2009 Spring Research Conference on Statistics in Industry and Technology

May 27 - 29, 2009

Executive Plaza Hotel
Coquitlam, BC, Canada

Contents

Sponsors

Spring Research Conference 2009 acknowledges the generous supports received from the following sponsors (in alphabetical order).

- American Statistical Association/Section on Physical and Engineering Sciences (ASA/SPES)
- Institute of Mathematical Statistics (IMS)
- JMP
- Mathematics of Information Technology and Complex Systems (MITACS)
- Pacific Institute for the Mathematical Sciences (PIMS)
- SAS Institute, Inc.
- Simon Fraser University (SFU)

Committees and Chairs

SRC Management Committee

Chair: Andrew J. Booker, *The Boeing Company*

Members: George Michailidis, *University of Michigan*
Peter Z.G. Qian, *University of Wisconsin-Madison*
Shane Reese, *Brigham Young University*
Jan Shi, *University of Michigan*
Vengazhiyil Roshan Joseph, *Georgia Institute of Technology*

SRC 2009 Organizing Committee

Chair: Boxin Tang, *Simon Fraser University*

Contributed Program Chair: Hongquan Xu, *UCLA*

Members: Derek Bingham, *Simon Fraser University*
Jiahua, Chen, *University of British Columbia*
Thomas M. Loughin, *Simon Fraser University*
William J. Welch, *University of British Columbia*

SRC 2009 Session Organizers and Chairs

Derek Bingham, *Simon Fraser University*
Jiahua, Chen, *University of British Columbia*
Joan Hu, *Simon Fraser University*
Sue Lewis, *University of Southampton*
Runze Li, *Pennsylvania State University*
Yufeng Liu, *University of North Carolina at Chapel Hill*
Jason L. Loeppky, *UBC-Okanagan*
Thomas M. Loughin, *Simon Fraser University*
Abhyuday Mandal, *University of Georgia*
Peter Z.G. Qian, *University of Wisconsin-Madison*
Jonathan Rougier, *University of Bristol*
David Steinberg, *Tel Aviv University*
Kenny Ye, *Albert Einstein College*
Xiangrong Yin, *University of Georgia*
William J. Welch, *University of British Columbia*

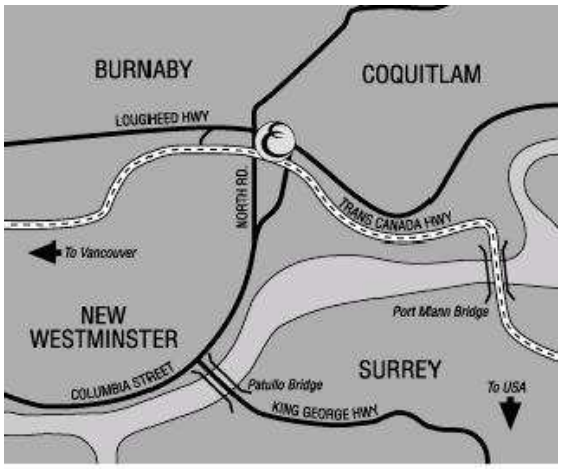
General Information

The conference venue is Executive Plaza Hotel & Conference Centre Coquitlam. In addition to its location and direction, this section will provide some general information including registration, transportation and parking, etc.

Conference Venue

The Executive Plaza Hotel Coquitlam is situated at the corner of Lougheed and North Road and is within walking distance to Lougheed Mall, shopping, dining, entertainment and recreation, and 25 minutes from Downtown Vancouver.

405 North Road
Coquitlam, B.C. V3K 3V9
Tel.: (604) 936-9399 — Fax: (604) 937-4577
Toll Free: 1-888-433-EXEC (3932)



Registration

Registration will be available in Room Aspen Birch during the Reception on Tuesday May 26 (5:30pm - 8:00pm). The registration desk will stay open in the Foyer of the Ballroom, from 8:00am to 5:00pm on Wednesday May 27 and from 9:00am to 5:00pm on Thursday May 28.

Transportation

The conference venue is easily accessible via the city skytrain. Its nearest station is Lougheed Town Center station, the red solid dot in the map of the skytrain below. For those who wish to visit Simon Fraser University, the bus route 145 in the Production Way-University station (one station west of the Lougheed station) can take you directly to the main campus. Skytrain is part of Vancouver's public transit system, Translink. The skytrain ticket is only good for 90 minutes after you purchase it. More detailed information such as fare pricing, fare zones, schedule and daypass is available on the website http://www.translink.bc.ca/.

Welcome Reception

The welcome reception will be held at Room Aspen Birch on Tuesday, May 26 from 5:30pm to 8pm. All conference attendees and companions are warmly invited and enjoy a social time with old and new friends.

Banquet

Banquet will be in Ballroom A-B on Thursday May 28 from 7:00pm to 9:00pm.

List of Sessions

Wednesday, May 27

09:00 - 10:00

Ballroom B-C	Plenary Session 1	Statistical Work in Nanomaterial Research

10:30 - 12:00

Ballroom A	Invited Session 1	Experimental Design
Ballroom B	Invited Session 2	Modern Statistical Machine Learning
Ballroom C	Contributed Session 1	Imputation, Estimation and Sampling Methods

13:30 - 15:00

Ballroom A	Invited Session 3	Fractional Factorial Designs
Ballroom B	Invited Session 4	Geosciences and Computer Models
Ballroom C	Contributed Session 2	Analysis of Counts, Degradation, Lifetime Data and Nonlinear Models

15:30 - 17:00

Ballroom A	Invited Session 5	The Randy Sitter Technometrics Session
Ballroom B	Contributed Session 3	Reliability and Quality

Thursday, May 28

09:00 - 10:00

Ballroom B-C	Plenary Session 2	The What, Why, and How of Multivariate Emulation

10:30 - 12:00

Ballroom A	Invited Session 6	Field Reliability: Analyses of Warranty Data
Ballroom B	Invited Session 7	Industrial Statistics
Ballroom C	Contributed Session 4	Fractional Factorial Designs

13:30 - 15:00

Ballroom A	Invited Session 8	Analysis of Industrial Experiments
Ballroom B	Invited Session 9	Drug Discovery
Ballroom C	Contributed Session 5	Modeling and Design of Experiments

15:30 - 17:00

Ballroom A	Invited Session 10	Computer Experiments
Ballroom B	Contributed Session 6	Optimal, Robust and Repeated Measurement Designs
Ballroom C	Contributed Session 7	Analysis of Spatial and Temporal Data

Friday, May 29

08:30 - 10:00

Ballroom A	Invited Session 11	Problems in Sampling Survey Practice
Ballroom B	Invited Session 12	Data Mining: dimension reduction and variable selection
Ballroom C	Contributed Session 8	Analysis of Computer Experiments and Variable Selection

10:30 - 12:00

Ballroom A	Invited Session 13	Design Theory
Ballroom B	Contributed Session 9	Statistical Methods in Networks, Finance and Pharmaceutical Industry

Conference Program Schedule

Tuesday, May 26, Day 0	
Time	**Room Aspen Birch**
17:30 - 20:00	Welcome Reception and Registration

Wednesday, May 27, Day 1			
Time	**Ballroom A**	**Ballroom B**	**Ballroom C**
08:30 - 09:00		Welcome Remarks	
09:00 - 10:00		Plenary Session 1 C.F. Jeff Wu	
10:00 - 10:30	Coffee Break		
10:30 - 12:00	Invited Session 1 Experimental Design	Invited Session 2 Modern Statistical Machine Learning	Contributed Session 1
12:00 - 13:30	Lunch Break		
13:30 - 15:00	Invited Session 3 Fractional Factorial Designs	Invited Session 4 Geosciences and Computer Models	Contributed Session 2
15:00 - 15:30	Coffee Break		
15:30 - 17:00	Invited Session 5 The Randy Sitter Technometrics Session	Contributed Session 3	

Wednesday, May 27, 8:30am - 9:00am

Welcome Remarks

Boxin Tang, Andrew J. Booker, Richard Lockhart

Wednesday, May 27, 9:00am - 10:00am

Plenary Session 1 - Chair: Boxin Tang

Statistical Work in Nanomaterial Research
C.F. Jeff Wu

Wednesday, May 27, 10:30am - 12:00pm

Invited Session 1: Experimental Design - Chair: Sue Lewis

Bin-based Sequential Design for Computer Experiments
Leslie M. Moore, Jason L. Loeppky & Brian J. Williams*

Optimal Designs for Multi-variable Spline Models
Dave Woods

Gaussian Process Models for Computer Experiments With Qualitative and Quantitative Factors
Huaiqing Wu , Peter Z.G. Qian & C.F. Jeff Wu*

Invited Session 2: Modern Statistical Machine Learning
Organizer: Yufeng Liu
Chair: Runze Li

Large Margin Hierarchical Classification
Xiaotong Shen, Huixing Wang & Wei Pan*

Robust Model-free Multiclass Probability Estimation
Hao Helen Zhang, Yichao Wu & Yufeng Liu*

Partial Correlation Estimation by Joint Sparse Regression Models
Ji Zhu, Jie Peng, Pei Wang & Nengfeng Zhou*

Contributed Session 1: Imputation, Estimation and Sampling Methods
Chair: Wilson Lu

Random Group Variance Estimators for Survey Data with Random Hot Deck Imputation
Qi Tang & Jun Shao*

Maximum Likelihood Estimation under a Hypothesis of Marginal Homogeneity for Data in a Square Contingency Table
Saumen Mandal & Ben Torsney*

Second Order Exchangeable Emulators to Assess Initial Condition Uncertainty
Leanna House, Michael Goldstein & Ian Vernon*

Accelerated Life Test Sampling Plans for the Weibull Distribution under Type I Progressive Interval Censoring with Random Removals
Chunyan Yang, Chang Ding & Siu-Keung Tse*

Wednesday, May 27, 1:30pm - 3:00pm

Invited Session 3: Fractional Factorial Designs - Chair: Jason L. Loeppky

D-optimal Designs with Minimum Aliasing
Bradley Jones & Chris Nachtsheim*

Optimal Design in High Throughput Screening
Xianggui Harvey Qu

D-optimal Two-level Orthogonal Arrays for Estimating Main Effects and Some Two-factor Interactions
Julie Zhou & Boxin Tang*

Invited Session 4: Geosciences and Computer Models
Organizer and Chair: Derek Bingham

Quantile Estimation in Climate Models
James Gattiker

Calibration and Prediction Problems in Catchment Scale Hydrology
Cari Kaufman

A Spatial Analysis of Regional Climate Model Ensembles
Stephan R. Sain

Contributed Session 2: Analysis of Counts, Degradation, Lifetime Data and Nonlinear Models - Chair: Andrew J. Booker

On Departure from Exponentiality When Estimating the Number of Component Processes
Subrata Kundu, Anup Dewanji & Tapan K. Nayak*

Inference with Censored Degradation Data
Yang Yang & Vijay Nair*

Parametric Inference for System Lifetime Data with Signatures Available
Hon Keung Tony Ng, J. Navarro & N. Balakrishnan*

Analysis of Transform-Both-Sides Nonlinear Regression Models
A.H.M. Mahbub Latif & Steven G. Gilmour*

