Co-expression structure and network analysis for deciphering disease mechanism

Lawrence Chan

Department of Health Technology and Informatics
Hong Kong Polytechnic University

Abstract

Gene expressions change due to the stress experienced by the cells. The changes are not necessarily significant but the relays of such form co-expression networks implicating the underlying molecular interactions and signaling cascades for adapting the stress.

This study aims to identify the difference in co-expression distributions between normal and neoplastic states and explain such structural difference through the scatter plot of co-expression levels in the two states on both genomic scale and specific functional gene sets.

The distributions and the corresponding scatter plot are called co-expression structures and galaxy in this project. Structural analysis will be applied to determine the difference in distributions and the optimal co-expression threshold for partitioning the co-expression galaxy into nine regions. The central region contains weakly co-expressed gene pairs. The surrounding eight regions contain normal-specific, neoplasm-specific, conforming and opposing strongly co-expressed gene pairs. This study also aims to explore how the regional difference in gene pair counts determines the structural difference between the two states on genomic scale.

Nonlinear error correction model with multiple regimes and multiple thresholds cointegration

N.H. Chan and Man Wang

Department of Statistics
The Chinese University of Hong Kong, Hong Kong

Abstract

Nonlinear error correction models (ECM) with multiple regimes have been widely used in finance and statistics. These models encompass the multiple threshold vector ECM as a special case. In this paper, the asymptotic properties of the least squares estimator of a nonlinear ECM with multiple regimes are established. For threshold cointegrated modeled of a threshold vector ECM, estimation procedures based on the least squares principle are examined. Both least squares and smoothed least squares estimations are studied and their asymptotic theories are established. In particular, the super-consistency of the least squares methods of the cointegration vector and the threshold parameters are developed. Simulation results confirm the theoretical findings.

Research supported in part by grants from HKSAR-RGC-GRF.

EM-test for finite mixture models

Jiahua Chen

Department of Statistics
University of British Columbia, Vancouver, BC, Canada V6T 1Z4

Abstract

In many scientific investigations, a population can often be divided into more homogeneous sub-populations. A finite mixture model is then a useful model. Inference on the mixing structure has been an important problem in various disciplines. Developing valid and effective statistical inference procedures on the mixing distribution has been technically challenging. Classical procedures often have sophisticated asymptotic properties which render them useless in applications. Recently, we invent a class of EM-tests that are advantageous in many respects. For a large number of finite mixture models, we have successfully designed corresponding EM-tests whose limiting distributions are easier to derive mathematically, simple for implementation in data analysis. The simulation indicates that the limiting distributions have good precision at approximating the finite sample distributions. A general procedure of selecting tuning parameters has also been developed.

Classification using function scraps

Peter Hall

Department of Mathematics & Statistics
The University of Melbourne, Parkville, VIC, 3010, Australia

Abstract

Function scraps are fragments of 'whole' curves that are, at least conceptually, conventional random functions defined over a common domain. In particular, a function scrap is observed only over a subinterval of the common domain, and the subintervals are generally different for different scraps. Functional data of this type are increasingly common. We shall suggest new methods for reconstructing the whole functions, and for classifying function scraps.

Concave group selection in high-dimensional models

Jian Huang

Department of Statistics and Actuarial Science
University of Iowa, Iowa City, IA 52242, USA

Abstract

Grouping structures arise naturally in many high-dimensional data analysis problems. In this talk, we present a class of group selection methods that respect such structures in parameter estimation and variable selection. These methods use the composition of a concave function and the Euclidean norm of the coefficients in each group as the penalty for selection and estimation. Under certain sparsity and regularity conditions, they possess an oracle property, meaning that with high probability they yield solutions that are equal to the oracle estimator under the unknown true model. This result holds even when the number of groups exceeds the sample size. We derive a group coordinate descent algorithm for computing the solution paths of group estimators. This algorithm takes advantage of the closed form expressions of the estimators for a single group model and is efficient in high-dimensional settings. We also discuss the applications of group selection in several statistical modeling and data analysis problems.

NMR peak picking through wavelet transform and volume filtering

Bing-Yi Jing

Department of Mathematics
Hong Kong University of Science & Technology

Abstract

Nuclear magnetic resonance (NMR) has been widely used as a powerful tool to determine the three-dimensional structures of proteins in vivo. However, the post-spectra stage of NMR structure determination usually involves a tremendous amount of time and expert knowledge, which includes peak picking, chemical shift assignment, and structure calculation steps. We propose WaVPeak, a fully automatic peak detection method. WaVPeak applies wavelet smoothing and then use volumes around its smoothed peaks to filter out the false peaks. WaVPeak can detect weak peaks which are the ones the NMR spectroscopists need the most help to deal with. Experimental results demonstrate that WaVPeak achieves better results than some alternatives. We will also discuss how to employ Benjamini-Hochberg procedure to determine how many peaks to select automatically.

The work is jointly conducted with Prof. Xin GAO and others.

Threshold regression for analysis of time-to-event data: with connection to proportional hazard model and applications

Mei-Ling Ting Lee

Department of Epidemiology and Biostatistics
University of Maryland at College Park

Abstract

The proportional hazards (PH) assumption required by the Cox model is not appropriate in some applications. Moreover, PH regression focuses mainly on hazard ratios and thus does not offer many insights into underlying determinants of survival. Threshold regression (TR) is an alternative methodology that is not built on consideration of hazards.

Threshold regression methodology is based on the concept that the degradation of a machine or a patient's health status follows a stochastic process. For engineering applications, the degradation can often be observed. For medical research, a patient's health status is a complex unobservable process. The onset of disease, or death, occurs when the process first reaches a failure threshold (i.e., a first hitting time). Instead of calendar time, analytical time (also called operational time) can be included in TR regression. The TR model is intuitive and does not require the proportional hazards assumption. It thus provides an important alternative for analyzing time-to-event data.

