Joint Meeting of

The 2011 Taipei International Statistical Symposium and

7th Conference of the Asian Regional Section of the IASC

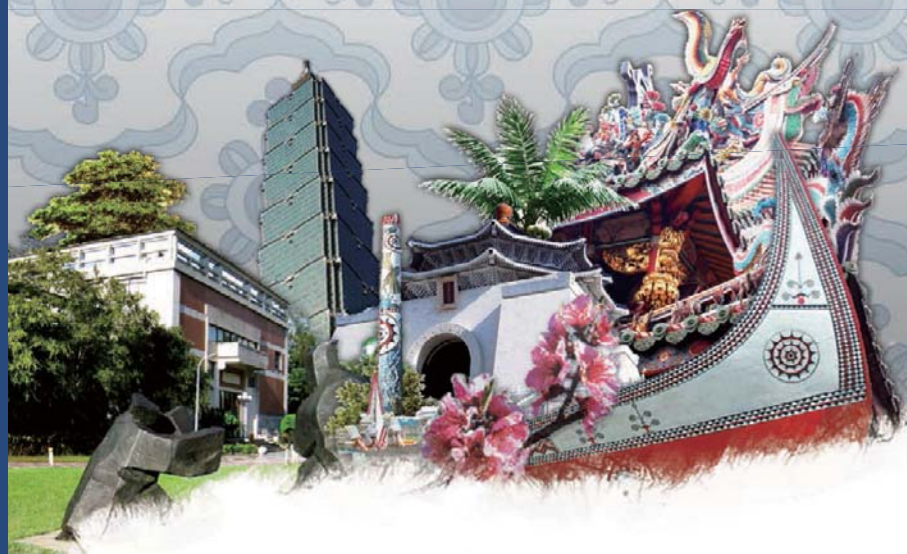
7th IASC-ARS

Σ joint 2011

Taipei Symposium

December 16 - 19, 2011

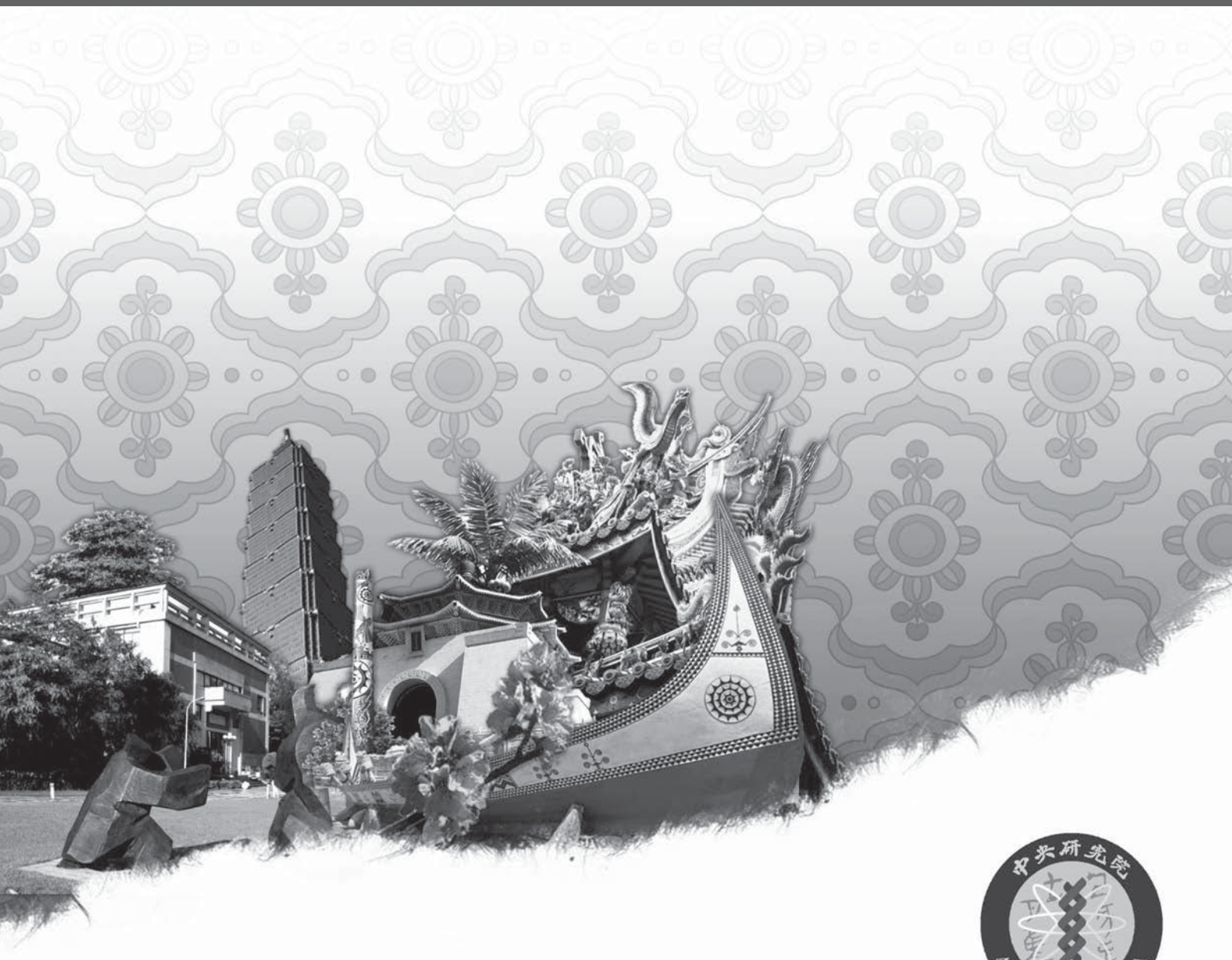
Academia Sinica, Taipei, Taiwan



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

中央研究院 ACADEMIA SINICA 1928

Abstract



December 19 (Monday)

Today's Highlights:

09:00 – 10:30	***Parallel Sessions 19a1 – 19a6***
10:50 – 11:40	***The Statistica Sinica Special Invited Session (II) by Tze Leung Lai***
11:40 – 12:00	***Closing Ceremony***
12:30 –	***Local Tour***

Abstract Pages:

Special Sessions:	SS Invited Session (II) Page 198

Session No.	19a1	19a2	19a3	19a4	19a5	19a6
Starting Page	198	200	201	203	205	206

The Statistica Sinica Special Invited Session (II)
December 19 (Monday), 10:50 - 11:40, HSS Center International Conference Hall
Chair: NgaiHang Chan

Dynamic Empirical Bayes Models and Their Applications to Finance and Insurance

Tze Leung Lai
Department of Statistics, Stanford University, U.S.A.

Empirical Bayes (EB) modeling has a long and celebrated history in statistical theory and applications. In particular, in insurance rate-making, linear EB models have provided a basic framework for credibility theory. However, evolutionary credibility models, in which the individual risk profiles evolve over time, have not used the EB framework but have relied on minimum-variance linear estimation via Kalman filtering. After a brief review of the literature, we propose a new dynamic EB modeling approach which provides flexible and computationally efficient methods for the analysis and prediction of economic panel data. This dynamic EB approach pools the cross-sectional information over individual time series to replace an inherently complicated hidden Markov model by a considerably simpler generalized linear mixed model. We apply dynamic EB modeling to provide a new approach to evolutionary credibility theory, to modeling loan defaults in finance, and to the well-known statistical problem of predicting baseball batting averages studied by Efron and Morris and recently by Brown.