Wednesday, May 27, 3:30pm - 5:00pm

Invited Session 5: The Randy Sitter Technometrics Session

Organizer: David Steinberg

Chair: Jonathan Rougier

Bayesian Design of Experiments for Nonlinear Models
Chris Gotwalt

Design and Analysis of Computer Experiments with Branching and Nested Factors
Ying Hung

Contributed Session 3: Reliability and Quality
Chair: Gemai CHEN

Data-Driven Reliability Estimation for Prognostics and Health Management
Aparna V. Huzurbazar

R(reliability)-Mark in Korea
Jaiwook Baik

Bias Correction When Using the Extreme Value approximation in Monte Carlo Estimation of High Reliability Systems
Peter W. Hovey & Brian Krilov*

Focus on Failure Avoidance and Risk Reduction through Variation Mode and Effect Analysis
Stefano Barone, Bo Bergman, Alexander Chakhunashvili & Per Johansson*

Thursday, May 28, Day 2			
Time	**Ballroom A**	**Ballroom B**	**Ballroom C**
09:00 - 10:00		Plenary Session 2 Jonathan ROUGIER	
10:00 - 10:30	Coffee Break		
10:30 - 12:00	Invited Session 6 Field Reliability: Analyses of Warranty Data	Invited Session 7 Industrial Statistics	Contributed Session 4
12:00 - 13:30	Lunch Break		
13:30 - 15:00	Invited Session 8 Analysis of Industrial Experiments	Invited Session 9 Drug Discovery	Contributed Session 5
15:00 - 15:30	Coffee Break		
15:30 - 17:00	Invited Session 10 Computer Experiments	Contributed Session 6	Contributed Session 7
19:00 - 21:00	Banquet Thomas M. LOUGHIN		

Thursday, May 28, 9:00am - 10:00am

Plenary Session 2 - Chair: Derek BINGHAM

The What, Why, and How of Multivariate Emulation
Jonathan Rougier

Thursday, May 28, 10:30am - 12:00pm

Invited Session 6: Field Reliability: Analyses of Warranty Data
Organizer and Chair: Joan Hu

Getting More from Warranty Data
Jeff Robinson

Estimation of Lifetime Parameters and Its Properties Using only Failure Information from Warranty Database
Kazuyuki Suzuki

When Can Relative Age and Usage Effects Be Ascertained from Warranty Data?
Jerry Lawless

Invited Session 7: Industrial Statistics - Chair: Peter Z.G. Qian

Designed Experiments with Limited Combinations of Factor Combinations
Christine M. Anderson-Cook & Timothy J. Robinson*

On t and EWMA t Charts for Monitoring Changes in the Process Mean
Gemai Chen

Impact of Fatigue Effect on Experimental Designs in Internet Based Discrete Choice Studies
William Li

Contributed Session 4: Fractional Factorial Designs
Chair: Arden Miller

Optimal Selection of Blocked Two-Level Fractional Factorial Designs
Weiming Ke

Creating Catalogs of Two-level Fractional Factorial Designs Based on the Criteria of Generalized Aberration
C. Devon Lin, Randy R. Sitter & Boxin Tang*

Semifolding Two-level Fractional Factorial Split-Plot Designs
Robert McLeod

Forms of Four-Word Indicator Functions with Implications to Two-Level Factorial Designs
Po Yang

Thursday, May 28, 1:30pm - 3:00pm

Invited Session 8: Analysis of Industrial Experiments
Organizer and Chair: Thomas M. Loughin

Taking a Closer Look at Optimal Screening Designs
Timothy J. Robinson & Christine M. Anderson-Cook*

Joint Model Selection of Location and Dispersion Effects in Unreplicated Factorial Experiments
Jorge E. Rodriguez & Thomas M. Loughin*

Lack-of-Fit Tests for Split-plot Response Surface Designs
Geoffrey G. Vining

Invited Session 9: Drug Discovery
Organizer and Chair: William J. Welch

ChemModLab: QSAR modeling and model-based searching
*Jacqueline M. Hughes-Oliver**, *Qianyi Zhang & Raymond Ng*

G-SELC: Optimization by Sequential Elimination of Level Combinations using Genetic Algorithms and Gaussian Processes
*Abhyuday Mandal**, *Pritam Ranjan & C.F. Jeff Wu*

Image Analysis and Structure/localization Relationships for Small Molecule Probes in Live Cells
Kerby Shedden & Fangyi Liu*

Contributed Session 5: Modeling and Design of Experiments
Chair: C. Devon Lin

Mixture-of-Mixture Experiments: A general modeling and design approach
*Lulu Kang**, *Vengazhiyil Roshan Joseph & William A. Brenneman*

Design of Experiments for Packet Communication Networks
Ben M. Parker

A Decision Theory Approach to the Choice of Factorial Design
Deepa Kapoor

Stochastic Optimization with Space-Filling Designs
Peter Z.G. Qian

Thursday, May 28, 3:30pm - 5:00pm

Invited Session 10: Computer Experiments
Organizer and Chair: Abhyuday Mandal

Optimization of a Laser-Assisted Micro-Grooving Process Using an Engineering-Statistical Approach
*Vengazhiyil Roshan Joseph**, *Ramesh K. Singh & Shreyes N. Melkote*

Deterministic Computer Models and GP as an Interpolator
Pritam Ranjan & Ronald Haynes*

Approximating a Computer Code via a Gaussian Process: A Practical Guide
*William J. Welch**, *Jason L. Loeppky & Jerome Sacks*

Contributed Session 6: Optimal, Robust and Repeated Measurement Designs - Chair: Dave Woods

A D-optimal Design for a Generalized Exponential Model Governing Growth of Nanowires
Li Zhu

Applications and Implementations of Continuous Robust Designs
Wai Kong Yuen & Xiaojian Xu*

Empirical Likelihood Based Control Charts
*Asokan Mulayath Variyath**

Repeated Measurement Designs under Subject Dropout
Shi Zhao & Dibyen Majumdar*

Contributed Session 7: Analysis of Spatial and Temporal Data
Chair: Huge Chipman

Large Gaussian Covariance Matrix Estimation with Markov Structures
Xinwei Deng

Nonparametric Estimation of Variogram and its Spectrum
Chunfeng Huang

Comparing Generalized Quasi-likelihood Approaches to Longitudinally Correlated Simulated Poisson Counts
N. Mamode Khan & V. Jowaheer*

Using Markov Chain Monte Carlo to Solve a Time-varying State-space Model for Magnetoencephalography Inverse Problem
Zhigang Yao

Thursday, May 28, 7:00pm - 9:00pm

Banquet

Thomas M. Loughin

Friday, May 29, Day 3			
Time	**Ballroom A**	**Ballroom B**	**Ballroom C**
08:30 - 10:00	Invited Session 11 Problems in Sampling Survey Practice	Invited Session 12 Data Mining: dimension reduction & variable selection	Contributed Session 8
10:00 - 10:30	Coffee Break		
10:30 - 12:00	Invited Session 13 Design Theory	Contributed Session 9	

Friday, May 29, 8:30am - 10:00am

Invited Session 11: Problems in Sampling Survey Practice
Organizer and Chair: Jiahua, Chen

Asymptotic Calibration Estimation Using Estimated Importance Sampling and Exponential Tilting
Jae-kwang Kim

Application of Experimental Designs in Survey Sampling
J.N.K. Rao

Efficient and Robust Model-Assisted Imputation for Covariate-Dependent Missing Responses
Jun Shao

Invited Session 12: Data Mining: dimension reduction and variable selection - Chair: Xiangrong Yin

On Dimension Folding of Matrix or Array Valued Statistical Objects
*Bing Li**, *Min Kyung Kim* & *Naomi Altman*

Regularization Parameter Selections via Generalized Information Criterion
Runze Li

Fast Simultaneous Dimension Reduction and Clustering: Viewing Data from Extremes
George Ostrouchov

Oracular Model Selection via Penalized Composite Quantile Regression
Hui Zou

Contributed Session 8: Analysis of Computer Experiments and Variable Selection - Chair: Pritam Ranjan

Fast Calibration of Complex Computer Models
Matthew T. Pratola

Kriging Analysis of Thermal Data for a Data-center Experiment
*Huizhi Xie**, *Lulu Kang*, *C.F. Jeff Wu*, *Emad Samadiani* & *Yogendra Joshi*

Reanalysis of Some Experiments via the Dantzig Selector with the Consideration of Interactions
Frederick K.H. Phoa

Sparse Covariance Thresholding for High-Dimensional Variable Selection with the Lasso
*Michael Yu Zhu**, *Jessie Jeng* & *John Daye*

Friday, May 29, 10:30am - 12:00pm

Invited Session 13: Design Theory - Chair: Kenny Ye

MDS-optimal Supersaturated Designs
Arden Miller

On Locally Optimal Designs for Generalized Linear Models with Group Effects
*John Stufken** & *Min Yang*

Two-Level Nonregular Designs Constructed via Quaternary Codes
Hongquan Xu

Contributed Session 9: Statistical Methods in Networks, Finance and Pharmaceutical Industry - Chair: Rob Easterling

Statistical Learning for Networks
Hugh Chipman

Computational and Statistical Aspects of Pricing Models
*Gracie-Xiaojuan Ma** & *Sergey Utev*

Stepwise confidence interval method for Identification of the Minimum Effective Dose
*Xiaoqing Tang** & *Jian Tao*

Statistical Aspects of Intermediate Precision Studies
*Balazs Banfai**, *Katalin Ganzler* & *Sandor Kemeny*

Abstracts of Plenary Talks

Plenary Session 1 Wednesday, May 27, 9:00am - 10:00am

Statistical Work in Nanomaterial Research

C.F. Jeff Wu, *Georgia Institute of Technology*

My talk consists of two parts. The first part concerns the accurate estimation of elastic modulus of certain nanomaterials (e.g., Zinc Oxide nanobelt), which is important in some applications. A recently proposed approach in the materials literature 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, which lead to bias in the estimation. In this work (joint with Deng, Mai, Joseph, Wang, to appear in Proc Nat Acad Sci), we propose a statistical model to account for these various 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 applications on real data. In the second part of the talk I 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 is 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 are proposed. Performance of the algorithm has been studied using experimental data on nanowire synthesis as well as standard test functions. A modification of the algorithm based on non-parametric fitting is proposed for random functions. (The second work is joint with Joseph and Dasgupta.)

Plenary Session 2 Thursday, May 28, 9:00am - 10:00am

The What, Why, and How of Multivariate Emulation

Jonathan Rougier, *University of Bristol*

Complex systems like climate are studied in computer experiments, in which large system-models are integrated forwards under specified boundary conditions to make predictions. Impressive though these models can be (and for climate they run on some of the largest computers in the world), their limitations introduce substantial uncertainty about actual behaviour, and quantifying and reducing this uncertainty is now a major research focus. The difficulty is that is is impractical to embed these models 'inside the loop' of a statistical calculation, which is what we would like to do to account for their parametric uncertainty. Instead, we create a stochastic representation of the model, termed an 'emulator', from a collection of model-runs, and then the emulator takes the place of the model inside the loop. Emulating a model (a complicated deterministic function) is challenging; particularly so with a large number of outputs, because of the quantity of relationships that must be accounted for simultaneously, which is the number of model-runs times number of model-outputs. A recent conjugate Bayesian approach, known as an Outer Product Emulator (OPE), provides a solution. The OPE is illustrated for a large-scale atmospheric model with functional output (TIE-GCM).

