2013 NCTS/ISRG Workshop on Recent Trends on Statistics
On Celebration of "2013: International Year of Statistics"

PROGRAM

Sunday, December 15, 2013

09:30 -10:00
Registration and Reception

10:00 -10:10
Opening Remarks: Dr. Winnie Li, Director of National Center of Theoretical Sciences

10:10 -12:10

Session I: SPC and Reliability
Chair: Ching-Shui Cheng, Academia Sinica

Speaker: N. Balakrishnan, McMaster University, Canada
"Differential Smoothing in the Bivariate Exponentially Weighted Moving Average Chart"

Speaker: T.H. Fan* and J. K Chou, National Central University
"Reliability Analysis of a Two-Component Series System with Bivariate Weibull Lifetime Distribution under Type-I Censoring Life Tests"

Speaker: Sheng-Tsaing Tseng*, Nan-Jung Hsu and Yi-Chiao Lin,
National Tsing-Hua University
"Joint Modeling of Laboratory and Field Data with Application to Warranty Prediction for Highly-Reliable Products"

12:10 – 13:30
Lunch

13:30 – 15:30

Session II: Modeling I
Chair: Henry H-S Lu, National Chiao Tung University

Speaker: Sang-Hoon Cho, Soongsil University, Seoul, Korea
"Correlated Gamma Variables under a Hierarchical Model Applied to

Statistical Inference in Microarray Experiments"

Speaker: Yi-Hau Chen, Academia Sinica

"Gene Selection for Survival Prediction under Dependent Censoring"

Speaker: Li-Shan Huang, National Tsing-Hua University

"Analysis of Deviance in Generalized Partial Linear Models"

15:30 – 15:50

Break

15:50 – 17:50

Session III : Modeling II
Chair: Shao-Wei Cheng, National Tsing Hua University

Speaker: Ming-Yen Cheng, National Taiwan University

"A New Test for One-Way ANOVA with Functional Data and Application to Ischemic Heart Screening"

Speaker: Shih-Hao Huang and Mong-Na Lo Huang*, National Sun Yat-Sen University

"Robust Designs for Probability Estimation in Binary Response Experiments"

Speaker: J.T. Gene Hwang, National Chung Cheng University

"Empirical Bayes Confidence Intervals for Individual Parameter among A Large Number of Parameters"

Monday, December 16

9:30 – 10:50

Session IV: Graphs and Edges
Chair: Regina Liu, Rutgers University

Speaker: Sahand Negahban, Yale University

"Iterative Ranking from Pair-wise Comparisons"

Speaker: Wei-Fang Niu and Henry Horng-Shing Lu*, National Chiao Tung University

"Community Detection by Hierarchical Graphical Modeling and Reversible Jump Markov Chain Monte Carlo"

10:50 – 11:10

Break

11:10-12:30

Session V: Challenges in Statistical Computing
Chair: Dennis Lin, Pennsylvania State University

Speaker: Jianjun Shi*, Georgia Institute of Technology
"An Adaptive Sampling Strategy for Online Monitoring of High-dimensional Streaming Data"
Speaker: Samuel Kou, Harvard University
"Statistical Computation in Protein Folding"

12:30 – 13:30
Lunch

13:30-15:30

Session VI: Design of Experiment & the Undesigned
Chair: Li-Shan Huang, National Tsing-Hua University

Speaker: Chia-Jung Chang, Pennsylvania State University
"Quantitative Characterization and Modeling Strategy of Nanoparticle Dispersion in Polymer Composites"
Speaker: Regina Liu, Rutgers University
"Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness"
Speaker: Dennis Lin, Penn State University
"*Big Data and Dimensional Analysis*"

ABSTRACTS

Differential Smoothing in the Bivariate Exponentially Weighted Moving Average Chart