[Tze Leung Lai, Department of Statistics, Stanford University, U.S.A.; lait@stat.stanford.edu]

19a1-Recent Advances in Analysis of Dependent Data
December 19 (Monday), 9:00 - 10:30, HSS 1st Conference Room
Organizer: Shu-Hui Yu
Chair: Shu-Hui Yu

19a1-1 Parametric and Nonparametric Analysis of Temporal Trend in Extreme Values with Applications to Wind Storm Losses and Temperature Data

Nader Tajvidi
Mathematical Statistics, Centre for Mathematical Sciences, Lund Institute of Technology, Sweden

A topic of major current interest in extreme-value analysis is the investigation of temporal trends. For example, the potential influence of 'greenhouse' effects may result in severe storms becoming gradually more frequent, or in maximum temperatures gradually increasing, with time. One approach to evaluating these possibilities is to fit, to data, a parametric model for temporal parameter variation, as well as a model describing the marginal distribution of data at any given

point in time. In this talk we discuss some parametric trend models and illustrate the methods by application to a dataset on windstorm losses in south of Sweden. We shall also discuss difficulties which might arise in formulating structural trend-models. Motivated by datasets on windstorm severity and maximum temperature, we suggest a nonparametric approach to estimating temporal trends when fitting parametric models to extreme values from a weakly-dependent time series. We illustrate the method through applications to time series where the marginal distributions are approximately Pareto, generalised-Pareto, extreme-value or Gaussian. We introduce time-varying probability plots to assess goodness of fit, we discuss local-likelihood approaches to fitting the marginal model within a window, and we propose temporal cross-validation for selecting window width. In cases where both location and scale are estimated together, the Gaussian distribution is shown to have special features that permit it to play a universal role as a 'nominal' model for the marginal distribution (Joint work with Peter Hall and Holger Rootzen).

[Nader Tajvidi, Mathematical Statistics, Centre for Mathematical Sciences, Lund Institute of Technology, Box 118, SE-221 00 Lund, Sweden; nader@maths.lth.se]

19a1-2 **Tuning Parameter Selection in High Dimensional Penalized Likelihood**

Yingying Fan
Cheng Yong Tang
Information and Operations Management Department, Marshall School of Business, University of Southern California, U.S.A.

How to appropriately select the tuning parameter is essential in regularized likelihood methods for high dimensional data analysis. For this purpose, we first examine the generalized information criterion (GIC) for generalized linear models when the dimensionality of covariates p increases exponentially with sample size n. To ensure consistently identifying the true model, we find that the penalty on the model complexity in GIC should diverge at the rate of some power of $log(p)$ depending on the tail probability behavior of the response variables. This reveals that using the AIC or BIC to select the tuning parameter may not be able to consistently identify the true model. We then propose to select the tuning parameter in the regularized likelihood methods by optimizing the GIC with appropriate model complexity penalty. We show that the proposed approach for tuning parameter selection consistently identifies the true model from the candidate models with probability tending to 1. We confirm the performance of the proposed procedure by numerical simulations and a gene expression data analysis.

[Yingying Fan, Information and Operations Management Department, Marshall School of Business, University of Southern California; fanyingy@marshall.usc.edu]

19a1-3 **A Note on Mean Squared Prediction Error under the Unit Root Model with Deterministic Trend**

Shu-Hui Yu
Institute of Statistics, National University of Kaohsiung, Taiwan, R.O.C.

Assume that observations are generated from the first-order autoregressive (AR) model with linear time trend and the unknown model coefficients are estimated by least squares. This paper develops an asymptotic expression for the mean squared prediction error (MSPE) of the least squares predictor in the presence of a unit root. As a by-product, we also obtain a connection between the MSPE and the growth rate of the Fisher information. The key technical tool used to derive these results is the negative moment bound for the minimum eigenvalue of the normalized Fisher information matrix.

[Shu-Hui Yu, Institute of Statistics, National University of Kaohsiung; shuhui@nuk.edu.tw]

19a2-Recent Advances in Statistical Model Selection
December 19 (Monday), 9:00 - 10:30, HSS 2nd Conference Room
Organizer: Ching-Kang Ing
Chair: Ching-Kang Ing

19a2-1 On Choosing the Size of Data Perturbation in Adaptive Model Selection

Hung Chen
Department of Mathematics, National Taiwan University, Taiwan, R.O.C.

