Joint Meeting of

The 2011 Taipei International Statistical Symposium and

7th Conference of the Asian Regional Section of the IASC

7th IASC-ARS

Σ joint 2011

Taipei Symposium

December 16 - 19, 2011

Academia Sinica, Taipei, Taiwan

http://joint2011.stat.sinica.edu.tw/

中央研究院 ACADEMIA SINICA 1928

Abstract

December 16 (Friday)

Today's Highlights:

09:00 – 10:20	***Tutorial (I) by David O. Siegmund***
10:40 – 12:00	***Tutorial (II) by David O. Siegmund***
13:00 – 13:20	***Opening Ceremony***
13:20 – 14:20	***Keynote Speech (I) by Peter Hall***
14:40 – 16:10	***Parallel Sessions 16a1 – 16a7***
16:30 – 18:00	***Parallel Sessions 16b1 – 16b7***

Abstract Pages:

Special Sessions:	Keynote Speech (I) Page 59

Session No.	16a1	16a2	16a3	16a4	16a5	16a6	16a7
Starting Page	59	61	62	63	65	68	70

Session No.	16b1	16b2	16b3	16b4	16b5	16b6	16b7
Starting Page	73	75	77	79	81	83	85

Keynote Speech (I)
December 16 (Friday), 13:20 - 14:20, HSS Center International Conference Hall
Speaker: Peter Hall
Chair: Genshiro Kitagawa

Isotonic Nonparametric Regression in the Presence of Measurement Error

Raymond J. Carroll
Aurore Delaigle
Peter Hall
Faculty of Science Department of Mathematics and Statistics, University of Melbourne, Australia

In a great many regression problems the explanatory variable, X, represents the value taken by a treatment, for example a dosage, and the conditional mean of the response, Y, is anticipated to be a monotone function of X. Indeed, if this regression mean is not monotone (in the appropriate direction) then the medical or commercial value of the treatment is likely to be significantly curtailed, at least for values of X that lie beyond the point at which monotonicity fails. Addressing these problems requires a method for testing the hypothesis that the regression mean is monotone, and, if the conclusion of the test is positive, a technique for estimating the mean response subject to the constraint that it is monotone. Methodology for solving these problems already exists, but it ignores the potential for errors in measuring X. In this talk we outline an approach that accommodates those errors, using statistical tilting.

[Peter Hall, Faculty of Science, Department of Mathematics and Statistics, University of Melbourne, Australia; halpstat@ms.unimelb.edu.au]

16a1-Innovative Statistical Methods for Translational Research in Clinical Trials
December 16 (Friday), 14:40-16:10, HSS 1st Conference Room
Organizer: Chen-Hsin Chen
Chair: Jen-Pei Liu

16a1-1 Sample Size Estimation in Phase III Clinical Trials

Kuang-Kuo G Lan
Janssen Pharmaceutical Companies of Johnson & Johnson Raritan, NJ, U.S.A.

The standard method for moving from estimates of effect size obtained in a Phase II trial to the sample size of a Phase III trial usually ignores the variability inherent in the point estimate from Phase II. Use of a realistic prior distribution instead of a fixed alternative is likely to increase the

sample size required for a Phase III trial. For example, the use of a normally distributed prior to express that uncertainty can lead to power that does not approach one as the sample size approaches infinity.

We examine normal, truncated normal, and gamma priors for treatment effect in Phase III studies, and demonstrate analytically an approach to approximating the power for a truncated normal prior. We also propose a simple compromise method that requires moderately larger sample size than the one derived from the fixed method.

[Kuang-Kuo Gordon Lan, 920 Route 202 Raritan, NJ 08869, USA; glan@its.jnj.com]

16a1-2 Right Medicines for Right Patients

Christy Chuang-Stein
Pfizer Inc, Kalamazoo Michigan, U.S.A.

New technologies and scientific discoveries are changing our strategies for disease management. Over the past 15 years, significant progress has been made in developing targeted therapies that effectively treat sub-populations of patients, based on the molecular characteristics of the target. These advancements have helped us enter the era of personalized medicines, with oncology leading the way. In this talk, I will discuss some recent stories in developing personalized medicines for non-small cell lung cancer. I will also highlight statistical challenges in our journeys to these successful medicines.

[Christy Chuang-Stein, 5857 Stoney Brook, Kalamazoo Michigan 49009, U.S.A.; Christy.j.chuang-stein@pfizer.com]

16a1-3 Sequential Treatment Selection in Early Phase Clinical Trials using a Prognostic Biomarker: A Simple SPRT Approach

Ying Kuen Cheung
Department of Biostatistics, Mailman School of Public Health, Columbia University, U.S.A.

When a prognostic biomarker that can be quickly assessed is available, frequent interim monitoring of a trial is ethically and practically appealing for clinicians. In this talk, I will focus on early phase treatment screening trial in which the objective is to select a treatment with a practically significant improvement upon an active control group, or to declare futility if no such treatment exists. I will propose a class of sequential selection boundaries that are easy to implement in a blinded fashion, and can be applied on a flexible and frequent monitoring schedule in terms of calendar time. Design calibration with respect to pre-specified levels of confidence is simple, and can be accomplished when the response rate of the control group is known only up to an interval.

[Ying Kuen (Ken) Cheung, Department of Biostatistics, Mailman School of Public Health, Columbia University, U.S.A.; yc632@columbia.edu]

16a2-Biostatistics & Other
December 16 (Friday), 14:40 - 16:10, HSS 2nd Conference Room
Organizer: Lianwen Zhao
Chair: Lianwen Zhao

16a2-1 Oracle Inequalities and Model Selection

Lianwen Zhao
Dept.of Statistics, Southwest Jiaotong University, Chengdu, China

Most of the works on model selection are based on a specific particular problem, such as fixed-design regression. In this talk the model selection problems are described by using a general contrast function. We investigate the conditions on these contrast functions suffice to derive oracle inequalities. By employing the empirical processes and concentration inequalities, the oracle risk bounds are also derived.

[Lianwen Zhao, Department of statistics, College of Mathematics, Southwest Jiaotong University, 610031 Chengdu, China; lwzhao@home.swjtu.edu.cn]

16a2-2 Bayes Lasso for Detecting Gene-gene Interaction in Case-control Study

Haitao Zheng
Qifa Yan
Southwest Jiaotong University, Chengdu, China

Method for detecting Gene-gene and gene environment interaction is not rich in GWAS based on SNP analysis, especially when many SNPs are involved in the model. We propose to use Bayes Lasso method to find those effects in Case-control study and the new approach is examined via simulation study. We also apply the method to a real data.

[Haitao Zheng, Southwest Jiaotong University, Chengdu, China; htzheng@gmail.com]

16a2-3 Adaptive Estimation and Model Selection in Stationary Sequence

Cheng Liu
Dept.of Statistics,Southwest Jiaotong University,Chengdu, China

The subject of this talk is autoregression and moving average modeling of a stationary, Gaussian discrete time processes, based on a finite sequence of observations. We adopt the nonparametric minimax framework and study how well the processes can be approximated by a finite order ARMA model. The effectiveness of the model selection is expressed by an oracle inequality, which compares the performance of the selected to that of the best possible choice.

[Cheng Liu, Department of statistics, College of Mathematics, Southwest Jiao-tong University, 610031, Chengdu, China; lccelia@126.com]

16a3-Applied Probability
December 16 (Friday), 14:40 - 16:10, HSS Media Conference Room
Organizer: Ting-Li Chen
Chair: Ting-Li Chen

16a3-1 On the Optimal Transition Matrix for MCMC Sampling

Ting-Li Chen
Wei-Kuo Chen
Chii-Ruey Hwang
Academia Sinica, Taiwan, R.O.C.

