The Symposium of Frontiers of Statistics and Data Sciences

25 & 26 June 2016, The Hong Kong Polytechnic University

Program and Abstracts

Contents

General Information

Organizing Committee

Conference Chair:
Prof. Heung Wong, The Hong Kong Polytechnic University
Committee Members
Dr. Xin Guo, The Hong Kong Polytechnic University
Dr. Binyan Jiang, The Hong Kong Polytechnic University
Dr. Catherine Liu, The Hong Kong Polytechnic University
Dr. Zhen Pang, The Hong Kong Polytechnic University
Dr. Xingqiu Zhao, The Hong Kong Polytechnic University

Contacting Information

Post, fax or email to:
Dr. Binyan Jiang
Department of Applied Mathematics
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
Fax: (852) 27646349
Email: by.jiang@polyu.edu.hk

Registration

June 24, 2016, 2:00 pm – 5:45 pm, TU717
June 25, 2016, 8:30 am – 9:00 am, Y301
June 26, 2016, 8:30 am – 9:00 am, Y301

Conference Venues

Plenary Sessions: Y301
Parallel Sessions: Y301, Y303, Y305, Y306

Social Events

Welcome Dinner
June 24, 2016, 18:00 – 20:00, V Cuisine
Group Photo
June 25, 2016, 10:00–10:10, Y301
Lunch
June 25, 2016, 12:30 – 14:00, 4/F Communal Building
Symposium Banquet
June 25, 2016, 18:00 – 20:00, Metropolis Harbour View Chinese Cuisine
Lunch
June 26, 2016, 12:00 – 14:00, Kong Chiu Lau

Closing Dinner
June 26, 2016, 18:00 – 20:00, V Cuisine

Program at a glance
Saturday, 25 June 2016

Venue: Y301 08:30–08:45
Registration
08:45–09:00 Opening Ceremony Master of Ceremonies: Xin Guo
Prof. Kwok-yin Wong, Associate Vice President, The Hong Kong Polytechnic University Prof. Kai Wang Ng, President, Hong Kong Statistical Society Prof. Heung Wong, Department of Applied Mathematics, The Hong Kong Polytechnic University
09:00–10:00
Plenary I : Tony Cai (p.5) *Chair: Catherine Chunling Liu*
10:00–10:30
Tea Break & Photo Session
10:30–11:30
Plenary II : Runze Li (p.5) *Chair: Zhen Pang*

Venue: Y301 *Chair: Jingchen Hu*	**Venue: Y303** *Chair: Xin Guo*	**Venue: Y305** *Chair: Zhen Pang*
11:40–12:05	11:40–12:05	11:40–12:05
Rui Song (p.15)	Kai W. Ng (p.14)	Jian Qing Shi (p.15)
12:05–12:30	12:05–12:30	12:05–12:30
Yiming Ying (p.19)	Qiang Wu (p.17)	Alan Welsh (p.17)

12:30–14:00 Lunch Break	
Venue: Y301 *Chair: Heung Wong*	**Venue: Y303** *Chair: Binyan Jiang*
14:00–14:25	14:00–14:25
Jinshan Liu (p.13)	Yu Cheng (p.8)
14:25–14:50	14:25–14:50
Quanxi Shao (p.14)	Jialiang Li (p.13)
14:50–15:15	14:50–15:15
Qiang Xia (p.17)	Liming Xiang (p.18)
	15:15–15:40
	Yuan Yao (p.18)
15:40–16:00 Tea Break	
Venue: Y301 *Chair: Frédéric Ferraty*	**Venue: Y303** *Chair: Yuan Yao*
16:00–16:25	16:00–16:25
Ana Colubi (p.9)	Kaoru Irie (p.10)
16:25–16:50	16:25–16:50
Zhengchu Guo (p.10)	Jingchen Hu (p.10)
16:50–17:15	16:50–17:15
Erricos John Kontoghiorghes (p.11)	Matt Wand (p.16)
17:15–17:40	17:15–17:40
Jae Chang Lee (p.12)	Benchong Li (p.12)

Sunday, 26 June 2016

Venue: Y301 9:00–10:00	
Plenary III : Nicholas P. Jewell (p.6) *Chair: Xingqiu Zhao*	
10:00–10:20	
Tea Break	
Venue: Y301 *Chair: Kaoru Irie*	**Venue: Y303** *Chair: Jianguo Sun*
10:20–10:45	10:20–10:45
Hao Chen (p.7)	Yi Li (p.13)
10:45–11:10	10:45–11:10
Zehua Chen (p.7)	Cheng Wang (p.16)
11:10–11:35	11:10–11:35
Xinyuan Song (p.15)	Lingsong Zhang (p.20)
11:35–12:00	11:35–12:00
Jin-Ting Zhang (p.19)	Hui Zou (p.21)
12:00–14:00 Lunch Break	
Venue: Y301 14:00–15:00	
Plenary IV : Jianqing Fan (p.6) *Chair: Binyan Jiang*	
15:00–15:20	
Tea Break	
Venue: Y301 *Chair: Catherine Chunling Liu*	**Venue: Y303** *Chair: Xingqiu Zhao*
15:20–15:45	15:20–15:45
Andreas Christmann (p.8)	Zhezhen Jin (p.11)
15:45–16:10	15:45–16:10
Frédéric Ferraty (p.9)	Shuangge Ma (p.14)
16:10–16:35	16:10–16:35
Linglong Kong (p.11)	Jianguo Sun (p.16)
16:35–17:00	16:35–17:00
Wenyang Zhang (p.20)	Ying Zhang (p.20)

Abstracts of Plenary Talks

Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity

Tony Cai
Department of Statistics, The Wharton School, University of Pennsylvania
Email: tcai@wharton.upenn.edu

Confidence sets play a fundamental role in statistical inference. In this paper, we consider confidence intervals for high dimensional linear regression with random design. We first establish the convergence rates of the minimax expected length for confidence intervals in the oracle setting where the sparsity parameter is given. The focus is then on the problem of adaptation to sparsity for the construction of confidence intervals. Ideally, an adaptive confidence interval should have its length automatically adjusted to the sparsity of the unknown regression vector, while maintaining a prespecified coverage probability. It is shown that such a goal is in general not attainable, except when the sparsity parameter is restricted to a small region over which the confidence intervals have the optimal length of the usual parametric rate. It is further demonstrated that the lack of adaptivity is not due to the conservativeness of the minimax framework, but is fundamentally caused by the difficulty of learning the bias accurately.