Model selection procedures via penalization often use a fixed penalty, such as AIC and BIC, to avoid choosing a model which fits a particular data set extremely well. As a correction for not including the variability induced in model selection, generalized degrees of freedom (GDF) is introduced in Ye (1998) as an estimate of model selection uncertainty that arise in using the same data for both model selection and associated parameter estimation. Built upon generalized degrees of freedom, Shen and Ye (2002) propose a data-adaptive complexity penalty. The estimate of GDF is then obtained by randomly perturbs (adds noise to) the output variable and re-runs the modeling procedure. In this talk, we address how to select the size of perturbation through unbiased risk estimate, Stein Lemma, and generalized derivative with variable selection in nested linear regression models and wavelet thresholding. We also comment its connection with little bootstrap and tiny bootstrap considered in Breiman (1992, 1995, and 1996).

[Hung Chen, Department of Mathematics, National Taiwan University; hchen@math.ntu.edu.tw]

19a2-2 Adaptive Minimax Estimation with Sparse l_q Constraints

Yuhong Yang
University of Minnesota, Minneapolis, U.S.A.

For high-dimensional linear regression, both l_0 and l_1 norms on the coefficients have been used for sparse modeling of the regression function. In this work, we identify the minimax rates of convergence for regression estimation under l_q constraints on the coefficients for $0 < q < 1$ for both random and fixed designs. Furthermore, our estimators based on model combination/selection are showed to simultaneously achieve the optimal rates over the whole range of $0 \le q \le 1$. Our results also permit model mis-specification. The work is joint with Zhan Wang, Sandra Paterlini and Fuchang Gao.

[Yuhong Yang, 313 Ford Hall, 224 Church Street, Minneapolis, MN 55455, U.S.A.; yyang@stat.umn.edu]

19a2-3 **Model Selection Principles in Misspecified Models**

Jinchi Lv
Jun S. Liu
University of Southern California, U.S.A.

Model selection is of fundamental importance to high-dimensional modeling featured in many contemporary applications. Classical principles of model selection include the Kullback-Leibler divergence principle and the Bayesian principle, which lead to the Akaike information criterion and Bayesian information criterion when models are correctly specified. Yet model misspecification is unavoidable in practice. In this paper, we propose a family of semi-Bayesian principles for model selection in misspecified models that bridge the two wellknown principles. We derive novel asymptotic expansions of the semi-Bayesian principles in misspecified generalized linear models, which give the new semi-Bayesian information criteria (SIC_γ). A specific form of SIC admits a natural decomposition into the negative maximum quasi-log-likelihood, a penalty on model dimensionality, and a penalty on model misspecification directly. Numerical studies demonstrate the advantage of the SIC methodology for model selection in both correctly specified and misspecified models.

[Jinchi Lv, University of Southern California; jinchilv@marshall.usc.edu]

19a3-Interface Special Invited Session: Visualization
December 19 (Monday), 9:00 - 10:30, HSS Media Conference Room
Organizer: Juergen Symanzik
Chair: Thomas Mathew

19a3-1 **Data Visualization and Its Interface to Statistical Models**

Adalbert F.X. Wilhelm
Jacobs University Bremen, Germany

The development and availability of efficient statistical software has also led to a wider distribution, a wider use and a broader presence of both exploratory statistical graphics and sophisticated, complex modelling approaches. However, there are only a few publications that link the two fields together. The graphical representation of data quite often requires the corresponding statistical modelling phase to yield practically significant results. On the other hand there are plenty of examples such as the famous Anscombe quartet based on four different data sets but resulting in one identical regression model, that show that statistical modelling without a corresponding visualization misses out on some fundamental features. It seems obvious that the one can't be reasonable performed without the other. Doing statistical modelling without a proper graphical representation of data and model is risky and problematic. Exploring the data graphically without the attempt to model them properly, usually falls short and leaves the analyst with isolated insights and anecdotes. The systematic approach of modelling combined with the flexible use of exploratory graphics combines the strengths of both fields and constitutes a powerful research tool. This talk will illustrate this by providing an eclectic tour through the modelling process and illustrating the potential applications of exploratory graphics in the various steps. We will focus on three main stages of modelling: visualization prior to the modelling to check data quality and model adequacy, during the modelling process to check for model assumptions and model quality and after the modelling process to enhance interpretation of the modelling parameters as well as comparing between competing models.