Hui-Ming Pai
National Taipei University, Taiwan, R.O.C.

Let X be a finite space and πßbe an underlying probability on X. For any real-valued function f defined on X, we are interested in calculating the expectation of f under π. Let X_0 , X_1 , ⋯, X_n , ⋯ be a Markov chain generated by some transition matrix P with invariant distribution πβ. The time average, $1/n\sum_{k=0}^{k=n-1} f(X_k)$, is a reasonable approximation to the expectation, E_π [$f(X)$]. Which matrix P minimizes the asymptotic variance of $1/n\sum_{k=0}^{n-1} f(X_k)$? The answer depends on f. Rather than a worst-case analysis, we will identify the set of P 's that minimize the average asymptotic variance, averaged with respect to a uniform distribution on f.

[Hui-Ming Pai, Department of Statistics, National Taipei University, Taiwan, R.O.C.; hpai@mail.ntpu.edu.tw]

16a3-2 On the Stochastic Heat Equations

Shang-Yuan Shiu
Institute of Mathematics, Academia Sinica, Taiwan, R.O.C.

Stochastic partial differential equations (S.P.D.E.) were introduced by J. Walsh in 1986. Nowadays, S.P.D.E. is a hot topic in probability. We consider the heat equation and add a perturbation which is a space-time white noise. We will discuss the moment estimations, fluctuation and intermittency. Results of more general setting will be given.

[Shang-Yuan Shiu, Institute of Mathematics, Academia Sinica, Taiwan, R.O.C.; shiu@math.sinica.edu.tw]

16a3-3 Generative Models for Image Analysis: Data Likelihood versus Feature Likelihood

Lo-Bin Chang
National Chiao Tung University, Taiwan, R.O.C.

A probabilistic grammar for the grouping and labeling of parts and objects, when taken together with pose and part-dependent appearance models, constitutes a generative scene model and a Bayesian framework for image analysis. This talk will focus on constructing generative scene models for the appearances of parts. I will propose a principle to model data distributions and propose a likelihood technique to learn the parameters of the models with a discussion on data likelihood and feature likelihood. To demonstrate the utility of the models, I will provide some experiments on sampling and image classification.

[Lo-Bin Chang, Department of Applied Mathematics, National Chiao Tung University, Hsinchu, Taiwan 300; R.O.C. lobin chang@math.nctu.edu.tw]

16a4-Genomics and Personalized Medicine
December 16 (Friday), 14:40 - 16:10, AC 1st Conference Room
Organizer: James J. Chen
Chair: Huey-Miin Hsueh

16a4-1 Application of Principal Component Analysis on Evaluation of Gene Signature's Clinical Association

Dung-Tsa Chen
Moffitt Cancer Center, U.S.A.

Ying-Lin Hsu
Department of Applied Mathematics and Institute of Statistics, National Chung Hsing University, Taiwan, R.O.C.

Evaluation of a gene signature is a challenging task. One important issue is how to integrate all the genes as a whole and develop an index score to represent the signature in order to test its association with clinical outcomes. Principal component analysis has been widely used in genome-wide expression studies. In this study, we highlight the state of the art of principal component analysis and demonstrate that the use of the first principal component is feasible to evaluate a gene signature for its clinical association. Examples of a set of cancer data are used for illustration to show the derived index score by the first principal component is effective to correlate with clinical outcomes (e.g., overall survival).

[Dung-Tsa Chen, 12902 Magnolia Drive Tampa, FL 33612, U.S.A.; Dung-Tsa.Chen@moffitt.org]

16a4-2 A Powerful Association Test of Rare Variants Using a Random-Effects Model

Kuang-Fu Cheng
Biostatistics Center, China Medical University, Taichung, Taiwan, R.O.C.

J.Y. Lee
Graduate Institute of Statistics, National Central University, Chungli, Taiwan, R.O.C.

Chun Li
Center for Human Genetics Research and Department of Biostatistics, Vanderbilt University, Nashville, U.S.A.

There is an emerging interest in sequencing-based association studies of multiple rare variants. Most association tests suggested in the literature involve collapsing rare variants with or without weighting. These tests often have good power performance when there are many causal variants in the analysis. However, the power of these tests can be dramatically reduced when many non-causal variants are included or when the range of variant frequency is wide with varied effect sizes. Recently, a variance-component score test, SKAT, was proposed to test for association between rare and common variants in a region and disease status that addresses these limitations. Although SKAT was shown to outperform most of the alternative rare-variant association tests, its applications and power might be restricted and influenced due to missing genotypes. As missing genotypes are common in sequencing studies, it is desirable to have a test that is not only more powerful than the alternative tests but also robust to the presence of missing genotypes. In this paper, we suggest a new method based on testing whether the fraction of causal variants in a region is zero. This can be achieved by using a random-effects model and a mixture distribution for the effect sizes. The new association test, T_{REM} , is based on a likelihood ratio statistic and a permutation procedure for p-value calculation. T_{REM} is easy to compute and allows for missing genotypes. We performed null

simulations with small to large number of non-causal rare variants and various genotype missing rates to study the type I error rates of the competing tests. Simulations with small to large proportion of non-causal variants in a region were also given to study power performance of the tests. Further, the impacts of the common variants, effect directionality, and missing genotypes were extensively investigated to study the robustness of the competing tests. The simulation results showed that T_{REM} was a valid test in terms of controlling type I error and had better power performance. In particular, T_{REM} was less sensitive to the inclusion of non-causal rare, common variants, or missing genotypes. When the effects were more consistent in the same direction, T_{REM} also had better power performance. Finally, an application to the Shanghai breast cancer study showed that rare variants at the FGFR2 gene were detected to be associated with the breast cancer by T_{REM} and SKAT, but not by other tests. However the results from the analysis for various sets of rare and common variants in this gene were more consistent, if T_{REM} was applied. This indicates that T_{REM} is a more robust test.

[Kuang-Fu Cheng, Biostatistics Center, School of Public Health, China Medical University, Taichung, Taiwan, R.O.C.; kfcheng@mail.cmu.edu.tw]

16a5-Machine Learning
December 16 (Friday), 14:40 - 16:10, AC 2nd Conference Room
Chair: Hong-Wei Chuang

16a5-1 Effect of Unlabeled Data and Labeling Strategy on Error Rate in Linear Discriminant Analysis

Keiji Takai
Data Mining Laboratory, Kansai University, Osaka, Japan

Kenichi Hayashi
Graduate School of Medicine, Osaka University, Osaka, Japan

It is widely said that partially labeled data analysis, the analysis based on labeled data with unlabeled data, improves the estimation precision or the classification performance. There are situations where we can choose some cases to be labeled in the unlabeled data. Then, two questions arise: (1) do/how unlabeled data contribute the label prediction? (2) are there better strategies to determine cases to be labeled? Our study investigates them in linear discriminant analysis. For practical comparison, we defined an efficiency of the error rate of complete-case analysis compared to that of available-case analysis in a similar way to O'Neil [J.Amer.Stat. Assoc.73(1978):821–826] and evaluated it for randomly labeled data and nonrandomly labeled data, respectively. Next, we compared available-case analysis with randomly labeled data to that with nonrandomly labeled data. The results imply that employing labeled data affects positively on prediction but the labeling strategy should be considered to achieve significant improvement.

[Keiji Takai, Kansai University 3-3-35, Yamate, Suita, Osaka 564-868 Japan; takaikeiji@gmail.com]

16a5-2 Visualization Tools for Feature Extraction of Order Fluctuations

Tomokazu Fujino
Fukuoka Women's University, Fukuoka, Japan

Yoshiro Yamamoto
Tokai University, Kanagawa, Japan

In this study, we propose statistical graphics for finding relationship of order fluctuation between two or more variables to the other factors based on the parallel coordinate plot. In addition, we develop the Web-based software for displaying and handling such graphics with interactive facilities, which can be generated by R scripts. After that we show some examples of practical data analysis using the software. This graphics can be used in the process of exploratory data analysis, especially in the marketing field.

