specifications for products such as microelectronics and pharmaceuticals change quickly, not leaving enough time to implement a more accurate model once it has been built and validated. However, mechanistic understanding will be encoded in the candidate models, such that they can be useful in process design, especially over a subset of the experimental space. One way to approach this problem is to use the set of candidate models, along with the available experimental data, to design the next experiment. By designing experiments to improve the candidate models' prediction variance at the predicted optimal operating point of a process, one can better understand the underlying phenomena and apply this knowledge to improve the candidate models. We apply concepts for model selection and spatial statistics to experimental design, for the microstructure design of metal oxides by a chemical vapor deposition process.

Spatial pattern and process: Assessing spatial performance of spatial models
Lance Waller

The topic is the use of spatial statistics to evaluate spatial simulation models with examples (primarily) from disease ecology, but I'll link these to more general aspects of linking mathematical models and statistical inference.

Statistical issues in the development of an automotive on-board diagnostics
Stephano Barone

The automotive on-board diagnostics (OBD) is a complex system whose aim is to monitor the state-of-health of another complex system, i.e. the vehicle engine. OBD aimed at the containment of polluting exhaust emissions is nowadays compulsory on every vehicle model introduced into the market. Increasingly stringent regulations and ever more demanding customers push from opposite sides towards a perfect working of OBD systems. Therefore manufacturers are globally interested in the high robustness and reliability of such critical systems.
The aim of this presentation is to give an insight into the main statistical problems arisen during an extensive research project carried out in collaboration with an Italian manufacturer. In particular three aspects will be evidenced: a) the modeling and the analysis of variation aimed at understanding the causes of variation of diagnostic indexes and prioritizing improvement actions; b) the robust calibration of OBD systems by opportunely combining physical and computer experiments; c) the treatment of truly unequally spaced time series for the statistical process control in the burn-in phase. Some generalizations and directions for future research will be illustrated.

Invited Session 9 – DOE Computer Experiments
Chair: Peter Qian

Data-Driven Nonstationary Modeling of Deterministic Computer Models
Max Morris

“Spatial” Gaussian stochastic processes are the basis for much of the statistical methodology recently developed for analyzing data from deterministic computer models. Unless substantial prior information is available about the computer model (or function), applications are typically based on stationary processes, or on processes with stationary variance structure and very simple spatial structure in the mean (e.g. polynomial). When the computer model has characteristics that are not typical of stationary process realizations, this can often result in undesirable behavior in the point predictors of unobserved model outputs and/or output prediction intervals that undercover their targets. In this talk, we introduce a family of function prediction procedures that take the general form of traditional conditional/posterior predictors based on stationary models, but that have the flexibility to behave in a nonstationary way according to the observed data. Demonstration based on test functions shows that these methods can result in predictions of substantially smaller root-mean-squared-error and closer-to-nominal prediction interval coverage than either Bayes or Empirical Bayes predictions based on stationary processes.

Orthogonal Nearly Latin Hypercube Designs
David Steinberg

Latin Hypercube (LHC) designs are one of the most popular choices for experiments run on computer simulators. As first proposed by McKay, Beckman and Conover in 1979, LHC designs guarantee that input factor settings are uniformly spread for each single factor, but rely on “random mating” to achieve good spread in high dimensions. In experiments with many factors, some pairs of factors typically have moderately high correlations and a number of schemes have been proposed to reduce the correlations. Steinberg and Lin derived a construction scheme that gives n-run LHC designs, with close to n orthogonal columns, but for very limited values of n. Here we extend that approach to a broad set of sample sizes. We are able to preserve the perfect orthogonality, but not the LHC property. However, the designs are “nearly” LHC's, in the sense that the univariate projection of each factor is close to uniform. The construction scheme relies on projection properties of Plackett-Burman designs and proves to be useful in the original class of designs, as well, for avoiding poorly covered low-dimensional projections. This is joint work with Dennis Lin.