[Adalbert F.X. Wilhelm, Jacobs University Bremen, School of Humanities and Social Sciences, Campus Ring 1, 28759 Bremen, Germany; a.wilhelm@jacobs-university.de]

19a3-2 Statistical Inference on Graphics

Heike Hofmann
Department of Statistics, Iowa State University, U.S.A.

Statistical graphics play a crucial role in exploratory data analysis, model checking and diagnosis. Formal visual methods for determining statistical significance of findings. This changed, when Buja et al. [2009] conceptually introduced protocols for formal tests of visual findings. We will take this one step further and compare the lineup protocol against classical statistical testing of the significance of regression model parameters. A human subjects experiment is conducted using simulated data to provide controlled conditions. Results suggest that the lineup protocol provides results equivalent to the uniformly most powerful (UMP) test and for some scenarios yields better power than the UMP test.

[Heike Hofmann, Department of Statistics, Iowa State University, U.S.A.; hofmann@iastate.edu]

19a3-3 Visualizing High Dimensional Data: Applying Graph Theory to Data Visualization

Wayne Oldford
Department of Statistics & Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada

In statistical data analysis, we are often looking for structure in high dimensional data. In three or fewer dimensions, our visual system is an important asset, as much (even unanticipated) structure can be recognized effortlessly when points can be plotted so few dimensions. Unfortunately, modern digital data sources (e.g. images, text documents, sound files,...) often have 100s or 1000s of dimensions. Even after formal dimension reduction methods have been applied, we are often faced with many more dimensions than three. In this talk, I will explore some visualization methods for high dimensional data. I will review and illustrate methods based on radial, parallel, and orthogonal coordinates. These three axis systems have different strengths and weaknesses. In all cases however, improvements may be had by casting the axis arrangement in a graph theoretic framework. I will explore the relevant graph theoretic representations and illustrate their use on real data sets (e.g. Frey image data). I will pay particular attention to the orthogonal axis system and show how graph traversal can be used to meaningfully navigate through high dimensional space. All software used is available from R packages PairViz and RnavGraph. This talk is based on joint work with Catherine Hurley, NUI Maynooth (Ireland) and Adrian Waddell, Waterloo (Canada).

[Wayne Oldford, Department of Statistics & Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada; rwoldford@uwaterloo.ca]

19a4-Time Series
December 19 (Monday), 9:00 - 10:30, AC 1st Conference Room
Organizer: Meihui Guo
Chair: Meihui Guo

19a4-1 On Recent Developments of Nonstationary and Long-Memory Time Series

Ngai Hang Chan
Department of Statistics, Chinese University of Hong Kong, Hong Kong

This talk aims at surveying some of the recent developments of nonstationary and long-memory time series. The underlying theme of recent endeavour arises from the consideration of the order of magnitude of the observed Fisher's information number. By using a simple AR(1) model, the so-called "SNoTE", it is shown how this number affects the nonstationary behaviour in a subtle but important way. The talk concludes with the discussion of some of the recent results involving negative moment bounds of the observed Fishers' information number and their applications to mult-step ahead prediction.

[Ngai Hang Chan, Department of Statistics, Chinese University of Hong Kong, Shatin, NT, Hong Kong; nhchan@sta.cuhk.edu.hk]

19a4-2 Doubly Constrained Factor Models: Estimation and Applications

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

Factor models have been widely used in recent years to improve the accuracy of forecasting when many explanatory variables are available. However, the models often encounter the difficulties of over-parameterization and factor interpretation. In this paper, we first consider constrained factor analysis to obtain a parsimonious factor model and propose likelihood ratio statistics to test the adequacy of factor constraints. Real and simulated examples are used to demonstrate the proposed analysis. In an application, we show that the constrained factor analysis can provide a deeper understanding of variations in monthly financial asset returns. We then extend the constrained models to the doubly constrained factor models by incorporating external information on both rows (e.g., subjects) and columns (e.g., variables) of a data matrix. Maximum likelihood estimates and likelihood ratio statistics of the proposed models are derived. Finally, we consider the applications of doubly constrained factor models in economics and finance. (This is a joint work with Ching-Wei Cheng and Ruey S. Tsay).

