# The 2012 Taipei International Statistics Workshop

1 - 3 May 2012

## Venue:

Room 440 Astrophysics and Mathematics Building, National Taiwan University

## Organising Committee:

Shu-Hui Chang (National Taiwan University)

Ming-Yen Cheng (National Taiwan University)

Chih-Kang Chu (National Dong Hwa University)

Byeong Park (Seoul National University)

Masanobu Taniguchi (Waseda University)

## Sponsors:

Mathematics Division, National Center for Theoretical Sciences (Taipei Office)

Mathematics Research Promotion Center

National Taiwan University

## Program Overview

| | Tuesday | Wednesday | Thursday |
|---|---|---|---|
| 9:30-10:00 | | Hwai-Chung Ho | Laurent Cavalier |
| 10:00-10:30 | | Masanobu Taniguchi | Henghsiu Tsai |
| 10:30-11:00 | *Registration* | *Break* | *Break* |
| 11:00-11:30 | Susan R. Wilson | Jaeyong Lee | Byeong Park |
| 11:30-12:00 | Ci-Ren Jiang | Guan-Hua Huang | Hiroaki Ogata |
| 12:00-12:30 | Chunming Zhang | Hwan Chung | Li-Shan Huang |
| 12:30-14:00 | *Lunch* | *Lunch* | *Lunch* |
| 14:00-14:30 | Hee-Seok Oh | Taeryon Choi | |
| 14:30-15:00 | Junichi Hirukawa | Feng Yao | |
| 15:00-15:30 | Cathy WS Chen | Ching-Kang Ing | |
| 15:30-16:00 | *Break* | *Break* | |
| 16:00-16:30 | Zudi Lu | Berwin Turlach | |
| 16.30-17:00 | Tomoyuki Amano | Jinfang Wang | |
| 17:00-17:30 | Alexandre Petkovic | Zhengjun Zhang | |

# Program

## Day 1, Tuesday 1 May

**Registration: 10:30-11:00**

**Morning session: 11:00-12:30**

11:00-12:30 Chair: Guan-Hua Huang

- 11:00 **Susan R. Wilson**
  Application of Mutual Information to Simultaneously Integrate Very Large Data Sets Containing Different Types of Variables
- 11:30 **Ci-Ren Jiang**
  Functional single index models for longitudinal data
- 12:00 **Chunming Zhang**
  Penalized Bregman divergence for large-dimensional regression and classification

**Lunch: 12:30-14:00**

**Afternoon session: 14:00-17:30**

14:00-15:30 Chair: Li-Shan Huang

- 14:00 **Hee-Seok Oh**
  Composite Quantile Periodogram
- 14:30 **Junichi Hirukawa**
  Ruin probabilities in time series premium model
- 15:00 **Cathy WS Chen**
  A Bayeisan Perspective on Backtesting Value-at-Risk Models

15:30-16:00 **Break**

16:00-17:30 Chair: Ming-Yen Cheng

- 16:00 **Zudi Lu**
  Estimating Nonlinear Spatial Quantile Regression: Theory and Application
- 16:30 **Tomoyuki Amano**
  Estimating function estimator for financial time series models

- 17:00 **Alexandre Petkovic**
  Robust Portfolio Estimation under Skew-Normal Return Processes

## Day 2, Wednesday 2 May

### Morning session: 9:30-12:30

9:30-10:30 Chair: Ci-Ren Jiang

- 9:30 **Hwai-Chung Ho**
  A resampling method for long-range dependent processes
- 10:00 **Masanobu Taniguchi**
  Jackknifed Whittle Estimators

10:30-11:00 **Break**

11:00-12:30 Chair: Shu-Hui Chang

- 11:00 **Jaeyong Lee**
  Bayesian regression based on principal components for high-dimensional data
- 11:30 **Guan-Hua Huang**
  Bayesian Inferences of Latent Class Models with an Unknown Number of Classes
- 12:00 **Hwan Chung**
  Analysis for latent stage-sequential process in early-onset drinking behaviors

