Nonparametric Statistics Workshop 2016
Integration of Theory, Methods and Applications

The University of Michigan, Ann Arbor, June 6-7, 2016

Abstracts

Block Models with Covariates: Likelihood Methods of Fitting

Bickel, Peter J. (bickel@stat.berkeley.edu)
University of California, Berkeley

Type: Plenary Talk

Abstract. We introduce block models with edge and block covariates, along the lines of Hoff, Handcock, Raftery (2002), specializing to covariate forms of the types proposed by Zhang, Levina, Zhu (2014) and generalizing that of Newman, Clauset(2015). We study maximum likelihood and mean field variational fitting for these methods along the lines of Celisse, Daudin, Pierre (2011) and B., Choi, Chang, Zhang (2013) and partly extend their results to the regime where the average degree tends to infinity faster than loglog(n). We show by example and simulation when mean field methods work and how they can be adapted to succeed when they fail. Co-authors: Purna Sarkar (U. of Texas, Austin), Soumendu Mukherjee (UC, Berkeley), Sharmodeep Bhattacharyya (Oregon State University) and David Choi (Carnegie Mellon University).

Keywords: Block models; Field variational fitting; Infinite average degree.

A Semiparametric Single-Index Risk Score Across Populations

Carroll, Raymond J (carroll@stat.tamu.edu)
Texas A&M University

Type: Plenary Talk

Abstract. We consider a problem motivated by issues in nutritional epidemiology, across diseases and populations. In this area, it is becoming increasingly common for diseases to be modeled by a single diet score, such as the Healthy Eating Index, the Mediterranean Diet Score, etc. For each disease and for each population, a partially linear single-index model is fit. The partially linear aspect of the problem is allowed to differ in each population and disease. However, and crucially, the single-index itself, having to do with the diet score, is common to all diseases and populations, and the nonparametrically estimated functions of the single-index are the same up to a scale parameter. Using B-splines with an increasing number of knots, we develop a method to solve the problem, and display its asymptotic theory. An application to the NIH-AARP Study of Diet and Health is described, where we show the advantages of using multiple diseases and populations simultaneously rather than one at a time in understanding the effect of increased milk consumption. Simulations illustrate the properties of the methods. Joint work with Shujie Ma, Yanyuan Ma, Yanqing Wanf and Eli Kravitz

Keywords: Nutritional epidemiology; Index-modeling.

A Bayesian Multivariate Functional Dynamic Linear Model

Ruppert, David (dr24@cornell.edu)
Cornell University

Type: Plenary Talk

Abstract. We present a Bayesian approach for modeling multivariate, dependent functional data. To account for the three dominant structural features in the data–functional, time dependent, and multivariate components–we extend hierarchical dynamic linear models for multivariate time series to the functional data setting. We also develop Bayesian spline theory in a more general constrained optimization framework. The proposed methods identify a time-invariant functional basis for the functional observations, which is smooth and interpretable, and can be made common across multivariate observations for additional information sharing. The Bayesian framework permits joint estimation of the model parameters, provides exact inference (up to MCMC error) on specific parameters, and allows generalized dependence structures. Sampling from the posterior distribution is accomplished with an efficient Gibbs sampling algorithm. We illustrate the proposed framework with two applications: (1) multi-economy yield curve data from the recent global recession, and (2) local field potential brain signals in rats, for which we develop a multivariate functional time series approach for multivariate time-frequency analysis. *This is based on a joint work with Daniel Kowal and David Matteson

Keywords: Hierarchical dynamic models; Multivariate dependent functional data.

From Functional to Neuroimaging Data and Beyond

Wang, Jane-Ling (wang@wald.ucdavis.edu)
University of California, Davis

Type: Plenary Talk

Abstract. Functional data are samples of random functions, which can be defined on a one or higher-dimensional domain and are intrinsically considered elements of an infinite dimensional space. Functional data analysis (FDA) analysis is concerned with the analysis of such data, and a traditional assumption has been the availability of an independent sample. This has been relaxed with the rise of next generation functional data with complex dependency structures, for which neuroimaging data provide many examples. In this talk, we discuss the pros and cons of the FDA approach and its applications to neuroimaging data. FDA is geared towards handling noise in the data and is very flexible due to its inherently nonparametric nature. We provide two examples of PET imaging analysis that demonstrate the benefits of employing FDA. A major challenge when applying FDA is the inverse problem that arises when inversion of a covariance operator is involved, as is the case in functional regression or functional canonical correlation. We show how to avoid inverse problems by developing alternative approaches that do not involve operator inverses and illustrate the potential of applying functional correlation in studies of functional connectivity of human brain. A new functional connectivity measure will also be presented and illustrated for fMRI data.