Abstracts of Invited Talks

Invited Session 1 — Wednesday, May 27, 10:30am - 12:00pm

Experimental Design

Bin-based Sequential Design for Computer Experiments

Leslie M. Moore*, *Los Alamos National Laboratory*, Jason L. Loeppky, *UBC-Okanagan* & Brian J. Williams, *Los Alamos National Laboratory*

An approach to adding sequential sets of simulation runs to improve emulator performance is described that extends the conceptual properties of orthogonal array based Latin hypercube sampling (OA-based LHS), in common use to select the initial set of simulation runs used for emulator construction. Computer models are used to gain insight to physical phenomena, augmenting information from or even replacing physical experiments in some situations. Deterministic computer models often are slow to run so use of a statistical emulator in place of the code is appealing. However, when limited data is available from an initial computer experiment, the opportunity to obtain additional simulation runs may help achieve stability in the predictions of the emulator. Distance based criteria have been used with much success in selection of an initial design for emulator construction, often incorporating an underlying structure such as an orthogonal array. Using the maximin distance criteria with a binning requirement for a batch of sequential points to be added in a spirit similar to OA-based LHS is proposed. Whether or not the initial design is an OA-based LHS, the values of the inputs can be binned into a specified number of groups such that the binned design is a subset of a factorial design. In the case of an OA-based LHS, the initial binned design is actually an orthogonal array. A sequential batch of design points is required to bin in such a way as to complement the initial binned design. Further, in the spirit of Latin hypercube sampling, the sequential batch of points are selected to fill in values in unrepresented strata of the marginal inputs defined by the number of runs. The resulting sequential sets of simulation runs have specified binning structure and dense values of marginal inputs. Additional modifications to this approach are suggested and open issues posed.

Optimal Designs for Multi-variable Spline Models

Dave Woods, *University of Southampton*

Many experiments in science and industry measure a continuous response that cannot be adequately described by a low-order polynomial regression model. For example, the response may exhibit multiple optima, or non-smooth or localised behaviour. In such circumstances, regression models that are constructed from a polynomial spline basis may prove effective. A spline is a piecewise polynomial with the pieces constrained to be continuous and usually smooth at the breakpoints or knots. Such models are inherently more flexible than single low-order polynomials while, for a given set of knots, retaining the model structure and interpretation of regression models. They have been successfully applied in a wide variety of application areas, including the automotive industry, chemistry and computer experiments.

In this talk, we consider the problem of finding optimal designs for experiments where are there two or more variables whose effects on the response are described by polynomial splines. When the location of the knots is unknown and must be estimated from the data, a nonlinear design problem results. We present theoretical results for the construction of Bayesian and maximin designs for additive spline models. For many applications, factors may interact. To accommodate this, designs are found numerically for models constructed from subsets of terms from the tensor product basis. The methods are illustrated throughout by examples.

Gaussian Process Models for Computer Experiments With Qualitative and Quantitative Factors

Huaiqing Wu*, *Iowa State University*, Peter Z.G. Qian, *University of Wisconsin-Madison* & C.F. Jeff Wu, *Georgia Institute of Technology*

Modeling experiments with qualitative and quantitative factors is an important issue in computer modeling. We propose a framework for building Gaussian process models that incorporate both types of factors. The key to the development of these new models is an approach for constructing correlation functions with qualitative and quantitative factors. An iterative estimation procedure is developed for the proposed models. Modern optimization techniques are used in the estimation to ensure the validity of the constructed correlation functions. The proposed method is illustrated with an example involving a known function and a real example for modeling the thermal distribution of a data center.

Invited Session 2 Wednesday, May 27, 10:30am - 12:00pm

Modern Statistical Machine Learning

Large Margin Hierarchical Classification

Xiaotong Shen*, Huixing Wang & Wei Pan, *University of Minnesota*

In this talk, I will present our recent research on hierarchical classification. In hierarchical

classification, class label is structured in that each label value corresponds to one non-root node in a tree, where the inter-class relationship is specified by directed paths of the tree. To leverage the inter-class relationship for enhancing the performance of flat classification, we propose a large margin method for single-path hierarchical classification, where only one path is permitted from the root to one node. This is based on a new concept of generalized margins with respect to a given hierarchy. Implementation of support vector machines and psi-learning will be discussed, in addition to numerical and theoretical examples.

Robust Model-free Multiclass Probability Estimation

Hao Helen Zhang*, Yichao Wu, *North Carolina State University* & Yufeng Liu, *University of North Carolina at Chapel Hill*

Classical approaches for multiclass probability estimation are typically based on regression techniques such as multiple logistic regression, or density estimation approaches such as LDA and QDA. These methods often make certain assumptions on the probability functions or on the underlying distributions of each subclasse. We propose a model-free procedure to estimate multiclass probabilities based on large-margin classifiers. The new estimation scheme is employed by solving a series of weighted large-margin classifiers and then systematically extracting the probability information from these multiple classification rules. A main advantage of the proposed probability estimation technique is that it does not impose any strong parametric assumption on the underlying distribution and can be applied for a wide range of large-margin classification methods. A general computational algorithm is developed for class probability estimation. Furthermore, we establish asymptotic consistency of the probability estimates. Both simulated and real data examples are presented to illustrate performance of the new procedure.

Partial Correlation Estimation by Joint Sparse Regression Models

Ji Zhu*, *University of Michigan*, Jie Peng, *University of California, Davis*, Pei Wang, *Fred Hutchinson Cancer Research Center* & Nengfeng Zhou, *University of Michigan*

In this talk, we propose a computationally efficient approach for selecting non-zero partial correlations under the high-dimension-low-sample-size setting. This method assumes the overall sparsity of the partial correlation matrix and employs sparse regression techniques for model fitting. We illustrate the performance of our method by extensive simulation studies. It is shown that our method performs well in both non-zero partial correlation selection and the identification of hub variables, and also outperforms two existing methods. We then apply our method to a microarray breast cancer data set and identify a set of "hub genes" which may provide important insights on genetic regulatory networks. Finally, we prove that, under a set of suitable assumptions, the proposed procedure is asymptotically consistent in terms of model selection and parameter estimation.

Invited Session 3 Wednesday, May 27, 1:30pm - 3:00pm

Fractional Factorial Designs

D-optimal Designs with Minimum Aliasing

Bradley Jones*, *JMP* & Chris Nachtsheim, *University of Minnesota*

In most experimentation the form of the model is not known in advance of experimentation. Generally, the researcher hopes that a simple model will be adequate but it may turn out that some higher order terms are necessary to fit the data. For example, in resolution III screening designs the researcher is betting that the main effects model will be explanatory. Yet, if there are active two-factor interactions, this usually leads to ambiguous results.

Optimal design algorithms require the specification of an a priori model. For many textbook design scenarios there are multiple nonisomorphic d-optimal designs. In such cases it is desirable to add a secondary criterion to further discriminate between competitive designs. We suggest a new optimality criterion and show through examples that it generates d-efficient designs with minimal aliasing with respect to some specified aliasing terms.

Optimal Design in High Throughput Screening

Xianggui Harvey Qu, *Oakland University*

High-throughput screening (HTS) is a large-scale process that screens hundreds of thousands to millions of compounds in order to identify potentially leading candidates rapidly and accurately. There are many statistically challenging issues in HTS. In this talk, I will focus the spatial effect in primary HTS. I will discuss the consequences of spatial effects in selecting leading compounds and why the current experimental design fails to eliminate these spatial effects. A new class of designs will be proposed for elimination of spatial effects. The new designs have the advantages such as all compounds are comparable within each microplate in spite of the existence of spatial effects; the maximum number of compounds in each microplate is attained, etc. Optimal designs are recommended for HTS experiments with multiple controls.

D-optimal Two-level Orthogonal Arrays for Estimating Main Effects and Some Two-factor Interactions

Julie Zhou*, *University of Victoria* & Boxin Tang, *Simon Fraser University*

Two-level orthogonal arrays (designs) are considered to jointly estimate all the main effects and some specified two-factor interactions. First we discuss the existence and construction of such designs. Theoretical results are obtained to provide a simple characterization of when such designs exist and how to construct them. Then we explore D-optimality of those designs. Interesting results are derived to find D-optimal designs. Examples will be given to illustrate the results.

Invited Session 4 Wednesday, May 27, 1:30pm - 3:00pm

Geosciences and Computer Models

Quantile Estimation in Climate Models

James Gattiker, *Los Alamos National Laboratory*

Recent work in computer models emphasizes the use of ensembles over uncertain parameters, and construction of emulators that can be used for detailed analysis of the model response. One of these analysis tasks is the assessing of the probability of events in the modeled system, e.g., the probability that the model response is below some value. Another goal is the related task of quantile estimation, i.e., what is a threshold such that there is a specified probability of exceeding the threshold. New work in this estimation task is demonstrated with an ensemble from an intermediate complexity model. The goal is the estimation of modeled future strength of the Meridional Overturning Circulation.

Calibration and Prediction Problems in Catchment Scale Hydrology

Cari Kaufman, *University of California, Berkeley*

Statistical models in environmental applications are making increasing use of the scientific knowledge represented in systems of partial differential equations describing the evolution of environmental processes over time. I will discuss statistical methodology for simultaneously 1) inferring unknown model parameters, and 2) predicting the behavior of the process. I will focus on modelling soil moisture fields under various climatic conditions, using a hierarchical model in which soil moisture is the process of interest. Its behavior is governed by a catchment-level hydrology model, and the observations are modeled according to a stochastic redistribution of total water.

A Spatial Analysis of Regional Climate Model Ensembles

Stephan R. Sain, *National Center for Atmospheric Research*

The North American Regional Climate Change Assessment Program (NARCCAP) seeks to examine the uncertainty in the output of regional climate models and projections of future climate and climate change. At the heart of the program is an ambitious experiment that seeks to use a number of regional climate models (RCMs) with boundary conditions supplied by different atmosphere-ocean general circulation models (GCMs) to produce a wide range of model output over North America. Our goal within this program is to develop statistical methodology to analyze this model output and assess and quantify the sources of uncertainty. To that end, we are developing a Bayesian hierarchical framework that is based upon a multivariate spatial model. This allows us to capture the complex distribution of the spatial fields produced by these regional climate models. Case studies will be presented based on an ensemble of regional climate model output over the western United States as well as an initial analysis of NCEP-driven regional model output associated with NARCCAP.

Invited Session 5 Wednesday, May 27, 3:30pm - 5:00pm

The Randy Sitter Technometrics Session

Bayesian Design of Experiments for Nonlinear Models

Chris Gotwalt , *SAS Institute*

I will present ideas for designing efficient experiments in nonlinear settings. The approach is general, and can be applied to both nonlinear regression models and generalized linear models with non-normal responses. In the former, the response Y is related to design factors X by a model of the form $Y = g(X;\theta) + \epsilon$, where the function g is nonlinear with respect to the parameter vector theta. In the latter, $E(Y) = \mu(X;\theta)$, with $h^(-1)[\mu(X;\theta)]$ a linear predictor in X, and Y having an exponential family distribution. Standard examples are logistic, probit and Poisson regression models.