Projection Test for High-Dimensional Mean Vectors with Optimal Direction

Runze Li
The Pennsylvania State University
Email: rli@stat.psu.edu

Testing the population mean is fundamental in statistical inference. When the dimensionality of a population is high, traditional Hotelling's T2 test becomes practically infeasible. In this paper, we propose a new testing method for high-dimensional mean vectors. The new method projects the original sample to a lower-dimensional space and carries out a test with the projected sample. We derive the theoretical optimal direction with which the projection test possesses the best power under alternatives. We further propose an estimation procedure for the optimal direction, so that the resulting test is an exact t-test under the normality assumption and an asymptotic chi-square test with 1 degree of freedom without the normality assumption. Monte Carlo simulation studies show that the new test can be much more powerful than the existing methods, while it also well retains Type I error rate. The promising performance of the new test is further illustrated in a real data example.

Counting Civilian Casualties–Statistics and Human Rights

Nicholas P. Jewell
University of California, Berkeley
Email: jewell@berkeley.edu

Civilian casualties are increasingly in the news as conflicts erupt at various regions of the globe. There is often confusion around the numbers of civilians who have been killed, particularly during the earliest stages of conflicts. Sometimes the range of estimates can cover several order of magnitudes, even decades after conflicts have ended, and thus estimates and claims can be hard to interpret especially when they are provided by advocacy groups on either side of a conflict. A related, but quite different human rights problem arises from the need to assess data on the number and patterns of child abductions and disappearances that occurred during El Salvador's civil war (approximately from 1978-1992). The need for accurate counts of civilian casualties (and other human rights events), particularly deaths, turns out to be a relatively modern phenomenon that has nevertheless attracted considerable scientific and political attention. I will discuss some historical background for casualty counts and include three current techniques that have been typically employed. These methods range from traditional demographic and epidemiological survey techniques, to multiple systems estimation that uses capture-recapture models, to attempts to provide full documentation including the modern technological approaches through crowdsourcing. I will also discuss trying to make casualty estimates in almost real time.

Guarding from Spurious Discoveries in High Dimension

Jianqing Fan
Department of Operations Research and Financial Engineering, Princeton University
Email: jqfan@princeton.edu

Many data-mining and statistical machine learning algorithms have been developed to select a subset of covariates to associate with a response variable. Spurious discoveries can easily arise in high-dimensional data analysis due to enormous possibilities of such selections. How can we know statistically our discoveries better than those by chance? In this paper, we define a measure of goodness of spurious fit, which shows how good a response variable can be fitted by an optimally selected subset of covariates under the null model, and propose a simple and effective LAMM algorithm to compute it. It coincides with the maximum spurious correlation for linear models and can be regarded as a generalized maximum spurious correlation. We derive the asymptotic distribution of such goodness of spurious fit for generalized linear models and L_1-regression. Such an asymptotic distribution depends on the sample size, ambient dimension, the number of variables used in the fit, and the covariance information. It can be consistently estimated by multiplier bootstrapping and used as a benchmark to guard against spurious discoveries. It can also be applied to model selection, which considers only candidate models with goodness of fits better than those by spurious fits. The theory and method are convincingly illustrated by simulated examples and an application to the binary outcomes from German Neuroblastoma Trials. This is a joint work with Wenxin Zhou.

Abstracts of Invited Talks

Change-point detection for locally dependent data

Hao Chen
University of California, Davis
Email: hxchen@ucdavis.edu

Local dependence is common in multivariate and object data sequences. We consider the testing and estimation of change-points in such sequences. A new way of permutation, circular block permutation with a randomized starting point, is proposed and studied for a scan statistic utilizing graphs representing the similarity between observations. The proposed permutation approach could correctly address for local dependence and make it possible the theoretical treatments for the non-parametric graph-based scan statistic for locally dependent data. We derive accurate analytic approximations to the significance of graph-based scan statistics under the circular block permutation framework, facilitating its application to locally dependent multivariate or object data sequences.

Simultaneous Feature Selection and Precision Matrix Estimation in High Dimensional Multivariate Regression Models

Zehua Chen
Department of Statistics and Applied Probability, National University of Singapore
Email: stachenz@nus.edu.sg

We consider multivariate regression models with a q-dimensional response vector, a p-dimensional feature space and a sample of size n. We deal with the problem of feature selection and precision matrix estimation of the models in the case that both q and p are large compared with the sample size n. In theoretical consideration, we allow them to diverge to infinity as n goes to infinity. We give a conditional formulation of the multivariate regression model and propose an iterated alternate method which alternates at each iteration between a feature selection step and a precision matrix estimation step. At the feature selection step, we use a sequential feature selection procedure called sequential Lasso (SLasso). At the precision matrix estimation step, we adopt the neighborhood detection approach and use a sequential scaled pairwise selection (SSPS) method. We will discuss the detailed algorithm of the iterated alternate method as well as its asymptotic properties. Simulation studies comparing the iterated alternate method with other available methods will be presented. An application to a real data set will be reported as well.

Association analysis of gap times with multiple causes

Xiaotian Chen, Yu Cheng*, Ellen Frank, David Kupfer
Department of Statistics, University of Pittsburgh
Email: yucheng@pitt.edu

We aim to close a methodological gap in analyzing durations of successive events that are subject to induced dependent censoring as well as competing-risk censoring. In the Bipolar Disorder Center for Pennsylvanians (BDCP) study, some patients who managed to recover from their symptomatic entry later developed a new depressive or manic episode. It is of great clinical interest to quantify the association between time to recovery and time to recurrence in patients with bipolar disorder. The estimation of the bivariate distribution of the gap times with independent censoring has been well studied. However, the existing methods cannot be applied to failure times that are censored by competing causes such as in the BDCP study. Bivariate cumulative incidence function (CIF) has been used to describe the joint distribution of parallel event times that involve multiple causes. To the best of our knowledge, however, there is no method available for successive events with competing-risk censoring. Therefore, we extend the bivariate CIF to successive events data, and propose nonparametric estimators of the bivariate CIF and the related conditional CIF. Moreover, an odds ratio measure is proposed to describe the cause-specific dependence, leading to the development of a formal test for independence of successive events. Simulation studies demonstrate that the estimators and tests perform well for realistic sample sizes, and our methods can be readily applied to the BDCP study.