Keywords: Functional connectivity; Functional data analysis.

Matched Bipartite Block Model with Covariates

Amini, Arash (aaamini@ucla.edu)
UCLA, Department of Statistics

Type: Invited Talk

Abstract. Community detection or clustering is a fundamental task in the analysis of network data. Many real networks have a bipartite structure which makes community detection challenging. Motivated by an application from biology, we consider a model which allows for matched communities in the bipartite setting, in addition to node covariates with information about the matching. We derive a simple fast algorithm for fitting the model based on variational inference ideas and show its effectiveness in simulations. This is joint work with Zahra Razaee and Jessica Li.
Keywords: Network analysis; Community detection; Bipartite networks; Matching; Variational inference.

Quadratic Estimation of Random Field Nnon-Stationarity

Anderes, Ethan (anderes@ucdavis.edu)
University of California at Davis

Type: Invited Talk

Abstract. More than a decade ago two physicists, Wayne Hu and Takemi Okamoto, invented a new estimator for measuring the dark matter distortion imprinted on the observations of the cosmic microwave background (which is a relic signal of the big bang). Their estimator, called the quadratic estimator, quickly became the state-of-the-art tool for detecting, measuring and mapping dark matter. From a spatial statistics perspective this estimator has some remarkable properties. In this talk I will present an analysis of the quadratic estimator in the context of both cosmology and in the larger context of estimating general random field non-stationarity. I will discuss a property of random fields we call, local invariant non-stationarity, which appears to be a sufficient condition for extending the remarkable properties of Hu and Okamoto's quadratic estimate to more general forms of non-stationarity.
Keywords: Spatial statistics; Non-stationarity; Quadratic estimate.

Divide and Conquer in Non-Standard Problems and the Super-Efficiency Phenomenon

Banerjee, Moulinath (moulib@umich.edu)
University of Michigan

Type: Invited Talk

Abstract. We study how the divide and conquer principle partition the available data into sub-samples, compute an estimate from each sub-sample and combine these appropriately to form the final estimator works in non-standard problems where rates of convergence are typically slower than $\sqrt{n}$ and limit distributions are non-Gaussian, with a special emphasis on the least squares estimator of a monotone regression function. We find that the pooled estimator, obtained by averaging non-standard estimates across the mutually exclusive sub- samples, outperforms the non-standard estimator based on the entire sample in the sense of point wise inference. We also show that, under appropriate conditions, if the number of subsamples is allowed to in- crease at appropriate rates, the pooled estimator is asymptotically normally distributed with a variance that is empirically estimable from the subsample-level estimates. Further, in the context of mono- tone function estimation we show that this gain in point wise efficiency comes at a price of the pooled estimator's performance, in a uniform sense (maximal risk) over a class of models worsens as the number of subsamples increases, leading to a version of the super-efficiency phenomenon. In the process, we develop analytical results for the order of the bias in isotonic regression, which are of independent interest.

Keywords: divide and conquer, non-standard problems, super-efficiency.

Network Modeling for Brain Functional Connectivity Analysis

Chen, Kehui (khchen@pitt.edu)
University of Pittsburgh

Type: Invited Talk

Abstract. In this talk, we will first introduce a feature adjusted stochastic block model to capture the impact of node features on the network links as well as to detect the residual community structure beyond that explained by the node features. The model is motivated by brain connectivity studies and has been successfully applied to a brain parcellation problem using resting-state fMRI data. In the second part, we will present a principal component analysis approach for the group analysis of brain functional connectivity data.

Keywords: Network community detection; Brain functional connectivity; Resting state fMRI.