A special feature of nonlinear models is that good designs depend on the unknown value of the parameter vector theta. Thus an efficient design must be robust with respect to the lack of exact knowledge about theta. A reasonable way to address this problem is to adopt a Bayesian approach, with a prior distribution assigned to theta. Our focus is on normally distributed priors, but we show that our method can be adapted easily to non-normal priors.

I will describe Bayesian design criteria and discuss methods for finding designs that achieve robust and efficient performance. This includes a very fast, yet accurate, quadrature scheme for integrating the design criteria.

I motivate the approach using several examples.

Design and Analysis of Computer Experiments with Branching and Nested Factors

Ying Hung , *Rutgers University*

In many experiments, some of the factors exist only within the level of another factor. Such factors are often called nested factors. A factor within which other factors are nested is called a branching factor. For example, suppose we want to experiment with two processing methods. The factors involved in these two methods can be different. Thus, in this experiment the processing method is a branching factor and the other factors are nested within the branching factor. Design and analysis of experiments with branching and nested factors are challenging and have not received much attention in the literature. Motivated by a computer experiment in a machining process, we develop optimal Latin hypercube designs and kriging methods that can accommodate branching and nested factors. Through the application of the proposed methods, optimal machining conditions and tool edge geometry are attained, which resulted in a remarkable improvement in the machining process.

Invited Session 6 Thursday, May 28, 10:30am - 12:00pm

Field Reliability: Analyses of Warranty Data

Getting More from Warranty Data

Jeff Robinson, *General Motors R&D Center*

Warranties are agreements between customers and manufacturer, but the data they generate are also a rich source of valuable information. This talk is a summary of issues and insights gained over 30 years in the auto industry. The accurate enumeration of the number of units at risk (to generate warranty claims) is a long-standing and obvious consideration, but experience suggests that it is a very important one. So is the proper accounting for any peculiar or special coverage, such as warranty claims filed by dealers prior to the sale of a vehicle. More recently new data sources, such as on-board diagnostics, when coupled with warranty data, are greatly expanding the potential realm of warranty data analysis. Finally, warranty analysis can be taken beyond the traditional activities of forecasting and product quality monitoring. One such area is service quality.

Estimation of Lifetime Parameters and Its Properties Using only Failure Information from Warranty Database

Kazuyuki Suzuki, *University of Electro-Communications*

This research discusses how to estimate lifetime parameters using warranty data which consist of only failure information. Usually, warranty data consists of only failure information, and non-failure information is not obtainable which makes estimation difficult. The sales data is also important as it contains time-in-service in calendar timescale for each non-failed product during the warranty plan. This paper discusses maximum likelihood estimation of lifetime parameters using warranty data along with sales data, and examines the precision of the estimators by the asymptotic variances obtained from Fisher Information Matrix. The practical consequence of this finding is that the proposed method produces estimators of the lifetime parameters with good precision for large sales amount.

When Can Relative Age and Usage Effects Be Ascertained from Warranty Data?

Jerry Lawless, *University of Waterloo*

The reliability of manufactured units or systems may depend on their age (time in service) and also on the intensity or rate with which they are used. For systems in which detailed records of usage are available, and where failures or other reliability-related events occur frequently, we can ascertain the relative effects of age and usage. However, when events are rare and usage is infrequently recorded, it is much harder to do this. This talk will consider the latter setting, with illustrations involving warranty claims for automobiles.

Invited Session 7 — Thursday, May 28, 10:30am-12:00pm

Industrial Statistics

Designed Experiments with Limited Combinations of Factor Combinations

Christine M. Anderson-Cook*, *Los Alamos National Laboratory* & Timothy J. Robinson, *University of Wyoming*

In some applications, the experimenter has limited options of what combinations of factors can be chosen for a designed experiment. For example, preliminary processing has created a collection of parts with known values of the explanatory variables. This sample is likely to be an uneven distribution of parts throughout the observed ranges with correlations existing between explanatory variables. Unlike traditional design of experiments in which correlation can be eliminated through the choice of an orthogonal design, this is not possible here. Only a subset of this larger sample can be tested to obtain responses for the analysis. We discuss some of the properties of a good design in this setting. Specifically, we consider precision of model parameter estimates, the issue of replication, ability to detect lack-of-fit, and several others. We also present strategies for selecting a subset of the parts which simultaneously maximizes multiple objectives associated with good designs. The methodology is presented through a case study based on a manufacturing example.

On t and EWMA t Charts for Monitoring Changes in the Process Mean

Gemai Chen, *University of Calgary*

The performance of X-bar chart is usually studied under the assumption that the process standard deviation is well estimated and does not change. This is, of course, not always the case in practice. We find that X-bar charts are not robust against errors in estimating the process standard deviation or changing standard deviation. In this talk we discuss the use of a t chart and an exponentially weighted moving average (EWMA) t chart to monitor the process mean. We determine the optimal control limits for the EWMA t chart and show that this chart has the desired robustness property.

Impact of Fatigue Effect on Experimental Designs in Internet Based Discrete Choice Studies

William Li, *University of Minnesota*

Internet based discrete choice surveys are a popular form of marketing research due to the wide availability of on-line respondents. These methods are usually cost efficient, but the fatigue of respondents can be a big concern. While efficient experimental designs have been proposed and used for such survey methods, none of existing designs directly take into account the impact of the fatigue effect. Our study was motivated by an internet based survey polling respondents on their choices of a glass supplier. We estimate the fatigue effect via a mixed multinomial logit model. We then study the impact of fatigue on

different experimental designs, using several penalty functions for the fatigue effect. Some recommendations are provided.

Invited Session 8 Thursday, May 28, 1:30pm - 3:00pm

Analysis of Industrial Experiments

Taking a Closer Look at Optimal Screening Designs

Timothy J. Robinson*, *University of Wyoming* & Christine M. Anderson-Cook, *Los Alamos National Laboratory*

The use of computer-generated experimental designs is quite prevalent in practice. When soliciting software packages to produce an optimal design, the user must provide the underling model form, factor levels and an optimality criterion. In the screening stage, the optimality criterion of choice is quite often the D-criterion in which the optimal design is the set of factor combinations which maximize the determinant of the information matrix. Since the D-criterion is a global summary of the entire design, it is sometimes difficult to know exactly how quickly performance changes as the efficiency drops. While the D-criterion is certainly appealing for a variety of reasons, a D-optimal design does not necessarily yield uniformly best precision for model parameter estimates. Also, competing designs which differ in terms of D-efficiencies may be only trivially different in terms of their abilities to detect underlying effects. In this seminar we re-visit optimal design selection at the screening stage of experimentation. Specifically, we take a closer look at D-optimality as it relates to power in hypothesis testing and as it relates to the distribution of precisions across the model parameter estimates. Along the way, we suggest some graphical approaches to aid in the choice of an optimal screening design.

Joint Model Selection of Location and Dispersion Effects in Unreplicated Factorial Experiments

Jorge E. Rodriguez* & Thomas M. Loughin, *Simon Fraser University*

In many settings identification of dispersion effects in factorial experiments can be as important as that of location effects. However, most of the procedures that have been proposed to identify dispersion effects in the absence of replication are based on the assumption that previous identification of location effects has been done correctly. Moreover, incorrect identification of location effects may impair subsequent identification of dispersion effects.

We demonstrate that joint identification of both location and dispersion effects will provide a more successful way of obtaining correctly both kinds of effects. A joint model is used that results in a single likelihood value, and hence combined location-dispersion models can be evaluated using information-based criteria. A search through the space of possible models is conducted via genetic algorithm. The proposed procedure is demonstrated on

several examples where existing methods have led to conflicting results.

Lack-of-Fit Tests for Split-plot Response Surface Designs
Geoffrey G. Vining, *Virginia Tech*

Many times in a designed experiment there are constraints that force certain factors to be much harder to change than others. Restricting the randomization is an appropriate approach to such an experiment and leads to a split-plot structure. The recent industrial statistics literature gives considerable attention to the design of split-plot response surface designs. One very important aspect to response surface methodology is the use of lack-of-fit tests, especially to determine when it is time to move from a first-order model to a second-order model. This paper develops the methodology for lack-of-fit tests within a split-plot structure for the case where the design used is ordinary least squares - generalized least squares equivalent. Several examples illustrate the methodology.

Invited Session 9 — Thursday, May 28, 1:30pm - 3:00pm

Drug Discovery

ChemModLab: QSAR modeling and model-based searching
Jacqueline M. Hughes-Oliver* & Qianyi Zhang, *North Carolina State University* & Raymond Ng, *University of British Columbia*

ChemModLab, written by the ECCR@NCSU consortium under NIH support, is a toolbox for fitting and assessing quantitative structure-activity relationships (QSARs). Its elements are: a cheminformatic component that computes five types of molecular descriptors for use in modeling; a set of sixteen statistical methods for fitting models; and methods for validating the resulting model. These sixteen statistical methodologies comprise a comprehensive collection of approaches. ChemModLab can produce eighty QSAR models that can be used individually or as the basis for ensembles. The first part of this presentation will introduce this web-accessible software.

The remainder of this presentation will focus on ensemble models, where output from many individual models are combined to yield an overall conglomerate model. Such methods have gained popularity in multiple areas of chemistry. The ensemble method Random Forests (RF) has been shown to be highly effective for predicting biological activity in many applications. RF is a family ensemble model because it uses base learners created from the same underlying mechanism, a recursive partitioning decision tree. While generally effective, RF can have poor performance when the training set is highly unbalanced. This is often the case for applications regarding QSARs, where the percent of active compounds can be very small. For such applications, we study the properties of family ensemble models

and make recommendations for obtaining improved performance.

G-SELC: Optimization by Sequential Elimination of Level Combinations using Genetic Algorithms and Gaussian Processes

Abhyuday Mandal*, *University of Georgia*, Pritam Ranjan, *Acadia University* & C.F. Jeff Wu, *Georgia Institute of Technology*

Identifying promising compounds from a vast collection of feasible compounds is an important and yet challenging problem in the 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 the SELC algorithm which was motivated by the 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 the 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.

Image Analysis and Structure/localization Relationships for Small Molecule Probes in Live Cells

Kerby Shedden* & Fangyi Liu, *University of Michigan*

Images of cells incubated with fluorescent small molecule probes can be analyzed to assess whether probes enter cells, and how the probes distribute within the cells that they enter. When images are available for a chemically diverse collection of probes, it becomes possible to ask whether particular chemical substructures – "optical address tags" – are associated with specific patterns of subcellular distribution. A number of interesting statistical questions arise in attempting to do this. Imaging artifacts must be identified and compensated for, distribution patterns must be quantified, and multivariate relationships between image properties and chemical structures must be discovered. I will discuss our recent work in this area, focusing on two issues: (i) how complex should image features be, and how many spatial scales are informative?; (ii) what information is contained in the variation in subcellular patterns observed across multiple cells exposed to the same probe?