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

19a4-3 Model Selection for Integrated Autoregressive Processes of Infinite Order

Ching-Kang Ing
Institute of Statistical Science, Academia Sinica, Taiwan, R.O.C.

Chor-Yiu Sin
Department of Economics, National Tsing Hua University, Taiwan, R.O.C.

Shu-Hui Yu
Institute of Statistics, National University of Kaohsiung, Taiwan, R.O.C.

We show that Akaike's information criterion (AIC) and its variants are asymptotically efficient in integrated autoregressive processes of infinite order (AR(1)). This result, together with its stationary counterpart established previously in the literature, ensures that AIC can ultimately achieve prediction efficiency in an AR(1) process, without knowing the integration order.

[Chor-Yiu Sin, Department of Economics, National Tsing Hua University, Taiwan, R.O.C.; cysin@mx.nthu.edu.tw]

19a5-Advances in Computational Biology
December 19 (Monday), 9:00 - 10:30, AC 2nd Conference Room
Organizer: Shwu-Rong Grace Shieh
Chair: Shwu-Rong Grace Shieh

19a5-1 On Automated Flow Cytometric Data Analysis

Geoff McLachlan
Department of Mathematics and Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia

Flow cytometry is widely used for single cell interrogation of surface and intracellular protein expression by measuring fluorescence intensity of fluorophore-conjugated reagents. We report some extensions to our previous work (Pyne et al., 2009, PNAS 106) that provides a procedure called FLAME (FLow analysis with Automated Multivariate Estimation) for automated high-dimensional flow cytometric analysis. The FLAME procedure, which has been incorporated with the GenePattern package of the Broad Institute, uses finite mixture models of heavy-tailed and asymmetric distributions in the form of the skew *t*-component distributions to identify and model cell populations in a flow cytometric sample. In our more recent work, we extend the component distributions in FLAME to include skew *t*-distributions without any restrictions on the form of skewness. A program is given using the EM algorithm that is able to fit these models exactly without need for Monte Carlo computation. It thus broadens the applications of flow cytometry to new biological and clinical problems.

[Geoff McLachlan, Department of Mathematics and Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia; g.mclachlan@uq.edu.au]

19a5-2 Statistical Learning Based on Distributions of Oligonucleotides in DNA Sequences

Probal Chaudhuri
Theoretical Statistics & Mathematics Unit, Indian Statistical Institute, India

I will start with some examples from evolutionary biology to demonstrate how statistical analysis of distributions of various oligonucleotides in DNA sequences can lead to important biological discoveries. The problems considered can be formulated as statistical learning problems that may be supervised, unsupervised or partially supervised in nature. The statistical analysis involves an interesting variable selection problem. This motivated developing certain probabilistic models for

DNA sequences, and those will be discussed.

[Probal Chaudhuri, Theoretical Statistics & Mathematics Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata700108, India; probal@isical.ac.in, probalchaudhuri@gmail.com]

19a5-3 Sequential Lasso for Feature Selection with Ultra-high Dimensional Feature Space

Zehua Chen
Shan Luo
Department of Statistics and Applied Probability, National University of Singapore, Singapore

We propose a novel approach, Sequential Lasso, for feature selection in linear regression models with ultra-high dimensional feature spaces. In this talk, we discuss the asymptotic properties of Sequential Lasso, especially, its selection consistency. Like other sequential methods, the implementation of Sequential Lasso is not limited by the dimensionality of the feature space. However, it has advantages over other sequential methods. The simulation studies comparing Sequential Lasso with other sequential methods and demonstrating the advantages of Sequential Lasso will be reported.

