2016 Joint Conference

ISI ISM ISSAS

Indian Statistical Institute, India(ISI)

Institute of Statistical Mathematics, Japan(ISM)

Institute of Statistical Science, Academia Sinica, Taiwan(ISSAS)

http://www3.stat.sinica.edu.tw/2016issas/

31th January - 2nd February

2nd Conference Room,
Humanities and Social Sciences Building,
Academia Sinica,
Taipei, Taiwan

2016 Joint Conference

ISI ISM ISSAS

Jan. 31 (Sunday) Poster Session

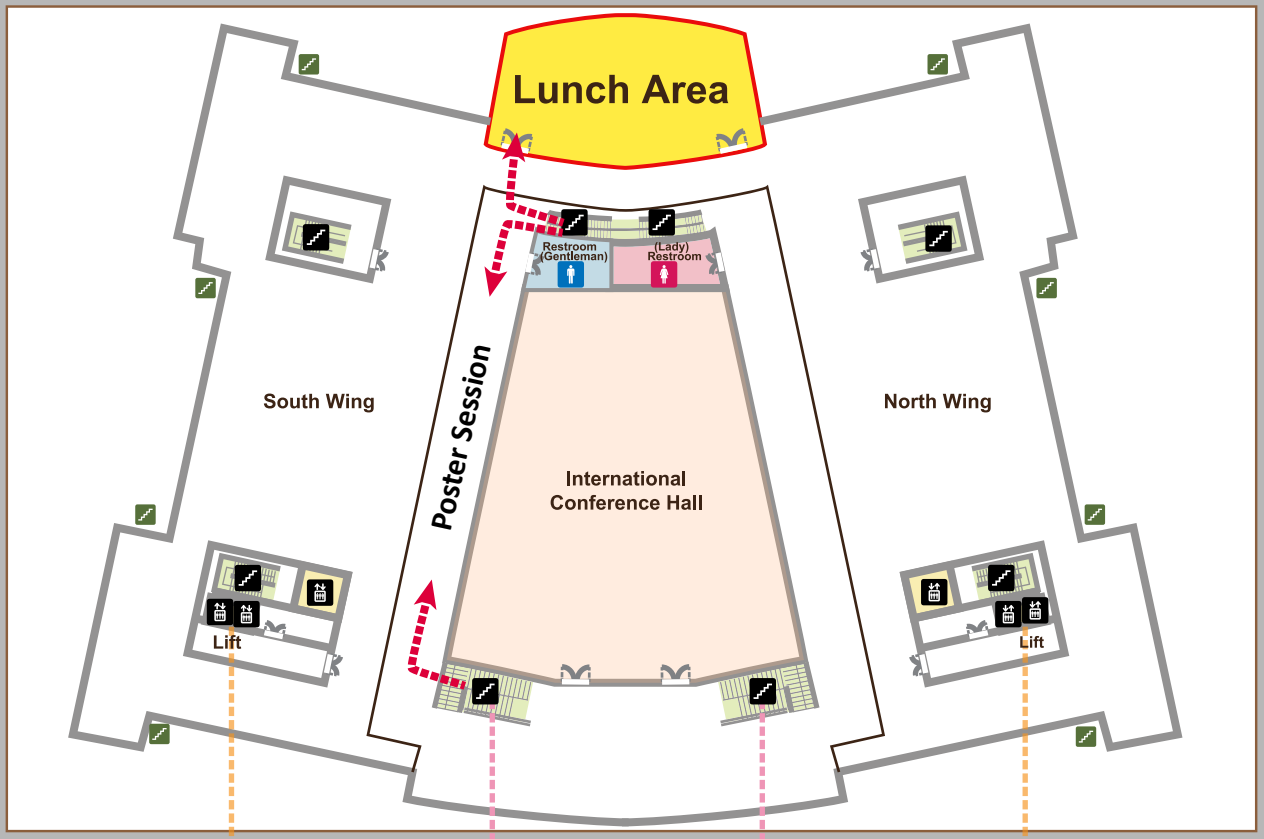
08:45-09:00	Registration		
09:00-09:05	Opening : Tso-Jung Yen		
09:05-10:15	Oral Presentation(1) (5 minutes per person by showing slides)		
	1. Andrei Akhmetzhanov Alteration of the active pathway as the resistance mechanism to targeted BRAF treatment in advanced melanoma: mathematical modeling framework	**2. Yuh-Chyuan Tsay** Logistic-AFT location-scale cure models for the relative survival with an application to HCV mono-infected patients	**3. Takao Kumazawa** Analysis of earthquake occurrence rates modulated by volcanic activities
	4. Peter Chang-Yi Weng Uncertainty quantification on linear-time system	**5. Jiun-Wei Liou** A statistical approach to diffusion tensor imaging analysis	**6. Yuta Tanoue** Credit risk outside operational base of Japanese regional banks
	7. Hung-Yin Chen Conditional tail expectation for integrated processes with stochastic volatility	**8. Tai-Chi Wang** Statistical methodologies for community detection in social networks	**9. Takayuki Kawashima** Robust sparse regression
	10. Wei-Cheng Hsiao On high-dimensional cross-validation	**11. Shih-Hao Huang** Optimal group testing designs for estimating prevalence with uncertain testing errors	**12. Tomoaki Imoto** A method for generating symmetric distributions on the circle and its application
10:15-10:35	Tea Break		
10:35-11:10	Oral Presentation(2) (5 minutes per person by showing slides)		
	13. Yuan-Lung Lin Experimental design with circulant property and its application to fMRI experiment	**14. Charlotte Wang** A rediscovery of dis-similarity measure with U-statistic for rare variant association	**15. Sara Kropf** Limiting distributions for outputs of automata
	16. Yu-Hsiang Cheng A nonparametric copula density estimator incorporating information on bivariate marginals	**17. Yi-Ran Lin** A flexible mover-stayer model for left-truncated and interval-censored data	**18. Ming-Chung Chang** Two-level minimum aberration designs under a conditional model with a pair of conditional and conditioned factors
11:10-11:20	Break		
11:20-12:20	Poster Presentation (Presenter stands next to the poster)		
12:20	Lunch		