Invited Session 10 Thursday, May 28, 3:30pm - 5:00pm

Computer Experiments

Optimization of a Laser-Assisted Micro-Grooving Process Using an Engineering-Statistical Approach

Vengazhiyil Roshan Joseph*, *Georgia Institute of Technology*, Ramesh K. Singh, *Indian Institute of Technology Bombay* & Shreyes N. Melkote, *Georgia Institute of Technology*

The objective of our study is to optimize a laser-assisted micro-grooving process. The process uses a laser beam to thermally soften the material thereby lowering the cutting forces and associated machine and tool deflections. However, the use of laser heating can produce a detrimental heat affected zone (HAZ) in the workpiece surface layers. Consequently, the laser and micro-grooving parameters need to be optimized in order to achieve the desired thermal softening effect while minimizing the formation of a HAZ in the material. Although thermal and force models for the process have been developed, they are quite expensive to evaluate and are not accurate due to various simplifying assumptions. This makes the optimization expensive and the results not reliable. We overcome these limitations using metamodels and engineering-statistical models. The optimization strategy is experimentally verified and shown to yield good results.

Deterministic Computer Models and GP as an Interpolator

Pritam Ranjan* & Ronald Haynes, *Acadia University*

For many deterministic computer simulators, the outputs are noiseless and the desired predictor is an interpolator of the observed data. Gaussian spatial process (GP) is commonly used to model such simulator outputs. Fitting a GP model to n data points requires inversion of an n x n correlation matrix R. This becomes a computational issue due to near-singularity of R if any pair of data points are close together in the input space. The popular approach to overcome near-singularity introduces over-smoothing of the data. In this talk, we will present a new iterative approach to construct a predictor that gives higher prediction accuracy. Several examples will be presented to illustrate that the proposed predictor converges to the GP interpolator.

Approximating a Computer Code via a Gaussian Process: A Practical Guide

William J. Welch*, *University of British Columbia*, Jason L. Loeppky, *UBC-Okanagan* & Jerome Sacks, *NISS*

Using a Gaussian Process to approximate the output of a computationally intensive computer code as a function of its inputs is now commonplace. The practitioner has to make several choices in the implementation, however. First, when are regression terms (beyond a simple constant mean parameter) useful? Secondly, the correlation function has to be specified. The power-exponential and Matern functions each appear widely in the literature. Is

there a practical difference in terms of predictive performance? Through simulation and example codes we will determine if and when these choices are crucial and give corresponding advice to the practitioner.

Invited Session 11 Friday, May 29, 8:30am - 10:00am

Problems in Sampling Survey Practice

Asymptotic Calibration Estimation Using Estimated Importance Sampling and Exponential Tilting

Jae-kwang Kim, *Ohio State University*

We consider the problem of parameter estimation with auxiliary information, where the auxiliary information often takes the form of known moments. Calibration estimation is a typical example of using the moment conditions in sample surveys.

Given the parametric form of the original distribution, we use the estimated importance sampling of Henmi et al (2007) to obtain an improved estimator. If we use the normal density to compute the importance weights, the resulting estimator takes the form of the exponential tilting estimator. The proposed estimator is shown to be asymptotically equivalent to the empirical likelihood estimator, but has some computational advantages. Results from a limited simulation study are presented.

Application of Experimental Designs in Survey Sampling

J.N.K. Rao, *Carleton University*

Some early uses of experimental designs and more recent applications are reviewed. Role of experimental designs in controlled sampling is appraised. Some new work on constructing balanced bootstrap replicates from stratified random samples is also reported. Role of embedded experiments in on-line surveys will also be discussed.

Efficient and Robust Model-Assisted Imputation for Covariate-Dependent Missing Responses

Jun Shao, *University of Wisconsin-Madison*

Missing data is an ubiquitous problem in medical and social science studies. Imputation is one of the most popular methods in dealing with missing data. The most commonly used imputation that makes use of covariates is regression imputation, where the regression model can be parametric, semiparametric, or nonparametric. Parametric regression imputation is efficient but it is not robust against the misspecification of the regression model. Although nonparametric regression imputation (such as nearest neighbor imputation and kernel regression imputation) is model free, it is not efficient, especially if the dimension of covariate vector is high (the well-known problem of curse of dimensionality). Assuming that the missing mechanism is covariate-dependent and the propensity function can be correctly

specified, we propose a regression imputation method that possess good efficiency and is robust against regression model misspecification. Furthermore, our method is valid as long as one of the regression model and propensity model is correct, which is referred to as the double robustness property by Robins, Rotnitzky, and Zhao (1994). We show that asymptotically the sample mean based on our imputation achieves the semiparametric efficient lower bound if both regression and propensity models are correctly specified. Simulation results show that the proposed method outperforms many existing methods for handling missing data, especially when the regression model is misspecified. As an illustration, an economic observational dataset is analyzed.

Invited Session 12 Friday, May 29, 8:30am - 10:00am

Data Mining: dimension reduction and variable selection

On Dimension Folding of Matrix or Array Valued Statistical Objects

Bing Li*, Min Kyung Kim & Naomi Altman, *The Pennsylvania State University*

We consider dimension reduction for regression or classification in which the predictors are matrix- or array-valued. This type of predictors arise when measurements are obtained for each combination of two or more underlying variables — for example, the voltage measured at different channels and times in electroencephalography data. For these applications it is desirable for the reduced predictor to preserve its array structure (e.g. time versus channel), but this cannot be achieved within the classical dimension reduction formulation. In this paper we introduce a dimension reduction method, to be called dimension folding, for matrix- or array-valued predictors that preserves the array structure. In an application of dimension folding to an electroencephalography data set, we correctly classified 86 out of 122 subjects as alcoholic or nonalcoholic based on their electroencephalography.

Regularization Parameter Selections via Generalized Information Criterion

Runze Li, *The Pennsylvania State University*

We apply the nonconcave penalized likelihood approach to obtain variable selections as well as shrinkage estimators. This approach relies heavily on the choice of regularization parameter, which controls the model complexity. In this paper, we propose employing the generalized information criterion, encompassing the commonly used Akaike information criterion (AIC) and Bayesian information criterion (BIC), for selecting the regularization parameter. Our proposal makes a connection between the classical variable selection criteria and the regularization parameter selections for the nonconcave penalized likelihood approaches. We show that the BIC-type selector enables identification of the true model consistently, and the resulting estimator possesses the oracle property in the terminology of Fan and Li (2001). In contrast, however, the AIC-type selector tends to overfit. However, following the loss efficiency of Li (1987) and Shao (1997), we further showed that under

appropriate conditions, AIC selector enjoys an asymptotic loss efficiency which BIC-type selectors do not possess. Our simulation results confirm these theoretical findings, and an empirical example is presented.

Fast Simultaneous Dimension Reduction and Clustering: Viewing Data from Extremes

George Ostrouchov, *Oak Ridge National Laboratory*

Extremely large data sets often have a complex structure that includes many clusters and high dimensionality. As effective dimension reduction can be hampered by the presence of several clusters, it is useful to partition the data into clusters before applying dimension reduction to each cluster. At the same time, clustering is often more easily done after dimension reduction. The difficulty of either of these approaches suggests that simultaneous dimension reduction and clustering is needed. I will describe a dimension reduction approach based on simple ideas from machine learning that use distances from extreme points. When combined with ideas from robust statistics, it can be used to provide a fast algorithm for simultaneous dimension reduction and clustering. While the algorithm is of linear computational complexity in the amount of data, distributed parallel processing can be used for data sets that do not fit into single processor memory or for data sets that are scattered across different locations.

Oracular Model Selection via Penalized Composite Quantile Regression

Hui Zou, *University of Minnesota*

The concept of model selection oracle introduced by Fan and Li (2001) characterizes the optimal behavior of a model selection procedure. In the sparse linear regression model case, the error distribution is often unspecified, and hence coefficient estimation and variable selection is routinely done in the penalized least squares (LS) framework. Oracle-like estimators such as the SCAD and the adaptive Lasso mimic the least-squares oracle. However, the least-squares oracle theory breaks down if the error variance is infinite. Moreover, for some non-normal error models with finite variance, the least-squares oracle estimator has unsatisfactory efficiency.

In this talk we introduce a new oracle estimator using a new regression technique called composite quantile regression (CQR). We employ the adaptive Lasso penalty to produce a penalized CQR estimator that mimics the CQR-oracle estimator. We show that the oracle model selection theory using the CQR oracle works beautifully even when the error variance is infinite. When the error variance is finite, CQR still enjoys great advantages in terms of estimation efficiency. We show that the relative efficiency of CQR-oracle compared to the least-squares oracle is greater than 0.864 regardless the error distribution. Moreover, the CQR-oracle could be much more efficient and sometimes arbitrarily more efficient than the least-squares oracle. The same conclusions hold when comparing a CQR-oracular estimator with a LS-oracular estimator.

Invited Session 13 — Friday, May 29, 10:30am - 12:00pm

Design Theory

MDS-optimal Supersaturated Designs

Arden Miller, *The University of Auckland*

A minimal dependent set (MDS) is a set of vectors that are linearly dependent but if any one of them is removed the resulting subset is independent. This talk will discuss the relationship between the minimal dependent sets of the column vectors of the design matrix for a 2-level supersaturated design and the resolvability of the design. It will introduce the concepts of MDS-resolution and MDS-aberration as criteria for comparing supersaturated designs. Results concerning supersaturated designs that have minimum MDS-aberration will be presented.

On Locally Optimal Designs for Generalized Linear Models with Group Effects

John Stufken*, *University of Georgia* & Min Yang, *University of Missouri*

Generalized linear models have long become a staple in statistical modeling and data analysis. While the literature contains also a number of noteworthy contributions to the selection of optimal designs for specific generalized linear models and specific optimality criteria, the optimal design problem has proven to be very difficult for obtaining general results. We present general results that extend those by Yang and Stufken (2009, Annals of Statistics 37, 518-541) to the situation where group effects are allowed to occur.

More precisely, our results cover generalized linear models in which the linear predictor part is of the form $b_0 + b_1 * x$ for a single predictor x, but where the parameters b_0 and b_1 are allowed to depend on one or more group effects, such as race, gender, ethnicity or other categorical variables, to allow for subject heterogeneity. The focus is on locally optimal designs under the A- and D-optimality criteria. We present theoretical results that ensure the existence of optimal designs of a relatively simple form. While these results greatly reduce the search for optimal designs, in the presence of multiple groups good computational tools are nevertheless indispensable to search for the best designs among the designs of the simple form. We present results based on a computational approach that is very successful at identifying support points and weights for optimal designs.

Two-Level Nonregular Designs Constructed via Quaternary Codes

Hongquan Xu, *UCLA*

The research of developing a general methodology for the construction of good nonregular designs has been very active in the last decade. A promising method is the use of quaternary codes, which are linear spaces over the ring of integers modulo 4. This talk explores the properties of two-level nonregular designs constructed via quaternary codes. Some theoretical results are obtained regarding the aliasing structure of such designs. Optimal

designs are constructed under the maximum resolution, minimum aberration and maximum projectivity criteria. These designs often have larger generalized resolution and larger projectivity than regular designs of the same size. Some of these designs indeed have generalized minimum aberration and maximum projectivity among all possible designs.