### Lunch: 12:30-14:00

### Afternoon session: 14:00-17:30

14:00-15:30 Chair: Yao-Hsiang Yang

- 14:00 **Taeryon Choi**
  Gaussian Process Partially Linear Regression Models
- 14:30 **Feng Yao**
  Stock Market Causal Relations between the U.S. and East Asian Countries
- 15:00 **Ching-Kang Ing**
  Predictor selection for non-negative autoregressive processes

15:30-16:00 **Break**

16:00-17:30 Chair: Henghsiu Tsai

- 16:00 **Berwin Turlach**
  On sampling from arbitrary copulae
- 16:30 **Jinfang Wang**
  Computation of Probabilistic Conditional Independence
- 17:00 **Zhengjun Zhang**
  Generalized Measures of Correlation for Asymmetry, Nonlinearity in Economic Study and Beyond

## Day 3, Thursday 3 May

### Morning session: 9:30-12:30

9:30-10:30 Chair: Jyh-Jen Horng Shiau

- 9:30 **Laurent Cavalier**
  Oracle inequalities for inverse problems
- 10:00 **Henghsiu Tsai**
  Inference of Seasonal Long-memory Time Series with Measurement Error

10:30-11:00 **Break**

11:00-12:30 Chair: Lu-Hung Chen

- 11:00 **Byeong Park**
  Some Generalizations of Varying Coeffcient Regression Models
- 11:30 **Hiroaki Ogata**
  Estimation with generalized empirical likelihood for multivariate stable distributions
- 12:00 **Li-Shan Huang**
  Local polynomial and penalized trigonometric series regression

### Lunch: 12:30-14:00

# Abstracts

### Estimating function estimator for financial time series models

Tomoyuki Amano
Faculty of Economics, Wakayama University

There have been proposed many financial time series models in order to represent behaviours of data in the finance and many reseachers have investigated these models.One of the most fundamental estimators for financial time series models is the conditional least squares estimator (CL estimator). CL estimator has two advantages; it can be calculated easily and it does not need the knowledge of the innovation. Hence CL estimator has been widely used. However Amano and Taniguchi (2008) showed CL estimator is not asymptotically optimal in general for ARCH model. On the other hand Chandra and Taniguchi (2001) constructed the optimal estimating function estimator (G estimator) for ARCH model based on Godambes optimal estimating function and showed G estimator is better than CL estimator in the sense of the sample mean squared error by simulation. In this talk we apply CL and G estimators to famous financial time series models (ARCH, GARCH, CHARN) and show G estimator is better than CL estimator in the sense of the efficiency theoretically. Furthermore we derive the condition of the asymptotical optimality of G estimator based on LAN.

### Oracle inequalities for inverse problems

Laurent Cavalier
Department of Mathematics, Universite Aix-Marseille

There exist many fields where inverse problems appear. Some examples are: astronomy (blurred images of the Hubble satellite), econometrics (instrumental variables), or medical image processing (X-ray tomography). These are problems where we have indirect observations of an object (a function) that we want to reconstruct, through a linear operator $A$. Due to its indirect nature, solving an inverse problem is usually rather difficult. We consider a sequence space model of statistical linear inverse problems where we need to estimate a function $f$ from indirect noisy observations. Let a finite set $\Lambda$ of linear estimators be given. Our aim is to mimic the estimator in $\Lambda$ that has the smallest risk on the true $f$, i.e. the oracle. This problem corresponds to a data-driven selection of the regularization parameter. Under general conditions, we show that this can be achieved by simple minimization of unbiased risk estimator (URE). The main result is a nonasymptotic oracle inequality that is shown to be asymptotically exact. This inequality can be also used to obtain sharp minimax adaptive results.