[Tomokazu Fujino, 1-1-1, Kasumigaoka, Higashi-ku, Fukuoka, Japan; fujino@fwu.ac.jp]

[Yoshiro Yamamoto, 4-1-1, Kitakaname, Hiratsuka, Kanagawa, Japan; yama@tokai-u.jp]

16a5-3 Text Mining Application to Internet Shopping Mall Customers' Reviews

Seok-Won Oh
Korea University, Chungnam, Korea

Seohoon Jin
Korea University, Chungnam, Korea

Usually text data doesn't allow us to apply general statistical analysis because of its atypical structure. Therefore preprocessing for structuralizing is needed to analyze text data. In this study, we applied text mining technique including preprocessing to internet shopping mall customers' reviews. Since term-document matrix has very high dimensionality as well as lots of zeros, singular value decomposition is used for reducing dimensionality of the data. SNA (Social Network Analysis) techniques are adopted for figuring out relationship between terms. Besides, customers' reviews are clustered by text clustering technique. Each cluster is characterized by representative terms. Clustering results can be coupled with customers' demographic information so as to understand customer's preference by demographic characteristics.

[Seohoon Jin, Department of Informational Statistics, Korea University, Jochiwoneup, Yeongi-gun, Cuhingnam, 339-700, Korea; seohoon@korea.ac.kr]

16a5-4 Analysis of Influence Pattern at Training Samples in Discrimination

Kuniyoshi Hayashi
Hiroshi Suito
Koji Kurihara
Graduate School of Environmental Science, Okayama University, Okayama, Japan

Diagnostics in statistical discriminant analysis have been proposed and studied in the field of statistics. The influence functions for discriminant score and misclassification probability have an important role in their assessments. On diagnostics in discriminant methods, there are two aspects. One is to detect large influential training samples for the result of analysis and the other is to detect the training samples that are directly connected with the improvement of prediction accuracy. Particularly, using the direction of perturbation at a training sample, we proposed an evaluation method for prediction. In addition, we have developed the unified framework on single-case and multiple-case diagnostics. In this study, we assume that there are mislabeled influential samples in training data. For such cases, we apply our approach to a discriminant method in pattern recognition and show the performance.

[Kuniyoshi Hayashi, 3-1-1 Tsushima-naka, Kita-ku, Okayama City, Okayama 700-8530, Japan; k-hayashi@ems.okayama-u.ac.jp]

16a5-5 Text Mining with Extraction of Similar Expression Patterns by Using Signed Bipartite Graph

M. Kitano
H. Yadohisa
Doshisha University, Kyoto, Japan.

One of the purposes of text mining is extracting different expressions with a same meaning as similar expression patterns. Some methods have been proposed to extract them. For example, Ueno, et al. [IPSJ. SIG. TR.1 (2004):169–176] considers words of similar co-occurrence relations as similar expression patterns. This method extracts subgraphs with high edge density from a bipartite graph that represents case frames which describes word dependencies on the verb. However, in existing methods, antonyms, which have similar co-occurrence relations and an opposite polarity, are regarded as similar expression patterns. For polarity of the word, Nasukawa and Kaneyama [IPSJ. SIG. TR.73 (2004):109–116] determines polarity by using context coherence. But, in this method, similar expression patterns cannot be considered. In this paper, we propose a method for extracting the similar expression patterns considering the polarity of the word. Specifically, polarized case frames are represented as a signed bipartite graph. Additionally, a new edge density index for extracting signed subgraphs is defined. In our approach, a polarity corresponding to the

case frames can be attached to the similar expression patterns and then, the patterns can be extracted with more informative concepts.

[Michiharu Kitano, Guraduate School of Culture and Information Science, Doshisha University, Kyoto, 610-0394, Japan.; dil0017@mail4.doshisha.ac.jp]

16a6-Clinical Trials (I)
December 16 (Friday), 14:40 - 16:10, AC 3rd Conference Room
Chair: Chi-Sheng Chang

16a6-1 **Semiparametric Stochastic Modeling for Epidemic Data**

Chia-Hui Huang
Institute of Statistical Science, Academia Sinica, Taiwan, R.O.C.

Epidemic models and statistical tools are developed to study the underlying mechanisms of the spread of an infectious disease. This development is motivated by the invasive Methicillin-resistant Staphylococcus Aureus (MRSA) infections in Children's Hospital intensive care units study, in which one of the aims is to find out whether or not hospital staff may become carriers in the transmission of infectious diseases. We propose a new approach for data arising from such a situation when there is a large number of independent small groups of correlated event times, and previous event occurrence may become part of risk factors for subsequent event occurrence. The latter makes the usual marginal models and frailty models not applicable. A dynamic hazards function is built to model the risk of susceptible individuals contracting a disease based on a data-driven approach. The regression parameters in this model consist of two parts, one relates how the hazards varies in response to the individual's explanatory variables in a multiplicate scale and the other one is the relative risk of being exposed to failures within its own cluster. Under this setup, the estimation of covariate effects and standard errors are carried out using a martingale approach. Related hypothesis testing on the contact effect is also developed, extensive simulation studies are conducted to access the performance of the proposed methods. Particular attention is paid to potential bias, which may be caused by discretized of failure times. The data set from Columbia University Children's Hospital is analyzed for MRSA which is caused by bacterial infection. And a summary report and conclusions are provided.

[Chia-Hui Huang, Institute of Statistical Science, Academia Sinica, Taiwan, R.O.C.; chuang2342 @stat.sinica.edu.tw]

16a6-2 **Sample Size Re-estimation without Breaking the Blind in Clinical Trial**

Chien-Hua Wu
Chung-Yuan Christian University, Taiwan, R.O.C.

Shu-Mei Wan
Lunghwa University of Science and Technology, Taiwan, R.O.C.

From a regulatory perspective it is important that the sample size recalculation is performed such that all patients involved in the study remain blinded and that the procedure does not inflating the type I error rate. Gould and Shih (1992) uses expectation-maximization (EM) algorithm to estimate individual group means and estimate variances for continuous case, but some controversy over appropriateness of EM by Friede and Kieser (2002, 2005) and Gould and Shih (2005). Given the controversy over appropriateness of EM algorithm, it would be useful to have an alternative approach that is simple to use yet versatile enough for a broad range of models. The proposed procedure is such tool to produce the sample sizes in the middle of the trial without breaking the blind. The proposed model is equivalent to the randomized response model introduced by Shih and Zhao (1997) if we have dichotomized outcomes of a two-arms study. It is also suitable to estimate the means and variances for continuous cases in three-arms study without the normality assumption. Simulation experiments are conducted to evaluate our proposed method. The unblind estimators for the population mean and population variance have smaller variances than that of the blind estimators in terms of bias and MSE's. The accuracies of means and variances depend on the proportion of a specific treatment for each stratum.

[Chien-Hua Wu, Chung-Yuan Chirstian University, Taiwan, R.O.C.; cwu@cycu.edu.tw]

16a6-3 **Futility Stopping in Clinical Trials: Theory and Applications**

Olivia Y. Liao
Pei He
Tze Leung Lai
Stanford University, Stanford, U.S.A.

Early stopping due to futility, or go/no-go decisions, during interim analysis has become an important feature of clinical trial designs. Current methods for futility stopping in the literature are mostly based on conditional power in conjunction with the theory of stochastic curtailment. They have certain drawbacks that have been noted in the literature. Herein we discuss a new approach of futility stopping in clinical trial designs that can overcome these difficulties, such as issues of maximum sample size and the unknown alternative. It protects type I error while not losing too much power. Simulation studies and theoretical analysis show the advantages of the approach in parametric, semi-parametric, and nonparametric problems, and in fixed sample size, group sequential, and time-sequential designs.