Regularized Kernel Methods With Special Emphasis on Additive Models

Andreas Christmann
Department of Mathematics, University of Bayreuth, Germany
Email: andreas.christmann@uni-bayreuth.de

Regularized kernel based methods, e.g. support vector machines, play an important role in modern nonparametric statistics, machine learning theory, and in approximation theory, see e.g. Vapnik (1998), Cucker and Zhou (2007), and Steinwart and Christmann (2008). Although often treated in a purely nonparametric manner, such kernel methods can also be used for additive models which are popular in semiparametric statistics. There are at least three reasons why additive models are important in applied statistics. Sometimes there is some prior knowledge to justify the application of an additive model. Sometimes the applied researcher is only interested in an additive model and not in a purely nonparametric model, because current laws imply that the estimator can be explained to non-experts and is not a black-box technique. Finally, additive models are less prone to the curse of high dimensionality than nonparametric methods.

The talk will discuss some recent results on consistency and robustness of regularized kernel based methods for independent or for dependent data as well as learning rates for regularized kernel based methods for additive models. These learning rates compare favourably in particular in high dimensions to recent results on optimal learning rates for purely nonparametric regularized kernel based methods using the Gaussian RBF kernel, if the assumption of an additive model is valid.

Keywords: RKHS; Additive Model; Learning rate; Robustness; Nonparametric; Semiparametric.

References

[1] Christmann, A. and Zhou, D.X. (2015). Learning rates for the risk of kernel-based quantile regression estimators in additive models. *Accepted, online first: Analysis and Applications.*

[2] Christmann, A. and Hable, R. (2012). Consistency of support vector machines using additive kernels for additive models. *Computational Statistics & Data Analysis* **56**, 854 – 873.

[3] Cucker, F. and Zhou D. X. (2007). *Learning Theory. An Approximation Theory Viewpoint.* Cambridge University Press. Cambridge.

[4] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines.* Springer. New York.

[5] Vapnik, V. N. (1998). *Statistical Learning Theory.* John Wiley & Sons. New York.

Hypothesis testing about the expected value of random intervals

Ana Colubi
University of Oviedo, Spain
Email: colubi@uniovi.es

Random intervals will be considered as natural interval data generating elements. A framework based on distances and the set-arithmetic useful to handle interval data from an 'ontic' perspective will be reviewed. The focus will be on the two main parameters within this framework: The Aumann expected value and the Frechet variability. The distance-based approach makes it natural to develop some inferences, e.g. asymptotic and bootstrap techniques for the one-sample, the 2-sample and the k-sample cases. However, the extensions of these classical tests are not enough to handle interval data, which are richer in some senses. To illustrate this, partial and full inclusion tests for the Aumann expectation will be developed and the theory behind will be discussed. Empirical results based on simulations and a case-study will also be considered for illustrative purposes.

Nonparametric regression on contaminated functional predictor with application to hyperspectral data

Frédéric Ferraty
Toulouse Mathematics Institute, France
Email: ferraty@math.univ-toulouse.fr

We propose to regress nonparametrically a scalar response Y on a random curve X when only a contaminated version X^* of X is observable at some measurement grid. To override this common setting, a kernel presmoothing step is achieved on the noisy functional predictor X^*. Afterthen, the kernel estimator of the regression operator is built using the smoothed functional covariate instead of the original corrupted ones. Rate of convergence is stated for this nested-kernel estimator with a special attention on high-dimensional setting (i.e the size of the measurement grid is much larger than the sample size). The proposed method is applied on simulated datasets in order to assess its finite-sample properties. Our methodology is further illustrated on a real hyperspectral dataset involving a supervised classification problem.

Learning Theory of Distributed Spectral Algorithms

Shaobo Lin, Zhengchu Guo*, and Ding-Xuan Zhou
School of Mathematical Sciences, Zhejiang University
Email: guozc@zju.edu.cn

Spectral algorithms have been widely used and studied in learning theory and inverse problems. In this talk, we consider distributed spectral algorithms for handling big data based on a divide-and-conquer approach. We present these distributed kernel-based learning algorithms in a regression framework including nice error bounds and optimal minimax learning rates achieved by means of a novel integral operator approach and a second order decomposition of inverse operators.

This is a joint work with Prof. Ding-Xuan Zhou and Dr. Shaobo Lin.

Dirichlet Process Mixture Models for Nested Categorical Data

Jingchen Hu*, Jerome P. Reiter, Quanli Wang
Vassar College, USA
Email: jihu@vassar.edu

We present a Bayesian model for estimating the joint distribution of multivariate categorical data when units are nested within groups. Such data arise frequently in social science settings, for example, people living in households. The model assumes that (i) each group is a member of a group-level latent class, and (ii) each unit is a member of a unit-level latent class nested within its group-level latent class. This structure allows the model to capture dependence among units in the same group. It also facilitates simultaneous modeling of variables at both group and unit levels. We develop a version of the model that assigns zero probability to groups and units with physically impossible combinations of variables. We apply the model to estimate multivariate relationships in a subset of the American Community Survey. Using the estimated model, we generate synthetic household data that could be disseminated as redacted public use files with high analytic validity and low disclosure risks.

Fox News Network Data Analysis: Bayesian Dynamic Modeling

Xi Chen, Kaoru Irie*, David Banks, Robert Haslinger, Jewell Thomas, and Mike West
Faculty of Economics, University of Tokyo
Email: ki22@stat.duke.edu

We propose a Bayesian approach to analyze data on Internet traffic flow among Fox News websites. The observations are time-varying counts (non-negative integers), so the straightforward application of existing Gaussian-type state-space models is not available. It is a Big Data problem, with many different types of articles, raising scalability issues. These features of the data motivate use of dynamic versions of count data models (Poisson-Gamma models and Multinomial-Dirichlet models), and lead to fitting an interpretable Gravity model that is an equivalent to two-way ANOVA. The conjugacy of this model enables use of Forward Filtering and Backward Sampling to obtain the posterior distributions. In addition, the Gravity model reveals the underlying structure of traffic networks across websites, allowing the detection of significant flows and flow dynamics, and enabling computational advertisers to better target their ad campaigns.

This is the joint work with Xi Chen, David Banks, Mike West (Duke), Robert Haslinger and Jewell Thomas (MaxPoint).

Analysis of data from Alzheimer's disease studies subject to censoring

Zhezhen Jin
Department of Biostatistics, Mailman School of Public Health, Columbia University, USA
Email: zj7@cumc.columbia.edu

In this talk, some statistical issues and methods arise in the analysis of data from Alzheimer's disease studies will be discussed. One is in the area of screening patients for the risk of developing Alzheimer's disease with a 40-item test by the standardized University of Pennsylvania Smell Identification Test (UPSIT) of olfactory function, which involves the issues of item selection and assessment. Another area to be discussed is on the analysis of health-related quality of life data that arise from Alzheimer's disease studies, in which patients' mental and physical conditions are measured over their follow-up periods and the resulting data are often complicated by subject-specific measurement times and possible terminal events associated with outcome variables.