### A Bayeisan Perspective on Backtesting Value-at-Risk Models

Cathy W.S. Chen[(1),*], Edward M.H. Lin[(1)], and Richard Gerlach[(2)]
[(1)]Department of Statistics, Feng Chia University, Taiwan
[(2)]Discipline of Business Analytics, University of Sydney, Australia

Many papers have proposed forecasting VaR (Value-at-Risk) measures in a Bayesian framework. However, when the authors deal with back-testing, most jump outside the Bayeisan framework, or use informal criteria, to compare VaR models. An important issue that arises in this context is how to evaluate the performance of VaR models/methods. It is desirable to have formal testing procedures for comparison, which do not necessarily require knowledge of the underlying model, or if the model is known, do not restrict attention to a specific estimation procedure. The motivation of the study is to propose back-testing based on the idea of Gaglianone et al. (2011 JBES) which evaluates VaR models via quantile regression, which does not rely solely on binary variables like violations. Using bivariate quadrature methods and the standard asymmetric Laplace quantile likelihood, we analytically estimate the Bayes factor in favour of the proposed models forecasts being accurate and independent. For the empirical study, we compare the performance of the proposed method, and another three non-Bayesian back-testing methods, to figure out which back-tests are the most reliable, and most suitable for model validation processes. The proposed Bayesian tests provide sound methods to assess the finite sample performance of a quantile model.

Keywords: Bayesian Hypothesis testing, MCMC, Value-at-Risk, quantile estimation.

## Gaussian Process Partially Linear Regression Models

Taeryon Choi
Department of Statistics, Korea University

A partially linear regression model is a semiparametric regression model that consists of parametric and nonparametric regression components in an additive form. In this paper, we consider statistical inference for partially linear regression models using Gaussian process priors. Specifically, a Gaussian process prior with a zero mean function and a suitably chosen covariance function is put on the nonparametric regression component in the partially linear regression model. The posterior inference is performed for estimation as well as hypotheses testing, i.e. selecting appropriate regression models based on Bayes factors. For estimation, posterior distributions are derived for unknown parameters, and numerical schemes are discussed to generate posterior samples from them. For hypotheses testing, we deal with model comparisons between parametric representation and semi/nonparametric ones. We illustrate empirical performance of the proposed model based on synthetic data and real data applications, and investigate the asymptotic issue on consistency.

## Analysis for latent stage-sequential process in early-onset drinking behaviors

Hwan Chung
Department of Statistics, Korea University

In longitudinal research on early-onset drinkers, a great deal of attention has been paid to the identification of subgroups of individuals who follow similar sequential patterns of drinking behaviors. However, research on the sequential development of drinking behavior can be challenging in part because it may not be possible to directly observe the particular drinking behavior stage at a given point in time. To address this difficulty, one can use a latent class analysis (LCA), which provides a set of principles for the systematic identification of homogeneous subgroups of individuals. In this work, we apply an LCA in an investigation of stage-sequential patterns of drinking behaviors among early-onset drinkers, using data from the 'National Longitudinal Survey of Youth 1997.' An LCA approach is used to sort different patterns of drinking behaviors into a small number of classes at each measurement occasion; and the class sequencing of early-onset drinkers over the entire set of time points is evaluated in order to identify two or more homogeneous early-onset drinkers who exhibit a similar sequence of class memberships over time. This approach uncovers four common drinking behaviors in early-onset drinkers over three measurements from early to late adolescence. The sequences of drinking behaviors can be grouped into three sequential patterns representing the most probable progression of early-onset drinking behaviors.

# Ruin probabilities in time series premium model

Takeyuki Suzuki

Junichi Hirukawa*
Department of Mathematics, Niigata University

We call subjects (ex. Insurance company) the portfolio. We are interested in the surplus process (risk process) of portfolio, which has a ruin possibility. The surplus process is the surplus if portfolio's all contracts have been finished at that time. Thus, if the surplus becomes negative, the portfolio becomes ruined. The claim amount process is a stochastic process, which is the total amount of claim. In this talk we consider the case that the premium process is also stochastic process, which is the total amount of premium. Namely, we deal with the case that both claim amount and premium are locally stationary processes.