New methods for Analyzing Partially Observed Functional Data

Delaigle, Aurore (A.Delaigle@ms.unimelb.edu.au)
University of Melbourne

Type: Invited Talk

Abstract. We consider analysis of functional data which are only partially observed. Often in such cases, the observed fragments of curves are supported on quite different intervals, in which case

standard methods of analysis cannot be used. We propose new approaches to analysing fragments of curves observed on different intervals. The techniques we suggest involve discretising the observed fragments, and then extending them outside the interval where they were observed. Using the same approach we can construct estimators of the mean and covariance functions, and, for example, deal with functional linear regression.
Keywords: incomplete functional data. fragments, Markov chains.

Statistical Inference for Stochastic Block Models

Lei, Jing (jinglei@andrew.cmu.edu)
Carnegie Mellon University

Type: Invited Talk

Abstract. We consider some basic inference problems for stochastic block models and its variants. First, we introduce a Network Cross-Validation method (NCV) for automatically determining the order of the model, and for selecting between the regular block models and degree corrected models. The NCV differs from the standard cross-validation and takes advantage of the special structure of SBM. We also observe that NCV with repetition can lead to significantly improved performance. Second, we apply recent results in random matrix theory to develop a goodness-of-fit test for stochastic block models. This test can be used to detect additional structures, such as mixed membership and degree heterogeneity, as well as to estimate the number of communities.
Keywords: stochastic block model; model selection; goodness-of-fit; cross-validation.

Interpretable Models for Prediction on Networks with Cohesion

Levina, Liza (elevina@umich.edu)
University of Michigan

Type: Invited Talk

Abstract. Prediction problems typically assume the training data are independent samples, but in many modern applications samples come from individuals connected by a network. For example, in adolescent health studies of risk-taking behaviors, information on the subjects' social networks is often available and plays an important role through network cohesion, the empirically observed phenomenon of friends behaving similarly. Taking cohesion into account in prediction models should allow us to improve their performance. Here we propose a regression model with a network-based penalty on individual node effects to encourage similarity between predictions for linked nodes, and show that it performs better than traditional models both theoretically and empirically when network cohesion is present. The framework is easily extended to other models, such as the generalized linear model and Cox's proportional hazard model. Applications to predicting levels of recreational activity and marijuana usage among teenagers based on both demographic covariates and their friendship networks are discussed in detail and demonstrate the effectiveness of our approach.

Joint work with Tianxi Li and Ji Zhu.

Keywords: networks; cohesion; penalized regression; prediction.

Generalized Quasi-Likelihood Ratio Tests for Semiparametric Analysis of Covariance Models in Longitudinal Data

Li, Yehua (yehuali@iastate.edu)
Iowa State University

Type: Invited Talk

Abstract. We model generalized longitudinal data from multiple treatment groups by a class of semiparametric analysis of covariance models, which take into account the parametric effects of time dependent covariates and the nonparametric time effects. In these models, the treatment effects are represented by nonparametric functions of time and we propose a generalized quasi-likelihood ratio test procedure to test if these functions are identical. Our estimation procedure is based on profile estimating equations combined with local linear smoothers. We find that the much celebrated Wilks phenomenon which is well established for independent data still holds for longitudinal data if a working independence correlation structure is assumed in the test statistic. However, this property does not hold in general, especially when the working variance function is mis-specified. Our empirical study also shows that incorporating correlation into the test statistic does not necessarily improve the power of the test. The proposed methods are illustrated with simulation studies and a real application from opioid dependence treatments.

Keywords: Analysis of variance; Functional data; Hypothesis testing; Kernel smoothing; Longitudinal data; Semiparametric.