Optimal Estimation for Quantile Regression with Functional Response

Linglong Kong
Dept of Math & Stat, University of Alberta
Email: lkong@ualberta.ca

Quantile regression with functional response and scalar covariates has become an important statistical tool for many neuroimaging studies. In this paper, we study optimal estimation of varying coefficient functions in the framework of reproducing kernel Hilbert space. Minimax rates of convergence under both fixed and random designs are established. We have developed easily implementable estimators which are shown to be rate-optimal. Simulations and real data analysis are conducted to examine the finite-sample performance. This is a joint work with Xiao Wang, Zhengwu Zhang, and Hongtu Zhu.

The estimation of the downdated general linear model

Erricos John Kontoghiorghes
Cyprus University of Technology and Birkbeck University of London, UK
Email: erricos@dcs.bbk.ac.uk

It is often computationally infeasible to re-estimate afresh a large-scale model when a small number of observations has been modified. Furthermore for a large data set out of core and recursive algorithms need to be developed. Within this context, a numerically stable algorithm is proposed to re-estimate the general linear model (GLM) after observations are deleted. This is known as the downdated-GLM problem. The new estimation algorithm updates the original GLM model with the imaginary deleted observations. This, results to an updated-GLM having imaginary values and a non-positive definite dispersion matrix which comprises complex covariance values. It is proven that the generalised least square estimator of the updated-GLM is the same as that of estimating afresh the downdated-GLM. The algorithm computes efficiently the equivalent estimator of the updated-GLM without performing any complex arithmetic and by using orthogonal transformations. The computational performance of the new algorithm is demonstrated. In addition the case of the GLM with singular dispersion matrix and the downdated seemingly unrelated regression model are addressed.

Where can we find “Statistics and Statisticians” in this rapid changing world?

Jae Chang Lee
Korea University, Korean
Email: jaeclee@korea.ac.kr

Statistics is a scientific discipline that deals research methodologies involving DATA. But the form and nature of DATA have been transforming and developing fast in many directions in the last few decades along with rapid change like digital technology. Statistics started with aims to confirm scientific hypothesis and estimate parameters in scientific models based on observed data with uncertainty statements. It was possible because of well-designed experiments and objective measurement capabilities. From the second half of last century we have been facing a deluge of “non-statistical data” that are typically “by-products” of public administration, health services, communications or business operations. They are also called “found data” recently to make it distinct from Statistical DATA. Using high speed computing power many new patterns of information and knowledge have been discovered. Complexity, size and variety of these data naturally attracted scientists for their explorative investigation and they have been very successful. Traditional statistics has been at the center of all these exploratory endeavors, but no one is ready to give a full credit to statistics. It was because statisticians never tried to promote effectively with “SEXY MARKETING” buzzwords like “DATA MINING” or “BIG DATA” for their tools and know-how. They simply watched them passed by letting others reinvent what they have done for a long time. With statements made by John Tukey (1963), Jeff Wu (1997), Bill Cleveland (2001) just to name a few we will examine what happened in this recent history. We also examine what should be the key roles of statistics in this changing world and what should be considered for the future of so-called “Data Science” to be beneficial for science and human development. A few Examples will be given for a better future of statisticians.

VC dimension induced by discrete Markov networks and Bayesian networks

Benchong Li*, Youlong Yang, Yang Li
School of Mathematics and Statistics, Xidian University, Xi’an 710126, P. R. China
Email: libc580@nenu.edu.cn

Markov networks and Bayesian networks are two popular models for classification. Vapnik-Chervonenkis dimension and Euclidean dimension are two measures of complexity of a class of functions. In this paper, we show that these two dimensional values of the concept class induced by a discrete Markov network are identical, and the value equals dimension of the toric ideal corresponding to this Markov network as long as the toric ideal is nontrivial. Furthermore, for a general Bayesian network, we show that dimension of the corresponding toric ideal offers an upper bound of Euclidean dimension.

Survival Impact Index and Ultrahigh-Dimensional Model-Free Screening with Survival Outcomes

Jialiang Li
Department of Statistics and Applied Probability, National University of Singapore, Singapore
Email: stalj@nus.edu.sg

Motivated by ultrahigh-dimensional biomarkers screening studies, we propose a model-free screening approach tailored to censored lifetime outcomes. Our proposal is built upon the introduction of a new measure, survival impact index (SII). By its design, SII sensibly captures the overall influence of a covariate on the outcome distribution, and can be estimated with familiar nonparametric procedures that do not require smoothing and are readily adaptable to handle lifetime outcomes under various censoring and truncation mechanisms.We provide large sample distributional results that facilitate the inference on SII in classical multivariate settings. More importantly, we investigate SII as an effective screener for ultrahigh-dimensional data, not relying on rigid regression model assumptions for real applications. We establish the sure screening property of the proposed SII-based screener. Extensive numerical studies are carried out to assess the performance of our method compared with other existing screening methods. A lung cancer microarray data is analyzed to demonstrate the practical utility of our proposals.

Classification with Ultrahigh-Dimensional Features

Yi Li
University of Michigan
Email: yili@med.umich.edu

Although much progress has been made in classification with high-dimensional features, classification with ultrahigh-dimensional features, wherein the features much outnumber the sample size, defies most existing work. This paper introduces a novel and computationally feasible multivariate screening and classification method for ultrahigh-dimensional data. Leveraging inter-feature correlations, the proposed method enables detection of marginally weak and sparse signals and recovery of the true informative feature set, and achieves asymptotic optimal misclassification rates. We also show that the proposed procedure provides more powerful discovery boundaries compared to those in Cai and Sun (2014) and Jin et al. (2009). The performance of the proposed procedure is evaluated using simulation studies and demonstrated via classification of patients with different post-transplantation renal functional types.

Bayesian Analysis of Multiple Thresholds Autoregressive Model

Jinshan Liu
Department of Mathematics, South China Agricultural University
Email: liujs58@yahoo.com.cn

Bayesian analysis of threshold autoregressive (TAR) model with various possible thresholds is considered. A method of Bayesian stochastic search selection is introduced to identify a threshold-dependent sequence with highest probability. All model parameters are computed by a hybrid Markov chain Monte Carlo (MCMC) method, which combines the Metropolis-Hastings (M-H) algorithm and Gibbs sampler. The main innovation of the method introduced here is to estimate the TAR model without assuming the fixed number of threshold values, thus is more flexible and useful. Simulation experiments and real data examples lend further support to the proposed approach.