2016 Joint Conference

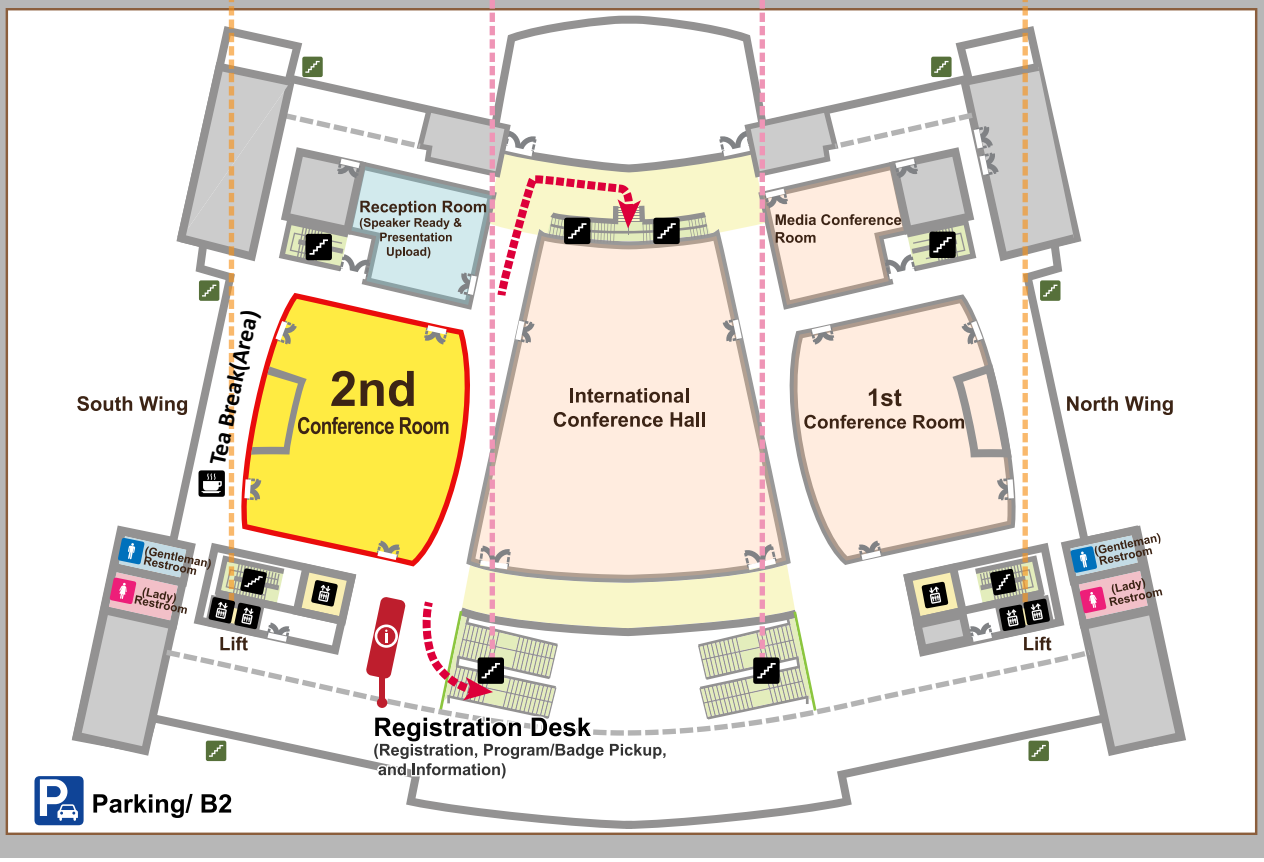
ISI ISM ISSAS

	Feb. 01 (Monday)	Feb. 02 (Tuesday)
08:50-09:00	**Opening : Ching-Shui Cheng**	
	Session 1 Chair: **Atanu Biswas**	Session 5 Chair: **Teppei Ogihara**
09:00-09:30	**Su-Yun Huang** Integrating multiple randomized sketches of SVD and PCA	**Satoshi Kuriki** Some distributions associated with the cone of positive semidefinite matrices and their applications
09:30-10:00	**Smarajit Bose** Robust Speaker Identification	**Antar Bandyopadhyay** A new approach to classical and modern urn models
10:00-10:30	**Hironori Fujisawa** The gamma-lasso: Robust estimation for sparse Gaussian graphical modeling	**Shogo Kato** A tractable and interpretable four-parameter family of unimodal distributions on the circle
10:30-11:00	**Tea Break**	
	Session 2 Chair: **Tso-Jung Yen**	Session 6 Chair: **Asish Sengupta**
11:00-11:30	**Ching-Kang Ing** On model selection from a finite family of possibly misspecified models	**Frederick Kin Hing Phoa** A new SOP for accurate and efficient community detection
11:30-12:00	**Teppei Ogihara** Parameter estimation for diffusion processes with high-frequency data	**Hisashi Noma** Quantifying indirect evidence in network meta-analysis
12:00-12:30	**Ashis SenGupta** Stress dependent strength models: Constructions and applications to HALT	**Kaushik Jana** Adjustment of bifurcated river flow measurements from historical data: Paving the way for an international accord
12:30-14:00	**Lunch**	
	Session 3 Chair: **Chen-Hsiang Yeang**	Session 7 Chair: **Ayan Basu**
14:00-14:30	**Atanu Biswas** Response adaptive allocation design for circular treatment outcomes	**Saurabh Ghosh** Association mapping of quantitative traits: Population stratification and count phenotypes
14:30-15:00	**Junji Nakano** A measure of dissimilarity between aggregated symbolic data with categorical variables	**Chen-Hsiang Yeang** Analysis of spatial-temporal gene expression patterns reveals dynamics and regionalization in developing mouse brain
15:00-15:30	**Hsin-Chou Yang** SMART: Statistical metabolomics analysis - an R tool	**Indranil Mukhopadhyay** Tight clustering for large datasets with an application to microarray data
15:30-16:00	**Tea Break**	
	Session 4 Chair: **Shogo Kato**	Session 8 Chair: **Hironori Fujisawa**
16:00-16:30	**Ayanendranath Basu** Statistical Inference based on the Bridge Divergences	**Masayo Yoshimori Hirose** An empirical Bayes confidence interval for high leverage area in small area estimation
16:30-17:00	**Yuta Koike** Quadratic variation estimation of an irregularly observed semimartingale with jumps and noise	**Tso-Jung Yen** Solving large scale penalized regression problems via parallel proximal algorithms
17:00-17:10		**Closing**

Humanities and Social Sciences Building (Floor 4)



Humanities and Social Sciences Building (Floor 3)



Poster session

08:45~12:20

January 31th, 2016

Alteration of the active pathway as the resistance mechanism to targeted BRAF treatment in advanced melanoma: mathematical modeling framework

Andrei R. Akhmetzhanov*, Chen-Hsiang Yeang

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

At the beginning of the twenty-first century the advanced stage melanoma was mostly untreatable. The situation has started to change since the arrival of a new drug *vemurafenib* in 2011. It targeted mutated BRAF observed in most melanomas. Unfortunately, the resistance to the drug eventually develops, and after initial regression, a tumor grows back with mutations impervious to the drug. Several studies such as (1,2) showed, however, that the use of state- and time-dependent therapies can significantly prolongate the tumor remission period. Hence, the optimization of the treatment regimes becomes of great importance.