### A resampling method for long-range dependent processes

Hwai-Chung Ho
Institute of Statistical Science, Academia Sinica

Studies on the resampling method on data of short-range dependence have been extensive. The corresponding work for the stationary long-range dependent (LRD) processes, however, remains in large part incomplete. The exiting gap in the theoretical development has become the major obstacle for the advancement of the statistical inference for LRD processes in time domain. Take the simple statistic of sample mean for example, the asymptotic distribution derived in the LRD setting is often time non-normal and can only be represented in the form of a stochastic integral whose percentiles are infeasible to be tabulated even numerically. A consistent ressampling method thus becomes particularly appealing when dealing with data exhibiting the long-range dependence. This talk considers a resampling method for the LRD processes which extends earlier results focusing on linear processes or functionals of Gaussian processes.

### Local polynomial and penalized trigonometric series regression

Li-Shan Huang
Institute of Statistics, National Tsing Hua University

We investigate the connections between local polynomial regression, mixed models, and penalized trigonometric series regression. Expressing local polynomial regression in a projection framework, we derive equivalent kernels for both the interior and boundary points. For interior points, it is shown that the asymptotic bias decreases as the order of polynomial increases. Then we show that, under some conditions, the local polynomial projection approach admits an equivalent mixed model formulation where the fixed effects part includes the polynomial functions. The random effects part in the representation is shown to be the trigonometric series asymptotically. The connections are extended to partial linear models and additive models. These results suggest a new smoothing approach using a combination of unpenalized polynomials and penalized trigonometric functions. We illustrate the potential usefulness of the new approach with real data analysis. This is a joint work with Professor Kung-Sik Chan at the University of Iowa.

## Bayesian Inferences of Latent Class Models with an Unknown Number of Classes

Guan-Hua Huang
Institute of Statistics, National Chiao Tung University

This paper focuses on analyzing data collected in situations where investigators use multiple discrete indicators as surrogates, for example, a set of questionnaires. A very flexible latent class model is used for analysis. We propose a Bayesian framework to perform the joint estimation of the number of latent classes and model parameters. The proposed approach applies the reversible jump Markov chain Monte Carlo to analyze finite mixtures of multivariate multinomial distributions. In the paper, we also develop a procedure for the unique labelling of the classes. We have carried out a detailed sensitivity analysis for various hyperparameter specifications, which leads us to make standard default recommendations for the choice of priors. Usefulness of the proposed method is demonstrated through computer simulations and a study on subtypes of schizophrenia using the Positive and Negative Syndrome Scale (PANSS).

## Predictor selection for non-negative autoregressive processes

Ching-Kang Ing
Institute of Statistical Science, Academia Sinica

Let observations be generated from a non-negative first-order autoregressive (AR) process. In both the stationary and unit root cases, we derive moment bounds and limiting distributions of an estimator of the AR coefficient that minimizes the ratios of two consecutive observations. These results enable us to provide an asymptotic expression for the mean squared prediction error (MSPE) of the corresponding predictor, named "minimum ratio predictor", of the next future observation. Based on this expression, we compare the performance of the minimum ratio predictor and the least squares predictor from the MSPE point of view. Our comparison reveals that the better predictor between these two predictors is determined not only by whether a unit root exists or not, but also by the behavior of the unknown error distribution near the origin, and hence cannot be identified in practice. To circumvent this difficulty, we calculate the accumulated prediction errors (APE) of these two predictors and choose the predictor with the smaller APE. We show that the selected predictor is asymptotically equivalent to the better predictor, thereby alleviating the above difficulty. Both real and simulated data sets are used to illustrate the proposed method. (This is joint work with Chiao-Yi Yang.)

## Functional single index models for longitudinal data

Ci-Ren Jiang
Institute of Statistical Science, Academia Sinica

A new single-index model that reflects the time-dynamic effects of the single index is proposed for longitudinal and functional response data, possibly measured with errors, for both longitudinal and time-invariant covariates. With appropriate initial estimates of the parametric index, the proposed estimator is shown to be n-consistent and asymptotically normally distributed. We also address the nonparametric estimation of regression functions and provide estimates with optimal convergence rates. One advantage of the new approach is that the same bandwidth is used to estimate both the nonparametric mean function and the parameter in the index. The finite-sample performance for the proposed procedure is studied numerically.