Promote similarity in integrative analysis

Shuangge Ma
Department of Biostatistics, Yale University, USA
Email: Shuangge.ma@yale.edu

For multiple high-dimensional problems, it is desired to conduct the integrative analysis of multiple independent datasets. Under a few important scenarios, it can be expected that the estimates of multiple datasets are "similar" in certain aspects, which may include magnitude, sparsity structure, sign, and others. The existing approaches do not have a mechanism promoting such similarity. In our study, we conduct the integrative analysis of multiple independent datasets. Penalization techniques are developed to explicitly promote similarity. The consistency properties are rigorously established. Numerical studies, including simulation and data analysis, show that the proposed approach has significant advantages over the existing benchmark.

Adequate sample size for using limiting distributions

Kai W. Ng
The University of Hong Kong
Email: kaing@hku.hk

In recent years, non-Bayesian inference of parameters is mostly based on limiting distributions as the sample size, or the number of observational time points in quantitative finance, "goes to infinity" whatever that means. As usual, such limiting theories in statistics do not provide any guideline on the adequate sample size for using the limiting distribution in actual applications, e.g. the blanket usage of normal distribution for Maximum Likelihood Estimator (MLE). Thus the researchers and even the journal referees all seem to act as if this question on sample size should not be asked at all. Thus the applied users happily follow suit in using the limiting distributions without any question on sample size in their own applications. This talk aims to serve a humble reminder with simple examples that the sample size is critical for using a limiting distribution in applications. It is because the convergence to the limiting distribution is NOT a uniform convergence over all the values of the same set of parameters, so that the required sample size can be vastly different being a function of the true parameter value which is not yet known. For the particular limiting normality of MLE, there is another reason adding weight to the requirement of adequate sample size, due to the subtle conflict of the *invariance principle* of limiting normality under a continuous one-to-one transform of MLE (i.e. under any re-parametization) against the *basic principle* of ordinary normality that a random variable (vector) having a normal distribution will not have normal distribution again after any of the foresaid transformations except the linear transform.

Big Data Challenges in Hydroclimate Research

Quanxi Shao
Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia
Email: Quanxi.Shao@csiro.au

Big data has been becoming a popular and hot research topic in many research fields and has attracted many industrial investments. It is undoubted that any big data product would be a multi-disciplinary effort. For our fellow statisticians, we need to identify our role in the big data game. Based on our experiences in the statistical applications in hydroclimate research, in this talk, I will share and present some thoughts with examples on where we can make significant contributions in the big data movement. The talk will also touch some of our newly planned research in other fields such as finance and computational system.

Nonlinear Mixed-effects Scalar-on-function Models and Variable Selection for Kinematic Upper Limb Movement Data

Jian Qing Shi
Newcastle University, UK
Email: jian.shi@newcastle.ac.uk

In this talk, I will discuss a nonlinear mixed-effects scalar-on-function regression model using a Gaussian process prior. This model is motivated from the analysis of movement data which are collected in our current joint project on assessing upper limbs' function after stroke. The talk will focus on a novel variable selection algorithm, namely functional least angle regression (fLARS), and demonstrate how the algorithm can be used to do variable selection from a very large number of candidates including both scalar and function-valued variables. Numerical results including simulation study and application to the big movement data will also be discussed.

Concordance-Assisted Learning for Estimating Optimal Individualized Treatment Regimes

Rui Song
Department of Statistics, North Carolina State University
Email: rsong@ncsu.edu

We propose a new concordance-assisted learning for estimating optimal individualized treatment regimes. We first introduce a type of concordance function for prescribing treatment and propose a robust rank regression method for estimating the concordance function. We then nd treatment regimes, up to a threshold, to maximize the concordance function, named prescriptive index. Finally, within the class of treatment regimes that maximize the concordance function, we need the optimal threshold to maximize the value function. We establish the convergence rate and asymptotic normality of the proposed estimator for parameters in the prescriptive index. An induced smoothing method is developed to estimate the asymptotic variance of the proposed estimator. We also establish the $n^{1/3}$-consistency of the estimated optimal threshold and its limiting distribution. In addition, a doubly robust estimator of parameters in the prescriptive index is developed under a class of monotonic index models. The practical use and effectiveness of the proposed methodology are demonstrated by simulation studies and an application to an AIDS data.

Analysis of Proportional Mean Residual Life Model with Latent Variables

Xinyuan Song
Chinese University of Hong Kong
Email: xysong@sta.cuhk.edu.hk

In this study, we propose a proportional mean residual life (MRL) model with latent variables to assess the effects of observed and latent risk factors on MRL function in the presence of right censoring. We employ a factor analysis model to characterize latent risk factors via multiple observed variables. We develop a borrow-strength estimation procedure, which incorporates the expectation-maximization algorithm and the corrected estimating equation approach. The asymptotic properties of the proposed estimators are established. Simulation shows that the performance of the proposed methodology is satisfactory. The application to the study of type 2 diabetes reveals insights into the prevention of end stage renal disease.

Sieve Maximum Likelihood Regression Analysis of Bivariate Interval-censored Failure Time Data

Jianguo Sun
Department of Statistics, University of Missouri-Columbia, USA
Email: sunj@missouri.edu

Interval-censored failure time data arise in a number of fields and many authors have discussed various issues related to their analysis. However, most of the existing methods are for univariate data and there exists only limited research on bivariate data, especially on regression analysis of bivariate interval-censored data. In this talk, we will discuss a class of semiparametric transformation models for the problem and for inference, a sieve maximum likelihood approach will be developed. The model provides a great flexibility, in particular including the commonly used proportional hazards model as a special case, and in the approach, Bernstein polynomials are employed. An illustrative example will also be discussed.

Semiparametric Mean Field Variational Bayes

Matt Wand
School of Mathematical and Physical Sciences, University of Technology Sydney
Email: matt.wand@uts.edu.au

Mean field variational Bayes is gaining popularity as an alternative to Markov chain Monte Carlo for approximate Bayesian inference. Even though mean field variational Bayes is less accurate, it has advantages such as speed, parallelisability and the possibility of real-time analyses of streaming data. This makes mean field variational Bayes attractive in particular applications, especially those where Markov chain Monte Carlo is not viable because of size and speed considerations. In this talk we discuss an important extension known as semiparametric mean field variational Bayes. A particular focus is numerical issues of semiparametric mean field variational Bayes.