Estimation and Inference in Generalized Additive Coefficient Models for Nonlinear Interactions with High-Dimensional Covariates

Ma, Shujie (shujie.ma@ucr.edu)
University of California, Riverside

Type: Invited Talk

Abstract. In this talk, I will introduce estimation and inference procedures we have proposed for the generalized additive coefficient models (GACM) when the dimension of the variables is high. The GACM has been demonstrated to be a powerful tool for studying nonlinear interaction effects of variables. Specifically, we propose a groupwise penalization based procedure to distinguish significant covariates for the large p small n setting. The procedure is shown to be consistent for model structure identification. Further, we construct simultaneous confidence bands for the coefficient

functions in the selected model based on a refined two-step spline estimator. We also discuss how to choose the tuning parameters. To estimate the standard deviation of the functional estimator, we adopt the smoothed bootstrap method. We conduct simulation experiments to evaluate the numerical performance of the proposed methods and analyze an obesity data set from a genome-wide association study as an illustration. This is joint work with Raymond Carroll, Hua Liang and Shizhong Xu.
Keywords: . Adaptive group lasso, bootstrap smoothing, gene-environment interaction, inference for high dimensional data, penalized likelihood.

Functional Single Index Model

Ma, Yanyuan (yanyuanma@gmail.com)
Penn State University

Type: Invited Talk

Abstract. To study the relation between a univariate response and multiple functional covariates, we propose a functional single index model that is semiparametric. The parametric part of the model integrates the linear regression modeling for functional data and the sufficient dimension reduction structure. The nonparametric part of the model further allows the response-index dependence or the link function to be unspecified. We use B-splines to approximate the coefficient function in the functional linear regression model part and reduce the problem to a familiar dimension folding model. We develop a new method to handle the subsequent dimension folding model by using kernel regression in combination with semiparametric treatment. The new method does not impose any special requirement on the inner product between the covariate function and the B-spline bases, and allows efficient estimation of both the index vector and the B-spline coefficients. The estimation method is general and applicable to both continuous and discrete response variables. We further derive asymptotic properties of the class of methods for both the index vector and the coefficient function. We establish the semiparametric optimality, which has not been done before in a semiparametric model where both kernel and B-spline estimation are involved.
Keywords: Dimension reduction; Dimension folding; Functional linear model; Kernel; Single index model..

Community Detection in Degree-Corrected Block Models

Ma, Zongming (zongming@wharton.upenn.edu)
University of Pennsylvania

Type: Invited Talk

Abstract. Community detection is a central problem of network data analysis. Given a network, the goal of community detection is to partition the network nodes into a small number of clusters,

which could often help reveal interesting structures. The present paper studies community detection in Degree-Corrected Block Models (DCBMs). We first derive asymptotic minimax risks of the problem for a misclassification proportion loss under appropriate conditions. The minimax risks are shown to depend on degree-correction parameters, community sizes, and average within and between community connectivities in an intuitive and interpretable way. In addition, we propose a polynomial time algorithm to adaptively perform consistent and even asymptotically optimal community detection in DCBMs. This is a joint work with Chao Gao, Anderson Y. Zhang and Harrison H. Zhou at Yale University.
Keywords: Clustering; Minimax rates; Network analysis; Spectral clustering; Stochastic block models.

Density Estimation with MCMC Samples

MacEachern, Steve (snm@stat.osu.edu)
The Ohio State University

Type: Invited Talk

Abstract. Markov chain Monte Carlo samplers are the main tool with which Bayesian models are fit. The posterior and predictive distributions are explored on the basis of these samples, both numerically and graphically. Graphical exploration is most commonly accomplished through classical kernel density estimation. Density estimation in this setting differs from traditional work in several ways, with one of the most important being the dependence in the sequence of data. This talk proposes a practical rule for adjusting traditional KDEs to account for dependence and shows its benefits relative to other commonly used strategies. If time permits, further implications of the work will be presented.

This is joint work with Hang Joon Kim (University of Cincinnati) and Yoonsuh Jung (University of Waikato)
Keywords: KDE; MCMC; Dependence; Sheather-Jones; Cross-validation.

Inference Methods with Constrained Splines

Meyer, Mary (meyer@stat.colostate.edu)
Colorado State University

Type: Invited Talk

Abstract. The partial linear model is considered where shape-constrained splines are used to estimate regression functions. Inference methods for the linear and non-linear terms are proposed.
Keywords: constrained splines; convex; monotone; hypothesis test.