Design of Computer Experiments: Does size matter?
Jason Loeppky

In recent years, virtual experiments implemented by a complex computer code or mathematical model are supplementing or even replacing physical experiments. The computer code mathematically describes the relationship between several input variables and one or more output variables. Often the computer models in question can be computationally demanding. Thus, direct evaluation of the code for optimization or validation is not possible in general. The general strategy is to build a statistical model to act a surrogate or an emulator of the true code. A long used rule of thumb for sample size takes a runs size that is 10 times the number of active dimensions. In this talk we investigate this rule of thumb for a variety of problems encountered in practice. In some cases we will show that increasing the sample size has a large effect on prediction quality and in other cases increasing the sample size has little to no effect. These issues will be demonstrated using a model for polar ice caps and a model for the ligand activation of a G-protein in yeast.

Contributed Session 7 – Medical Applications
Chair: Scott Grimshaw

Choosing the Number of Repeats when Characterizing an Isotope Identification Algorithm
Thomas Burr, Michael Hamada

The performance of radio-isotope identification (RIID) algorithms using gamma spectroscopy is a subject of increasing importance. For example, deployed sensors at locations that screen for illicit nuclear material rely on isotope identification to resolve innocent nuisance alarms arising from naturally occurring radioactive material. Recent and upcoming data collections for RIID testing consist of repeat measurements for each of several scenarios to test RIID algorithms. Efficient allocation of measurement resources requires an appropriate number of repeats for each scenario. Here we consider the potential merit of augmenting real repeats with realistic synthetic repeats in the context of choosing an appropriate number of repeats. Because synthetic data can rarely capture all relevant details of real data, this leads to the well-known bias-variance tradeoff in the context of RIID performance estimation. A Bayesian approach (with a vague prior so that results are similar to a corresponding non-Bayesian approach) considers the variability in the sample mean. A second approach ignores it. In both approaches, our results suggest that using, for example, 60 simulated spectra based on an estimated count rate mean from approximately 10 real repeats will result in similar-quality inference based on all 60 real repeats.

Kendall's Tau Type Rank Statistics in Genome Data
Moon Su Kang

High-dimensional data models abound in genomics studies, where often inadequately small sample sizes create impasses for incorporation of standard statistical tools. Conventional assumptions of linearity of regression, homoscedasticity and (multi-) normality of errors may not be tenable in many such interdisciplinary setups. In this study, Kendall's tau-type rank statistics are employed for statistical inference, avoiding most of parametric assumptions to a greater extent. The proposed procedures are compared with Kendall's tau statistic based ones. Applications in microarray data models are stressed.

Different Methods, Similar Approaches: Geostatistical Analysis and Sparse Principal Component Analysis in Testing the Activation of fMRI Time Series
Jun Ye

Functional Magnetic Resonance Imaging (fMRI) is a relatively new non-invasive technique for studying the workings of the active human brain. Due to the high noise level in fMRI data experiments, many clustering techniques have aimed at cross-correlation between fMRI time series and the experimental protocol signal. In this paper, we use geostatistical methods and sparse principle component analysis in clustering and demonstrate their uses in fMRI data analysis. Our results show that both techniques can effectively identify regions of similar activations. In the analysis of brain imaging data, using the empirical autocorrelation function offers an important advantage over existing correlation approaches. Unlike conventional correlation methods, the proposed geostatistical method does not require prior knowledge about the reference function. Our analysis also demonstrates that using sparse principal component analysis may increase the precision of the dimension reduction step compared to the traditional screening via 2-sample t-test by more accurately capturing the main patterns of the voxels. SPCA produces comparable clustering results to K-means "correlation" method. The analysis also provides evidence that masking the brain can change the clustering results. Although masking the brain is a common dimension reduction step, we show that it is not necessarily effective and may result in less convincing inference, especially when the data has "head motion" evident around the edges of the brain and has minor noise inside the brain.