Among many routes to escape the treatment, cancer cells are capable to switch between distinct gene programs (3). As such, MAPK and NF-(kappa)B pathways, mutually inhibiting each other, have high expression profiles in melanomas. For example, mutated BRAF causes high activation of MAPK pathway, whereas an alternate NF-(kappa)B pathway is suppressed and remain passive. However, the situation can be reversed, when the treatment is applied and the MAPK pathway is inhibited by the drug.

In our work we develop the theoretical framework for modeling tumor growth and consequent treatment application. The dynamics are present by a stochastic system with two different scales: first, the birth-death branching process, and second, the expression levels of MAPK and NF-(kappa)B pathways. The former process is intermingled by update events of the latter process, so that the cell division and cell death probabilities are determined by activation level of the pathways. Since the drug (BRAF inhibitor) suppresses only the MAPK pathway, leaving the NF-(kappa)B pathway unaffected, we investigate what the characteristic switching times between MAPK and NF-(kappa)B pathways, and how it depends on the dosage of the drug.

We find that optimization of the dosing strategy even for one drug agent significantly improves the treatment outcome. We show that a constantly administrated drug with a maximum dosage is much less efficient than the use of intermittent treatment regimes with moderate drug dosages. This also gives a plasible effect to lower the toxic effects of the drug.

Our findings encourage the development and adoption of personalized precision medicine, since the current instrinsic states of tumor cells determine the choice of the treatment strategy in question.

Logistic-AFT location-scale cure models for the relative survival with an application to HCV mono-infected patients

Yuh-Chyuan Tsay

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

In population-based studies, relative survival is the ratio of the observed survival from a diseased group of individuals to the expected survival from a comparable group in the general population, and it provides a measure of the excess mortality rate of the disease under study based on the estimation of the observed mortality of the diseased group in comparison with the background mortality of the general population. Due to enhancement of medical diagnostic technology, patients may be diagnosed early and treated adequately with a better management of medical conditions, and then results in a cured fraction of the patients. In this case, cure occurs when the observed age-specific mortality of the diseased group returns to the same level as the background age-specific mortality of the general population. Equivalently, the relative survival tails off to a plateau such as the excess mortality rate approaches zero after some time point. It motivates us to extend the logistic-AFT location-scale mixture regression models (Chen et al., 2013) to estimate the relative survival with cure attributable to a disease for left truncated and interval censored data. Simulation studies were conducted to demonstrate the validity of the proposed estimation procedure. The method is applied to analyze the mortality of HCV infection patients in the Taiwan National Health Insurance Research Database as an illustration.

[This is a joint work with Chen-Hsin Chen, Yu-Tseng Chu and Mei-Shu Lai]

Analysis of earthquake occurrence rates modulated by volcanic activities

Takao Kumazawa[1*], Yosihiko Ogata[1,2], Kazuhiro Kimura[3], Kenji Maeda[3], Akio Kobayashi[3]

[1]*Institute of Statistical Mathematics, Japan*
[2]*The University of Tokyo, Japan*
[3]*Japan Meteorological Agency, Japan*

Abstract

Near the eastern coast of Izu peninsula is an active submarine volcanic region in Japan, where magma intrusions have been observed many times. The forecast of earthquake swarm activities and eruptions are serious concern particularly in nearby hot spring resort areas. It is well known that temporal durations of the swarm activities have been correlated with early volumetric strain changes at a certain observation station of about 20 km distance apart. Therefore the Earthquake Research Committee (2010) investigated some empirical statistical relations to predict sizes of the swarm activity. Here we looked at the background seismicity rate changes during these swarm periods using the non-stationary ETAS (Epidemic Type Aftershock Sequence) model (Kumazawa and Ogata, 2013, 2014), and have found the followings. The modified volumetric strain data, by removing the effect of earth tides, precipitation and coseismic jumps, have significantly higher cross-correlations to the estimated background rates of the ETAS model than to the swarm rate-changes. Specifically, the background seismicity rate synchronizes clearer to the strain change by the lags around half a day. These relations suggest an enhanced prediction of earthquakes in this region using volumetric strain measurements. Hence we propose an extended ETAS model where the background rate is modulated by the volumetric strain data. We have also found that the response function to the strain data can be well approximated by an exponential functions with the same decay rate, but that their intersects are inversely proportional to the distances between the volumetric strain-meter and the onset location of the swarm. Our numerical results by the same proposed model show consistent outcomes for the various major swarms in this region.

Uncertainty quantification on linear-time system

Peter Chang-Yi Weng*, Frederick Kin Hing Phoa

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

Uncertainty quantification (UQ) is a new and hot branch of computer experiment, which provides a calibration on parameter estimates in a computer model when small perturbations exist. In this poster, we study the linear-time invariant system described by discretizing the partial differential equations. Such system has many important applications such as circuits, signal processing, spectroscopy, control theory and many others. In order to understand the random errors, we add some perturbation to the parameters in a continuous-time linear system that involves some uncertainties. When we choose the optimal control to minimize a cost function, the continuous-time Riccati differential equation is derived, and the solutions represent the approximations of the corresponding random process. We provide sufficient conditions for the existence of the stabilized solution of the stochastic algebraic Riccati equation. Some numerical simulation of the stochastic linear-time system are presented.

A statistical approach to diffusion tensor imaging analysis

Jiun-Wei Liou

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

We discuss the statistical analysis of diffusion tensor imaging (DTI). Conventionally, the fractional anisotropy (FA) index is used as a summary statistic for water molecular diffusion in the brain white matter in DTI analysis. Variance of the FA statistic is an important tool for normalization and comparison between research findings from different hospitals. We derive the variance estimate for the FA statistic and show its use in a real data example. DTI data are spatially correlated, based on which test statistics cannot be evaluated using the conventional t or F distributions by assuming independent sample observations. The surrogate data distribution can be generated using the phase-randomization procedure, which uses original spectra with randomly permuted spatial phases to preserve the original correlated structure in the surrogate data. We also show the use of the phase-randomization approach for thresholding the FA statistic.

Credit risk outside operational base of Japanese regional banks

Yuta Tanoue

Institute of Statistical Mathematics, Japan

Abstract

The depopulation in provincial areas causes the decline of these area's economies in recent years in Japan. Accordingly, regional banks have been advancing into areas outside their original operational bases, especially into major cities and urban regions. Since amount of lending to areas outside the original operational base is increasing and such lending's uncertainty is being high, the lendings are expected to significantly affect the credit risk of the regional bank's lending portfolios. Therefore, we analyze the effects of the lending areas (the local region, outside the local region, and Tokyo) on the credit risk of the regional bank's lending portfolios. Using five regional bank's credit data, (1) we describe the fundamental statistics of each area, (2) we develop credit risk estimation models to analyze the effect of lending area flag parameters. These analysis indicate that (1) lending area variables affect the credit risk, (2) default is less likely to occur outside the local region and in Tokyo, and (3) LGD is small in the local region compare as outside the local region and in Tokyo.