Dynamic Modeling of Longitudinal Snippets

Müller, Hans-Georg (hgmueller@ucdavis.edu)
University of California, Davis

Type: Invited Talk

Abstract. Longitudinal data are often plagued with sparsity of time points where measurements are available. The functional data analysis perspective provides an effective and flexible approach to address this problem for the commonly studied case where measurements are sparse but their times are randomly distributed over an interval. Here we consider a different scenario of data snippets. These are very short stretches of longitudinal measurements, which arise in accelerated longitudinal studies. For each subject the stretch of available data is much shorter than the time frame of interest. An added challenge is introduced if a meaningful time proxy such as time since disease onset is not available. We approach this problem through conditional quantile trajectories for monotonic processes that arise as solutions of a dynamic system and discuss consistent estimates and an application to shrinking brain volumes in Alzheimer's patients. This talk is based on joint work with Matt Dawson, UC Davis.

Keywords: Functional Data Analysis; Longitudinal Studies; Sparse Data; Conditional Quantile Trajectories.

Scalable and Consistent Variable Selection for High Dimensional Logistic Regression

Narisetty, Naveen (naveen@illinois.edu)
University of Illinois at Urbana Champaign

Type: Invited Talk

Abstract. Within the framework of Bayesian computation, we provide a novel variable selection method for logistic regression that adapts to both the sample size n and the number of potential covariates p with desirable features. We propose a Gibbs sampler called "Skinny Gibbs" whose computational complexity grows only linearly in p, but it attains the property of strong model selection consistency even in the cases of p > n. In contrast with the standard Gibbs sampler, Skinny Gibbs is much more scalable to high dimensional problems, both in memory and in computational feasibility. We compare our proposed method with several leading variable selection methods through a simulation study to show that our proposed approach selects the correct model with higher probabilities than existing methods while being computationally appealing.

Keywords: Model Selection, High Dimensional Data, Bayesian Computation.

Restricted Strong Convexity and Weak Submodularity

Negahban, Sahand (sahand.negahban@yale.edu)
Yale University

Type: Invited Talk

Abstract. We connect high-dimensional subset selection and submodular maximization. Our results extend the work of Das and Kempe (2011) from the setting of linear regression to arbitrary objective functions. This connection allows us to obtain strong multiplicative performance bounds on several greedy feature selection methods without statistical modeling assumptions. This is in contrast to prior work that requires data generating models to obtain theoretical guarantees. Our work shows that greedy algorithms perform within a constant factor from the best possible subset-selection solution for a broad class of general objective functions. Our methods allow a direct control over the number of obtained features as opposed to regularization parameters that only implicitly control sparsity.

Joint work with: Ethan R Elenberg elenberg@utexas.edu, Rajiv Khanna khanna.rajiv84@gmail.com, Alexandros Dimakis dimakis@austin.utexas.edu.

Keywords: high-dimensional statistics; submodular optimization; sparsity.

Shape-Restricted Survey Estimators

Opsomer, Jean (jopsomer@mac.com)
Colorado State University

Type: Invited Talk

Abstract. Many variables in surveys follow natural orderings that should be respected in the final estimates. For instance, the U.S. National Compensation Survey estimates mean wages for many job categories, and these mean wages are expected to be non-decreasing according to job level. In this type of situation, isotonic regression can be applied to give constrained estimators satisfying the monotonicity. We combine domain estimation and the pooled adjacent violators algorithm to construct new design-weighted constrained estimators. Under mild conditions on the sampling design and the population, we obtain the asymptotic properties of the estimator. Simulation results also demonstrate improved point estimators and confidence intervals for domain means using linearization-based and replication-based variance estimation compared to survey estimators that do not incorporate the constraints.

Joint work with Mary Meyer and Jiwen Wu.

Keywords: Constrained estimation; Design-based estimation; Isotonic regression.