Identifying Active Region for the Brain via Social Network Modeling in Multiple Subjects
Ana Moura Bargo

Functional Magnetic Resonance Imaging (fMRI) is a relatively new and non-invasive procedure that uses magnetic resonance imaging (MRI) techniques to measure metabolic changes that result in activation in a region of the brain. In this paper, we propose an innovative technique, which utilizes a social network model, to predict the active brain voxels for single and multiple subjects. For single subject analysis, Bayesian data analysis techniques are employed to identify the structure of the social network. The parameters of the associated models are estimated by Pseudolikelihood (PLKHD) and Markov Chain Monte Carlo (MCMC) methods, both of which lead to consistent estimates of activation probabilities. The conditions for which phase transition, which can be defined as the existence of several local conditional probabilities, as well as conditions to guarantee non-degeneracy of the system, which happens when the value of the sufficient statistic falls close to the boundary of the convex hull of the social network, are identified. Multiple subject analysis is more complex because of the between-subject variability, which is larger than within-subject one (Handwerker et al., 2004). For multiple subject analysis, social network structure is determined via a variation of Principal Component Analysis (PCA) and Canonical Covariance Analysis (CCA), combined with a 2-D test for white noise.

May 20, 2008 :: Tuesday :: 13:00 to 14:30

Invited Session 10 - Technometrics
Chair: David Steinberg

Statistical Methods for Fighting Financial Crimes
Agus Sudjianto

Financial crimes affect millions of people every year, and financial institutions must employ methods to protect themselves and their customers. Fraud and money laundering are two common types of crimes, and there has been extensive research to develop algorithms to detect these crimes. Detection algorithms face a large set of challenges: financial crimes are rare events, which leads to severely imbalanced classes; criminals deliberately attempt to conceal the nature of their actions, resulting in severe class overlapping; and they quickly change strategies over time in order to trick current detection techniques. In some cases, legal constraints and investigation delays make it impossible to actually verify suspected crimes in a timely manner, resulting in class mislabeling. In addition, the volume and complexity of financial data requires not only algorithm effectiveness, but also execution and retraining efficiency. We discuss some of the classic statistical techniques that have been applied, as well as more recent machine learning and data mining algorithms. Our focus is to introduce the subject and to provide a survey of broad classes of methodologies accompanied by illustrative examples of applications. When required by practical applications, we also suggest some improvements to the existing methodologies.

Statistical Challenges in Modern Biosurveillance
Galit Shmueli

Modern biosurveillance is the monitoring of a wide-range of pre-diagnostic and diagnostic data for the purpose of enhancing the ability of the public health infrastructure to detect, investigate, and respond to disease outbreaks. Statistical control charts have been a central tool in classic disease surveillance and have also migrated into modern biosurveillance. However, the new types of data monitored, the processes underlying the time series derived from these data, and the application context all deviate from the industrial setting for which these tools were originally designed. Assumptions of normality, independence, and stationarity are typically violated in syndromic time series; target values of process parameters are time-dependent and hard to define; data labeling is ambiguous in the sense that outbreak periods are not clearly defied or known. Additional challenges arise such as multiplicity in several dimensions, performance evaluation, and practical system usage and requirements. Our focus is mainly on the monitoring of time series for early alerting of anomalies to stimulate investigation of potential outbreaks, with a brief summary of methods to detect significant spatial and spatiotemporal case clusters. We discuss the different statistical challenges in monitoring modern biosurveillance data, describe the current state of monitoring in the field, and survey the most recent biosurveillance literature.

Contributed Session 8 - Reliability
Chair: Ben Haaland

Prediction Intervals for Remaining Life of Power Transformers Based on Left Truncated and Right Censored Lifetime Data
Yili Hong, William Meeker

Electrical transmission is an important part of the US energy industry. There are approximately 150,000 power transmission transformers in service in the US. Un-expected failures of transformers can cause large unnecessary economic loss. Thus, prediction of remaining life of transformers is an important issue for the owners of these assets. Lifetime data for transformers are hard to collect because transformer lifetimes extend over several decades. In the date set that we analyze in this paper, there is complete installation and failure data for unites put into service after 1980. Although the data set contains many units that were installed before 1980, there is no information about units that were installed and then failed before 1980. Hence, the data are left truncated and right censored. a parametric lifetime model is proposed to estimate the lifetime distribution of individual transformers. A statistical procedure is developed for computing a prediction interval for remaining life for individuals based an age-adjusted distribution. While these particular intervals are too wide to be directly useful, the quantitative information does provide a potentially useful ranking. These ideas are also extended to provide useful predictions and prediction intervals for the number of failures, over a range of time, for the overall population of transformers now in service.