On dimensionality effects in linear discriminant analysis for large dimensional data

Cheng Wang
Department of Mathematics, Shanghai Jiao Tong University
Email: chengwang@sjtu.edu.cn

We study the asymptotic results of linear discriminant analysis (LDA) in large dimensional data where the observation dimension p, is of the same order of magnitude as the sample size n. Roughly speaking, we know when $p/n \to 0$, LDA is an “good" classifier which means the empirical misclassification error tends to the theoretical one and if $p/n \to y \in (0,1)$, we should pay some price for estimating the means and covariance matrix. The explicit theoretical results about dimensionality effects in LDA will be derived in this work. We also study the dimensionality effects of the regularized LDA (RLDA). Specially, we get the asymptotic distribution of the misclassification error using recent results in random matrix theory. Based on these results, a scale adjusted classifier will be suggested to the classical LDA to handle data with un-equal sample sizes. Finally, simulations will be conducted to support these results.

Fitting misspecified linear mixed models

Alan Welsh
Mathematical Sciences Institute, The Australian National University
Email: Alan.Welsh@anu.edu.au

Linear mixed models are widely used in a range of application areas, including ecology and environmental science. We study in detail the effects of fitting the two-level linear mixed model with a single explanatory variable that is misspecified because it incorrectly ignores contextual effects. In particular, we make explicit the effect of (the usually ignored) within-cluster correlation in the explanatory variable. This approach produces a number of unexpected findings. (i) Incorrectly omitting contextual effects affects estimators of both the regression and variance parameters not just, as is currently thought, estimators of the regression parameters and the effects are different for different estimators. (ii) Increasing the within cluster correlation of the explanatory variable introduces a second local maximum into the log-likelihood and REML criterion functions which eventually becomes the global maximum, producing a jump discontinuity (at different values) in the maximum likelihood and REML estimators of the parameters. (iii) Standard statistical software such as SAS, SPSS, STATA, lmer (from lme4 in R) and GenStat often returns local rather than global maximum likelihood and REML estimates in this very simple problem. (iv) Local maximum likelihood and REML estimators may fit the data better than their global counterparts but, in these situations, ordinary least squares may perform even better than the local estimators, albeit not as well as if we fit the correct model.

Bias Correction for Regularization Regression and its Applications

Qiang Wu
Department of Mathematical Sciences, Middle Tennessee State University
Email: qwu@mtsu.edu

We propose an approach to reduce the bias of ridge regression and regularization kernel network. Compared to the original regularization algorithms, the bias corrected algorithms have smaller bias but larger variance When applied to a single data set, the bias corrected algorithms have comparable learning performance to the uncorrected algorithms. When applied to incremental learning with block wise streaming data or in the divide-and-conquer method, the bias corrected algorithms are more efficient due to bias reduction. This is verified by theoretical characterizations and simulation studies.

Determining the Number of Factors for High-dimensional Time Series

Qiang Xia
Department of Mathematics, South China Agricultural University
Email: xiaqiang@scau.edu.cn

In this paper, we suggest a new method of determining the number of factors in factor modeling for high-dimensional stationary time series. When the factors are of different degrees of strength, the eigenvalue-based ratio method of Lam and Yao (2012) needs a two-step procedure to estimate the number of factors. As a modification of the method of Lam and Yao (2012),Âăhowever, ourÂă method only need a one-step procedure for the determination of the number of factors.Âă The resulted estimator is obtained simply by minimizing the ratio of the contribution of two adjacent eigenvalues. Some asymptotic results are also developed for the proposed method. The finite sample performance of the method is well examined by some Monte Carlo simulations and a real data analysis.

Semiparametric regression analysis of clustered semi-competing risks data

Liming Xiang
School of Physical and Mathematical Sciences, Nanyang Technological University
Email: LMXiang@ntu.edu.sg

Semi-competing risks data often arise in clinical studies in which a subject can experience both non-terminal and terminal events. The time to the intermediate non-terminal event (e.g. relapse) is subject to dependent censoring by the terminate event (e.g., death) but not vice versa. Typically times to the two events are correlated. In many applications, subjects may also be nested within cluster, such as patients in a multi-center study, leading to possible association among event times due to unobserved shared factors across subjects. We seek to conduct marginal regressions and joint association analysis for clustered time-to-event data with semi-competing risks. We propose a semiparametric modeling framework where a copula model is used for the joint distribution of the nonterminal and terminal event times, and marginal distributions for the two event times are modeled by Cox proportional hazards regressions with a shared frailty to incorporate association of event times within the same cluster. We develop a nonparametric maximum likelihood estimation procedure for estimating unknown baseline cumulative hazards and parameters in the model. The estimators are shown to be consistent and asymptotically normal. An extensive simulation study evidences that our inferential procedure performs very well with finite sample sizes. The practical utility of the method is illustrated with a real data set.

End-point sampling

Yuan Yao*, Wen Yu, Kani Chen
Department of Mathematics, Hong Kong Baptist University, Hong Kong
Email: yaoyuan@hkbu.edu.hk

Retrospective sampling designs, including case-cohort and case-control designs, are commonly used for failure time data in the presence of censoring. In this talk, we propose a new retrospective sampling design, called end-point sampling, which improves the efficiency of the case-cohort and case-control designs. The regression analysis is conducted using the Cox model. Under different assumptions, the maximum likelihood approach as well as the inverse probability weighting approach is developed respectively to estimate the regression parameters. The resulting estimators are proved to be consistent and asymptotically normal. Simulation and real data studies show favorable evidence for the proposed design in comparison with the existing ones.

Stochastic Online AUC Maximization

Yiming Ying*, Siwei Lyu, and Longyin Wen
Department of Mathematics and Statistics, State University of New York at Albany, NY, USA
Email: yying@albany.edu

Online learning has been actively studied due to its high efficiency to big data. However, most of such studies focus on the misclassification error or prediction accuracy. AUC (Area Under the ROC Curve) is a widely used performance measure for binary classification. This measure is particularly suitable for applications where the labels are imbalanced. A specific challenge in developing online AUC maximization algorithms is that the objective function is usually defined over a *pair* of training examples from opposite classes. In this talk, I will present a new online learning algorithm for AUC maximization, which only needs to pass the data once. Existing online AUC algorithms have expensive space and time complexities which are *quadratic* $\mathcal{O}(d^2)$ where d is the dimensionality of the data. In contrast, our online algorithm has a *linear* space and per-iteration complexity $\mathcal{O}(d)$. I will also present the theoretical results about the convergence of the proposed algorithm and discuss the relationship with the standard stochastic gradient descent. Encouraging experimental results will be presented. This talk is based on a joint work with Siwei Lyu and Longyin Wen from SUNY Albany.