Conditional tail expectation for integrated processes with stochastic volatility

Hwai-Chung Ho[1,2], Hung-Yin Chen[1*], Henghsiu Tsai[1]

[1] *Institute of Statistical Science, Academia Sinica, Taiwan*

[2]*National Taiwan University, Taiwan*

Abstract

The present paper investigates the conditional tail expectation (CTE) for I(1) processes of returns following a general class of multivariate stochastic volatility model. We propose a non-parametric consistent estimate of CTE. The estimate is easy to implement, and the long-run variance of the estimate's limiting normal distributions is derived explicitly. Monte Carlo experiments are conducted to demonstrate the superiority of our approach in terms of coverage ratios for confidence intervals. Results on the estimation of CTE for the long-horizon returns of the S&P 500 index and other indices are also presented.

Keywords: Conditional Tail Expectation; I(1) process; Stochastic Volatility Model; asymptotic normality

Statistical methodologies for community detection in social networks

Tai-Chi Wang*, Frederick Kin Hing Phoa

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

In this poster, we demonstrate a new innovation for community detection in social networks consisting of four statistical methodologies. First, the “network screening” procedure is executed for quickly partitioning a large-scale network into several *n*-cliques, which are defined as the largest geodesic distance between any two nodes is no greater than *n*, by the graph coloring techniques. Then, we develop a scan statistic for detecting statistically significant communities with structure, attribute, and both structure and attribute clustering patterns. When a community is detected, it is desired to statistically test its center and its cluster range. The focus centrality measure is developed to determine whether a node is the truly important central node. Since the scan statistic can only detect communities with a specific shape, we use the swarm intelligence based (SIB) approach to fine-tune irregular communities and construct a hierarchical structure of communities based on a partition criterion derived from an exponential random graph model (ERGM) with multiple communities. Following the above procedures, we can glimpse the community structure in a social network.

Robust sparse regression

Takayuki Kawashima*, Hironori Fujisawa

Institute of Statistical Mathematics, Japan

Abstract

We consider a robust sparse regression. The parameter is estimated by minimizing the objective loss function based on the gamma-divergence with *L1* penalty. The robustness and sparseness are provided by the gamma-divergence and *L1* penalty, respectively. But, the parameter estimation algorithm is not easy to be constructed, because two complicated functions, gamma-divergence and *L1* penalty, are included in the objective loss function. We construct the parameter estimation algorithm using an MM algorithm to easily deal with two complicated functions. The parameter estimation algorithm obtained here is expressed as a closed form and monotonically decreases the objective loss function at each step. We also consider the outliers on the explanatory variables as well as the response variables. The model density for the explanatory variables is assumed to be independent to easily deal with high-dimensional model. It works like the conventional LASSO with standardized explanatory variables.

On high-dimensional cross-validation

Wei-Cheng Hsiao

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

Cross-validation (CV) is one of the most popular methods for model selection. By splitting n data points into a training sample of size n_c and a validation sample of size n_v with $n_c/n_v \to 1$ and $n_c \to \infty$, Shao (1993) showed that subset selection based on CV is consistent in a regression model of p candidate variables with $p << n$. However, in the case of $p >> n$, not only does CV's consistency remain undeveloped, but subset selection is also practically infeasible. In this paper, we fill this gap by using CV as a backward elimination tool for eliminating variables that are included by high-dimensional variable screening methods possessing sure screening property. By choosing an n_v such that n_v/n converges to 1 at a rate faster than the one in Shao's (1993) paper, we establish the consistency of our selection procedure. We also illustrate the finite-sample performance of the proposed procedure using Monte Carlo simulation.

Optimal group testing designs for estimating prevalence with uncertain testing errors

Shih-Hao Huang

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

We construct optimal designs for group testing experiments where the goal is to estimate the prevalence of a trait using a test with uncertain sensitivity and specificity values. Using optimal design theory, we show that the optimal design for accurately estimating the prevalence along is to have three different group sizes with unequal frequencies. We also compare performances of competitive designs and provide insights on how the unknown sensitivity and specificity of the test affect the quality of the prevalence estimator in a group testing setting.

Keywords: D-optimality; Ds-optimality; Group testing; Sensitivity; Specificity

[Work done jointly with M.-N. L. Huang, K. Shedden, and W. K. Wong]

A method for generating symmetric distributions on the circle and its application

Tomoaki Imoto

Institute of Statistical Mathematics, Japan

Abstract

We provide a method for generating symmetric circular distributions using discrete distributions on integers, symmetric about 0. Some well-known circular distributions can be generated through the proposed method. As an application of the method, a modified von Mises distribution, symmetric and possibly bimodal, is proposed. The distribution is constructed from difference of two random variables which follow a discrete Charlier series distribution. The p th cosine moment is expressed in terms of the probabilities that the difference of the two independent and identically distributed random variables takes values 0 and p. The illustrative example of the proposed distribution is given in this talk.

[Joint work with K. Shimizu]

Experimental design with circulant property and its application to fMRI experiment

Yuan-Lung Lin[1*], Frederick Kin Hing Phoa[1], Ming-Hung Kao[2]

[1]*Institute of Statistical Science Academia Sinica, Taiwan*

[2]*Arizona State University,USA*

Abstract

Experimental designs have been widely used for cost-efficiency. Orthogonal arrays are commonly used to study the effects of many factors simultaneously, but they do not exist in any sizes. Recently, orthogonal arrays with circulant property receive great attention and are applied to experiments in many fields, such as functional magnetic resonance imaging (fMRI). fMRI is a pioneering technology for studying brain activity in response to mental stimuli. Efficient fMRI experimental designs are important for rendering precise statistical inference on brain functions, but a systematic construction method for this important class of designs does not exist. In this poster, we propose an innovative and unified construction method for efficient, if not optimal, fMRI designs via circulant almost orthogonal arrays (CAOAs). Since circulant Hadamard matrices, that can also be viewed as circulant orthogonal arrays of symbols two and strength two, have been conjectured nonexistence, CAOAs are considered.

We characterize this new class of efficient designs and propose a systematic construction via a newly invented algebraic tool called Complete difference sets (CDS). We not only prove the equivalence relation of CDS and CAOA, but also construct many classes of CAOAs with very high efficiency.

Finally, we apply these efficient CAOAs to fMRI experiments, demonstrating that our constructed designs have better properties than the traditional designs in terms of cost-efficiency.