Cluster Analysis of Longitudinal Profiles with Subgroups

Qu, Annie (anniequ@illinois.edu)
University of Illinois at Urbana-Champaign

Type: Invited Talk

Abstract. In this paper, we cluster profiles of longitudinal data using a penalized regression method. The novelty of our approach is that we allow longitudinal patterns from each subject to be unique. Specifically, we identify clusters by applying a pairwise-grouping penalization to the corresponding parameters associated with the nonparametric B-spline models, and therefore distinguish clusters based on different patterns of the predicted longitudinal curves. One advantage of the proposed method is that we do not need to pre-specify the number of clusters; instead it is selected automatically through a model selection criterion. Our method is also applicable for unbalanced data where different subjects could have different time points of measurements. To implement the proposed method, we develop an alternating direction method of multipliers (ADMM) algorithm which has the convergence property. In theory, we establish the consistency properties. In addition, we show that our method outperforms the existing competitive approaches in our numerical studies.
Keywords: ADMM, longitudinal data, minimax concave penalty, model selection, nonparametric spline method.

Bayesian Nonparametric Methods for Structured Sequential Data

Sarkar, Abhra (abhra.stat@gmail.com)
Duke University

Type: Invited Talk

Abstract. We are developing a broad array of novel statistical frameworks for analyzing complex sequential data sets. Our research is primarily motivated by a collaboration with neuroscientists trying to understand the neurological, genetic and evolutionary basis of human communication using bird and rodent models. The data sets comprise structured sequences of syllables or 'songs' produced by animals from different genotypes under different experimental conditions. The primary goals are to elucidate the roles of different genotypes and experimental conditions on animal vocalization behaviors and capabilities and also to learn complex serial dependency structures and systematic patterns in the vocalizations. We are developing novel statistical methods based on first and higher order Markovian dynamics that help answer these important scientific queries. The methods have appealing theoretical properties and practical advantages and are of very broad utility, with applications not limited to analysis of animal vocalizations.
Coauthor: David B. Dunson
Keywords: Animal Vocalizations, Bayesian Nonparametrics, Markov Chains, Sequential Data.

Estimation of a Two-component Mixture Model with Applications to Multiple Testing

Sen, Bodhisattva (bodhi@stat.columbia.edu)
Columbia University

Type: Invited Talk

Abstract. We consider estimation and inference in a two component mixture model where the distribution of one component is completely unknown. We develop methods for estimating the mixing proportion and the unknown distribution nonparametrically, given i.i.d. data from the mixture model. We use ideas from shape restricted function estimation and develop ”tuning parameter free” estimators that are easily implementable and have good finite sample performance. We establish the consistency of our procedures. Distribution-free finite sample lower confidence bounds are developed for the mixing proportion. We discuss the connection with the problem of multiple testing and compare our procedure with some of the existing methods in that area through simulation studies.
Keywords: Cramer-von Mises statistic, isotonic regression, local false discovery rate, lower confidence bound.

Estimation and Inference for High-Dimensional Kernel-Penalized Regression

Shojaie, Ali (ashojaie@uw.edu)
University of Washington

Type: Invited Talk

Abstract. Kernels are widely used to incorporate external biological information into statistical analyses. We will discuss a general framework for kernel-penalized regression, for incorporating external information into high-dimensional regression. Within this framework, the estimate of regression coefficients is obtained via the joint eigenproperties of multiple similarity matrices, or kernels. We will then present an inference framework to test for association of individual features with the outcome, when external information is incorporated in the form of a network representing similarities among covariates. We will discuss the power properties of the proposed framework for high-dimensional regression, as well as numerical results on various simulated and real data sets. This talk is based on joint work with Sen Zhao (UW) and Tim Randolph (FHCRC).
Keywords: kernel machine regression, similarity network, high-dimensional inference.

A Computationally Efficient Approach to Non-Parametric Density Estimation in the Presence of Measurement Error

Staudenmayer, John (jstauden@math.umass.edu)
University of Massachusetts, Amherst

Type: Invited Talk

Abstract. We consider the problem of estimating a density when the observed data are contaminated with measurement error. We take a Bayesian approach and model the density of interest

with a possibly infinite mixture of normals using a Dirichlet process mixture model. This talk's contribution is to develop a variational approximation (VA) approach to estimation and inference for this model. We conduct a simulation study to compare the VA approach to a Monte Carlo Markov Chain (MCMC) approximation to the posterior and deconvoluting kernel (DK) approach. The VA approach is nearly identical to MCMC in terms of accuracy and precision, and it is much faster to compute. While the DK also can be computed quickly, it is relatively inaccurate and imprecise. The methods are also illustrated on a dataset where error-prone accelerometers (fitness-tracker type devices) are used to estimate how active people are. This is joint work with Yue Chang, a PhD candidate at the University of Massachusetts, Amherst.
Keywords: Bayesian; deconvolution; Dirichlet process; physical activity; variational approximation.