[Tze Leung Lai, 390 Serra Mall Sequoia Hall, Stanford University, Stanford CA94305, U.S.A.; lait@stanford.edu]

16a6-4 A Consistency Approach to Evaluation of Biosimilar Products

Hsiao-Hui Tsou
Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Miaoli County, Taiwan, R.O.C.

Recently, biosimilars have attracted much attention from sponsors and regulatory authorities while patients on early biological products will soon expire in the next few years. The European Medicines Agency (EMA) of the European Union (EU) has published a guideline on similar biological medicinal products for approval of these products [*Available at: http://www.ema.europa.eu/docs/en GB/ document library/Scientific guideline/2009/09/WC500003517.pdf*]. Based on the foundational principles of EMA guideline, biosimilars are expected to be similar, not identical, to the innovator biologics they seek to copy. In this paper, we develop a consistency approach for assessment of similarity between a biosimilar product and the innovator biologic. Method for sample size determination for conducting a clinical trial to assess the biosimilar product is also proposed. A numerical example is given to illustrate applications of the proposed approach in different scenarios.

[Hsiao-Hui Tsou, Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, 35 Keyan Road, Zhunan, Miaoli County 350, Taiwan, R.O.C.; tsouhh@nhri.org.tw]

16a7-Mathematical Statistics
December 16 (Friday), 14:40 - 16:10, AC 4th Conference Room
Chair: Hung-Yi Lu

16a7-1 Abstract Tube Associated with a Perturbed Polyhedron and Multidimensional Normal Probability Calculation

Satoshi Kuriki
The Institute of Statistical Mathematics, Tokyo, Japan

Tetsuhisa Miwa
National Institute for Agro-Environmental Sciences, Tsukuba, Japan

Anthony J. Hayter
University of Denver, Denver, U.S.A.

Let K be a closed convex polyhedron defined by a finite number of linear inequalities. In this study, we investigate abstract tubes (Naiman and Wynn, 1997, Ann. Statist.) associated with K. In particular, we focus on the case when K is perturbed lexicographically in outer direction.

An algorithm for constructing the abstract tube by means of linear programming and its implementation are discussed. Using the abstract tube for perturbed K combined with the recursive integration technique devised by Miwa, Hayter and Kuriki (2003, JRSS, B), we show that the multidimensional normal probability for a polyhedral region K can be efficiently calculated. Abstract tubes and distribution functions for the studentized range statistics are exhibited as numerical examples.

[Satoshi Kuriki, The Institute of Statistical Mathematics, 10-3 Midoricho, Tachikawa, Tokyo 190-8562, Japan; kuriki@ism.ac.jp]

16a7-2 Test Statistics, Orbits, and Optimization

Hsih-Chia Hsieh
Hsing-Kuo University, Taiwan, R.O.C.

Pei-Gin Hsieh
National Chung-Cheng University, Taiwan, R.O.C.

Thomas-Sun Hsieh
Sun-Yat Sen University, Taiwan, R.O.C.

The objective is to optimize a region; general equilibrium policy is the unique smooth orbit of global and local consistency, transforming the correlation, controlling and altering the direction of potential natural forces. Method is that dynamic quadratic regression (or non-deterministic polynomials) yields general equilibrium and spectrum. Outcomes are that the data accuracy may follow the Gaussian distribution, $x(t) \sim N(0, \sigma)$, or the non-Gaussian distribution, $x(t) \sim N(x^*, 0)$ at time t. Equilibrium x^* is the turning point, the minimum and maximum in policy and welfare effect size, and is located for all time or ages. Conflicts and errors vanish if the convergence probability $p \rightarrow 1$ is the criterion of acceptability, controllability, and incentive-compatibility for unique budget coverage, cooperation and competition in interdiscipline problems. The target is that equilibrium is the smooth orbit of flows for long-time existence in the interaction of n-codimensional spheres. The curse of dimensionality (D) is the small sample size with a large number of conflicts or unknown parameters, $n < t$. The equilibrium solution is the prediction and a unique, unbiased, consistent, and efficient test statistic, and is a boundary and fractional optimization, classifying normality and irrationality.

[Hsih-Chia Hsieh, Hsing-Kuo University, R.O.C.; hsihchia.hsieh@gmail.com]

[Pei-Gin Hsieh, National Chung-Cheng University, R.O.C.; actpgh@ccu.edu.tw]

[Thomas-Sun Hsieh, Sun-Yat Sen University, R.O.C.; actpgh@ccu.edu.tw]

16a7-3 On the Two-step Estimator for the Shape Parameter of the Gamma Distribution

Yoshiji Takagi
Nara University of Education, Nara, Japan

Our purpose is to construct some desirable and practical estimators for the shape parameter of the gamma distribution. Takagi [Statis. Prob. Letter (in press)] shows that the maximum likelihood estimator (MLE) is second-order inadmissible under every loss function, and provides a method of constructing some second-order admissible estimators by the bias-adjustment of the MLE under a given loss function. However, the likelihood equation involves the gamma function and its logarithmic derivative, so that it is not so easy to obtain the maximum likelihood estimate numerically. So, we introduce a twostep estimator which is asymptotically equivalent to the MLE up to the second-order and apply the method of construcing second-order admissibile estimators in Takagi by substituting the two-step estimator instead of the MLE.

[Yoshiji Takagi, Takabatake-cho, Nara, Japan; takagi@nara-edu.ac.jp]

16a7-4 Predicting Survival Outcomes Based on Compound Covariate Method under Cox Proportional Hazard Models with Microarrays

Takeshi Emura
Graduate Institute of Statistics, National Central University, Taiwan, R.O.C.

Predicting survival outcomes from high-dimensional microarray data, e.g., gene expression values, is a current focus of statistical and medical research. We study a methodology based on the compound covariate method, which is performed under univariate Cox proportional hazard models. The compound covariate method is a simple method and has been used in microarray studies. However, few papers discuss its statistical properties and comparative performance with existing methods. We demonstrate via simulations and real data analysis that the compound covariate method generally competes well with ridge regression and Lasso methods, well known and well-studied methods for predicting survival outcomes with microarrays. Furthermore, we propose to refine the compound covariate method by incorporating multivariate likelihood information under multivariate Cox models. The new proposal borrows information contained in both the univariate and multivariate Cox regression estimators, and it can further improve the ability of the compounding covariate method. We show that the new proposal has a theoretical justification from a statistical large sample theory, and it is naturally interpreted as a Shrinkage-type estimator, a popular class of estimators in statistical literature. Two microarray dataset, the non-small-cell lung

cancer data and the Dutch breast cancer data, are used for illustration. This is joint work with Dr. Yi-Hau Chen and Hsuan-Yu Chen from Institute of Statistical Science, Academia Sinica.

[Takeshi Emura, Graduate Institute of Statistics National Central University, Taiwan, R.O.C.; emura@stat.ncu.edu.tw]

16a7-5 Reconstruction of Individual Patient Data for Meta-analysis via Bayesian Approach

Yusuke Yamaguchi
Wataru Sakamoto
Shingo Shirahata
Division of Mathematical Science, Graduate School of Engineering Science, Osaka University, Japan

Masashi Goto
Biostatistical Research Association, NPO, Japan

In recent meta-analysis, methods based on individual patient data (IPD), which should be measured in each trial, have been attracted attention. However, most of these are often difficult to implement because of some issues associated with IPD collection. For this background, we suggested a method based on simulated IPD (SIPD), in which pseudo-IPD are reconstructed by statistical simulation using available summary statistics, and then more flexible patient-specific statistical models are applied to the SIPD (Yamaguchi et al. [Proc. of 58th ISI. (2011): CPS035-04]). Here, we focus on sampling scheme of the SIPD, which is one of the most important procedures in our method. Yamaguchi et al. used an algorithm which is known as Poor Man's data augmentation by Wei and Tanner [JASA. 85 (1990): 699-704]; however, it is possible to generate more proper SIPD by taking into account uncertainty in estimating parameters with Bayesian approaches. This is inferred from the principle of multiple imputation in statistical analysis with missing data (Rubin [JASA. 81 (1996): 473-489]).