A rediscovery of dis-similarity measure with *U*-statistic for rare variant association

Charlotte Wang[1,3*], Jung-Ying Tzeng[2], Chuhsing Kate Hsiao[3]

[1]*Institute of Statistical Science, Academia Sinica, Taiwan*

[2]*North Carolina State University, USA*

[3]*National Taiwan University, Taiwan*

Abstract

Missing heritability has been an important issue in genome-wide association studies (GWAS). One possible explanation is the failure in detecting the contribution from rare variants whose effects may be large but cannot be identified successfully with existing genotyping techniques or statistical approaches. Similarity-based measures can capture, if designed properly, the differences between biomarkers. Despite the fact that many similarity-based association tests have been proposed for common variants, the utilization for rare variants has not been considered. Here in this study we proposed a *U*-statistic-based test with Hamming distance metric that compares the dis-similarities of SNP genotypes between subjects in the disease and non-disease group. If the SNP-set is not associated with the disease of interest, then the similarities of SNP genotypes between the disease and non-disease group should be large. We also discuss the statistical properties of the test statistic. Simulation results and applications showed that the performance of the proposed test works well regardless of effect directions, percentage of causal SNPs on the analytic region, sample sizes, and the case-to-control ratios, when compared with SKAT and other *U*-statistic-based tests.

Limiting distributions for outputs of automata

Sara Kropf

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

We consider sequences defined as the sum of the output of an automaton. This is a generalization of automatic sequences, the sum-of-digits function and other digital sequences. We asymptotically analyze these sequences when the input of the transducer is a random integer in [0, N).
Depending on properties of the automaton, the sequence is asymptotically normally distributed. We give the expected value and the variance of the sequence, including the main term and the periodic fluctuation in the second order term. We further investigate properties of this periodic fluctuation.

[This is joint work with Clemens Heuberger and Helmut Prodinger]

A nonparametric copula density estimator incorporating information on bivariate marginals

Yu-Hsiang Cheng

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

Copulas have been used in many financial applications recently, such as in the evaluation of the value at risk or the problem of asset allocation. Using a copula allows one to describe the full dependence structure of a multivariate distribution. In this study, we propose a nonparametric method to estimate a copula density when some bivariate marginal information is available. We have established the consistency of the estimator. The results of several simulation studies will be presented.

[This presentation is based on joint work with Dr. Tzee-Ming Huang and Dr. Su-Yun Huang]

A flexible mover-stayer model for left-truncated and interval-censored data

Yi-Ran Lin[1*], Wei-Hsiung Chao[2], Chen-Hsin Chen[1,3]

[1]*Institute of Statistical Science, Academia Sinica, Taiwan*

[2] *National Dong Hwa University, Taiwan*

[3]*National Taiwan University, Taiwan*

Abstract

In epidemiology study, multi-state models have been widely used in analysis of the dynamic disease progression. In some scenarios, a fraction of the study subjects may be non-susceptible to a specific disease event. A mover-stayer model was firstly proposed for multi-state regression modeling by considering a mixture cohort combining the movers (for those who will proceed to subsequent disease states) with the stayers (for those who permanently remain in the initial state since the study entry). By viewing the observed disease progressions as a mixture of finite Markov processes, we generalize it to a mover-stayer model which allows for study subjects having various disease progressions with probabilities of staying in some subsequent state. The proposed model is developed for longitudinal data in the presence of left truncation, interval-censoring, and possible right censoring. The maximum likelihood estimation procedure is implemented with the Newton-Raphson method.

In the analysis of the longitudinal follow-up data from the REVEAL-HBV study which is a community-based cohort study carried out in seven townships of Taiwan, we pursue the risk evaluation of viral load elevation and associated liver disease/cancer with hepatitis B virus (HBV). A four-state mover-stayer model taking account of subjects' different ages at study entry is developed to assess the multistage, multifactorial process of the liver disease progression starts from chronic hepatitis B to cirrhosis or hepatocellular carcinoma. The proposed regression analysis method can also be applied to the analysis and interpretation for studying other diseases in research of epidemiology and biobanks in the future.

Keywords: longitudinal data, multi-state model, mover-stayer model, finite mixture model, left truncation, interval censoring, right censoring

Two-level minimum aberration designs under a conditional model with a pair of conditional and conditioned factors

Ming-Chung Chang

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

Two-level factorial designs are considered under a conditional model with a pair of conditional and conditioned factors. Such a pair can arise in many practical situations. With properly defined main effects and interactions, an appropriate effect hierarchy is introduced under the conditional model. A complementary set theory as well as an efficient computational procedure, supported by a powerful recursion relation, are developed to implement the resulting design strategy, leading to minimum aberration designs. This calls for careful handling of many new and subtle features of the conditional model as compared to the traditional one.

Some key words: Bias; Complementary set; Effect hierarchy; Model robustness; Orthogonal array; Regular design; Universal optimality; Wordlength pattern.

[This is joint work with Rahul Mukerjee and C. F. Jeff Wu]

Oral sessions

08:50~17:10

February 1st-2nd, 2016

Integrating multiple randomized sketches of SVD and PCA

Su-Yun Huang

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

The computation of singular value decomposition (SVD) and principal component analysis (PCA) is expensive when the size of the matrix becomes large. Instead of processing a full scale of the matrix, a randomized sketch is to work on a much smaller matrix, which is the full matrix multiplied by a random projection matrix to a lower dimensional subspace. In this talk, we will present an integration algorithm to combine results from multiple randomizations. This integration algorithm is based on a Kolmogorov-Nagumo type average. It consists of iterative steps of lifting orthogonal matrices (which are points on Stiefel manifold) to its tangent space, taking average on the tangent space, and then map back to the manifold. We will demonstrate the numerical performance and also give theoretic results of the asymptotic behavior of the integrated SVD.

[joint work with David Chang, Hung Chen, Ting-Li Chen, Chen-Yao Lin and Weichung Wang]

Robust speaker identification

Smarajit Bose

Indian Statistical Institute, India

Abstract

A novel solution to the speaker identification problem is proposed through minimization of the statistical divergence between the (hypothetical) probability distribution (g) of feature vectors from the test utterance and the probability distributions of the feature vector corresponding to the speaker classes. This approach is made more robust to the presence of outliers, through the use of suitably modified versions of the standard divergence measures. Three such measures were considered – the Likelihood Disparity, the Hellinger distance and the Pearson chi-square distance. The proposed approach was motivated by the observation that, in the case of the Likehood Disparity, when the empirical distribution function is used to estimate g, it becomes equivalent to maximum likelihood classification with Gaussian Mixture Models (GMMs) for speaker classes, a highly effective approach proposed by Reynolds (1995). Significant improvement in classification accuracy is observed under this approach on the benchmark speech corpus NTIMIT and a new bilingual speech corpus NISIS. Moreover, the ubiquitous principal component transformation, by itself and in conjunction with the principle of classifier combination, is found to enhance the performance further.