## Bayesian regression based on principal components for high-dimensional data

Jaeyong Lee
Department of Statistics, Seoul National University

The Gaussian sequence model can be obtained from the high-dimensional regression model through principal component analysis. It is shown that the Gaussian sequence model is equivalent to the original high-dimensional regression model in terms of prediction. Under a sparsity condition, we investigate the posterior consistency and convergence rates of the Gaussian sequence model. In particular, we examine two different modeling strategies: Bayesian inference with and without covariate selection. For Bayesian inferences without covariate selection, we obtain the consistency results of the estimators and posteriors with normal priors with constant and decreasing variances, and James-Stein estimator; for Bayesian inference with covariate selection, we obtain convergence rates of Bayesian model averaging (BMA) and median probability model (MPM) estimators, and the posterior with variable selection prior. Based on these results, we conclude that variable selection is essential in high-dimensional Bayesian regression. A simulation study also confirms the conclusion. The methodologies are applied to a climate prediction problem.

### Estimating Nonlinear Spatial Quantile Regression: Theory and Application

Zudi Lu
School of Mathematical Sciences, The University of Adelaide

Spatial data, which are collected at different sites on the surface of the earth, arise in various areas of research, including econometrics, epidemiology, environmental science, image analysis, oceanography and soil science, etc. Numerous applications of spatial models and important developments in the general area of spatial statistics under linear correlation structures have been widely investigated in the literature.

In this talk, I will first review some of the recent developments in exploring nonlinear spatial neighbouring effects that my coauthors and I have done. In particular, I will introduce some progress in estimating nonlinear spatial neighbouring effects from the quantile perspective. In order to reduce the ‘curse of dimensionality that nonparametric spatial analysis often suffers from, we developed a general robust framework for estimation of semiparametric functional (varying)-coefficient quantile regression with spatial data. The local M-estimators of the unknown functional-coefficient functions are proposed by using local linear approximation, and their asymptotic distributions are then established under weak spatial mixing conditions allowing the data processes to be either stationary or non-stationary with spatial trends. Application to the soil data set is demonstrated with interesting findings that go beyond traditional conditional mean regression analysis.

## Estimation with generalized empirical likelihood for multivariate stable distributions

Hiroaki Ogata
School of International Liberal Studies, Waseda University

The generalized empirical likelihood (GEL) method are considered to estimate the parameters of the multivariate stable distribution. The multivariate stable distributions are widely applicable as they can accommodate both skewness and heavy tails. We treat the spectral measure, which summarizes scale and asymmetry, by discretization. In order to estimate all the model parameters simultaneously, we apply the estimating function constructed by equating empirical and theoretical characteristic functions. The efficacy of the proposed GEL method is demonstrated in Monte Carlo studies. An illustrative example involving daily returns of market indexes is also included.

## Composite Quantile Periodogram

Hee-Seok Oh
Department of Statistics, Seoul National University

Quantile periodogram recently developed by Li (2011) provides rich information of the frequency of signal, compared to a single estimate of the mean frequency. It has been derived by adopting a quantile function that replaces the least square loss function in the harmonic regression procedure. However, it is difficult to find a specific quantile that identifies a hidden frequency of time series. In addition, it may not be efficient to consider all quantiles in the interval between 0 and 1. In this study, we propose a data-adaptive composite quantile periodogram using a weighted linear combination of quantile periodogram. The main advantage of the proposed method is that it does not require prior knowledge of the signal. Furthermore, the proposed periodogram is extended to a penalized version. Simulation studies and real data examples demonstrate significant improvements in the quality of the periodogram.