[Yusuke Yamaguchi, 1-3 Machikaneyama-cho, Toyonaka, Osaka 560-8531, Japan; yamaguti @sigmath.es.osaka-u.ac.jp]

16b1-Shrinkage in Statistics and Machine Learning
December 16 (Friday), 16:30 - 18:00, HSS 1st Conference Room
Organizer: Chen-Hai Andy Tsao
Chair: Chen-Hai Andy Tsao

16b1-1 Post Selection Statistical Inference and Empirical Bayes Confidence Intervals.

J.T. Gene Hwang
Cornell University and National Chung Cheng University

Modern statistical applications often involve selection. Statistical inference after selection is becoming increasingly important. Naive statistical inference causes severe bias especially in the large *p* small *n* (i.e. large population size and small sample size) scenario. We shall review some examples to illustrate the advantage of Empirical Bayes procedures including confidence intervals. If properly constructed, they can provide valid statistical analysis. The intervals can be short. They can turn the ”curse of dimensionality” into ”blessing of dimension- ality”. Also they can reduce the selection bias. If time allows, we shall report results relating to false coverage rate (FCR), which parallels false discovery rate (FDR) for testing hypotheses problems. In the modern applications including microarry data analysis to which our intervals were applied to, selections as well as the large *p* and small *n* scenario are often encountered.

[J.T. Gene Hwang, Department of Mathematics, Cornell University, U.S.A.; hwang@math.cornell.edu]

16b1-2 A Bayesian Rating System Using W-Stein’s Identity

Ruby Chiu-Hsing Weng
Department of Statistics, National Chengchi University, Taiwan, R.O.C.

For Internet games, large online ranking systems are muchneeded. We propose a Bayesian approximation method, based on a variant of Stein’s identity (1981), to obtain online ranking algorithms with simple analytic update rules. Experiments on game data show that the accuracy of our approach is competitive with state of the art systems such as TrueSkill, but the running time as well as the code are much shorter. We also compare our method with Glicko rating system, designed for rating chess players.

[Ruby Chiu-Hsing Weng, Department of Statistics, National Chengchi University, Taiwan, R.O.C.; chweng@nccu.edu.tw]

16b1-3 Dimension Reduction for Change Point Detection with Application to fMRI

John Aston
CRiSM, Dept of Statistics, University of Warwick, UK

Functional Magnetic Resonance Imaging (fMRI) provide ways to examine the brain in-vivo. The data sets consist of dense spatial and temporal information and thus can be considered as spatial functional data recorded over time. To assess the stationary in the data, change point detection in sequences of functional data is examined where the spatial functional observations are temporally dependent and where the distributions of change points from multiple subjects is required. Of particular interest is the case where the change point is an epidemic change (a change occurs and then the observations return to baseline at a later time). The special case where the covariance can be decomposed as a tensor product is considered with particular attention to the power analysis for detection. This is of interest in the application to fMRI, where the estimation of a full covariance structure for the three-dimensional image is not computationally feasible. It is found that use of basis projections such as principal components for detection of the change points can be optimal in situations where PCA is traditionally thought to perform badly. (Joint work with Claudia Kirch, Karlsruhe Institute of Technology)

[John Aston, CRiSM, Dept of Statistics, University of Warwick, Coventry, CV47AL, U.K.; j.a.d.aston@warwick.ac.uk]

16b2-Multiple Endpoints in Clinical Trials
December 16 (Friday), 16:30 - 18:00, HSS 2nd Conference Room
Organizer: Toshimitsu Hamasaki
Chair: Chin-Fu Hsiao

16b2-1 Testing Multiple Endpoints in Complex Clinical Trial Designs

H.M. James Hung
Food and Drug Administration, U.S.A.

In many disease areas, designs of pivotal clinical trials are increasingly complex. For assessing cardiovascular risks in a clinical program, multiple trials may be jointly analyzed to assess a mortality endpoint whereas each trial is planned to assess a different endpoint. For assessing a rare safety event, multiple trials may be jointly analyzed to assess collectively for sufficient study power and consistency across trials. In another case, a single trial may be conducted to assess a major adverse clinical event and a symptom endpoint by splitting the trial into two trials. Active controlled designs with or without a placebo arm and adaptive designs are also complex with many difficult problems for testing endpoints. This paper will present the challenges of the conventional statistical inference frameworks and stipulate a number of approaches to the multiplicity problems associated with testing multiple endpoints under such designs in confirmatory clinical trials.

[H.M. James Hung, Food and Drug Administration, U.S.A.; HsienMing.Hung@fda.hhs.gov]

16b2-2 Utility and Pitfalls of Endpoint Selection with an Adaptive Design

Sue-Jane Wang
Office of Biostatistics, Office of Translational Sciences Center for Drug Evaluation and Research, U.S. Food and Drug Administration, U.S.A.

A clinical research program for drug development often consists of a sequence of clinical trials. Adaptive designs are not infrequently proposed for use. There are many design choices depending on the study objectives. In this work we stipulate the multiplicity issues with adaptive designs encountered in regulatory applications. In particular, we elucidate the utility of endpoint selection with an adaptive design in exploratory adaptive trial versus in confirmatory adaptive trial. For confirmatory adaptive design clinical trials, controlling studywise type I error and type II error is of paramount importance. For exploratory adaptive trials, we define the probability of correct selection of design features, e.g., endpoint, effect size, and the probability of correct decision for drug development. We will assert that maximizing these probabilities would be critical to determine whether the drug development program continues or how to plan the confirmatory trials if the development continues.

[Sue-Jane Wang, Office of Biostatistics, Office of Translational Sciences Center for Drug Evaluation and Research, U.S. Food and Drug Administration, U.S.A.; suejane.suejane.wang@fda.hhs.gov]

16b2-3 Sample Size Determination in Clinical Trials with Multiple Co-Primary Endpoints

Sozu Takashi
Department of Biostatistics, Kyoto University School of Public Health, Japan

Tomoyuki Sugimoto
Department of Mathematics Science, Hirosaki University Graduate School of Science and Technology,Japan

Toshimitsu Hamasaki
Department of Biomedical Statistics, Osaka University Graduate School of Medicine, Japan

In most comparative clinical trials, the clinical efficacy of a test treatment is characterized by a set of possibly correlated outcomes because patients responses to the treatment may comprise several different aspects. Traditionally, a single outcome is selected as the primary variable and is used as the basis for the design, interim and final analyses of the trial. However, recent clinical and pharmaceutical drug development settings suggested the need for additional challenges to

multiple endpoints because the assessment of a treatment with use of a single endpoint may not always provide a comprehensive picture of all the treatments benefits for the subjects entire experience of a disease. Therefore, more than one outcome is viewed as key, with the aim at 1) all being sufficient for proof of efficacy, or 2) at least one being sufficient for proof of efficacy with a prespecified ordering (or no ordering) of outcomes. Having such multiple endpoints creates difficulties for statisticians in handling multiplicity in the design and analysis of clinical trials. As well as incorporating a predefined strategy for statistical analysis into a trial protocol at the design stage, statisticians should consider how to provide an appropriate sample size corresponding to the statistical analysis. In this presentation, we describe the power and sample size calculation for comparative clinical trials with multiple correlated outcomes to be evaluated as primary variables. In general, there are two major strategies for statistical testings with such multiple outcomes to be evaluated: one is to provide the statistical significance results in favor of test treatment compared with control treatment, for all of the outcomes; the other is to show the statistical significance result for at least one of the outcomes. The former needs no adjustment for type I error for statistical testing and each null hypothesis should be rejected at the same significant level as statistical significance is required for all of the outcomes. However, type II error increases as the number of outcomes to be tested increases. This means that type II error needs to be controlled carefully in the design stage of a clinical trial, especially in planning sample size. On the other hand, the latter requires adjustment for type I error as statistical signif- icance tests need to be controlled adequately. We will discuss the sample size determination for two strategies separately, paying more attention to the former one.