Linear Hypothesis Testing for Heteroscedastic One-Way MANOVA with High-Dimensional Data

Jin-Ting Zhang
National University of Singapore
Email: stazjt@nus.edu.sg

In recent decades, with rapid development of data collecting technologies, high-dimensional data become increasingly prevalent. Much work has been done for hypotheses on mean vectors, especially for high-dimensional two-sample problems. Rather than considering a specific problem, we are interested in a general linear hypothesis testing (GLHT) problem in heteroscedastic one-way MANOVA for high-dimensional data, which include many existing hypotheses about mean vectors as special cases. Several existing methodologies on this important GLHT problem impose strong assumptions on the underlying covariance matrix so that the null distributions of the associated test statistics are asymptotically normal. In this paper, we propose a simple and adaptive L2-norm based test for the above GLHT problem. For normal data, we show that the null distribution of our test statistic is the same as that of a chi-square type mixture which is generally skewed. We give a sufficient and necessary condition such that the null distribution of our test statistic is asymptotically normal. However, this condition is not always satisfied in real data analysis. To overcome this difficulty, we propose to approximate the distribution of our test statistic using the well-known Welch-Satterthwaite chi-squared-approximation. The ratio-consistent estimators of the associated parameters are obtained. The asymptotic and approximate power of our new test is also investigated. The methodologies are then extended for non-normal data. Two simulation studies and a real data application are presented to demonstrate the good performance of our new test.

On Imbalanced and High Dimensional Low Sample Size Classification

Lingsong Zhang
Department of Statistics, Purdue University
Email: lingsong@purdue.edu

In this talk, we will address some challenges and issues related to classifying imbalanced high dimensional low sample size data sets. We will discuss both binary classifiers and multi-category classification methods. Two popular methods, Support Vector Machine (SVM) and Distance Weighted Discrimination (DWD) will be used as examples. Novel classification methods that possess the merits of both methods are proposed. We show that the new classifier inheres the merit of DWD, and hence, overcomes the data-piling and overfitting issue of SVM. On the other hand, the new method is not subject to imbalanced data issue which was a main advantage of SVM over DWD. Several theoretical properties, including Fisher consistency and asymptotic normality of the DWSVM solution are developed. We use some simulated examples to show that the new method can compete DWD and SVM on both classification performance and interpretability.

Model and Feature Selection in a Class of Semiparametric Models

Wenyang Zhang
Department of Mathematics, University of York, UK
Email: wenyang.zhang@york.ac.uk

Model selection is an old song in statistics. With the surge of high dimensionality in recent years, people start to play it with a new tune —- the penalised likelihood method. In this talk, I am going to investigate a class of semiparametric models where the number of potential explanatory variables grows much fast than the sample size. I am going present a new penalised likelihood procedure which selects the important features and identifies the correct model simultaneously. I will explore the effectiveness of the penalty part in the proposed procedure, and present a new way to put penalty. Asymptotic properties will be presented to justify the proposed methodology. I will also show the performance of the proposed procedure when sample size is finite by simulation studies. Finally, I will illustrate the application of the proposed method by a real data example.

Nonparametric Inference with Misclassified Competing Risks Data

Ying Zhang
Department of Biostatistics, Indiana University, USA
Email: yz73@iu.edu

Competing risks data play an important role in medicine, epidemiology and public health. However, a frequent complication in biomedical research is that ascertainment of the causes of event is subject to error and this can lead to seriously biased inference. To deal with this issue when misclassification probabilities are not a-priory known, a double-sampling design can be adopted to randomly select a small sample of subjects with terminal event and then using a gold standard, which could be a very expensive outcome ascertainment procedure, to determine the true cause. Based on this additional information and a parametric model for the probability of the misclassification on the causes, we develop a closed form nonparametric pseudolikelihood estimator (NPMPLE) of the cause-specific cumulative incidences and we show that the estimator is uniformly consistent and converges weakly to a tight zero-mean Gaussian random field. We conduct simulation studies to justify the validity of the proposed method. Finally, the method is applied to a motivating example from a large HIV/AIDS study in Sub-Saharan Africa to evaluate the PEPFAR-funded HIV healthcare programs, where serious death under-reporting results in classifying deceased patients as being disengagers from HIV care.

Another Look at Distance Weighted Discriminant

Hui Zou
School of Statistics, University of Minnesota, USA
Email: zouxx019@umn.edu

Distance weighted discrimination (DWD) is a modern margin-based classifier with an interesting geometric motivation. Despite many recent papers on DWD, DWD is far less popular compared with the support vector machine, mainly due to computational and theoretical reasons. In this work, we greatly advance the current DWD methodology and its learning theory. We propose a novel efficient algorithm for solving DWD, and our algorithm can be several hundred times faster than the existing state-of-the-art algorithm based on the second order cone programming (SOCP). In addition, our algorithm can handle the generalized DWD, while the SOCP algorithm only works well for a special DWD but not the generalized DWD. Furthermore, we formulate a natural kernel DWD in a reproducing kernel Hilbert space and then establish the Bayes risk consistency of the kernel DWD using a universal kernel such as the Gaussian kernel. This result solves an open theoretical problem in the DWD literature. We compare DWD and the support vector machine on several benchmark data sets and show that the two have comparable classification accuracy, but DWD equipped with our new algorithm can be much faster to compute than the support vector machine.

List of Participants

Tony CAI
Department of Statistics, The Wharton School, University of Pennsylvania
tcai@wharton.upenn.edu

Hao CHEN
Department of Statistics, University of California, Davis
hxchen@ucdavis.edu

Zehua CHEN
Department of Statistics and Applied Probability, National University of Singapore
stachenz@nus.edu.sg

Yu CHENG
Department of Statistics, University of Pittsburgh
yucheng@pitt.edu

Andreas CHRISTMANN
Department of Mathematics, University of Bayreuth, Germany
andreas.christmann@uni-bayreuth.de

Ana COLUBI
University of Oviedo, Spain
colubi@uniovi.es

Jianqing FAN
Department of Operations Research and Financial Engineering, Princeton University
jqfan@princeton.edu

Frédéric FERRATY
Toulouse Mathematics Institute, France
ferraty@math.univ-toulouse.fr

Xin Guo
Department of Applied Mathematics, The Hong Kong Polytechnic University
x.guo@polyu.edu.hk

Zhengchu GUO
School of Mathematical Sciences, Zhejiang University
guozc@zju.edu.cn

Jingchen HU
Vassar College, USA
jihu@vassar.edu

Kaoru IRIE
Faculty of Economics, University of Tokyo
ki22@stat.duke.edu

Nicholas P. JEWELL
University of California, Berkeley
jewell@berkeley.edu

Binyan JIANG
Department of Applied Mathematics, The Hong Kong Polytechnic University
by.jiang@polyu.edu.hk