The gamma-lasso: robust estimation for sparse gaussian graphical modeling

Kei Hirose[1], Hironori Fujisawa[2*]

[1]*Osaka University, Japan*

[2]*Institute of Statistical Mathematics, Japan*

Abstract

Gaussian graphical modeling has been widely used to explore various network structures, such as gene regulatory networks and social networks. We often use a penalized maximum likelihood approach with the *L1* penalty for learning a high-dimensional graphical model. However, the penalized maximum likelihood procedure is sensitive to outliers. To overcome this problem, we introduce a robust estimation procedure based on the gamma-divergence. The parameter estimation procedure is constructed using the Majorize-Minimization algorithm, which guarantees that the objective function monotonically decreases at each step. An extensive simulation study showed that our procedure performed much better than the existing methods when the contamination rate was large. We provide two real data examples to illustrate the usefulness of our proposed procedure.

On model selection from a finite family of possibly misspecified models

Ching-Kang Ing

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

Consider finite parametric models.

In many practical situations, we are often faced with the fundamental problem of selecting a model from a finite family of candidate models, none of which is necessarily the true data generating process (DGP). Although existing literature on model selection is quite vast, the above problem does not seem to have received much attention. In fact, model selection problems are usually classified into two categories according to whether the true DGP is included among the family of candidate models. The first category assumes that the true DGP belongs to the candidate family, and the objective of model selection is simply to choose this DGP with probability as high as possible.

The second category assumes that the true DGP is not one of the candidate models. In this case, one primary objective is to choose the model that has the best prediction capability. However, most existing model selection criteria can only perform well in at most one category, and hence when the underlying category is unknown, the choice of selection criteria becomes a serious point of contention. More seriously, none of them has addressed the fundamental problem mentioned above. In this article, we propose a misspecification-resistant information criterion (MRIC) to overcome this difficulty under the fixed-dimensional framework, which requires that the family of candidate finite-parametric models is fixed, independent of the sample size. We prove the asymptotic efficiency of MRIC regardless of whether the true DGP belongs to the candidate family or not.

We also illustrate MRIC's finite-sample performance using Monte Carlo simulation.

Parameter estimation for diffusion processes with high-frequency data

Teppei Ogihara

Institute of Statistical Mathematics, Japan

Abstract

In the study of portfolio risk management of financial assets, it is significant to estimate a variance-covariance matrix of security prices. We study maximum-likelihood-type estimation and its asymptotic behaviors for security prices modeled by parametric diffusion processes with high-frequency observations. In particular, we focus on two problems on analysis of high-frequency data, that is, nonsynchronous observations and the presence of observation noise called market microstructure noise. We construct maximum-likelihood-type estimator of parameters under these problems, and study its asymptotic mixed normality and asymptotic efficiency.

Stress dependent strength models: constructions and applications to HALT

Ashis SenGupta

Indian Statistical Institute, India

Abstract

Strength of a system in general depends inversely on the stress it is subjected to. Probability models for these systems are constructed, where strength and/or stress may be defined on the positive half or the entire real-line. The requirement of negative correlation or dependency gives rise to some interesting and challenging problems with both the construction of as well as the inference for such probability models. Some dependency results are derived for these models. Inference procedures are then developed for the situation where such models are applied to Highly Accelerated Life Testing (HALT) problems. Challenges for extensions of such models to the multivariate case are discussed. Some further applications of the proposed models are given and illustrated by several real-life examples.

Response adaptive allocation design for circular treatment outcomes

Atanu Biswas

Indian Statistical Institute, India

Abstract

Ethics is often maintained by assigning a larger number of subjects to the better (or best) treatment in phase III of a clinical trial. Considering ethical aspects, a number of allocation designs are developed for continuous and binary treatment outcomes. However, if the response is circular in nature, the definition of a better treatment differs from that under the linear response and hence the already developed designs lack appropriateness. In the current work, we propose an invariant allocation function for circular treatment outcomes together with the response adaptive route for practical implementation. We study the resulting design both theoretically and numerically with the help of extensive simulation. The procedure is also illustrated in the light of a real life example on cataract surgery.

[This is based on a joint work with Rahul Bhattacharya and Taranga Mukherjee]

A measure of dissimilarity between aggregated symbolic data with categorical variables

Junji Nakano[1*], Nobuo Shimizu[1], Yoshikazu Yamamoto[2]

[1]*The Institute of Statistical Mathematics, Japan*

[2]*Tokushima Bunri University, Japan*

Abstract

We often have huge amount of individual data with many categorical variables. As they are too huge, it is impossible to see each individual data closely. One way to handle them is to divide them into meaningful groups and compare these groups. Symbolic data analysis provides statistical methods to treat such groups.

We consider that individual data is expressed by dummy variables to describe categorical variables, and use first and second order sample moments of variables to describe a group. We call them as “aggregated symbolic data”. Clearly first order sample moments are marginal distributions of each categorical variable, and second order sample moments are given by Burt matrix whose submatrix is a contingency table for a pair of categorical variables. Note that Burt matrix also contains marginal distributions in it.

We define the dissimilarity between two Burt matrices. We can consider that numbers in cells of Burt matrix are distributed with multinomial distributions. If two groups are similar, probabilities are same for two Burt matrices. If two groups are different, probabilities should be different. Therefore, we can test the null hypothesis that the two groups are same against the alternative hypothesis that two groups are different by a likelihood ratio test statistic using probabilities of multinomial distributions, and use the test statistics as a measure of dissimilarity between two aggregated symbolic data. It is not difficult to decompose the dissimilarity into several meaningful components.

SMART: Statistical Metabolomics Analysis – an R Tool

Hsin-Chou Yang

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

Metabolomics data provides unprecedented opportunities to decipher metabolic mechanisms through studying hundreds to thousands metabolites. Data quality issues and complex batch effects in a metabolomics study must be tackled by statistical analysis properly. This study was aimed to develop an integrated analysis tool for metabolomics studies to streamline the whole analysis flow from the initial data preprocessing to downstream association analysis. We developed SMART (Statistical Metabolomics Analysis - an R Tool) which can analyze different formats of input files, visualize various types of data features, implement peak alignment, carry out quality control for samples and peaks, explore batch effects, and perform association analysis. A pharmacometabolomics study of antihypertensive medication was conducted and data were analyzed by using SMART. Neuromedin *N* was identified as an important metabolite significantly associated with an antihypertensive medication, angiotensin converting enzyme inhibitor, in the metabolome-wide association analysis [$p = 7.30\times10^{-5}$ in an analysis of covariance (ANCOVA) with an adjustment for unknown latent groups and $p = 2.13\times10^{-4}$ in an ANCOVA with an adjustment for hidden substructures].