On Proximity of Weighted and Exponential Distributions
Broderick Oluyede

Weighted distributions occur naturally in a wide variety of settings including applications in economics and actuarial finance, reliability, biometry, survival analysis, quality control and renewal theory to mention a few areas. In renewal theory, the residual lifetime has a limiting distribution that is a weighted distribution with the weight function equal to the reciprocal of the failure rate function. In importance sampling, weighted or transformed distributions are used to improve the efficiency of the Monte Carlo simulations. In this talk, stochastic relations and closure results for weighted distributions including proportional hazards models (PHM) are presented. Exponential approximations to the class of increasing failure rate (IFR) and decreasing failure rate (DFR) weighted distributions including proportional hazards models are obtained. These include approximations via equilibrium, residual life and length-biased exponential distributions. Also presented are bounds and moment-type inequalities for weighted life distributions. Some applications and examples are presented.

Monte Carlo Conditioning on a Sufficient Statistics, with Application in Reliability
Bo Lindqvist

Many statistical procedures are based on conditional distributions given sufficient statistics. The clue is that such distributions do not depend on unknown parameters. In certain cases one is able to derive the conditional distributions analytically. In general, however, it is impossible to arrive a tractable explicit expressions. The obvious solution is then to try Monte Carlo simulations. This talk is about such a simulation approach for computing conditional expectations given sufficient statistics, with emphasis on applications in reliability analyses.

Enhanced Monte Carlo Estimation of High Reliability Systems
Peter Hovey

Safely critical systems generally require high levels of reliability. For example, aircraft turbine engines must be designed to achieve an extremely high reliability. Current design strategies are focused on achieving a specific probability of failure for the engine. Traditional Monte Carlo techniques require excessive computing time because of the complexity of the finite element calculations that determine when a failure occurs and the large number of trials required to estimate a probability that is close to 1. A new method for analyzing Monte Carlo results based on extreme value theory is discussed that significantly decreases the number of simulations that are required, thus increasing computation speed.

May 20, 2008 :: Tuesday :: 15:00 to 16:30

Invited Session 11 – Analysis of Supersaturated Design
Chair: Brad Jones

Analyzing Supersaturated Designs via Model Selection Methods - Do They Work?
William Li

In the area of supersaturated designs, most attention has been given on the construction of the efficient supersaturated designs in the literature. On the analysis of supersaturated designs, while many traditional model selection methods are applicable, it has been warned in the literature that these methods should only be used with caution.
We compare several model selection methods, including all-subset, stepwise, LASSO, and sparse sliced inverse regression. The simulation results show that probability of picking true effects depends on the chose supersaturated design, the correlation structure of the design, and the magnitude of the coefficients. We conclude with the analysis of a real application of the supersaturated designs in the financial services industry.

Supersaturated Designs: Our Results Aren't Significant
Robert Mee

Whether using forward selection or all-subsets regression, it is common to select models from supersaturated designs that explain a very large percentage of the total variation in a response. The naïve p-values one sees for the selected model can persuade the user that in fact the included factors are clearly active. We show how permutation procedures may be used to more appropriately ascertain statistical significance both for the overall model and for individual coefficients. We illustrate the methods for several examples. We also show how the power for detecting an active effect decreases as the number of factors in the supersaturated design increases.

Invited Session 12 – Nanotechnology
Chair: Tirthankar Dasgupta

Sequential Minimum Energy Designs for Synthesis of Nanostructures
Tirthankar Dasgupta

The talk will discuss the development of an experimental design methodology, tailor-made to address the unique phenomena associated with nanostructure synthesis. A sequential space filling design called Sequential Minimum Energy Design (SMED) is proposed for exploring best process conditions for synthesis of nanowires. The SMED is a novel approach to generate designs that are model independent, can quickly carve out regions with no observable nanostructure morphology, allow for the exploration of complex response surfaces, and can be used for sequential experimentation. A unique feature of this technique lies in the fact that it originates from a combination of statistical theory and fundamental laws of physics. The basic idea has been developed into a practically implementable algorithm for deterministic functions, and guidelines for choosing the parameters of the design have been proposed. Performance of the algorithm has been studied using experimental data on nanowire synthesis as well as standard two dimensional and higher dimensional test functions. A modification of the algorithm based on non-parametric smoothing has been proposed for random functions.