[Sozu Takashi, Department of Biostatistics, Kyoto University School of Public Health, Japan; sozu.takashi.4s@kyoto-u.ac.jp]

16b3-Copula and Its Applications
December 16 (Friday), 16:30 - 18:00, HSS Media Conference Room
Organizer: Ping He
Chair: Ping He

16b3-1 A New Algorithm to Generate Random Numbers with Some Classical Probability Distributions

Ping He
Mathematical school, Southwest Jiaotong University, Chengdu, China

As a basic stochastic model, random walk plays an important role in the development of probability and stochastic processes. Galton board experiment illustrates normal distribution using a simple random walk. In this paper, we study a kind of special one-dimension random walk. The rule of random walk is described as follow. Starting from a particular point, the probability of every

unit time to a left or right is 1/2 , moving step changes over time based on certain rules and only depends on the current state (which is different from all other researches on random walk). We will illustrate some classical probability distributions by special constructed random walks. These special constructed random walks can be use to generate random numbers following corresponding distribution. We construct a series of new algorithms that can generate random numbers following exponential distribution, Reyleigh distribution, Cauchy distribution, uniform distribution, Weibull distribution, extreme value distribution, respectively.

[Ping He, Mathematical school, Southwest Jiaotong University, Chengdu, China, 610031; heping@home.swjtu.edu.cn]

16b3-2 Dependent Analysis Based on Variable Structure Copula Model

Qin Wang
School of Mathematics, Southwest Jiaotong University, Chengdu 610031, China

In the paper, based on Spearman's rho, a method by which a Copula model with the characteristics of variable structure can be diagnosed is put forward at the first time. Though Monte Carlo Simulation (MC) technology, the validity of the diagnostic method is verified. By using of the staged modeling technique and Spearman's rho, it is analyzed whether or not the dependency structure between Shanghai and Shenzhen stock markets possesses the characteristics of variable structure. The results confirm that the dependent relationship of variables possessed the characteristics of variable structure can be captured by using Spearman's rho. The diagnostic method using Spearman's rho has a certain degree of superiority.

[Qin Wang, School of Mathematics, Southwest Jiaotong University, Chengdu 610031, China; qinyuer7311@163.com]

16b3-3 Copula Reliability Models for Typical Unrepairable System Involving Failure Correlation

Jiayin Tang
Southwest Jiaotong University, Chengdu, China

For the correlation structure existed in the multiple failure modes of component, we present a new theoretical method for reliability calculation involving failure correlation in mechanical systems. The static and dynamic calculation models based on Copulas theory were given. Since *n*-degree integral operation was substituted by *n*-degree difference operation, the calculation is simplified greatly. The reasonableness of Copulas reliability models was verified, and the problem of determining the correlation degree was solved successfully, thus the precision was ensured. In

addition, the Copula theory verified the dynamic boundary continuity of reliability associated with correlation parameter, and the estimation method for the parameter of correlation degree was given. Finally, the results of a practical case study show the effectiveness and accuracy, by means of comparison with other methods.

[Jiayin Tang, Southwest Jiaotong University, Chengdu, China; tangjiayin@home.swjtu.edu.cn]

16b3-4 The Regime Switching Character on the Volatility Spillover Effect between Stock Market and Bond Market

Lu Wang
Southwest Jiaotong University, Chengdu, China

There are different volatility spillover relationships between stock market and bond market with the lack of research on regime switching. After smoothing correlation coefficient is used to express the strength of volatility spillover, the markov switching ARMA(1,1) is selected to describe the regime switching character, whichs significance is test by LR test.Then, the trend of the strength of volatility spillover is forecast by probability extrapolation. The results show that the regime switching is asymmetric and the duration of positive correlation state is longer and so on.

[Lu Wang, Department of Statistics,Southwest Jiaotong University, Chengdu, China; wiwilwang@163.com]

16b4-Genomic Statistics
December 16 (Friday), 16:30 - 18:00, AC 1st Conference Room
Organizer: Hsin-Chou Yang
Chair: Hsin-Chou Yang

16b4-1 Functional Logistic Regression and Genome Wide Association Studies

Richard M Huggins
Sabrina Rodrigues
Department of Mathematics and Statistics, University of Melbourne. Australia

Functional regression allows the modelling of the relationship between a response and a functional covariate and as such has applications to genome wide association studies (GWAS). We examine the application of these methods to a GWAS on drug resistance in epilepsy. The results are compared with more traditional approaches.

[Richard Huggins, Department of Mathematics and Statistics, University of Melbourne, VIC 3010. Australia; r.huggins@ms.unimelb.edu.au]

16b4-2 Estimation of Odds Ratios of Genetic Variants for the Secondary Phenotypes Associated with Primary Diseases

Sanjay Shete
The University of Texas, M. D. Anderson Cancer Center, HOUSTON, U.S.A.

Genetic association studies for binary diseases are designed as case-control studies. At the time of case-control collection, information about secondary phenotypes is also collected. To study the secondary phenotypes, investigators use standard regression approaches. However, using the secondary phenotype as an outcome variable in a case-control study might lead to a biased estimate of ORs. This is because the secondary phenotype is associated with the primary disease of interest; therefore, individuals with (case subject) and without (control subject) the secondary phenotype are not sampled following the principle of a case-control design. We demonstrate that such analyses will lead to a biased estimate of OR. We propose new estimating equations-based approaches to provide a more accurate OR estimate of genetic variants associated with the secondary phenotype for both un-matched and frequency-matched association studies. Applying our new method to smoking intensity as a secondary phenotype reduced the bias in odds ratio estimation.

[Sanjay Shete, Department of Epidemiology, Unit 1340, The University of Texas M. D. Anderson Cancer Center 1155 Pressler Blvd., CPB4.3628 Houston, TX77030, U.S.A.; sshete@mdanderson.org]

16b4-3 Statistical Method for Detecting Disease-Associated Deletions Using Case-Control Genome-wide SNP Genotype Data

Chih-Chieh Wu
Department of Epidemiology, Division of Cancer Prevention & Population Sciences, MD Anderson Cancer Center, Houston, Texas, U.S.A.

Deletion copy number variations (CNVs) are frequently observed in patients with microdeletion syndromes and certain neuron-developmental disorders. We extended genome-wide association studies (GWAS) to deletion CNV detection for discovering additional common genetic variants that influence susceptibility. We have developed a method for detecting disease-associated deletion variants using high-density SNP genotype data. It first detects statistically significant evidence of a deletion at individual SNPs for SNP-by-SNP analyses, and then we combine the information from multiple neighboring SNPs for cluster analyses. GWAS are designed to discover individual disease-associated SNPs; in contrast, methods for detecting deletion CNV were generally designed to find small deleted chromosomal segments. Our method has proven to be useful and robust in the presence of linkage disequilibrium. We applied this method and analyzed the high-density 550K SNP genotype data from a GWAS of rheumatoid arthritis and glioma each to detect common

deletion CNVs that influence susceptibility.

