Poster Session

Workshop on Analysis of High-Dimensional and Functional Data in Honor of Peter Hall on the occasion of his 60th birthday

1. **Hung Chen** *(Department of Mathematics, National Taiwan University, Taipei, Taiwan)*

   Title : On validity of adaptive penalty selection with variable selection over the nested linear regression.

   Abstract : Model selection procedures often use a fixed penalty, such as Mallows' C_p, to avoid choosing a model which fits a particular data set extremely well. These procedures are often devised to give an unbiased risk estimate when a particular chosen model is used to predict future responses. As a correction for not including the variability induced in model selection, generalized degrees of freedom 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. In this article, we evaluate the validity of such an approach on model selection of linear regression when the set of candidate models satisfies nested structure and includes true model. It is found that the performance of such an approach is even worse than Mallows' C_p on the probability of correct selection. However, this approach coupled with proper selection of the range of penalty or tiny bootstrap proposed in Breiman (1995) can increasing the probability of correct selection but does not achieve model selection consistency.

2. **Kehui Chen** *(Department of Statistics, University of California, Davis)*

   Title : Conditional distribution modeling for functional covariates.

   Abstract : This study is motivated by the conditional growth charts problem, where conditional quantiles of a scalar response are of interest when predictors take values in a functional space. The proposed method aims at estimating conditional distribution functions under a generalized functional regression framework. The uniform convergence rate of the estimated coefficient function is established. The good performance of the proposed method in comparison with other methods, both for sparsely and densely observed functional covariates, is demonstrated through simulations and an application to growth curves, where the proposed method can, for example, be used to assess the entire growth pattern of a child by relating it to the predicted quantiles of adult height.

3. **Senke Chen** *(Department of Statistics, University of California, Davis)*

   Title : Semi-parametric tests for equality of means of single-location and multiple-location processes.

   Abstract : We develop a framework for testing for equality of the means of multiple random processes, with an application to ozone data analysis. We consider two main settings: (I) testing for equality of multiple processes, and, (II) testing for equality across sets of random processes with spatial dependence. In the single location case, L^2 loss is used to measure the divergence between multiple mean functions in a semi-parametric regression framework. We develop a bootstrap test procedure that can be applied to temporally dependent data with ARMA-type serially correlation. Consistent estimators

of the parameters can be achieved under both the null and a range of alternative hypothesis under mild regularity conditions. In order to test "local" features in the time series, we also propose a wavelet-domain version of the test that allows for testing equality at specific time-scales of interest.

In the multiple location setting, a space-time model is proposed to capture both the temporal and spatial dependence. The estimators are obtained under the dynamic linear model framework based on the likelihood function. A similar bootstrap test procedure with L^2 loss test statistic is carried out in this case. We conclude with some simulation studies and an application to testing for model-data agreement with multiple sources of ozone data.

4. **Zongyin John Daye** (Department of Biostatistics and Epidemiology, University of Pennsylvania)

   Title : High-dimensional heteroscedastic regression with an application to eQTL data analysis.

   Abstract : Heteroscedasticity is a common consideration in the development of statistical methodology for applications in practice. In this presentation, we will introduce the high-dimensional heteroscedastic regression (HHR) that allows for nonconstant error variances in high-dimensional estimation and model selection. Our method provides a unified framework for the incorporation of heteroscedasticity arising from predictors explanatory of variability, outliers, and data from varying sources. We demonstrate the presence of heteroscedasticity in and apply our method to an expression quantitative trait loci (eQTLs) study. Our results identify eQTLs that are associated with gene expression variations and demonstrate improved prediction errors and model selection.

5. **Oleksandr Gromenko** *(Department of Mathematics, Utah State University)*

   Title : Nonparametric estimation in small data sets of spatially indexed curves with application to temporal trend determination.