Abstracts of Contributed Talks

Contributed Session 1 Wednesday, May 27, 10:30am - 12:00pm
Imputation, Estimation and Sampling Methods

Random Group Variance Estimators for Survey Data with Random Hot Deck Imputation

Qi Tang* & Jun Shao, *University of Wisconsin - Madison*

Random hot deck imputation is often applied to survey data with nonresponse. One of the popular methods for variance estimation without nonresponse is the random group method, which has to be adjusted when it is applied to imputed data. One such kind of adjustment is re-imputing nonresponses in each random group. To reduce the computational complexity due to re-imputation, a shortcut random group method has been considered in some applications. The main purpose of this talk is to show how to correctly implement a shortcut random group method so that it produces asymptotically unbiased and consistent variance estimators for estimated population totals can be obtained. We also show that the random group method with re-imputation produces asymptotically unbiased and consistent variance estimators. As a special case of our general result, the random group variance estimator for the case of no nonresponse is asymptotically unbiased and consistent, a result that has not been documented although the random group method is frequently used in applications.

Maximum Likelihood Estimation under a Hypothesis of Marginal Homogeneity for Data in a Square Contingency Table

Saumen Mandal*, *University of Manitoba* & Ben Torsney, *University of Glasgow*

We consider an optimization problem in which the aim is to find an optimizing probability distribution or measure. There are a variety of problems in statistics, which demand the calculation of such probability distributions. Optimal regression design is a particular example. Other examples include parameter estimation and stratified sampling. In particular, we consider a problem of determining maximum likelihood estimates under a hypothesis of marginal homogeneity for data in a square contingency table. This is an example of an optimization problem with respect to variables which should be nonnegative and satisfy

several linear constraints. To construct the optimizing distributions, we consider a class of multiplicative algorithms, indexed by a function (f) which depends on the derivatives of the likelihood function. The function (f) satisfies some conditions and may depend on one or more free parameters. These iterations neatly satisfy the constraints of the problem. We consider an example of data for which marginal homogeneity is of interest, namely, a grading of the unaided distance vision of each eye of men and women (Placket, 1974, The analysis of categorical data, Griffin, London). In conclusion, the properties of the algorithm and some results will be reported and discussed.

Second Order Exchangeable Emulators to Assess Initial Condition Uncertainty

Leanna House*, Michael Goldstein & Ian Vernon, *Virginia Tech*

We address the uncertainty of deterministic computer models that rely on both input parameters and initial conditions. We refer to such models as semi-deterministic. Purely deterministic computer models either do not have an initial condition or fix (without error bounds) the value for the initial condition so that the same output will result from one set of input parameter values, even when the model is implemented multiple times. Semi-deterministic models however, allow the condition to vary, and thus have the potential to produce more than one result per input. When multiple outcomes per input are present, current approaches rely primarily on summary statistics (e.g., mean and variance per input), and apply standard deterministic model uncertainty analysis approaches. However, inferences based solely on such statistics require implicitly strong assumptions which we are unwilling to make. Thus, we introduce the notion of latent computer model outcomes which correspond to the results of the semi-deterministic model when using the appropriate, but unknown, initial condition for the physical system of interest. The goal for this paper is to make inferences about the latent model given a sequence of realized semi-deterministic model evaluations. We consider the sequence elements to be second order exchangeable and use Bayes linear methods to assess the posterior expectation and variance of the latent model given the realised evaluations. We demonstrate our methods using semi-deterministic results from a galaxy formation model called Galform that relies on initial specifications of dark matter.

Accelerated Life Test Sampling Plans for the Weibull Distribution under Type I Progressive Interval Censoring with Random Removals

Chunyan Yang*, Chang Ding & Siu-Keung Tse, *City University of Hong Kong*

This paper considers the design of accelerated life tests (ALT) sampling plans under Type I progressive interval censoring with random removals (PICR). We assume that the lifetime of products follows a Weibull distribution. Two levels of constant stress higher than the use condition are used. The sample size and the acceptability constant which satisfy given levels of producer's risk and consumer's risk are found. In particular, the optimal stress level and the allocation proportion are obtained by minimizing the generalized asymptotic

variance of the maximum likelihood estimators of the model parameters. Furthermore, for validation purposes, a Monte Carlo simulation is conducted to assess the true probability of acceptance for the derived sampling plans.

Contributed Session 2 Wednesday, May 27, 1:30pm - 3:00pm

Analysis of Counts, Degradation, Lifetime Data and Nonlinear Models

On Departure from Exponentiality When Estimating the Number of Component Processes

Subrata Kundu*, *George Washington University*, Anup Dewanji, *Indian Statistical Institute Calcutta* & Tapan K. Nayak, *George Washington University*

In recent past, estimating the number of component processes in a system of superimposed IID renewal processes has gained some interest with its wide ranging applications in software reliability and other areas . Formally, there are an unknown number ν of IID renewal processes each having a renewal distribution given by, say F_θ, involving parameter θ, which may be a vector. We observe only the superimposition of these ν component processes up to a prefixed time, say τ. If an event from a particular process occurs before time τ, then this process is detected and further events from this process by time τ can be identified with the process. In software reliability application, this identification is done by a mechanism known as recapture debugging. If no event from a particular process occurs by time τ, then this process is not detected. We, therefore, observe the number of detected processes. The objective is to estimate ν (which is the number of bugs in software reliability application) based on the above observation. Most work in this area assume Poisson processes leading to exponential distribution for the renewal time. Much work has been done under the parametric set up. However, the method is iterative involving extensive computation and the results are naturally sensitive to the model assumption on the renewal distribution. One may be keen to know about the effect on the estimate of ν due to departure of the renewal distribution from the assumed model and whether the simple estimate based on exponential distribution can be used regardless of the true renewal distribution. Our results demonstrate that violation from the exponential distribution can result in serious bias in the estimate of ν. In this work, we propose a simple nonparametric method for the estimation of ν, which seems to perform well in terms of bias and asymptotic normality. We derive asymptotic distribution of our estimator and use simulation to show the effectiveness of the proposed estimator.

Inference with Censored Degradation Data

Yang Yang* & Vijay Nair, *University of Michigan*

Degradation data are commonly measured over time to investigate the physical deterioration of products. We have been involved in a project to analyze distress indices which are recorded to evaluate the condition of road pavements. The goal of the project was to

determine the effect of various pavement design parameters and predict performance and the need for repairs over time. The database was very incomplete with missing observations as well as left and right censoring. In this paper, we discuss methods for analyzing this type of censored degradation data and for making predictions at the individual device level. Unlike the estimation of failure time distributions, nonparametric estimation appears difficult, so we focus on Gaussian models with a time series structure. Maximum likelihood estimation, use of EM algorithm, and connections to GEE models are discussed. The results are illustrated on real pavement degradation data.

Parametric Inference for System Lifetime Data with Signatures Available

Hon Keung Tony Ng*, *Southern Methodist University*, J. Navarro, *University of Murcia* & N. Balakrishnan, *McMaster University*

In this talk, the statistical inference of the lifetime distribution of component based on observing the system lifetimes with signature available is discussed. A general proportional hazard rate model for the lifetime of the components is considered, which includes some commonly used lifetime distributions. Different estimation methods for the proportional parameter are discussed. Monte Carlo simulation study is used to compare the performance of these estimation methods and recommendations are made based on these results.

Analysis of Transform-Both-Sides Nonlinear Regression Models

A.H.M. Mahbub Latif* & Steven G. Gilmour, *Queen Mary, University of London*

Nonlinear regression models have been widely used in different stages of the drug development process, e.g., in the early pre-clinical stage the Michaelis-Menten model is used to identify enzymes involved in the metabolism of novel compound entities. Identification of such enzymes at that early stage of drug development could be useful in avoiding adverse drug reactions which may be observed when the drug is in the market. Fitting nonlinear models using the maximum likelihood (ML) method of estimation requires the assumption that the additive error terms of the nonlinear model follow a normal distribution. When the normality assumption is violated, a transformation technique is often used in practice. Transforming both sides of the model has advantages over transforming only the response, e.g. interpretations of the original model parameters remain the same which is important because most nonlinear models have a scientific justification. Fitting a transform-both-sides (TBS) nonlinear model is computationally more complex than fitting the corresponding nonlinear model. In this talk, we will discuss a method, which we call the Anova method, for fitting TBS nonlinear model. The Anova method considers a linear treatment model to estimate the transformation parameter and then the TBS nonlinear model is estimated with a known transformation parameter. The Anova method is computationally simpler than the commonly used ML method for fitting the TBS nonlinear model. Considering the Michaelis-Menten model as an example of nonlinear model, we will show the results of a simulation study for comparing ML and Anova methods where

the Box-Cox transformation is used for transforming both sides of the Michaelis-Menten model. Simulation results show that the Anova method can estimate the parameter of the TBS Michaelis-Menten model with less bias and can provide a more precise estimate of the pure error compared to the ML method.

Contributed Session 3 Wednesday, May 27, 3:30pm - 5:00pm

Reliability and Quality

Data-Driven Reliability Estimation for Prognostics and Health Management

Aparna V. Huzurbazar, *Los Alamos National Laboratory*

Prognostics and Health Management (PHM) is increasingly important for understanding and managing today's complex systems. PHM develops the capability to make decisions about maintenance, based on prognostics information, resources, and operational conditions. This talk will present background on PHM, the suite of tools currently available to engineers working in this area the potential contribution that statistical tools and methods can provide. We will discuss the requirements for the types of data most-suited to PHM work and how to incorporate this information into a data-driven, decision making framework for PHM. For us, a primary application of PHM is to predict the future life of the system using data from a variety of sources such as vibration, degradation, and so on. This talk presents how to incorporate such information into a decision making framework for PHM.

R(reliability)-Mark in Korea

Jaiwook Baik, *Korea National Open University*

Korean economy has been increasing since the industrialization during 1960's. But recent problem was that as the trade increased, import from Japan or other advanced countries increased also, especially complex or sophisticated materials and components. Therefore, Korean government has been driving R(reliability)-Mark program since 2000 in order to localize critical materials and components. In this talk, I will introduce how R-Mark program operates and how it relates to statistics.

Bias Correction When Using the Extreme Value approximation in Monte Carlo Estimation of High Reliability Systems

Peter W. Hovey* & Brian Krilov, *University of Dayton*

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 presented previously showed better precision, but a significant bias. Some methods for correcting the bias are discussed.

Focus on Failure Avoidance and Risk Reduction through Variation Mode and Effect Analysis

Stefano Barone*, Bo Bergman, Alexander Chakhunashvili & Per Johansson, *Chalmers University of Technology*

Variation Mode and Effect Analysis (VMEA) is a quality improvement tool initially thought to help product development engineers focusing on variation. The method was inspired by the wide use of Failure Mode and Effect Analysis (FMEA) in business and industry and the increased attention on robust design. However, FMEA is based on the concept of failure while VMEA is developed on the concept of variation, i.e. it implies a step further toward the awareness of variation and its implications in terms of risk and failures. VMEA helps identifying, scrutinizing and measuring the sources of variation and the way they channel through and impact on important characteristics of the system under study. Applied systematically, it provides a solid basis for achieving system robustness. Three VMEA procedures have been set up for use at different system maturity levels. Several applications have been made over the last five years showing the usefulness of the tool for both product and process (here including also service) improvement. This presentation is aimed at giving an overview of the VMEA method and its possible ways/areas of use in the entire product/process development.