[Chih-Chieh Wu, Department of Epidemiology, Division of Cancer Prevention & Population Sciences, MD Anderson Cancer Center, Houston, Texas, U.S.A.; ccwu@mdanderson.org]

16b5-Computational Neuroscience
December 16 (Friday), 16:30 - 18:00, AC 2nd Conference Room
Organizer: Michelle Liou
Chair: Shuen-Lin Jeng

16b5-1 An Introduction to Mathematical Models of Neuronal Networks

Chin-Yueh Liu
Department of Applied Mathematics, National University of Kaohsiung, Taiwan, R.O.C.

Mathematical models (including statistical and computational models) of neuronal networks play an important role to elucidate fundamental mechanisms of neural computations via analyzing or efficiently simulating the rich dynamics of network activities resulting from complicated interactions among neurons. In this talk, I will first give a short survey of some famous models in history, which include Wilson-Cowan model and population density model, etc. With the advent of new experimental techniques such as multiple electrode recording, however, these models suffer their difficulty and limitation in explaining results from these new experimental techniques. This leads to the second part of my talk that I will introduce recent progress and challenge to develop higher order models of neuronal networks.

[Chin-Yueh Liu, Department of Applied Mathematics, National University of Kaohsiung, Taiwan, R.O.C.; cyliu1025@nuk.edu.tw]

16b5-2 Mismatch Negativity: Interaction between Cortical Networks and Serotonin Transporter Genotypes

Arthur C. Tsai
Institute of Statistical Science, Academia Sinica, Taiwan, R.O.C.

Olga V. Sysoeva
Institute of Higher Nervous Activity and Neurophysiology, Russian Academy of Sciences, Moscow

Electrophysiological measurement of sensitivity to stimulus changes in the auditory cortex, namely the mismatch negativity (MMN), is designed to investigate automatic neuronal responses to

event changes that occur close enough in time. There has been an extensive literature on research into the MMN related cognitive deficits in patients of brain dysfunction, aggression and anxiety disorders. By examining association between gene polymorphism and patterns of brain activity, researchers have directly investigated the genetic basis of between-subject differences in MMN. However, as disentangling concurrent processes underlying MMN is still a question to be resolved, and the location and network contributions to cerebral processing remain elusive. In this study, the sensitivity of the carriers of genotypes related to levels of 5-HT transmission to external stimulation without brain conscious control was examined through the MMN paradigm. The single-trial EEGs were analyzed by both electromagnetic spatiotemporal Independent component analysis (EMSICA) and reproducibility of event-related spectral perturbation (ERSP) analysis. Our approach is particularly oriented to analyzing single-trial EEGs for each individual subject. An integration of genetic and neurobiological evidence as well as statistical modeling of cortical networks allows examining different hypotheses in the theory of informational synthesis and promoting new knowledge on integrative mechanisms underlying MMN.

[Arthur C. Tsai, Institute of Statistical Science, Academia Sinica, Taiwan, R.O.C.; arthur@stat.sinica.edu.tw]

16b5-3 **Recursive Kernel Function Approximation for Bio-signal Clustering**

Jiann-Ming Wu
Chun-Chang Wu
Po-Yu Huang
Institute of Applied Mathematics, National Dong Hwa University, Taiwan, R.O.C.

This work explores learning MLPotts (multi-layer Potts perceptrons) networks for recursive kernel function approximation and bio-signal clustering. An iterative approach is proposed to partition many-channel observations to K clusters by tracking K recursive kernel functions. All temporal patterns according to their fitness to K recursive kernel functions, each being realized by an MLPotts network, can be partitioned to K clusters. The common recursive kernel function embedded within temporal patterns of the same cluster is approximated by learning an MLPotts network subject to paired data simultaneously sampled from many-channel observations by moving widows. Two primitive operations respectively for exclusive membership evaluation and common recursive kernel function extraction are iteratively executed for signal clustering. Learning an MLP (multi-layer perceptrons) network is shown a special case of learning an MLPotts network. By numerical simulations, the proposed iterative approach is shown effective and reliable for signal clustering and hill-valley temporal pattern classification. Its potential applications for EEG quantization, spike sorting and MRI time series clustering will be also addressed.

[Jiann-Ming Wu, Institute of Applied Mathematics, National Dong Hwa University, Taiwan, R.O.C.; jmwu@mail.ndhu.edu.tw]

16b6-Statistical Computing
December 16 (Friday), 16:30 - 18:00, AC 3rd Conference Room
Chair: Chih-Yao Chang

16b6-1 Split Variable Selection for Decision Trees with Ranking Data

Yu-Shan Lin
Yu-Shan Shih
Department of Mathematics, National Chung Cheng University, Taiwan, R.O.C.

A variable selection method for constructing decision trees with ranking data is proposed. It utilizes conditional independence tests based on loglinear models for contingency tables. Compared with other selection methods, our method is computationally more efficient. Moreover, our method is relatively unbiased and powerful in selecting the correct split variables. Simulation results and a real data study are given to demonstrate the strength of our method.

[Y.-S. Shih, National Chung Cheng University; yshih@math.ccu.edu.tw]

16b6-2 Variable Selection in Principal Components Analysis of Qualitative Data Using the Accelerated ALS Algorithm

Masahiro Kuroda
Yuichi Mori
O*kayama University of Science, Okayama, Japan*

Masaya Iizuka
Okayama University, Okayama, Japan

Michio Sakakihara
Okayama University of Science, Okayama, Japan

Principal components analysis (PCA) is a popular dimension-reducing tool that replaces the variables in a data set by a smaller number of derived variables. In PCA on a larger number of variables, the resultant principal components may not be easy to interpret. To give a simple interpretation of principal components, we select a subset of variables that best approximates all the variables. When applying PCA to qualitative data, we need the quantification of the data using the alternating least squares (ALS) algorithm. Then, the computation time has been a big issue so far. To reduce the computation time of the ALS algorithm, we use the vE accelerated ALS algorithm of Kuroda et al (2011) [CSDA 55, (2011): 143-153]. Moreover, we improve the vϵ acceleration algorithm adding a re-starting process for reducing both of the number of iterations and the

computation time.

[Masahiro Kuroda, 1-1 Ridaicho, Kitaku, Okayama, Japan; kuroda@soci.ous.ac.jp]

16b6-3 Calculation Methods and Graphical Expression of Two-step Cluster Analysis

Takako Korikawa
School of Science, Graduate School of Tokai University, Japan

Yoshiro Yamamoto
School of Science, Tokai University, Japan

To get some information of clusters for a huge data set, we usually use non-hierarchical cluster analysis, like *k*-means method. But we sometime want to know the relation of each cluster. Moreover we are interested in a small size cluster for royal customer that is not come out by *k*-means method of small *k*. One solution for these purposes is two-step cluster analysis. At the first step of this method, we partition into a high number of sub-clusters. Each sub-clusters are very close together. Next in the second step, we implement hierarchical clus- tering for sub-clusters from first step. In this way we obtain a hierarchy and we can decide the optimal number of clusters. We examine effective computational method and several methods of transformation for variables in second step. We also introduce visualization techniques for results of this method.

[Takako Korikawa, School of Science, Graduate School of Tokai University, Japan; 0bsfm004 @mail.tokai-u.jp]

[Yoshiro Yamamoto, School of Science, Tokai University, Japan; yama@tokai-u.jp]

16b6-4 King-Werner Method for EM Algorithm

Michio Sakakihara
Masahiro Kuroda
Okayama University of Science, Okayama, Japan

EM algorithm is a statistical parameter estimation method for variety problems since the iterations is convergence to an estimated parameter stably. In order to seep up convergence, there are many trials has been presented. Louis [*J.Roy. Statist. Soc.Ser.* B44 (1982):226V233] proposed the EM iteration with the Newton-Raphson method for M-step. It is worthy to consider some higher order nonlinear solution method for the EM iteration. In this report, we discuss King-Werner iterative method of which convergence order CO(KM) is $1 + \sqrt{\ }$ > CO(Newton-Raphson), for the EM algorithm.