The Focused Information Criterion for a Mixture Cure Model

Van Keilegom, Ingrid (ingrid.vankeilegom@uclouvain.be)
Universit catholique de Louvain / Institute of Statistics

Type: Invited Talk

Abstract. In many situations in survival analysis, it may happen that a fraction of the subjects under study will never experience the event of interest : they are considered to be cured. The mixture cure model is a common regression model in survival analysis that takes this feature into account. It supposes that the population consists of a mixture of two sub-populations, the cured ones and the non-cured ones, and it supposes a logistic model for the probability of being cured. For the non-cured sub-population we suppose a Cox proportional hazards regression model. We are interested in doing variable selection in this model using the focused information criterion (FIC). Of interest is therefore the asymptotic distribution of the estimators of the parameters and of the baseline hazard in a mixture cure model under local misspecification. Once this asymptotic distribution is obtained, the MSE can be used to guide selection of variables to be included in the logistic and Cox proportional hazards parts of the model. The method is illustrated by means of data regarding a UK financial institution. This is joint work with Gerda Claeskens.
Keywords: survival analysis; variable selection; local misspecification; cure models.

Nonparametric Statistics for Unlabeled Tree Objects

Wang, Haonan (wanghn@stat.colostate.edu)
Colorado State University

Type: Invited Talk

Abstract. In this talk, we consider a set of unlabeled tree objects with topological and geometric properties. Our motivating example is a data set of neurons from various brain regions. For each data object, two curve representations are developed to characterize its topological and geometric aspects. We further define the notions of topological and geometric medians as well as quantiles based on both representations. In addition, we take a novel approach to define the Pareto medians and quantiles through a multi-objective optimization problem. In particular, we study two different objective functions which measure the topological variation and geometric variation respectively. Analytical solutions are provided for topological and geometric medians and quantiles, and in general, for Pareto medians and quantiles, the genetic algorithm is implemented. The proposed methods are demonstrated in a simulation study and are also applied to analyze a real data set of pyramidal neurons from the hippocampus. Coauthor: Ela Sienkiewicz

Keywords: data object, genetic algorithm, multi-objective optimization, object oriented data, tree-structured data.

Comparing Large Covariance Matrices under Weak Conditions on the Dependence Structure

Wang, Lan (wangx346@umn.edu)
University of Minnesota

Type: Invited Talk

Abstract. Comparing large covariance matrices has important applications in modern genomics, where scientists are often interested in understanding whether relationships (e.g., dependencies or co-regulations) among a large number of genes vary between different biological states. We propose a computationally fast procedure for testing the equality of two large covariance matrices when the dimensions of the covariance matrices are much larger than the sample sizes. A distinguishing feature of the new procedure is that it imposes no structural assumptions on the unknown covariance matrices. Hence the test is robust with respect to various complex dependence structures that frequently arise in genomics. We prove that the proposed procedure is asymptotically valid under weak moment conditions. As an interesting application, we derive a new gene clustering algorithm which shares the same nice property of avoiding restrictive structural assumptions for high-dimensional genomics data. Using an asthma gene expression dataset, we illustrate how the new test helps compare the covariance matrices of the genes across different gene sets/pathways between the disease group and the control group, and how the gene clustering algorithm provides new insights on the way gene clustering patterns differ between the two groups. (Joint work with Jinyuan Chang, Wen Zhou and Wen-Xin Zhou)

Keywords: Gene clustering; High dimension; Hypothesis testing; Parametric bootstrap; Sparsity.