N. Balakrishnan
Department of Mathematics and Statistics, McMaster University
bala@univmail.cis.mcmaster.ca

Abstract: The multivariate exponentially weighted moving average (MEWMA) control chart proposed by Lowry et al. (1992) has become one of the most widely used charts to monitor multivariate processes. Its simplicity, combined with its high sensitivity to small and moderate process mean jumps, is at the core of its appeal. Lowry et al. (1992) advocated equal smoothing of each quality variable unless there is an a priori reason to weight quality characteristics differently. However, one may have situations where differential smoothing may be justified. For instance: (a) departures in process mean may be different across quality variables, (b) some variables may evolve over time at a much different pace than other variables, and (c) the level of correlation between variables could vary substantially. Here, we assess the performance of the differentially smoothed MEWMA chart. The case of two quality variables (BEWMA) is discussed in detail. A bivariate Markov-chain method that uses conditional distributions is developed for average run-length (ARL) calculations. The proposed chart is shown to perform at least as well as Lowry et al. (1992)'s chart and noticeably better in many mean-jump directions. Comparisons with the recently introduced double smoothed BEWMA chart and the use of univariate charts for the independent case, it shows that the proposed differentially smoothed BEWMA chart has superior performance.

Quantitative Characterization and Modeling Strategy of Nanoparticle Dispersion in Polymer Composites

Chia-Jung Chang

Department of Industrial Engineering, Pennsylvania State University

cuc28@engr.psu.edu

Abstract: Nanoparticle dispersion plays a crucial role in the mechanical properties of polymer nanocomposites. TEM/SEM images are commonly used to represent nanoparticle dispersion without further quantifications on its properties. Therefore, there is a strong need to develop a quantitative measure to effectively describe nanoparticle dispersion from a TEM/SEM image. In

this talk, I will cover an effective modeling strategy to characterize nanoparticle dispersion states among different locations of a nanocomposite surface. An engineering-driven inhomogenous Poisson random field is proposed to represent the nanoparticle dispersion at nanoscale. The model parameters are estimated through the Bayesian MCMC technique to overcome the challenge of limited amount of accessible data due to the time consuming sample collection process. The TEM images taken from nano-silica/epoxy composites will be used to support the proposed methodology. The research strategy and framework are generally applicable to other nanocomposite materials.

Gene Selection for Survival Prediction under Dependent Censoring

Yi-Hau Chen
Institute of Statistical Science, Academia Sinica
yhchen@stat.sinica.edu.tw

Abstract: Dependent censoring arises in biomedical studies when the survival outcome of interest is censored by competing risks. In survival data with microarray gene expressions, gene selection or gene filtering based on the univariate Cox regression analyses has been used extensively in medical research, which however, is only valid under the independent censoring assumption. In this talk, we first consider a copula-based framework to investigate the bias caused by dependent censoring on univariate gene selection. Then, we utilize the copula-based dependence model to develop an alternative gene selection procedure. Simulations show that the proposed procedure outperforms the existing method when dependent censoring is indeed present. The non-small-cell lung cancer data is analyzed to demonstrate the usefulness of our proposal.

A New Test for One-Way ANOVA with Functional Data and Application to Ischemic Heart Screening

Ming-Yen Cheng
Department of Mathematics, National Taiwan University
cheng@math.ntu.edu.tw

Abstract: We propose and study a new global test for the one-way ANOVA problem in functional data analysis. The test statistic is taken as the maximum value of the usual pointwise F-test statistics over the interval the functional responses are observed. A nonparametric bootstrap method is employed to approximate the null distribution of the test statistic and to obtain an estimated critical value for the test. The asymptotic random expression of the test statistic is derived and the asymptotic power is studied. In particular, under mild conditions, the proposed test asymptotically has the correct level and is root-n consistent in detecting local alternatives. Via some simulation studies, it is found that in terms of both level accuracy and power, the proposed test outperforms the Globalized Pointwise F test of Zhang and Liang (2013) when the functional data are highly or moderately correlated, and its performance is comparable with the latter otherwise. An application to an ischemic heart real dataset suggests that, after proper manipulation, resting electrocardiogram signals can be used as an effective tool in clinical ischemic heart screening, without the need of further stress tests as in the current standard procedure.