In this talk, we discuss the connections between these two regression methodologies. A case demonstration is used to highlight the greater understanding of scientific foundations that TR can offer in comparison to PH regression. Applications will also be demonstrated.

On buffered time series models

Wai Keung Li

Department of Statistics and Actuarial Science
University of Hong Kong

Abstract

We extend the classical threshold models via the regime switching mechanism mimicking a climatological example. It leads to a new model, which is called the buffered threshold model since there is a buffer zone for the model to switch regimes. This paper concentrates on the self-exciting buffered threshold autoregressive (BAR) model, and a sufficient condition is given for the geometric ergodicity of the two-regime BAR process. The conditional least squares estimation is considered for the BAR model, and its asymptotic properties including the strong consistency and the asymptotic distributions are also derived. Monte Carlo experiments give further support to the methodology developed for the new model, and two empirical examples demonstrates the importance of the BAR model.

Variable selection and estimation in generalized linear models with the seamless L_0 penalty

Xihong Lin

Department of Biostatistics
Harvard University

Abstract

We propose variable selection and estimation in generalized linear models using the seamless (SELO) penalized likelihood approach. The SELO penalty is a smooth function that very closely resembles the discontinuous L0 penalty. We develop an efficient algorithm to fit the model, and show that the SELO-GLM procedure has the oracle property in the presence of a diverging number of variables. We propose a Bayesian Information Criterion (BIC) to select the tuning parameter. We show that under some regularity conditions, the proposed SELO-GLM/BIC procedure consistently selects the true model. We perform simulation studies to evaluate the finite sample performance of the proposed methods. Our simulation studies show that the proposed SELO-GLM procedure has a better finite sample performance than several existing methods, especially when the number of variables is large and the signals are weak. We apply the SELO-GLM to analyze a breast cancer genetic dataset to identify the SNPs that are associated with breast cancer risk.

Likelihood ratio tests for the structural change of an AR(p) model to a threshold AR(p) model

Ke Zhu and Shiqing Ling

Department of Mathematics
Hong Kong University of Science and Technology

Abstract

This paper considers the likelihood ratio (LR) test for the structural change of an AR model to a threshold AR model. Under the null hypothesis, it is shown that the LR test converges weekly to the maxima of a two-parameter vector Gaussian process. Using the approach in Chan and Tong (1990) and Chan (1991), we obtain a parameter-free limiting distribution. This distribution is novel and its percentage points are tabulated via a Monte Carlo method. Simulation studies are carried out to access the performance of the LR test in the finite sample and a real example is given.

Nonparametric estimation for dependent interval-censored failure time data

Jianguo Sun

Department of Statistics
University of Missouri, Columbia, Missouri 65211, USA

Abstract

Nonparametric estimation of a survival function is one of the most commonly asked questions in the analysis of failure time data and for this, a number of procedures have been developed under various types of censoring structures (Kalbfleisch and Prentice, 2002). In particular, several algorithms are available for interval-censored failure time data with independent censoring mechanism (Sun, 2006; Turnbull, 1976). In this talk, we discuss the interval-censored data where the censoring mechanism may be related to the failure time of interest and some procedures are investigated.

An alternative GARCH-in-mean model

Heung Wong

Department of Applied Mathematics
The Hong Kong Polytechnic University

Abstract

We generalize the semi-parametric GARCH-in-mean model of Christensen, Dahl and Iglesias (2012). The generalized model allows one to take the asymmetric factor into account when describing the conditional volatility. The improved estimation adopts the local polynomial approximation, which is of more flexibility as compared to the original local constant approximation. Under some regularity conditions, it can be shown that the parametric estimator is consistent. Simulations show that the estimation works well and empirical studies justify the model generalization.

Semiparametric transformation models under biased sampling schemes

Zhiliang Ying

Department of Statistics
Columbia University, New York, NY 10027, USA

Abstract

We propose a unified estimation method for semiparametric linear transformation models under general biased sampling schemes. The new estimator is obtained from a set of counting process-based unbiased estimating equations, developed through a general weighting scheme that offsets the sampling bias. It is asymptotically normal with a closed-form formula for the limiting variance whose plug-in estimator is consistent. We demonstrate the unified approach through the special cases of truncation, length-bias, case-cohort design among others. Simulation studies and applications to real data sets are presented.

This is joint work with Jane Paik Kim, Wenbin Lu and Tony Sit.

Robust rank correlation based screening

Lixing Zhu

Department of Mathematics
Hong Kong Baptist University

Abstract

Independence screening is a variable selection method that uses a ranking criterion to select significant variables, particularly for statistical models with nonpolynomial dimensionality or "large p, small n" paradigms when p can be as large as an exponential of the sample size n. In this paper, we propose a robust rank correlation screening (RRCS) method to deal with ultra-high dimensional data. The new procedure is based on the Kendall τ correlation coefficient between response and predictor variables rather than the Pearson correlation of existing methods. The new method has four desirable features compared with existing independence screening methods. First, the sure independence screening property can hold only under the existence of a second order moment of predictor variables, rather than exponential tails or alikeness, even when the number of predictor variables grows as fast as exponentially of the sample size. Second, it can be used to deal with semiparametric models such as transformation regression models and single-index models under monotonic constraint to the link function without involving nonparametric estimation even when there are nonparametric functions in the models. Third, the procedure can be largely used against outliers and influence points in the observations. Last, the use of indicator functions in rank correlation screening greatly simplifies the theoretical derivation due to the boundedness of the resulting statistics, compared with previous studies on variable screening. Simulations are carried out for comparisons with existing methods and a real data example is analyzed.