A Statistical Approach to Quantifying the Elastic Deformation of Nanomaterials
Xinwei Deng

Accurate estimation of elastic modulus of certain nanomaterials such as Zinc Oxide nanobelt is important in many applications. A recently proposed approach was to estimate elastic modulus from a force-deflection model based on the continuous scan of a nanobelt using an Atomic Force Microscope tip at different contact forces. However, the nanobelt may have some initial bending and it may shift or deform during measurement leading to bias in the estimation. In this work we propose a statistical model to account for these various possible errors. The proposed approach can automatically detect and remove the systematic errors and therefore, can give an accurate and precise estimate of the elastic modulus. The advantages of the approach are demonstrated through the application on several data sets.

Contributed Session 9 – Statistical Computing II
Chair: Robert Parody

Sequential Quasi-Monte Carlo Sampling Construction and Applications
Ben Haaland, Peter Qian

Sequences of nested, extendable space filling designs are useful in statistics, computer science, and numerical analysis. Results for nesting and extending a class of designs with good uniformity properties, the (t,m, s)-nets and (t, s)-sequences, are presented. Applications and examples in numerical integration, design and modeling for computer experiments with multiple levels of accuracy and expense, and stochastic programming are given.

Feature Selection for Imbalanced Learning
Minghua Zhu, Zhonglin Lin, Wenjun Yin, Jin Dong

Feature selection has been an active research area in data mining and machine learning for a long time. In this paper, we propose a novel feature selection method that is designed to tackle the problem of learning from imbalanced data. Imbalanced data sets exhibit skewed class distributions in which almost all instances are belonging to one or more larger classes and far fewer instances belonging to a smaller, but usually more interesting class. The balance of training set is an underlying assumption for most learning systems, thus the performance of these systems can be influenced greatly when learn from imbalanced data. More specifically, models trained from imbalanced data sets are biased towards the majority classes and intended to ignore the minority but interested classes. Our feature selection methods are aimed to improve the prediction accuracy towards the minority class, while the existing feature selection methods are designed to improve the overall accuracy. The main advantage of our method is that we combine most of the known statistical information that have been used separately in existing feature selection methods. For each specific data set, we assign a weight value for each of the statistical information. We search for the near optimal weight value through the learning process of neural network. Extensive experiment using both real-world and artificial data demonstrated that our method is both effective and efficient.

Penalized Linear Methods for Estimation and Variable Selection in Index Models
Jian Zhang, Michael Zhu

Index models are among the most popular semi-parametric models in econometrics and statistics, mainly due to their flexibility in handling the unknown link function and resistance to the curse of dimensionality. In this work, we study the penalized linear methods for index estimation and variable selection, in particular the least squares methods with L1 and L2 penalties. These two methods are known as Lasso and ridge regression. We will show that the penalized linear methods can successfully estimate the index parameters and select the relevant variables. Theoretical results and real and simulation examples will be reported. The developed methods can be used in industrial applications where high dimensional and nonlinear data are prevalent.

A One-step Transformation Method to Generate a Random Vector
Wheyming Song, Wenchi Chiu, Dave Goldsman, Boray Huang

We present an efficient one-step transformation procedure for generating correlated random variables with any arbitrary marginal distribution. We first generate a correlated uniform vector as the base vector and then apply an inverse transformation method to generate the target vector.