   Abstract : Spatially indexed functional data are space-time data which have very rich temporal component and fairly limited spatial one. Such data can be modeled as a sum of the mean function and a functional error term. We propose fully nonparametric estimators for both the mean function and the space-time covariance structure. Nonparametric modeling of the space-time covariances is surprisingly simple, much faster than those previously proposed and less sensitive to computational errors. We develop a technique for the estimation a global trend in spatially indexed functional data and assessing its significance. We apply our methodology to global ionosonde records collected over forty years to test the hypothesis of ionospheric cooling.

6. **Kyunghee Han** *(Department of Statistics, Seoul National University)*

   Title : Regularization methods for functional linear regression based on principal component analysis.

   Abstract : In this paper, we consider regularization techniques for functional linear regression model. Based on functional principal component analysis (FPCA), we impose sparseness inducing penalty structures on regression coefficient function to explore how the techniques improve prediction performance of the model. The motivating idea is

to construct a statistical strategy of the choice of PC components for regressors, and consequently, to overcome ill-posed problem such as multicollinearity between functional covariates under deep expansion of FPCA.

7. **Xinge Jessie Jeng** *(Department of Biostatistics and Epidemiology, University of Pennsylvania)*

   Title : Simultaneous discovery of rare and common segment variants.

   Abstract : The identification of recurrent variants based on a large set of samples is one of the central issues in population-scale genomic data analysis. A bottleneck of identifying recurrent variants is the lack of an adaptive information pooling procedure which can automatically adjust to the unknown carrier's proportions and optimally identify both rare and common recurrent variants. We specifically consider the identification of recurrent DNA copy number variants (CNVs) in Neuroblastoma patients. It is likely that both rare and common CNVs cooperate to increase the risk of the disease. We developed the Proportion Adaptive Segment Selection (PASS) procedure, which can optimally and simultaneously identify both rare and common CNVs. The proposed method is statistically rigorous and computationally efficient. Theory, simulation, and real applications with validation are presented to demonstrate the performance of the method.

8. **Myung Hee Lee** *(Department of Statistics, Colorado State University)*

   Title : On the border of extreme and mild spiked models in the HDLSS framework.

   Abstract : The asymptotic behavior of Principal Component (PC) is examined in HDLSS context. In the spiked covariance model for HDLSS asymptotics where the dimension goes to infinity while the sample size is fixed, a few largest eigenvalues are assumed to grow as the dimension increases. The rate of the growth is crucial as the asymptotic behavior of the sample PC directions changes dramatically, from consistency to strong inconsistency, at the boundary of the extreme and the mild spiked covariance models. We study the HDLSS asymptotic behavior of PCs at the boundary spiked model and observe that they show intermediate behavior between the extreme and the mild spiked models.

9. **Chong Liu** *(Department of Mathematics, Boston University)*

   Title : Functional factor analysis for periodic remote sensing data.

   Abstract : We present a new approach to factor rotation for functional data. This is achieved by rotating the functional principal components toward a predefined space of periodic functions designed to decompose the total variation into components that are nearly-periodic and nearly-aperiodic with a predefined period. We show that the factor rotation can be obtained by calculation of canonical correlations between appropriate spaces which make the methodology computationally efficient. Moreover, we demonstrate that our proposed rotations provide stable and interpretable results in the presence of highly complex covariance. This work is motivated by the goal of finding interpretable sources of variability in gridded time series of vegetation index measurements obtained from remote sensing, and we demonstrate our methodology through an application of factor rotation of this data.

10. **Ahyeon Park** *(Department of Statistical Science, University College London)*

    Title : Functional data analysis of stratospheric ozone profiles over time.

Abstract : The discovery of ozone depletion is a concern due to its detrimental effects on human beings. Therefore researchers have paid substantial attention to the investigation of ozone trends. Along this concern, we study the effects of covariates of yearly trends and season, as well as atmospheric transport on ozone variations. We first construct the time series of ozone profiles (January 1978 to December 2009) through an appropriate basis system. To capture the primary modes of variation we perform a functional principal component analysis including a penalty term to dampen the excessive variation. The scores corresponding to the eigenfunctions are used for further analysis. Generalized additive models with mixed effects are employed to study the effects of the covariates on the profiles. The novel inclusion of a complex variance structure in the models enhances the fit.