[Michio Sakakihara, Ridai-cho 1-1, kitaku Okayama 700-0005, Japan; sakaki@mis.ousa.c.jp]

16b6-5 Some Investigations on Rao-Blackwellization of Metropolis-Hastings Algorithm

Wen-Da Lo
Chai-Wei Wang
Department of Mathematics, National Chung Cheng University, Chia-Yi, Taiwan, R.O.C.

Importance sampling is a variance reduction technique of simulations for the analysis of integrals and expectations. In this article, we aim to address importance sampling to Markov chain Monte Carlo, especially Metropolis-Hastings algorithm. In Metropolis-Hastings algorithm, we derive that the relationship between the importance weight and the expected number of a given candidate sample reserved in the produced Markov chain sequence. According to this result, we construct Metropolis-Hastings importance sampling (MHIS) through estimating unknown importance weights to improve the performance of crude Metropolis-Hastings estimator. We compare our method to the original Rao-Blackwellization principle for Metropolis-Hastings schemes in Casella and Robert (1996) and the modification method in Douc and Robert (2011). We illustrate their performances on simple examples and a real data set.

[Chai-Wei Wang, 168, University Rd., Min-Hsiung, Chia-Yi, 621 Taiwan, R.O.C.; ccu.src@gmail.com]

16b7-Mathematical Statistics (II)
December 16 (Friday), 16:30 - 18:00, AC 4th Conference Room
Chair: Tsiu-Ling Chen

16b7-1 On Compatibility of Discrete Conditional Distributions: A Graph-theoretic Approach

Shih-Chieh Chen
National Chengchi University, Taipei, Taiwan, R.O.C.

For two discrete random variables X and Y taking on values $x_1, \cdots, x_I$ and $y_1, \cdots, y_J$, respectively, a putative conditional model for the joint distribution of X and Y consists of two $I \times J$ matrices representing the conditional distributions of X given Y and of Y given X. We say that two conditional distributions (matrices) A and B are compatible if there exists a joint distribution of X and Y whose two conditional distributions are exactly A and B. We present new versions of necessary and sufficient conditions for compatibility of (finite) discrete conditional distributions via a graph-theoretic approach. Moreover, we show that there is a unique joint distribution for two

given compatible conditional distributions if and only if the corresponding graph is connected. (This is joint work with Professor Yi-Ching Yao.)

[Shih-chieh Chen, Mailing address: 6, Lane 19, Section 1, Dian Rd, Dian 106, Taipei, Taiwan, R.O.C.; 94354504@nccu.edu.tw]

16b7-2 Does the Tail Asymmetry Really Matter? An Analysis Based on a Threshold Extreme Value Model

Mike K.P. So
Ka Shing Chan
The Hong Kong University of Science and Technology, Hong Kong

The purpose of this research is to test if financial returns demonstrate tail asymmetry after accounting for the conditional heteroskedasticity. We proposed a threshold extreme value distribution (TEVD) and use it as an error distribution of a GARCH model. TEVD is constructed based on the peak-over-threshold approach of extreme value theory. It combines two generalized Pareto distribution for two tails and a truncated unimodal distribution for the bulk for flexible tail modeling. A Bayesian approach is adopted to estimate unknown parameters by using Markov Chain Monte Carlo (MCMC) and to establish the test of tail asymmetry via model selection analysis. Five Bayesian model selection techniques are investigated and their performance is evaluated using both simulated and real data. We confirm that all of the equity indices we considered demonstrate strong evidence on tail asymmetry in the recent decade. We also study the performance of our approach in forecasting value-at-risk with various holding periods. Compared with conventional GARCH-*t* model and symmetric version of our model, the asymmetric model performs superiorly for long holding periods or extreme quantiles for both tails.

[Raymond K.S. Chan, The Hong Kong University of Science and Technology, Hong Kong; imkschan@ust.hk]

16b7-3 Variable Selection through Random Subsets

Mike K. P. So
Department of ISOM, Hong Kong University of Science and Technology
Raymond W. M. Li
Hong Kong University of Science and Technology

Variable selection method in high dimensional data space poses several challenges to applications: individual ranking focuses only marginal effect of independent variables while subset evaluation becomes infeasible when the number of variables is large. The core idea from Chernoff (2009,

Annals of Applied Statistics) tries to solve the problem by randomly selecting many small subsets of independent variables, performing backward elimination on each subset and averaging the ranking among the subsets. However we find their method imposes several restrictions that may not be suitable for applications. We generalize their method by allowing independent variables to be of different types, both continuous and discrete, and evaluate them with a likelihood-related criterion. Simulation study reveals that our method performs well in various high-dimensional situations with manageable computational cost. We believe that the relaxed assumptions allow this method to be effective in applications like bioinformatics or news analytics. Empirical illustration is given using high-frequency finance data.

[Raymond W. M. Li, Hong Kong University of Science and Technology; wmli@stu.ust.hk]

16b7-4 Adjustment of Posterior Probability for Oversampling When Target Is Rare

Jong Hoo Choi
S. H. Hwang
N. Y. Yi
Department of Information & Statistics, Korea University, Korea

When an event of target variable is rare, a widespread strategy is to build a model on the sample that disproportionally over-represents the events, that is over-sampled. Using the data over-sampled from the original data set, the predicted values would be biased; however, it can be easily corrected to represent the population. In this study, we investigate into the relationship between the proportion of rare event on a data-mart and the model performance using real world data of a Korean credit card company. Also, we use the offset method and weighted method for adjusting of posterior probability for over-sampled data. Finally, we compare the performance of the methods using real data sets.

[J. H. Choi, Department of Information & Statistics, Korea University, Jochiwon-eup, Yeongi-gun, Chungnam 339-700, Republic of Korea.; jhchoi@korea.ac.kr]

16b7-5 Moment Bounds and Multistep Prediction of Linear Processes

Ngai Hang Chan
Chinese University of Hong Kong

Shih-Feng Huang
National University of Kaohsiung, Taiwan, R.O.C.

Ching-Kang Ing
Academia Sinica, Taiwan, R.O.C.

A uniform moment bound of the inverse Fisher's information matrix of a general linear process is established in this article. This bound is applied to derive the moment convergence of the normalized least squares estimate of the underlying linear process, which also includes the long-memory case. Based on this bound, an asymptotic expression for the corresponding mean squared prediction error (MSPE) is obtained. An important and intriguing application considered in this article is to establish and compare the asymptotic expressions of the multistep MSPEs of the least squares predictors for ARMA, *I(d)* and ARFIMA models. These asymptotic expressions not only offer means to assess the multistep prediction errors, but also explicitly demonstrate how the multistep MSPE manifests with the model complexity and the dependent structure of the underlying process, thereby shedding light about multistep prediction for general linear processes. Numerical findings are also conducted to verify the theoretical results.

[Shih-Feng Huang, Department of Applied Mathematics, National University of Kaohsiung, 700, Kaohsiung University Rd., Nanzih District, 811. Kaohsiung, Taiwan, R.O.C.; huangsf@nuk.edu.tw]

7th IASC-ARS Σ joint 2011 Taipei Symposium

Institute of Statistical Science, Academia Sinica

Asian Regional Section of the IASC

DGBAS, Executive Yuan, R.O.C.

National Science Council, R.O.C.

The Chinese Institute of Probability and Statistics

中國統計學社 Chinese Statistical Association