May 20, 2008 :: Tuesday :: 16:30 to 17:30

Plenary Session 2
Chair: Kwok Tsui

Nanogenerators: from designed nanomaterials synthesis to energy harvesting
Zhong Lin Wang

Developing novel technologies for wireless nanodevices and nanosystems are of critical importance for in-situ, real-time and implantable biosensing, environmental monitoring, defense technology and even personal electronics. It is highly desired for wireless devices and even required for implanted biomedical devices to be self-powered without using battery. Therefore, it is essential to explore innovative nanotechnologies for converting mechanical energy (such as body movement, muscle stretching), vibration energy (such as acoustic/ultrasonic wave), and hydraulic energy (such as body fluid and blood flow) into electric energy that will be used to power nanodevices without using battery. We have demonstrated an innovative approach for converting nano-scale mechanical energy into electric energy by piezoelectric zinc oxide nanowire (NW) arrays. By deflecting the aligned NWs using an array of conductive atomic force microscopy (AFM) tips in contact mode, the energy that was first created by the deflection force and later converted into electricity by piezoelectric effect has been measured for demonstrating nano-scale power generator. The operation mechanism of the electric generator relies on the unique coupling of piezoelectric and semiconducting dual properties of ZnO as well as the elegant rectifying function of the Schottky barrier formed between the metal tip and the NW. Based on this mechanism, we have recently developed DC nanogenerator driven by ultrasonic wave in bio-fluid, which is a gigantic step towards applications in practice. This presentation will introduce the basic principle of the nanogenerator and the key role played by designed growth of NW arrays for determining the performance of the nanogenerators.

May 20, 2008 :: Tuesday :: 19:00 to 21:30

Conference Banquet
Chair: Paul Kwam

Statistical Reflections from the Downhill Side of Hubbert's Peak
Max Morris

May 21, 2008 :: Wednesday :: 8:30 to 10:00

Invited Session 13 – Traditional DOE
Chair: Hongquan Xu

Multi-stratum Fractional Factorial Designs
Ching-Shui Cheng

There has been a lot of recent work on industrial experiments with multiple error terms including, e.g., split-plot designs, blocked split-plot designs, strip-plot designs, etc. In this talk, I will present a general and unified approach to the selection and construction of such designs.

On determining support points of locally optimal designs for nonlinear models
John Stufken

We develop new tools for identifying the support points of locally optimal designs for nonlinear models. In contrast to the commonly used geometric approach, we use an approach based on algebraic tools. The general results can be applied to often used models, including logistic, probit, double exponential and double reciprocal models for binary data, a loglinear Poisson regression model for count data, and the Michaelis-Menten model. The approach works both with constrained and unconstrained design regions and is relatively easy to implement. This is joint work with Min Yang, University of Missouri-Columbia.

Existence and construction of two-level orthogonal arrays for estimating main effects and some specified two-factor interactions
Boxin Tang

This paper considers two-level orthogonal arrays that allow joint estimation of all main effects and a set of prespecified two-factor interactions. We obtain some theoretical results, which provide a simple characterization of when such designs exist and how to construct them if they do exist. General as well as concrete applications of the results are discussed.

Contributed Session 10 – Design of Experiments IV
Chair: Broderick Oluyede

Nonparametric Bootstrap Inference on the Improvement in a Response Surface
Robert Parody

Quadratic response surface methodology often focuses on finding the levels of some (coded) predictor variables $\mathbf{x} = (x_1, x_1, \ldots, x_k)$ that optimize the expected value of a response variable y. Often the experimenter starts from some best guess or "control" combination of the predictor variables (usually coded to $\mathbf{x}=\mathbf{0}$) and performs an experiment varying them in a region around this center point. Parody and Edwards (2007) simulated a critical point to provide simultaneous confidence intervals for the improvement in E(y) at **x** as compared to **0**, $\delta(\mathbf{x}) = E(y \mid \mathbf{x}) - E(y \mid \mathbf{0})$ for all **x** within a specified distance of **0**. This technique holds for only 2nd-order rotatable designs and under the normal model assumptions. The approach given here will generalize this technique for use with all designs using a nonparametric bootstrap based on a pivot to create an approximate simulated critical point to provide these confidence intervals. The only assumption on the errors is that they are independent with mean vector is zero and covariance matrix $\sigma^2\mathbf{I}$. This still allows for the use of least squares to estimate the regression parameters. An example is presented illustrating this approach. A simulation study is presented to check coverage probabilities.