Correlated gamma variables under a hierarchical model applied to statistical inference in microarray experiments

Sang-Hoon Cho
Department of Statistics and Actuarial Science, Soongsil University, Korea
sanghcho@ssu.ac.kr

Abstract: The recent advances in high-throughput genomic, transcriptomic and proteomic technologies have posed great challenges in statistical analysis. In microarray experiments, the dimension of gene variables p often far exceeds the sample size n. As a small number of replicates are available, dimension reduction is indispensable for statistical inference in detecting genes with altered expression among tens of thousands of candidates.
In this talk, we propose a simple two-stage hierarchical model in which the first stage accounts for gene-specific random departures from latent mean expression levels and the second stage accounts for correlated mean expression levels from a common random effect distribution. The hierarchical structure

together with the assumption of a gene-specific constant coefficient of variation facilitates parameter estimation, effectively reducing the parameter space in high dimension. Also, our proposed model accounts for the empirical distributions of gene expression that are marginally extreme with a jointly concentrated pattern in two dimensional display. For gene-specific inference, the empirical Bayes mode in closed form is derived based on the posterior distribution. By simulation, we test our approach and apply it to a few real data sets for illustration.

Reliability Analysis of a Two-Component Series System with Bivariate Weibull Lifetime Distribution under Type-I Censoring Life Tests

T.H. Fan and J. K Chou
Graduate Institute of Statistics, National Central University
thfan@stat.ncu.edu.tw

Abstract: A series system fails if any of the components fails. These components are all from the same system, hence they may be correlated. In this paper, we consider a series system of two components which has a bivariate Marshall-Olkin Weibull lifetime distribution under Type I censoring scheme in life testing. It is often to include masked data in which the component that causes failure of the system is not observed. We apply the maximum likelihood approach via EM algorithm along with the missing information principle to estimate the standard errors of the MLE. Statistical inference on the model parameters and the mean lifetimes and the reliability functions of the system as well as components are derived. The proposed method is confirmed by simulation study.

Statistical computation in protein folding

Samuel Kou
Department of Statistics, Harvard University
kou@stat.harvard.edu

Abstract: Predicting the native structure of a protein from its amino acid sequence is a long standing problem. A significant bottleneck of computational prediction is the lack of efficient sampling algorithms to explore the configuration space of a protein. In this talk we will introduce a sequential Monte Carlo method to address this challenge: fragment regrowth via energy-guided sequential sampling (FRESS). The FRESS algorithm combines statistical learning (namely, learning from the protein data bank) with sequential sampling to guide the computation, resulting in a fast and effective exploration of the configurations. We will illustrate the FRESS algorithm with both lattice protein model and real proteins.

Analysis of Deviance in Generalized Partial Linear Models

Li-Shan Huang
Institute of Statistics, National Tsing-Hua University
lhuang@stat.nthu.edu.tw

Abstract: We develop nonparametric analysis of deviance tools for generalized partial linear models based on local polynomial fitting. Assuming a canonical link, we propose expressions for both local and global analysis of deviance, which admit an additivity property that reduces to ANOVA decompositions in the Gaussian case. Chi-square tests based on integrated likelihood functions are proposed to formally test whether the nonparametric term is significant. Simulation results are shown to illustrate the proposed chi-square tests. The methodology is applied to German Bundesbank Federal Reserve data. This is a joint work with Professor Wolfgang Härdle at the Humboldt University.

Robust Designs for Probability Estimation in Binary Response Experiments

Shih-Hao Huang and Mong-Na Lo Huang
Department of Applied Mathematics, National Sun Yat-sen University
lomn@math.nsysu.edu.tw

Abstract: In this work, robust design problems for estimation of the response probability curve under binary response experiments with model uncertainty consideration is inves- tigated. A minimax type of model robust design criterion, called the Weighted Bias optimum (WB-optimum in short), is proposed based on minimization of the maximum of the weighted squared probability bias function under two rival mod- els. The corresponding design issues are investigated and results under the above design criterion for given rival models with several commonly seen symmetric links are presented.

Keywords: Compromise weight functions; Equal oscillation; Logit model; Probit model; Minimax optimality; Model discrimination; WB-optimum.