## Some Generalizations of Varying Coeffcient Regression Models

Byeong U. Park
Department of Statistics, Seoul National University

In this talk we consider generalizations of the varying coefficient regression model proposed by Hastie and Tibshirani (1993). In the classical varying coefficient regression model, the covariates are divided into two groups and the model contains only interaction terms between the two groups. In our model we abstain from the division of the covariates into two groups and we allow interaction terms between all covariates. This broadens the field of applications of varying coefficient models. We discuss optimal rates for the estimation of nonparametric components of the model and we show that these rates can be attained by sieve and penalized least squares estimators. Furthermore, we give a detailed asymptotic distribution theory for kernel-type estimators that are given as the solution of a system of nonlinear integral equations. This talk reports in particular on the results in Lee, Mammen and Park (2012).

## Robust Portfolio Estimation under Skew-Normal Return Processes

Alexandre Petkovic
Department of Applied Mathematics, Waseda University

In this paper, we study issues related to the optimal portfolio estimators and the local asymptotic normality (LAN) of the return process under the assumption that the return process has an moving average representation with skew normal innovations. The paper consists of two parts. In the first part we discuss the influence of the skewness parameter of the skew-normal distribution on the optimal portfolio estimators. Based on the asymptotic distribution of the portfolio estimator for a non-Gaussian dependent return process, we evaluate the influence of the skewness parameter on the asymptotic variance. In the second part of the paper, we assume that the MA coefficients and the mean vector of the return process depend on a lower-dimensional set of parameters. Based on this assumption, we discuss the LAN property of the return's distribution when the innovations follow a skew-normal law. The influence of the skweness on the central sequence of LAN is evaluated both theoretically and numerical

### Jackknifed Whittle Estimators

Masanobu Taniguchi
Department of Applied Mathematics, Waseda University

Studies of genetic contributions to risk can be family- based such as the case-parents design , or population-based, such as the case-control design. Both provide powerful inference regarding associations between genetic variants and risks, but both have limitations. The case-control design requires identifying and recruiting appropriate controls to avoid the problem such as population stratification. On the other hand, the availability of parental genotypes can pose a problem for using case-parents design, especially when the disease of interest has a late age of onset. To improve the efficiency of the later design, a popular approach is to reconstruct the missing genotypes from the genotypes of their offspring and correct the biases resulting from reconstruction. In this paper, we show that two or more unrelated family studies for the same genetic marker can also be combined to improve the efficiency of the association tests. Simulation results confirm this method to be reasonable.

## On sampling from arbitrary copulae

Berwin Turlach
School of Mathematics and Statistics, The University of Western Australia

For some families of copulae efficient algorithms exist for sampling from these distributions. In this talk, while concentrating on extreme-value copulae, we will present a general, black box method for sampling from arbitrary copulae. As opposed to most of the existing sampling methods, the proposed methodology does not require that the corresponding density function is explicitly known for (exact) sampling from an arbitrary copulae. In fact, for (approximate) sampling, the density function does not need to exists and, consequently, the proposed methodology facilitates sampling from non-differentiable distributions. Various issues that affect the proposed method, such as the dimensionality and smoothness of the copula, will be discussed and numerical results will be presented.

### Inference of Seasonal Long-memory Time Series with Measurement Error

Henghsiu Tsai
Institute of Statistical Science, Academia Sinica

We consider the estimation of Seasonal Autoregressive Fractionally Integrated Moving Average (SARFIMA) models in the presence of additional measurement error by maximizing the Whittle likelihood. We show that the spectral maximum Whittle likelihood estimator is asymptotically normal, and study its finite-sample properties through simulation. We illustrate by simulation that ignoring measurement errors may result in incorrect inference. Hence, it is pertinent to test for the presence of measurement error, which we do by developing a likelihood ratio (LR) test within the framework of Whittle likelihood. We derive the non-standard asymptotic null distribution of this LR test. Finite sample properties of the LR test both under the null and the alternative are examined by simulations. The efficacy of the proposed approach is illustrated by a real-life example. (This is a joint work with Heiko Rachinger.)

### Computation of Probabilistic Conditional Independence

Jinfang Wang
Department of Mathematics and Informatics, Chiba University

Probabilistic conditional independence is a fundamental concept in a number of areas in statistical sciences including statistical causal inference. The author has recently attempted a universal algebraic approach for studying the relations concerning probabilistic conditional independence. In this talk, I will outline this approach; in particular I will show how a particular probabilistic conditional independence relation can be ‘calculated from a set of other such relations using this algebraic approach.