Robust Confidence Intervals for the Bernoulli Parameter
Chia-Jung Chang, Wheyming Song

Despite the simplicity of the Bernoulli process, developing good confidence-interval procedures for its parameter---the probability of success p---is deceptively difficult. The binary data yield a discrete number of successes from a discrete number of trials, n. This discreteness results in actual coverage probabilities that oscillate with the n for fixed values of p (and with p for fixed n). After reviewing and expanding upon some known confidence-interval procedures and a family of point estimators for p, we discuss ideas for developing an adjusted standard interval, including ways to improve choice for the Student-t distribution's degrees of freedom and ways to improve the standard-error estimator. The procedures we propose are combinations of our proposed adjusted standard interval and Wilson's procedure. The proposed procedures improve over their competitors in terms of the small amplitudes of oscillation of the coverage probability and the robustness of the sample size required to guarantee that the coverage probabilities stay above 0.92 for any $0.01 < p < 0.99$, without increasing either the expected interval width or the standard deviation of the interval width.

Discriminant Analysis with Interval Constriants (DAIC)
Heeyoung Kim

Efficient processing of high-dimensional data streams in manufacturing calls upon discriminant analysis with interval constraints. A statistical model is introduced in this paper. We propose an estimator which is consistent, can be efficiently computed, and provide satisfactory results on real data sets. We name our method discriminant analysis with interval constraints (DAIC) and demonstrate its application in tonnage data analysis for stamping processes.

The Hybrid Biased-Bootstrap
Mihai Giurcanu

In this research we propose a hybrid non-parametric bootstrap procedure for problems where the “naive" uniform bootstrap fails. This procedure is defined as a natural extension of a hybrid parametric bootstrap procedure to non-parametric problems through the biased-bootstrap. The main idea of our method is to first construct a pseudoparametric family of weighted empirical distributions associated with the model. Then, a generic biased-bootstrap resample is obtained by re-sampling from a weighted empirical distribution that corresponds to a particular estimator of the parameter. We apply this procedure to three problems and show that the proposed hybrid biased-bootstrap procedure provides consistent (in probability) distribution estimates indeed. In a small simulation study, we illustrate the performance of the proposed methods which confirm the theoretical results.

May 21, 2008 :: Wednesday :: 10:30 to 11:30

Plenary Session 3
Chair: Jeff Wu

Quantified Margins of Uncertainty: Error bars for Simulation Based Design
James Glimm

For a number of reasons, advanced engineering design is often based, in the first instance, on simulation. The Edisonian method of build and test is losing in a contest of practicality, some times due to time to market pressures, sometimes due to cost issues, and sometimes due to the infeasibility of experimental design methods. A clear sign of the importance given to simulation based design decisions is the increasing importance assigned to simulation error bars. Campaigns for verification and validation (V&V) are now commonplace, and test for whether the mathematical equations have been solved correctly numerically (verification) and whether the mathematical equations accurately describe the physical problem to be solved (validation). Following the V&V campaigns is an uncertainty quantification (UQ) assessment, which attempts to determine error bars for simulations, as due to numerical approximations, physical modeling approximations, or limited or inaccurate data. Finally, there is a concern with quantified margins of uncertainty, which are a translation of the traditional engineering safety margins into this modern simulation based technology. These attempt to assure design robustness in terms of a distance between a "safe" design point and possible failed design points, with allowances for all above mentioned uncertainties and errors. Plainly, these issues raise many statistical issues, in addition to the obvious physical, numerical, and engineering ones. Design of experiments, principal component analysis, Monte Carlo methods, Bayesian methods and ANOVA come to mind. Stochastic methods, such as stochastic partial differential equations, arise naturally. Sometimes, standard methods are sufficient, but for difficult problems it is very likely that extensions and even far reaching extensions will be required. In this talk we will present the statistical concepts which underlie modern simulation error analysis, and we will illustrate some of the difficulties by drawing on the presenter's experience in dealing with these issues. As a member of a scram jet design team in collaboration with the Stanford Turbulence Center, we will draw on examples, such as for turbulent combustion, where new ideas from statistics may be needed.

Alphabetical Index of Participants

W

X

Y

Z