Empirical Bayes confidence intervals for individual parameter among a large number of parameters

J.T. Gene Hwang
Department of Mathematics, National Chung Cheng University
jtghwang@gmail.com

Abstract: In modern applications, we often involved in a small *n* and large *p* problems. Empirical Bayes approaches are becoming very important in such scenarios. To provide statistical inference, it is crucial to develop statistical intervals. We examined in the basic normal setting the well-known empirical Bayes interval of Carl Morris published in the 1983 issue of *JASA*. It has been argued by Morris that his interval has good Bayes coverage probabilities which are at least (1-α) with respect to any prior distribution in a certain class of normal distributions. Morris defined such an interval as (1-α) empirical Bayes confidence interval. It has been long thought that Morris interval is a (1-α) empirical Bayes confidence interval and its coverage probability has also been thought to be on the conservative side.

In this talk, we shall demonstrate some numerical studies showing that this is not so. We also show that a more basic version of Morris' interval has coverage probability converging to zero for certain variance settings as *p* becomes large, a result that shocked us. We discuss other intervals having good Bayes coverage

probabilities numerically and whose basic version do not seem to suffer this zero asymptotic Bayes coverage probability phenomenon.

This is joint work with Jia-Chiun Pan (Chung Cheng University) and Zhigen Zhao (Temple University).

BIG data and Dimensional Analysis

Dennis K.J. Lin
University Distinguished Professor
Department of Statistics, The Pennsylvania State University
dkl5@psu.edu

The ability to parse information, faster and deeper, is allowing researchers to understand the world in a way they could only dream about before. In this talk (first portion), the common understanding about BIG data will be discussed. Its challenges and impacts to Statistics will be addressed. In particular, we are interested in what types of Statistics are needed for BIG data era. On the other hand, Dimensional Analysis (DA) is a fundamental method in the engineering and physical sciences for analytically reducing the number of experimental variables prior to the experimentation. The method is of great generality. In this talk (second portion), an overview/introduction of DA will be given. Some initial ideas on using DA for BIG data will be discussed. Future research issues will be proposed.

Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness

Regina Liu
Department of Statistics, Rutgers University,
rliu@stat.rutgers.edu

Abstract: Meta-analysis has been widely used for synthesizing evidence from multiple studies for common hypotheses or parameters of interest. However, it has not yet been fully developed for incorporating heterogeneous studies, which arise often in applications due to different study designs, populations or outcomes. For heterogeneous studies, the parameter of interest may not be estimable for certain studies, and in such a case, these studies are typically excluded from conventional meta-analysis. The exclusion of part of the studies can lead to a non-negligible loss of information. We propose a meta-analysis approach for heterogeneous studies by combining confidence density functions derived from the individual studies, hence referred to as the CD approach. It includes all the studies in the analysis and makes use of all information, direct as well as indirect. Under a general likelihood inference framework, this new approach is shown to have several desirable properties, including: i) the CD approach is asymptotically as efficient as the maximum likelihood approach using individual participant data (IPD) from all the studies; ii) unlike the IPD analysis, it suffices to use summary statistics to carry out the CD approach. Individual-level data are not required; and iii) the CD approach is robust against misspecification of the working covariance structure of the parameter estimates. Besides its own theoretical significance, the last property also substantially broadens the applicability of the CD approach. These properties of the CD approach are also illustrated by data simulated from a randomized clinical trials setting and by aircraft landing data collected by the Federal Aviation Administration. Overall, one obtains an unifying approach for combining summary statistics, subsuming many of the existing meta-analysis methods as special cases.

This is joint work with Dungang Liu (Yale University) and Minge Xie (Rutgers University).

Community Detection by Hierarchical Graphical Modeling and Reversible Jump Markov Chain Monte Carlo

Wei-Fang Niu and Henry Horng-Shing Lu*
Institute of Statistics, National Chiao Tung University
henryhslu@gmail.com

Abstract: This study proposes a hierarchical graphical model for clustering under an unknown number of communities. This model is designed for sparse graphs in which the edge rate is low and only a few edges provide significant information about the community structure. In addition to ordinary communities having high within-community edge rates and low between-community edge rates, special communities, like hubs, are allowed and characterized by a random variable/vector. The proposed model provides a parsimonious approach for exploring the structure of the graph data. Since this is a trans-dimensional problem, the reversible jump Markov Chain Monte Carlo (MCMC) procedure is applied to the parameter estimation for the model. Synthetic and real datasets are used to illustrate the performance of the model and estimation method.