Contributed Session 4 Thursday, May 28, 10:30am - 12:00pm

Fractional Factorial Designs

Optimal Selection of Blocked Two-Level Fractional Factorial Designs

Weiming Ke, *South Dakota State University*

Blocked two-level fractional factorial designs are very useful for efficient data collection in industrial experiments and other areas of scientific research. In most experiments, in addition to the main effects, some two-factor interactions are also meaningful and need to be estimated. In this article, a method is proposed and studied for selecting the best blocked two-level fractional factorial designs when some of the two-factor interactions are included in the model. How to search for best designs using this method is discussed and some results for designs of 8 and 16 runs are presented.

Creating Catalogs of Two-level Fractional Factorial Designs Based on the Criteria of Generalized Aberration

C. Devon Lin*, *Queen's University*, Randy R. Sitter & Boxin Tang, *Simon Fraser University*

We consider the problem of constructing good two-level fractional factorial designs, both regular and nonregular. The criteria of G and G2 aberration are used to rank designs. A general design structure is utilized to provide a solution to this practical, yet challenging, problem. Based on this design structure, we develop an efficient algorithm for obtaining a collection of good designs based on the aforementioned two criteria. Finally, we present some results for designs of 24, 32, and 40 runs obtained from applying this algorithmic approach.

Semifolding Two-level Fractional Factorial Split-Plot Designs

Robert McLeod, *University of Winnipeg*

Foldover designs have recently been advocated as a useful follow-up strategy for de-aliasing low-order effects in fractional factorial split-plot experiments. One argument against the use of foldover designs is that they are "degree-of-freedom inefficient". Semifoldover designs, which use only one-half of the runs of a foldover, have been explored in the fractional factorial setting but are considered sparingly in the fractional factorial split-plot design context. In this talk we consider the construction of split-plot semifoldovers and investigate the impact on the semifoldover selection process due to the presence of whole-plot and subplot factors. We observe that by semifolding resolution IV split-plot designs it is often possible to de-alias as many two-factor interactions as a foldover. If prior knowledge concerning the significance of low-order effects is available we also demonstrate that semifolding resolution III split-plot designs may be an attractive option.

Forms of Four-Word Indicator Functions with Implications to Two-Level Factorial Designs

Po Yang, *DePaul University*

Indicator functions are new tools to study fractional factorial designs. In this paper, we study indicator functions with four words and provide possible forms of the indicator functions and explain their implications to two-level factorial designs.

Contributed Session 5 Thursday, May 28, 1:30pm - 3:00pm

Modeling and Design of Experiments

Mixture-of-Mixture Experiments: A general modeling and design approach

Lulu Kang* & Vengazhiyil Roshan Joseph, *Georgia Institute of Technology* & William

A. Brenneman, *The Proctor and Gamble Company*

In mixture-of-mixture experiments, major components are defined as the components which themselves are mixtures of some other components, called minor components. In other mixture-of-mixture experiments, components are divided into different categories. Each category is called a major component, and the components within a major component are minor components. The special structure of the mixture-of-mixture experiment makes the design and modeling approaches different from a typical mixture experiment. In this paper, we propose a new modeling approach called the major-minor model. It overcomes some of the limitations of the multiple-Scheffe model, which has been studied and used for mixture-of-mixture experiments. We also provide a design strategy. Some examples are given to illustrate the methodology. We apply the proposed design and modeling approach to a mixture-of-mixture experiment conducted to formulate a new Pringles potato crisp.

Design of Experiments for Packet Communication Networks

Ben M. Parker, *Queen Mary, University of London*

We present interdisciplinary work to investigate measurement of packet communication networks, in which a user's information is transmitted from a source to a receiver by sending small chunks of information, called packets, according to some transmission protocol. We wish to find out some statistic of the network, such as traffic rate, load, or bandwidth. Often we cannot measure this statistic directly, perhaps due to security or logistical constraints. Data are instead gathered by sending survey packets, called active probes, into the network, and by making measurements on the probe packets. By sending packets into a potentially congested network we increase the congestion; we alter the experiment by performing it. Therefore we have conflicting aims: to gather the most information from our experiment, whilst taking as few readings as possible. We can regard this as a classical experiment where we take observations on a system, but we have a limited number of observations. Our design problem is then to find the optimal times to take measurements by sending probes into the network. Our technique models the evolution of the network as a Markov chain. We show that there is an optimal set of times at which to measure (probe) the network, and we demonstrate a method to find this optimal set of times for a general Markov chain. Packet communication networks make up the vast majority of global information exchange, and surprisingly there has been little statistical research on measuring them optimally. Our approach presents one method to address this. Although we use packet communication networks as an example, our method generalises for finding optimal times for repeated measurements on any systems that evolves according to a Markov chain.

A Decision Theory Approach to the Choice of Factorial Design

Deepa Kapoor, *Queen Mary, University of London*

The statistical tools available for designing experiments usually choose a design which is optimal, either in terms of a variance criterion such as D-efficiency, or in terms of estima-

bility criteria such as resolution and abberation, given the number of factors, the number of levels of each factor and the number of runs available. However, in practice these are not given but are to be decided by the experimenter and this is usually done in an informal manner by comparing what can be obtained from various designs of different sizes and then choosing a design based on what is sensible. Also, the usual optimal design criteria and utility functions do not provide the correct criteria for experimenters to answer certain questions, such as which design will maximise yield or correctly identify active factors. A decision theory approach is explored, which allows a more informed choice of design to be made when a fraction of a full factorial design is to be chosen. Prior distributions on the sizes of effects will be taken into consideration, and a design (including the number of runs, factors and levels) can then be chosen from a candidate set of designs using a utility function relevant to the objectives of the experiment. The general philosophy will be described and some specific examples presented.

Stochastic Optimization with Space-Filling Designs

Peter Z.G. Qian, *University of Wisconsin-Madison*

Space-filling designs, i.e., designs where the points are uniformly distributed, were invented and are getting more and more traction in computer experiments. Stochastic optimization problems, or called optimization under uncertainty problems, emerge in areas as varied as machine learning, statistics, control theory and structural engineering. Simple Monte Carlo sampling is usually used for the construction of sample approximations in solving such optimization problems. This talk will report some recent advances in empowering the Sample Average Approximation method with space-filling designs. Our theoretical and numerical results clearly demonstrate significant advantages of using such designs for solving stochastic optimization problems.

Contributed Session 6 Thursday, May 28, 3:30pm - 5:00pm

Optimal, Robust and Repeated Measurement Designs

A D-optimal Design for a Generalized Exponential Model Governing Growth of Nanowires

Li Zhu, *Harvard University*

A lot of research in modern-day nanotechnology is devoted to estimation of growth models of nanowires. Here, an effort is made to construct an optimal experimental design for estimation of parameters of a complex curve characterizing nanowire growth that is partially exponential and partially linear. The criterion of D-optimal design, which maximizes the determinant of the information matrix, is utilized to find the optimal design. A locally D-optimal design for the non-linear growth model is obtained by using a novel geometric approach. Further, a sequential algorithm is proposed for obtaining the D-optimal design. The convergence of the sequential algorithm to the D-optimal design is demonstrated using

Monte-Carlo simulations. Some guidelines for the choice of initial design are also proposed.

Applications and Implementations of Continuous Robust Designs

Wai Kong Yuen* & Xiaojian Xu, *Brock University*

In the literature of constructing robust designs, many resulting designs turned out to have densities. In practice, an exact design should tell the experimenter what the support points are and how many subjects should be allocated to each of these points. In particular, we consider a practical situation in which the number of support points allowed is constrained. We discuss two intuitive approaches, which motivate a new implementation scheme that minimizes the loss function based on the Kolmogorov and Smirnov distance between an exact design and the optimal design having a density. We present two case studies: one to illustrate the application and implementation of a robust design constructed for a real life dose-response experiment and the other to demonstrate the advantage of our optimal method for an example in the literature. Additionally, we perform some simulation studies to compare the efficiencies of the exact designs obtained by our optimal implementation and those by other commonly used implementation methods.

Empirical Likelihood Based Control Charts

Asokan Mulayath Variyath*, *Memorial University of Newfoundland*

After the introduction of six sigma methodology, control charts are widely to monitor the process product quality characteristics. The most commonly used Shewhart control chart for monitoring the process mean assume that the characteristic under study follow normal distribution. But in many situations, the quality characteristics of interest such as strength, ovality etc follows asymmetrical distribution. Since the distributional assumption is not valid, the Shewhart control chart may give a wrong signal that process is in control even though the process is out of control (and vice versa). In these situations, a natural choice is a control chart procedure which does not need any distributional assumptions. We propose empirical likelihood based control chart to monitor the process mean, irrespective of the distribution of the quality characteristic. Empirical likelihood (EL) is a non-parametric method which has similar properties of the parametric likelihood, which are widely used for parameter estimation (survey sampling, survival analysis etc) and construction of confidence intervals. We propose to monitor the EL ratio statistic to control the process mean. This proposed EL based control chart has few hurdles due to small subgroup size. Due to small sample size, EL ratio statistic may not follow the chi-square distribution, which is the backbone of the construction of the confidence interval. Also due to small sample size, the EL ratio statistic is not well defined in many cases since the overall mean does not fall inside the convex hull of the subgroup samples. We propose to use the adjusted EL proposed by Variyath (2006) and Chen, Variyath and Abraham (2008) in this situation. We use bootstrap method to construct the empirical distribution of the EL ratio statistic. We have carried out some simulation studies (univariate case) assuming different distributions such

as Normal, Exponential, etc. When the assumed distribution is Normal, the performance of the EL based control chart is comparable with standard Shewhart control chart. When assumed distribution is non-Normal, EL based control chart has better performance than the Shewhart control chart and most commonly used Bootstrap control chart in detecting the shift in the process.

Repeated Measurement Designs under Subject Dropout

Shi Zhao*,*Xerox Innovation Group* & Dibyen Majumdar, *University of Illinois at Chicago*

Crossover designs have been heavily used in various practices, ranging from psychology and human factor engineering to medical and agricultural applications. A carefully designed crossover study will produce a wealth of information and the parameters of interest can be estimated with high precision. This expectation, however, is based on a critical assumption that the experiment will yield all planned observations. Yet in many studies that involve human subjects, such as clinical trials, it is very common that subjects drop out of the study prior to the completion. In this paper, we try to address two major questions that concerned by a crossover design under subject dropout. We suppose that a well designed experiment is implemented, while "well designed" stands for the possession of nice statistical properties, including universally optimality or high efficiency. The first question asks: what statistical properties are maintained in the design at the end of the experiment observed? We answer it by proposing UBRMDs Balanced for Loss. We can show that the minimal designs of that proposed UBRMD are universally optimal in a fairly large design class; and are highly efficient in the whole design class. The second question concerns that whether and how we can measure the goodness of a planned design in terms of precision loss due to subject dropouts. If the answer is yes, as guidance, it may benefit the practitioners to choose a "robust" design with minimal precision loss. In order to answer the second question, we calculate two quantities for a planned design: the expected precision loss and the maximum precision loss. It should be noted that all above results are derived under a model assuming the subject effect is fixed and errors are i.i.d. Besides this model, there are two more realistic models considered for this problem. One is the model with a random subject effect and i.i.d errors; and the other is the one with a fixed subject effect and autoregressive errors.