11. **Matthew Reimherr** *(Department of Statistics, University of Chicago)*

    Title : Predictability of shapes of intraday price curves.

    Abstract : We develop a statistical framework, based on functional data analysis, for testing the hypothesis of the predictability of shapes of intraday price curves. We derive a test statistics based on signs of the scores of the functional principal components. We establish its asymptotic properties under the null and alternative hypotheses, and demonstrate via simulations that it has excellent finite sample properties. A small empirical study shows that the shapes of the intraday price curves of large US corporations are not predictable.

12. **Rhonda van Dyke** *(Division of Biostatistics and Epidemiology, University of Kentucky)*

    Title : Mixtures of shape-invariant models for biomarker classification.

    Abstract : A shape invariant model for functions $f_1, \ldots, f_n$ specifies that each individual function f_i can be related to a common shape function g through the relation $f_i(t) = a_i g(c_i t + d_i) + b_i$. We consider a mixture model that allows multiple shape functions $g_1, \ldots, g_K$, where each f_i is a shape invariant transformation of one of those g_k. We present an MCMC algorithm for fitting the model using Bayesian Adaptive Regression Splines (BARS) and discuss some of the computational difficulties that arise in application. The method is illustrated using intestinal current measurement data collected from a cystic fibrosis biomarker research trial, where the groups of functions may indicate different levels of conductance regulation in subgroups of patients.

13. **Wen Wen Tao** *(Department of Statistics, University of California, Davis)*

    Title : Inferring stochastic dynamics from functional data.

    Abstract : In most current data modeling for time-dynamic systems, one works with a pre-specified differential equation and attempts to estimate its parameters. This often leads to models that either do not fit the data well as the presumed underlying mechanisms that the model reflects are not well understood, or do not provide good approximations to the actual mechanisms. In addition, deterministic models rarely provide satisfactory fits to phenomena that are inherently stochastic in nature, since the dynamics vary across subjects or experiments.

    In the case of functional data, we demonstrate that we can directly obtain information about underlying dynamic systems, by deriving the differential equation from observing

many realizations of the trajectories that they generate. Assuming only that the dynamics are described by a first order nonlinear differential equation with a random component, we obtain data-adaptive dynamic equations from the observed data via a simple smoothing-based procedure.

We prove consistency and introduce diagnostics to ascertain the fraction of variance that is explained by the deterministic part of the equation. This approach is shown to yield useful insights into the time-dynamic nature of human growth.

14. **Xiaoke Zhang** *(Department of Statistics, University of California, Davis)*

    Title : Time-varying additive models for longitudinal data.

    Abstract : Additive model is an effective dimension reduction approach that provides flexibility to model the relation between a response variable and key covariates. The literature is largely developed to scalar response and vector covariates. In this paper, more complex data is of interest, where both the response and covariates may be functions. A functional additive model is proposed together with a new smooth backfitting algorithm to estimate the unknown regression functions, whose components are time-dependent additive functions of the covariates. Due to the sampling plan, such functional data may not be completely observed as measurements may only be collected intermittently at discrete time points. We develop a unified platform and efficient approach that can cover both dense and sparse functional data and the needed theory for statistical inference. The oracle properties of the component functions are also established.

15. **Xi Zhang** *(Department of Statistics, Utah State University, Logan)*

    Title : Functional prediction of intraday cumulative returns.

    Abstract : We define cumulative intraday returns and consider their prediction from such returns on a market index. We model these returns as curves in a function space. We propose several functional regression models which can be viewed as extensions of the Capital Asset Pricing Model to intraday returns defined as curves. After deriving parameter estimates and prediction functions for these models, we compare their prediction errors by application to cumulative intraday returns of large US corporations. We find that complex functional regression models do not perform better than a simple model. In particular, we find that modeling error dependence does not improve forecasts.