Zhezhen JIN
Department of Biostatistics, Mailman School of Public Health, Columbia University, USA
zj7@cumc.columbia.edu

Linglong KONG
Dept of Math & Stat, University of Alberta
lkong@ualberta.ca

Erricos John KONTOGHIORGHES
Cyprus University of Technology and Birkbeck University of London, UK
erricos@dcs.bbk.ac.uk

Jae Chang LEE
Korea University, Korean
jaeclee@korea.ac.kr

Benchong LI
School of Mathematics and Statistics, Xidian University, Xi'an 710126, P. R. China
libc580@nenu.edu.cn

Jialiang LI
Department of Statistics and Applied Probability, National University of Singapore, Singapore
stalj@nus.edu.sg

Runze LI
The Pennsylvania State University
rli@stat.psu.edu

Yi LI
University of Michigan
yili@med.umich.edu

Catherine LIU
Department of Applied Mathematics, The Hong Kong Polytechnic University
catherine.chunling.liu@polyu.edu.hk

Jinshan LIU
Department of Mathematics, South China Agricultural University
liujs58@yahoo.com.cn

Shuangge MA
Department of Biostatistics, Yale University, USA
Shuangge.ma@yale.edu

Kai W. NG
The University of Hong Kong
kaing@hku.hk

Zhen PANG
Department of Applied Mathematics, The Hong Kong Polytechnic University
zhen.pang@polyu.edu.hk

Quanxi SHAO
Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia
Quanxi.Shao@csiro.au

Jian Qing SHI
Newcastle University, UK
jian.shi@newcastle.ac.uk

Rui SONG
Department of Statistics, North Carolina State University
rsong@ncsu.edu

Xinyuan SONG
Chinese University of Hong Kong
xysong@sta.cuhk.edu.hk

Jianguo SUN
Department of Statistics, University of Missouri-Columbia, USA
sunj@missouri.edu

Matt WAND
School of Mathematical and Physical Sciences, University of Technology Sydney
matt.wand@uts.edu.au

Cheng WANG
Department of Mathematics, Shanghai Jiao Tong University
chengwang@sjtu.edu.cn

Alan WELSH
Mathematical Sciences Institute, The Australian National University
Alan.Welsh@anu.edu.au

Heung WONG
Department of Applied Mathematics, The Hong Kong Polytechnic University
heung.wong@polyu.edu.hk

Qiang WU
Department of Mathematical Sciences, Middle Tennessee State University
qwu@mtsu.edu

Qiang XIA
Department of Mathematics, South China Agricultural University
xiaqiang@scau.edu.cn

Liming XIANG
School of Physical and Mathematical Sciences, Nanyang Technological University
LMXiang@ntu.edu.sg

Yuan YAO
Department of Mathematics, Hong Kong Baptist University, Hong Kong
yaoyuan@hkbu.edu.hk

Yiming YING
Department of Mathematics and Statistics, State University of New York at Albany, NY, USA
yying@albany.edu

Jin-Ting ZHANG
National University of Singapore
stazjt@nus.edu.sg

Lingsong ZHANG
Department of Statistics, Purdue University
lingsong@purdue.edu

Wenyang ZHANG
Department of Mathematics, University of York, UK
wenyang.zhang@york.ac.uk

Ying ZHANG
Department of Biostatistics, Indiana University, USA
yz73@iu.edu

Xingqiu ZHAO
Department of Applied Mathematics, The Hong Kong Polytechnic University
xingqiu.zhao@polyu.edu.hk

Hui ZOU
School of Statistics, University of Minnesota, USA
zouxx019@umn.edu

Maps

Figure 1: Campus map

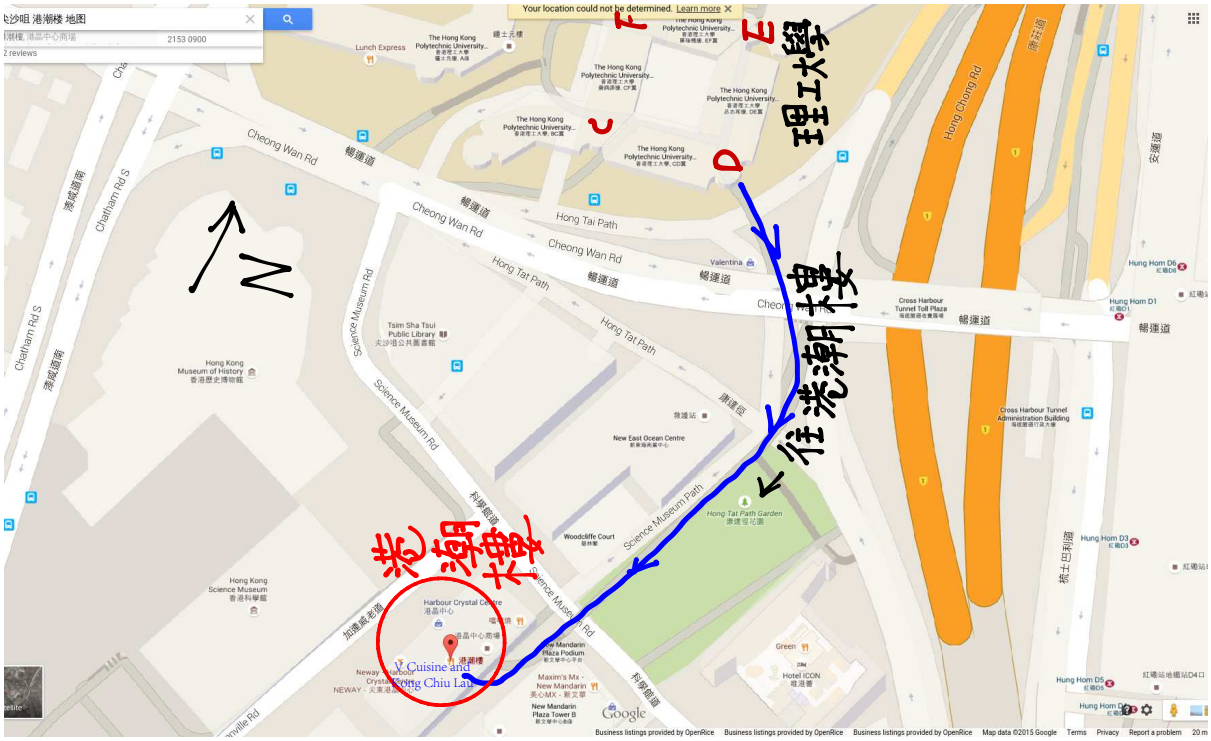

Figure 2: Map to V Cuisine and Kong Chiu Lau