Iterative Ranking from Pair-wise Comparisons

Sahand Negahban
Department of Statistics, Yale University
sahand.negahban@yale.edu

Abstract: The question of aggregating pairwise comparisons to obtain a global ranking over a collection of objects has been of interest for a very long time: be it ranking of online gamers (e.g. MSR's TrueSkill system) and chess players, aggregating social opinions, or deciding which product to sell based on transactions. In most settings, in addition to obtaining ranking, finding 'scores' for each object (e.g. player's rating) is of interest for understanding the intensity of the preferences.

In this talk, we explore an iterative rank aggregation algorithm for discovering scores for objects (or items) from pairwise comparisons. The algorithm has a natural random walk interpretation over the graph of objects with an edge present between a pair of objects if they are compared; the scores turn out to be the stationary probability of this random walk. To understand the efficacy of the method we consider the popular Bradley-Terry- Luce (BTL) model in which each object has an associated score which determines the probabilistic

outcomes of pairwise comparisons between objects. We bound the finite sample error rates between the scores assumed by the BTL model and those estimated by our algorithm. In particular, the number of samples required to learn the score well with high probability depends on the structure of comparison graph – when the Laplacian of the comparison graph has constant spectral gap, e.g. pairs chosen at random for comparison, this leads to order-optimal dependence on the number of samples. We demonstrate our theoretical results through experimental verification

An adaptive sampling strategy for online monitoring of high-dimensional streaming data

Jianjun Shi
Department of ISYE, Georgia Institute of Technology
jianjun.shi@isye.gatech.edu

Abstract: With the rapid development of sensing technology, sensors that generate real-time high-dimensional streaming data have been widely used in a variety of manufacturing systems. However, effective monitoring of such data is a challenging task for quality improvement due to the high computational cost, complex fault patterns, small signal-to-noise ratio, and resource constraints. Currently, there is a lack of efficient statistical process control (SPC) methodologies to address these challenges.

This research aimed at developing scalable and adaptive methodologies for online monitoring of high-dimensional streaming data. Specifically, we propose a systematic and scalable adaptive sampling strategy with only partial information available at each acquisition time for online high-dimensional process monitoring. The developed adaptive sampling strategy discussed in this article includes a broad scope of applications: (1) when only a limited number of sensors is available; (2) when only a limited number of sensors can be in "ON" state in a fully deployed sensor network; and (3) when only partial data streams can be analyzed at the fusion center due to limited transmission and processing capabilities even though the full data streams have been acquired remotely. A monitoring scheme of using the sum of top-r local statistics is developed and named as "TRAS" (Top-r based Adaptive

Sampling), which is scalable and robust in detecting a wide range of possible shifts in all directions. Two properties of this proposed method are also investigated. Case studies are performed on a hot forming process and a real solar flare process to illustrate and evaluate the performance of the proposed method.

This is joint work of Kaibo Liu at University of Wisconsin-Madison, and Yajun Mei at Georgia Institute of Technology.

Joint Modeling of Laboratory and Field Data with Application to Warranty Prediction for Highly-Reliable Products

Sheng-Tsaing Tseng, Nan-Jung Hsu, Nan-Jung Hsu and Yi-Chiao Lin
Institute of Statistics, National Tsing-Hua University
sttseng@stat.nthu.edu.tw

Abstract: Warranty policy is a useful tool for the manufacturers to compete with others. When a product fails during the warranty period, manufacturers shall provide their consumers with free charge service for the products. Therefore, the return rate prediction during the warranty period is an important decision problem for manufacturers. Recently, several research works focus on using the laboratory testing data of the product to predict its return rate during the warranty period. However, the results are very restricted to a single-product with continuous measurements for laboratory data and there is no suitable approach for the case of multiple-product with progressive-stress "go/ no go" laboratory data. Especially when the laboratory test has very limited resulting in few or even no failures, this study proposes a "Hierarchical Structure" together with empirical Bayesian inference to incorporate all failure information so that the parameters in the laboratory failure model can be estimated efficiently. The advantage of the proposed method is that even for zero-failure data, this method still provides us a robust estimation for the process parameters, and gives us a precise prediction on the return rate.