[Joint work with Yu-Jen Liang, Yu-Ting Lin, Chia-Wei Chen, Chien-Wei Lin, Kun-Mao Chao, and Wen-Harn Pan]

Statistical inference based on bridge divergences

Ayanendranath Basu

Indian Statistical Institute, India

Abstract

M-estimators offer simple robust alternatives to the maximum likelihood estimator. Much of the robustness literature has focused on the problems of location, location-scale, and regression estimation rather than on estimation of general parameters. The density power divergence (DPD) and the logarithmic density power divergence (LDPD) provide two different classes of competitive M-estimators (obtained from divergences) in general parametric models which contain the maximum likelihood estimator as a special case. In each of these families the robustness of the estimator is obtained as a density power down weighting of outlying observations. In this paper we present a generalized family of divergence incorporating the above two classes; this family provides a smooth bridge between the DPD and LDPD measures. Some of the intermediate divergences which provide a good compromise between the two families and help clarify some of the concerns raised by the previous authors.

[This is based on a joint work with Arun Kumar Kuchibhotla]

Quadratic variation estimation of an irregularly observed semimartingale with jumps and noise

Yuta Koike

The Institute of Statistical Mathematics, Japan

Abstract

The estimation of the quadratic variation of a semimartingale observed at a high-frequency is considered. High-frequency financial data are often modeled by discrete observations of a semimartingale, and the quadratic variation can be seen as a measure of the volatility of the corresponding asset, so its estimation has attracted attention in financial econometrics recently. In this talk, the situation where the observation data are contaminated by microstructure noise and the observed semimartingale is allowed to have jumps is considered, and the estimation of the entire quadratic variation is discussed. In such a situation, a pre-averaged version of the realized variance estimator is considered as a natural estimator for the quadratic variation. This talk presents the asymptotic mixed normality of this estimator under the situation where the observation times show irregularity. In particular, the result shows that some standard methods for constructing confidence intervals of the estimator, which are derived under the regular sampling assumption, are still valid in many irregular sampling settings.

Some distributions associated with the cone of positive semidefinite matrices and their applications

Satoshi Kuriki

The Institute of Statistical Mathematics, Japan

Abstract

Let A be a standard Gaussian random matrix in the space Sym(n) of n x n symmetric (or Hermitian) matrices. Let PD(n) be the cone of positive semidefinite matrices in Sym(n). In this talk, we derive the distribution of the squared distance between the random matrix A and the cone PD(n). This distribution appears as the null distribution of the likelihood ratio criterion for testing multivariate variance components. In real and complex normal population cases, the distributions are mixtures of chi-square distributions with weights expressed in terms of the Pfaffian and the determinant, respectively. Moreover, when the size n of the matrix goes to infinity, by modifying Johansson's (1998) central limit theorem for eigenvalues of random matrices, the limiting distribution is proved to be Gaussian. This property of limiting Gaussianity was conjectured in previous literature (e.g., Amemiya, Anderson and Lewis, 1990).

(Joint work with Tomoyuki Shirai and Trinh Khanh Duy of Kyushu University)

A new approach to classical and modern urn models

Antar Bandyopadhyay

Indian Statistical Institute, India

Abstract

In this talk, we will introduce a new approach for studying generalized Pólya urn schemes with balanced replacement mechanisms. We will present a representation of such an urn scheme in terms of a Markov chain associated with the replacement scheme and show that most of the classical asymptotic results may be derived easily from the representation. We will also show that the representation is valid for certain non-standard generalizations with of Pólya urn schemes with infinitely many colors. We will then derive many interesting and new results for such models.

[This is based on a joint work with Debleena Thacker]

A tractable and interpretable four-parameter family of unimodal distributions on the circle

Shogo Kato[1*], M. C. Jones[2]

[1]*Institute of Statistical Mathematics, Japan*

[2]*The Open University, UK*

Abstract

On the circle, as on the line, families of unimodal distributions with parameters controlling location, scale or concentration, skewness and, in some appropriate sense, kurtosis, are useful for robust modelling. Although numerous such families now exist on the line, fewer exist on the circle.

In this talk we present a family of four-parameter distributions for circular data by taking a new approach. Properties of the proposed family include: unimodality; a simple characteristic function and tractable density and distribution functions; interpretable parameters individually measuring location, concentration, skewness and kurtosis, respectively; a wide range of skewness and "kurtosis"; some submodels including the wrapped Cauchy and cardioid distributions; closure under convolution and multiplication by certain constants; straightforward parameter estimation by both method of moments (suitable for smaller samples and moderate parameter values) and maximum likelihood. We will show that our new proposal compares favourably with some of the current four-parameter unimodal families on the circle. Finally, an illustrative application of the proposed model is given.

A new SOP for accurate and efficient community detection

Frederick K. H. Phoa

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

Community is one of the most important features in social networks. There were many traditional methods in the literature to detect communities in network science and sociological studies, but few were able to identify the statistical significance of the detected communities. Even worse, these methods were computationally infeasible for networks with large numbers of nodes and edges. In this talk, we introduce a new SOP for detecting communities in a social network accurately and efficiently. It consists of four main steps. First, a screening stage is proposed to roughly divide the whole network into communities via complement graph coloring. Then a likelihood-based statistical test is introduced to test for the significance of the detected communities. Once these significant communities are detected, another likelihood-based statistical test is introduced to check for the focus centrality of each community. Finally, a metaheuristic swarm intelligence based (SIB) method is proposed to ne tune the range of each community from its original circular setting. Some famous networks are used as empirical data to demonstrate the process of this new SOP.

Quantifying indirect evidence in network meta-analysis

Hisashi Noma

Institute of Statistical Mathematics, Japan

Abstract

Network meta-analysis has enabled comprehensive synthesis of evidence concerning multiple treatments and their simultaneous comparisons based on both direct and indirect evidence. A fundamental pre-requisite of network meta-analysis is consistency of evidence that is obtained from different sources, particularly whether direct and indirect evidence are in accordance with each other or not, and how they contribute to the final estimates. We develop a novel frequentist framework and methods for quantifying indirect evidence and decomposing direct and indirect evidence, as well as testing procedures for evaluating their inconsistency. One of the synthesized estimators concerning indirect evidence is obtained by the Lindsay's composite likelihood method based on a partial likelihood that excludes the conventional direct comparison likelihood from the total likelihood. We show that the composite likelihood estimator has complete information for the indirect evidence. In addition, we propose another indirect comparison estimator that is naturally obtained from the decomposition formulae and that can be interpreted as an exact decomposed estimator from the direct evidence. Using these methodologies, we can also assess the degree of consistency and contribution rates of direct and indirect evidence to the overall estimate. These methods enable intuitive and quantitative interpretations of indirect evidence for the entire network as well as sensitivity analyses for assessing the influences of potentially inconsistent direct comparisons. Applications to a network meta-analysis of 12 new-generation antidepressants are provided.