[Zehua Chen, Department of Statistics and Applied Probability, National University of Singapore, Singapore; stachenz@nus.edu.sg]

19a6-Statistics in Finance
December 19 (Monday), 9:00 - 10:30, AC 3rd Conference Room
Organizer: Mike Ka Pui So
Chair: Wai-Sum Chan

19a6-1 Stable Mixture GARCH Models

Marc Paolella
Swiss Banking Institute, University of Zurich, Switzerland

A new model class for univariate asset returns is proposed which involves the use of mixtures of stable Paretian distributions, and readily lends itself to use in a multivariate context for portfolio selection. The model nests numerous ones currently in use, and is shown to outperform all its special cases. In particular, an extensive out-of-sample risk forecasting exercise for seven major FX and equity indices confirms the superiority of the general model compared to its special cases and other competitors. An improved method (in terms of speed and accuracy) is developed for the computation of the stable Paretian density. Estimation issues related to problems associated

with mixture models are discussed, and a new, general, method is proposed to successfully circumvent these. Because of the tractability of the stable Paretian characteristic function, the model is straightforwardly extended to support portfolio selection using expected shortfall as the downside risk measure by using an independent component analysis framework.

[Marc Paolella, Swiss Banking Institute, University of Zurich, Switzerland; marc.paolella@bf.uzh.ch]

19a6-2 Bayesian Analysis of Time-Varying Parameter Vector Autoregressive Model with the Ordering of Variables

Toshiaki Watanabe
Hitotsubashi University, Tokyo, Japan

This paper applies the time-varying parameter vector autoregressive model to the Japanese economy. The both parameters and volatilities, which are assumed to follow a random-walk process, are estimated using a Bayesian method with MCMC. The recursive structure is assumed for identification and the reversible jump MCMC is used for the ordering of variables. The empirical result reveals the time-varying structure of the Japanese economy and monetary policy during the period from 1981 to 2008 and provides evidence that the introduction of zero interest rate policy may have changed the order of variables.

[Toshiaki Watanabe, 2-1 Naka, Kunitachi, Tokyo 186-8603; watanabe@ier.hitu.ac.jp]

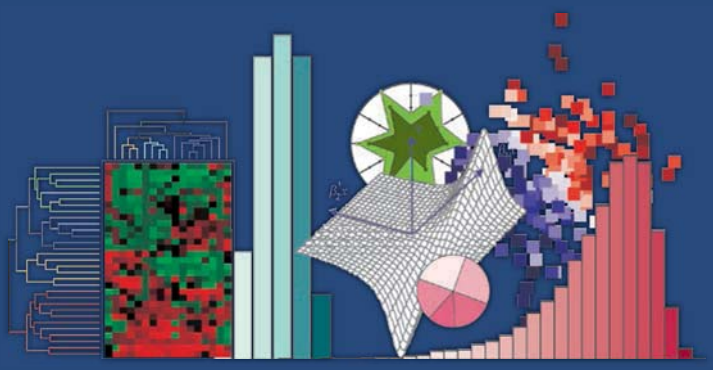
19a6-3 Realized Stochastic Volatility with Leverage and Long Memory

Yasuhiro Omori
University of Tokyo, Tokyo, Japan

The daily return and the realized volatility are simultaneously modeled in the stochastic volatility model with leverage and long memory. In addition to the stochastic volatility model with leverage for the daily returns, ARFIMA process is jointly considered for the realized volatilities. Using a state space representation of the model, we estimate parameters by Markov chain Monte Carlo methods. Model comparison with similar realized stochastic volatility models with short memory is conducted by computing the marginal likelihood.

[Yasuhiro Omori, Faculty of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-Ku, Tokyo 113-0033, Japan; omori@e.u-tokyo.ac.jp]

7th IASC-ARS Σ joint 2011 Taipei Symposium



Institute of Statistical Science, Academia Sinica

Asian Regional Section of the IASC

DGBAS, Executive Yuan, R.O.C.

National Science Council, R.O.C.

The Chinese Institute of Probability and Statistics

中國統計學社 Chinese Statistical Association