Efficient Estimation of Partially Linear Models for Spatial Data over Complex Domains

Wang, Lily (lilywang@iastate.edu)
Iowa State University

Type: Invited Talk

Abstract. We consider the estimation of partially linear models for spatial data distributed over complex domains. We use bivariate splines over triangulations to represent the nonparametric component on an irregular two-dimensional domain. The proposed method is formulated as a constrained minimization problem that does not require constructing finite elements or locally supported basis functions. Thus, it allows an easier implementation of piecewise polynomial representations of various degrees and various smoothness over an arbitrary triangulation. Moreover, the constrained minimization problem is converted into an unconstrained minimization via a QR decomposition of the smoothness constraints, which leads to a penalized least squares method to estimate the model. The estimators of the parameters are proved to be asymptotically normal under some regularity conditions. The estimator of the bivariate function is consistent, and its rate of convergence is also established. The proposed method enables us to construct confidence intervals and permits inference for the parameters. The performance of the estimators is evaluated by two simulation examples and by a real data analysis.
Keywords: Bivariate splines; Penalty; Semiparametric regression; Spatial data; Triangulation.

Optimal Estimation for Quantile Regression with Functional Response

Wang, Xiao (wangxiao@purdue.edu)
Purdue University

Type: Invited Talk

Abstract. 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 suitable for the big data setting and shown to be rate-optimal. Simulations and real data analysis are conducted to examine the finite-sample performance. This is a joint work with Zhengwu Zhang, Linglong Kong, and Hongtu Zhu.
Keywords: Big data; Functional regression; Minimax rate; Quantile regression.

Adaptive Prediction in Additive Models

Zhang, Cun-Hui (czhang@stat.rutgers.edu)
Rutgers University

Type: Invited Talk

Abstract. We consider penalized estimation in additive models with a large number of mixed components including univariate linear effects, group effects and nonparametric effects of one or

several variables. A prediction error bound, derived under a restricted eigenvalue or compatibility condition, provides rate optimality for the penalized estimator in various settings. In nonparametric additive models, the prediction error bound yields existing and new results under different smoothness and sparsity conditions. An adaptive estimator is constructed to unify these and some non-convex rate optimal methods. This is joint work with Zhiqiang Tan.
Keywords: Additive model; Reproducing kernel Hilbert space; Model selection; Prediction; Adaptive estimation..

Statistical and Computational Guarantees of Lloyd's Algorithm

Zhou, Harrison (huibin.zhou@yale.edu)
Yale University

Type: Invited Talk

Abstract. The initializer for the Lloyd's Algorithm is not required to be consistent, but needs to be better than random guess. After an order of log(n) iteration steps we show that the algorithm achieves the asymptotic efficiency. A consequence of this result is to improve some work in literature for Gaussian mixture models, community detection, and crowdsourcing.
Keywords: Lloyd's Algorithm; Optimality; Mixtures of Gaussians; Community Detection; Crowdsourcing..

Errors-in-Variables Models with Dependent Measurements

Zhou, Shuheng (shuhengz@umich.edu)
University of Michigan

Type: Invited Talk

Abstract. I will discuss an errors-in-variables model where the covariates in the data matrix are contaminated with random noise. This model is significantly different from those analyzed in the literature in the sense that we allow the measurement error for each covariate to be dependent across observations. Such error structures appear in the science literature, for example, when modeling the trial-to-trial fluctuations in response strength shared across a set of neurons. We provide theory, real-data examples and simulation evidence showing that we can recover the regression coefficients for a class of errors-in-variables problems. This is joint work with Mark Rudelson.
Keywords: Errors-in-variables regression; high dimensional data.

Tensor Partition Mixture Models for Heterogeneous Functional Prediction

Zhu, Hongtu (hzhu5@mdanderson.org)
UT MD Anderson Cancer Center

Type: Invited Talk

Abstract. We propose a tensor partition mixture modeling framework for efficiently predict a scalar response as a function of imaging covariates, that usually take the form of tensors. Our TPMM is developed to address the key features of imaging data: relatively low signal to noise ratio, spatially clustered effect regions, the tensorial structure of imaging data, and heterogeneous effect regions across subjects. We illustrate TPMM with simulations and a real data application based on the Alzheimer's disease neuroimaging initiative (ADNI) study. To gain a deep understanding of TPMM, we investigate the theoretical properties of TPMM under a general setting. This is based on a joint work with Michael Miranda and Lian Heng.

Keywords: Tens partition; mixture models; functional data; prediction; heterogeneous effect.