Adjustment of bifurcated river flow measurements from historical data: paving the way for an international accord

Kaushik Jana

Indian Statistical Institute, India

Abstract

In this paper, we consider an estimation problem arising in the measurement of bifurcated flow of the Teesta, a trans-boundary river flowing through India and Bangladesh. The location of measurement is an Indian barrage, located upstream of the international border, where a part of the flow is diverted from the main stream to a canal. The flows through the two channels are measured indirectly from the height of water on the upstream side and the dimensions of the control structures. The computational formula involves a hydrological constant used as a multiplier, which depends on the control structures used in the two channels. It appears that incorrect multipliers are currently used in the computational formula, and consequently the measured flows through the main stream and the canal are affected by different and unknown distortion factors. Our goal is to estimate the ratio of the distortion factors, so that the measurements of the flows can be brought to a common scale. This problem has important implications on sharing of the river water between India and Bangladesh. We present a model with carefully considered assumptions. The model permits identification of the ratio of the distortion factors, and it permits diagnostic tests for validation of the assumptions. We provide a consistent method for estimating the desired ratio and study the small sample performance of the method through a simulation study. Analysis of the data shows that the estimated ratio is about 0.76. Further analysis shows that adjustment of the historical measurements through this estimated factor substantially reduces the observed discrepancy between average aggregate flows before and after the diversion commenced. Similar adjustment of emerging measurements would help the governments of India and Bangladesh to effectively implement and monitor any water sharing agreement.

[This is based on a joint work with Debasis Sengupta and Kalyan Rudra]

Association mapping of quantitative traits: population stratification and count phenotypes

Saurabh Ghosh

Indian Statistical Institute, India

Abstract

Clinical end-point traits are often defined in terms of quantitative precursors (e.g., systolic blood pressure for hypertension, fasting glucose levels for Type 2 Diabetes, etc.). It has been argued that analyzing the quantitative precursors instead of the binary clinical end-points may be statistically a more powerful strategy to map genes that modulate the underlying complex trait. It is now well established that population stratification can result in spurious association findings in genetic case-control studies. Thus, it is of interest to evaluate the adverse effects of population stratification on the analyses of quantitative traits. The two popular statistical tests of association for quantitative traits using population level data are ANOVA and Kruskal-Wallis. We have theoretically shown that neither genetic heterogeneity nor phenotypic heterogeneity alone can affect the false positive rate of either of the tests. However, if the data comprise subpopulations with unequal allele frequencies at the marker locus of interest as well as different phenotypic means or distributions, the rate of false positives will be elevated. We have also carried out simulations under different genetic models and probability distributions of quantitative traits to assess the extent of increase in the rate of false positives in the presence of population stratification.

For psychiatric disorders, traits such as symptom counts often serve as endophenotypes of interest for understanding the genetic basis of the clinical end-point trait. Since such traits are discrete in nature, it may not be optimal to use ANOVA or the Kruskal Wallis test to detect association. For population level data, we propose a Poisson regression approach that computes the likelihood of the count phenotype conditional on an additive allele count at a SNP. For family-based data involving trios with at least one heterozygous parent at a SNP, we use a similar Poisson regression model conditional on two indicator variables: the marker allele transmitted by the heterozygous parent and the marker allele transmitted by the other parent. We evaluated the performance of the proposed Poisson regressions using extensive simulations. We applied our method to analyze an endophenotype defined as the number of externalizing symptoms associated with alcoholism using data generated in the Collaborative Study On the Genetics of Alcoholism (COGA) project. We found significant evidence of association in the class 1 alcohol dehydrogenase subunit ADH1C in the 4q22.3 region.

[This is based on joint works with Tanushree Haldar and Hemant S Kulkarni]

Analysis of spatial-temporal gene expression patterns reveals dynamics and regionalization in developing mouse brain

Chen-Hsiang Yeang

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

Allen Brain Atlas (ABA) provides a valuable resource of spatial/temporal gene expressions in mammalian brains. Despite rich information extracted from this database, current analyses suffer from several limitations. First, most studies are either gene-centric or region-centric, thus are inadequate to capture the superposition of multiple spatial-temporal patterns. Second, standard tools of expression analysis such as matrix factorization can capture those patterns but do not explicitly incorporate spatial dependency. To overcome those limitations, we proposed a computational method to detect recurrent patterns in the spatial-temporal gene expression data of developing mouse brains. We demonstrated that regional distinction in brain development could be revealed by localized gene expression patterns. The patterns expressed in the forebrain, medullary and pontomedullary, and basal ganglia are enriched with genes involved in forebrain development, locomotory behavior, and dopamine metabolism respectively. In addition, the timing of global gene expression patterns reflects the general trends of molecular events in mouse brain development. Furthermore, we validated functional implications of the inferred patterns by showing genes sharing similar spatial-temporal expression patterns with Lhx2 exhibited differential expression in the embryonic forebrains of Lhx2 mutant mice. These analysis outcomes confirm the utility of recurrent expression patterns in studying brain development.

Tight clustering for large datasets with an application to microarray data

Indranil Mukhopadhyay

Indian Statistical Institute, India

Abstract

This article is aimed to propose a practical and scalable version of tight clustering algorithm, originally introduced by Tseng and Wong (2005). This algorithm provides tight and stable relevant clusters as output leaving a set of points as noise or scattered points that would not go into any cluster. However the computational limitation to achieve this precise target of tight clusters prohibits it from being used for large microarray gene expression data or any other large datasets, which are common nowadays. We propose a modified and scalable version of tight clustering method that is applicable to a dataset of very large size. With extensive simulation study and a real gene expression data analysis we present the validity of our proposed algorithm.

[This is based on a joint work with Bikram Karmakar, Sarmistha Das, Sohom Bhattacharya, and Rohan Sarkar]

An empirical Bayes confidence interval for high leverage area in small area estimation

Masayo Yoshimori Hirose

Institute of Statistical Mathematics, Japan

Abstract

It is well-known that second-order empirical Bayes confidence interval is reliable for small area inference in terms of its length for a large number of areas. Nevertheless, another type confidence interval not based on empirical Bayes theory could be more reliable when the number of areas is not large. Even in such a case, Yoshimori and Lahiri (2014) produces a new second-order empirical Bayes confidence interval ensuring a smaller length than that of another type confidence interval under the Fay-Herriot model. However, this interval still has the disadvantage that it is hard to utilize for high leverage area. In this presentation, we suggest an alternative confidence interval for high leverage area. Incidentally, it also reduces the computer burden.

Solving large scale penalized regression problems via parallel proximal algorithms

Tso-Jung Yen

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

We develop an algorithm for solving regression estimation problems involving structured l_{0}-norm penalty functions. This algorithm incorporates the ideas of the proximal gradient method and iterative hard-thresholding. It carries out numerical computation by decomposing the task into several sub-tasks that can be solved separately in parallel. It obtains updates for parameters by using closed form representations for proximal operators of structured l_{0}-norm penalty functions. It is scalable in terms of sample size or the number of parameters, and is able to be implemented under a data parallelism framework. We demonstrate performance of the algorithm by conducting several simulation studies.