## Application of Mutual Information to Simultaneously Integrate Very Large Data Sets Containing Different Types of Variables

Susan Wilson
Australian National University and University of New South Wales

New statistical methods are needed to analyze the very large-scale quantities of different types of data that increasingly are being produced by recent technological developments. For example, in a complex disorder study such data may include discrete measures, such as (ordered) categorical data from single nucleotide polymorphism (SNP) chips, continuous measures such as produced by gene expression arrays, as well as clinical measures. We have developed a novel exploratory approach that is suitable for such data based on mutual information. This allows us to create a single information matrix including all types of data, namely continuous versus continuous, discrete versus discrete and continuous versus discrete, to be used for clustering and network inference. Examples will be given. This is joint research with PhD student Chris Pardy.

## Stock Market Causal Relations between the U.S. and East Asian Countries

Feng Yao
Faculty of Economics, Kagawa University

In this paper the central and classical questions in economics, the relationships of cause and effect between non-stationary multiple economic time series are discussed. We apply the Wald test of one-way effect causal measure presented by Yao & Hosoya (2000) and Yao (2007) to the analysis of causal relationships of the spotlighted stock market composite indices of the United States and East Asian countries in time domain and frequency domain. Based on error correction model for daily observations in the mentioned stock markets before financial crisis 2008, we showed the stock market causal characterizations between the United States and Japan, China, Korea, Hong Kong as well as Taiwan. Furthermore, the long-run and short-run causal relations are also discussed in view of the Wald test of local one-way effect.

## Penalized Bregman divergence for large-dimensional regression and classification

Chunming Zhang
Department of Statistics, University of Wisconsin

Regularization methods are characterized by loss functions measuring data fits and penalty terms constraining model parameters. The commonly used quadratic loss is not suitable for classification with binary responses, whereas the loglikelihood function is not readily applicable to models where the exact distribution of observations is unknown or not fully specified. We introduce the penalized Bregman divergence by replacing the negative loglikelihood in the conventional penalized likelihood with Bregman divergence, which encompasses many commonly used loss functions in the regression analysis, classification procedures and machine learning literature. We investigate new statistical properties of the resulting class of estimators with the number $p_n$ of parameters either diverging with the sample size $n$ or even nearly comparable with $n$, and develop statistical inference tools. It is shown that the resulting penalized estimator, combined with appropriate penalties, achieves the same oracle property as the penalized likelihood estimator, but asymptotically does not rely on the complete specification of the underlying distribution. Furthermore, the choice of loss function in the penalized classifiers has an asymptotically relatively negligible impact on classification performance. We illustrate the proposed method for quasilikelihood regression and binary classification with simulation evaluation and real-data application.

## Generalized Measures of Correlation for Asymmetry, Nonlinearity in Economic Study and Beyond

Zhengjun Zhang
Department of Statistics, University of Wisconsin

Applicability of Pearson's correlation as a measure of explained variance is by now well understood. One of its limitations is that it does not account for asymmetry in explained variance. Aiming to obtain broad applicable correlation measures, we use a pair of r-squares of generalized regression to deal with asymmetries in explained variances, and linear or nonlinear relations between random variables. We call the pair of r-squares of generalized regression generalized measures of correlation (GMC). We present examples under which the paired measures are identical, and they become a symmetric correlation measure which is the same as the squared Pearson's correlation coefficient. As a result, Pearson's correlation is a special case of GMC. Theoretical properties of GMC show that GMC can be applicable in numerous applications and can lead to more meaningful conclusions and decision making. In statistical inferences, the joint asymptotics of the kernel based estimators for GMC are derived and are used to test whether or not two random variables are symmetric in explaining variances. The testing results give important guidance in practical model selection problems. The efficiency of the test statistics is illustrated in simulation examples. In real data analysis, we present an important application of GMC in explained variances and market movements among three important economic and financial monetary indicators.