Contributed Session 7 Thursday, May 28, 3:30pm - 5:00pm

Analysis of Spatial and Temporal Data

Large Gaussian Covariance Matrix Estimation with Markov Structures

Xinwei Deng, *Georgia Institute of Technology*

Covariance matrix estimation for a large number of Gaussian random variables is a challenging yet increasingly common problem. A fact neglected in practice is that the random variables are frequently observed with certain temporal or spatial structures. Such a prob-

lem arises naturally in many practical situations with time series and images as the most popular and important examples. Effectively accounting for such structures not only results in more accurate estimation but also leads to models that are more interpretable. In this work, we propose shrinkage estimators of the covariance matrix specifically to address this issue. The proposed methods exploit sparsity in the inverse covariance matrix in a systematic fashion so that the estimate conforms to models of Markov structure and is amenable for subsequent stochastic modeling. The present approach complements the existing work in this direction that deals exclusively with temporal orders and provides a more general and flexible alternative to explore potential Markov properties. We show that the estimation procedure can be formulated as a semi-definite program and efficiently computed. We illustrate the merits of these methods through simulation and the analysis of a real data example.

Nonparametric Estimation of Variogram and its Spectrum

Chunfeng Huang, *Indiana University*

In the study of isotropic intrinsically stationary spatial processes, a new nonparametric variogram estimator is proposed through its spectral representation. The spectrum estimation is formulated in terms of solving a regularized inverse problem. A numerical implementation is presented through quadratic programming. We demonstrate our method in a simulation study and a dataset of temperature changes over America.

Comparing Generalized Quasi-likelihood Approaches to Longitudinally Correlated Simulated Poisson Counts

N. Mamode Khan* & V. Jowaheer, *University of Mauritius*

Counts are usually modeled using the Poisson distribution. In this paper, our focus is on the efficient estimation of regression parameters for a longitudinal Poisson regression model. So far, in literature, the Generalized estimating equations (GEE) were used to analyze such models but they may be yielding inefficient regression estimates as the correlation models may be misspecified. Recently, researchers have developed a general autocorrelation structure that can handle such limitations and can at the same time provide more efficient regression estimates than the existing GEEs. We are proposing another technique that does not require the estimation of any correlation parameters as they are often regarded as nuisance parameters and can still provide more efficient regression estimates than the existing methods. Initially, we will consider stationary longitudinal responses based on fixed time covariates and on the second instance we take into account Poisson regression models where the covariates are time-dependent. In this paper we will compare the techniques based on simulation of longitudinal Poisson counts that are generated through AR(1), MA(1) and Equicorrelation structures.

Using Markov Chain Monte Carlo to Solve a Time-varying State-space Model for Magnetoencephalography Inverse Problem

Zhigang Yao, *University of Pittsburgh*

Magnetoencephalography(MEG) is an imaging technique used to measure the magnetic signals outside the head produced by the electrical activity inside the brain. The MEG inverse problem, localizing the electrical source from the magnetic signal measurements, is ill-posed, that is, there are an infinite number of correct solutions. The common source localization methods assume the source does not vary with time and do not provide estimates of the variability of the fitted model. In this paper, we consider time-varying sources in the MEG inverse problem and model the time evolution using a state space model. Based on our new model, we investigate the inverse problem by finding the posterior source distribution given the multiple channels of observations at each time rather than fitting fixed source estimates. Our new model is more realistic than common models and estimates of the variation of the sources. The computational challenge happens when the number of states is very high and the slow convergence of the posterior distribution occurs when the source states are highly variable or the data likelihood is very flat. First, we change the regular Gibbs sampler by adding a conditional prior of the source proposal with a single/block move to find the posterior distribution for each state of the source. Second, to improve the computational efficiency and control the sampling variation for each state of the source, we develop an improved algorithm based on sequential importance sampling, where we use the local resampling scheme to avoid weight estimation for each source state. Third, to further improve the performance, we propose a dynamic sequential importance sampling with improved resampling; We throw away a significant amount of unnecessary weights and reduce dimension for each source state and also use a new weight function. These three algorithms are applied to simulated data, unlike the usual MCMC sampling scheme, these new methods converge.

Contributed Session 8 Friday, May 29, 8:30am - 10:00am

Analysis of Computer Experiments and Variable Selection

Fast Calibration of Complex Computer Models

Matthew T. Pratola, *Simon Fraser University*

Computer models enable scientists to investigate real-world phenomena in a virtual laboratory using computer experiments. Recently, statistical calibration enabled scientists to incorporate field data. However, the practical application is hardly straightforward. For instance, large and non-stationary computer model output is not well addressed. We present a computationally efficient approach using a criterion that measures discrepancy between the computer model and field data. One can then construct empirical distributions for the parameters and perform sequential design. The strength of this approach is its simple computation using existing algorithms. Our method also provides good parameter estimates

for large and non-stationary data.

Kriging Analysis of Thermal Data for a Data-center Experiment

Huizhi Xie*, Lulu Kang, C.F. Jeff Wu, Emad Samadiani & Yogendra Joshi, *Georgia Institute of Technology*

A data center is an integrated facility housing multiple-unit servers providing application services or management for data processing. The cooling mechanism of a data center is very crucial since the temperature of the facilities must be maintained at an acceptable level in order for them to function properly. To design an efficient cooling mechanism, computer simulations based on Computational Fluid Dynamics/Heat Transfer (CFD/HT) are used to study the relationship between the temperature and the factors like inlet velocity, heat load distribution of the racks and so on. They are, however, very time-consuming and do not support the continuous improvement of the design of data centers. Thus meta-models are desired as a compromise between computing speed and prediction accuracy. Proper Orthogonal Decomposition (POD) is a popular method to build meta-models in the mechanical engineering literature. However, it needs many observations and is not flexible in modeling the nonlinearity of the temperature with respect to the factors. To overcome these difficulties, we propose to use Kriging as the modeling approach. The merits of the proposed method are illustrated through the application to a data center problem.

Reanalysis of Some Experiments via the Dantzig Selector with the Consideration of Interactions

Frederick K.H. Phoa, *University of California, Los Angeles*

Fractional factorial designs are widely used experimental plans for cost reductions in scientific researches and investigations. Traditionally, Plackett-Burman and related designs have been used in screening experiments for identifying important main effects. However, there are plenty of practical situations where some interactions are significant and cannot be ignored. We reanalyze data from some chemical experiments via the Dantzig Selector with the consideration of interactions. We identify important interactions in each experiment and confirm these results with both a frequentist and a Bayesian approach. We demonstrate that ignoring interactions can both miss important factors and detect spurious factors.

Sparse Covariance Thresholding for High-Dimensional Variable Selection with the Lasso

Michael Yu Zhu*, Jessie Jeng & John Daye, *Purdue University*

An integral topic in statistical research is variable selection for high-dimensional applications, where the number of predictors p is large relative to the number of observations n. High-dimensional data often come from observational studies, in which predictors are considered to be random. One of the most popular methods for variable selection is the

lasso, which has been successful in many applications. However, under the 'large p small n' scenario, the lasso is not very satisfactory due to excessive variability and rank deficiency of the sample covariance matrix. These limitations can be mitigated when the covariance matrix is known to be sparse. Covariance sparsity is a natural assumption in many applications such as gene microarray analysis, image processing, etc., in which a large number of predictors are independent or weakly correlated with each other. In this paper, we propose to apply generalized covariance-thresholding to stabilize and improve the performances of the lasso under the covariance sparsity assumption. We call this procedure the covariance-thresholded lasso. We establish consistency results that relate the sparsity of the covariance matrix with variable selection and modify the LARS algorithm for our method. Finite sample performances are examined using simulation and real-data examples. Results indicate that our method can improve upon the lasso, adaptive lasso, and elastic net in prediction accuracy and variable selection for high-dimensional applications, particularly when $n \ll p$.

Contributed Session 9 Friday, May 29, 10:30am - 12:00pm

Statistical Methods in Networks, Finance and Pharmaceutical Industry

Statistical Learning for Networks

Hugh Chipman, *Acadia University*,

Data arising from networks can be rich and complex. For example, the 'header' information in email messages specifies a sender, recipient(s) and transaction time. Each email address can be viewed as a node in a graph, and email as transactions along the edges. The availability of large databases of such transactional network data presents challenging opportunities for the development of new models. I will discuss a specific model for email transactions, capable of grouping together individual on the basis of who they communicate with. Connections will be made with the mixed membership stochastic blockmodel of Airoldi et. al. (2008).

Computational and Statistical Aspects of Pricing Models

Gracie-Xiaojuan Ma* & Sergey Utev, *University of Nottingham*

Based on the analysis of the real market data, a simple and realistic model for share price is suggested. Monte Carlo simulations show that our model gives higher share option's payoff then the corresponding Black-Scholes model. Other properties such as the homogeneity and the embedding problem (a la Elfving, Kingman, Rosenthal and which seems to be an NP-hard problem) are considered for our data as well.

Stepwise confidence interval method for Identification of the Minimum Effective Dose

Xiaoqing Tang* & Jian Tao, *Shaoyang University*

Now we extend one method into a sequence of binomial data, propose a stepwise confidence

interval method for toxicity study and can identify a minimum effective dose. The first one is based on the well-known conditional confidence intervals for odds ratio, and the other one comes from Santner (1993), "small-sample confidence intervals for the difference of two success probabilities", and it produces exact intervals, through employing our method, all the declarations can be guaranteed to be correct with a probability higher than $100(1-\alpha)\%$. That is, the error rate is properly controlled.

Statistical Aspects of Intermediate Precision Studies

Balazs Banfai*, Katalin Ganzler & Sandor Kemeny, *Budapest University of Technology and Economics*

During validation of analytical methods the verification of precision is one of the most important steps. The levels of precision reflect the contribution of different sources to the overall variability. Repeatability expresses the analytical variability under identical conditions in a short interval of time. Intermediate precision contains further random effects within a laboratory, e.g. different days, operators or equipment. Reproducibility reflects the uncertainty between laboratories. In an intermediate precision study, the various uncertainty sources are evaluated in a random factor analysis of variance model. In these studies the degrees of freedom for the variance of repeatability are high enough to assess its suitability. The certain estimation of the variance for intermediate precision is harder, because the degrees of freedom can be low enough to cause a broad confidence interval. In a simple case with one source of variance (e.g. the effect of days), a one-way random model is used, usually with about 6 repetitions on 3 days, meaning df=15 for the repeatability but only df=2 for the intermediate precision. For the purpose of validation the repeatability and intermediate precision can be handled together. We can ask the question if the overall variability is in accordance with the requirements. In this case a conservative estimation is calculated for the variability and the confidence interval is compared to the allowed range. With this method the tolerable variability is assured in the case of acceptance. If the assessment is rejected, a more thorough investigation must be performed with elevated number of repetitions. The theoretical background of the different methods is discussed and the consequences are shown on analytical examples from the pharmaceutical industry. The principal role of the number of repetitions is discussed as well. The method is demonstrated on more complicated studies and additional sources of variability.

Index