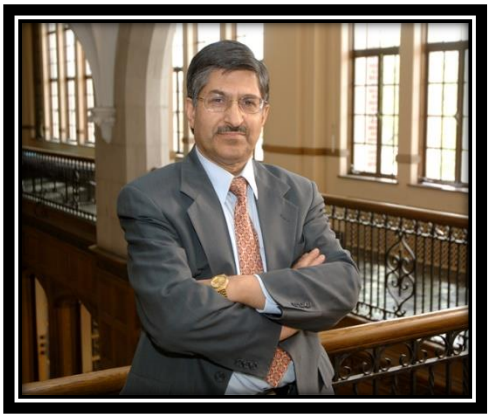
Michigan State Symposium on Mathematical Statistics and Applications

From Time Series and Stochastics, to Semi- and Non-Parametrics, and to High-Dimensions Models

In Honor of Hira L. Koul's Legacy

Jack Breslin Student Events Center
Michigan State University
September 14 – 16, 2018



The 2018 Michigan State Symposium on mathematical statistics and applications will be held at MSU from Sep 14-16, 2018. It is a conference designed around the scientific legacy of Prof. Hira L. Koul, who was a member of MSU's department of statistics and probability for decades. The scientific areas which will be represented at the conference are all connected, often quite directly, to the work, which Prof. Koul has produced. Topics covered will include: Semi- and non-parametric foundations of data science; Asymptotic theory of efficient and adaptive estimation; Inference for high-dimensional data; Inference for long-memory and other stochastic processes; Nonlinear Time Series analysis with applications to econometrics and finance; Robust multivariate methods; Survival analysis and its applications; and Sequential estimation and design. With ample time built into the schedule for discussions, the conference will give participants opportunities to engage in emerging and fruitful cross-group collaborations. It will bring together established and aspiring researchers from around the country and abroad, to explore frontiers of mathematical statistics.

Thank you to Scientific Committee
Chair: Weixing Song (Kansas State University)
Dennis Gilliland (Michigan State University)
Taps Maiti (Michigan State University)
Vince Melfi (Michigan State University)
James Stapleton (Michigan State University)
Frederi Viens (Michigan State University)

Special Thanks for their generous donation in support of the Symposium
Hongwen Guo and Zhongjun Ge
Tao He
Linyuan Li
Xiaoqing Zhu

Thank you to our sponsors:

National Science Foundation

Institute of Mathematical Statistics

Oluwakemi Ajayi
University of KwaZulu-Natal
South Africa

Introduction to Survival Analysis: Design and Analysis

In a competitive world which is rapidly changing, with increasing complexity across the financial services industries, causing banks, investment management and insurance firm to face a diverse array of challenges and concerns. To achieve economic development, Financial Service Industry plays a crucial role in sustainable development by relying on the use of information to evaluate financial services. In order to ensure that Finance Industry continues to recover from the crisis and strengthen risk management through effective decision-making on regulations and cost reduction. A financial service provider uses a model builder's data mining ability and skill to discover which customers are likely to switch to another provider. There is a claim that customers in a particular service group are 50% likely to migrate, which contradicts the company's previous conventional knowledge. To ensure that a model builder has truly identified something new in an SPSS model, a company analyst runs a number of statistical analyses and displays the Kaplan Meier survival curves. These two survival curves are compared statistically by testing the null hypothesis that there is no difference in survival among the two service groups against the appropriate alternative. The null hypothesis is statistically tested by another test known as a log-rank test. The statistical significance of the discovered pattern is measured with the use of survival analysis output to help understand the length of customer relationships with the company before they migrate. Plots from the Kaplan Meier survival function curve compares the two groups to show the difference. Cumulative survival proportion appears to be much higher in one group (A) compared to another (B). To determine if there are differences in the survival distribution between the groups, a log-rank test is performed. The associated p-value is compared at 5% level of significance. Summarized report and a brief description of the statistical analysis procedure for the Kaplan Meier survival curve and output are discussed. The report informs the company on how to wisely invest resources to address migrating individuals identified by SPSS modeler and provides information to decision-makers.

Cuneyt Akcora
Department of Computer Science
University of Texas at Dallas
USA

Understanding Cryptocurrency Price Formation from Time Series of Local Blockchain Graph Features

Over the last couple of years, digital cryptocurrencies and the Blockchain technology that forms their basis have witnessed a flood of attention. With the emergence and rapid adoption of Blockchain and the associated cryptocurrencies understanding the network dynamics behind Blockchain technologies has emerged as an important research direction. Unlike other financial networks such as stock and currency trading, blockchains have the entire time series of interaction graph accessible to the public. This facilitates a thorough analysis of the network data with a time series approach. A natural question to ask is whether the network dynamics of a cryptocurrency impact its price in dollars. We show that on the one hand, time series of standard global graph features such as degree distribution are not enough to capture the network dynamics that impact the underlying cryptocurrency price. In contrast, multiple time series of persistent topological homologies can explain higher level interactions among nodes in Blockchain graphs and can be used to build more accurate price prediction models.

Fatimah Alshahrani
Department of Statistics and Probability
Michigan State University
USA

Investigating Mean-Reverting Processes with Gumbel Noise

Extreme Value Theory (EVT) deals with studying "rare events" by taking either the maxima or minima of observations. We applied EVT to the annual maximum for sea level data provided by the Actuaries Climate Index, (ACI). Our analysis shows that the annual maxima follow the Type I Generalized Extreme Value Distribution, which is known as the Gumbel distribution. The time series of monthly records, as described by ACI is said to follow the Gumbel distribution. Initial investigation shows the de-trended data has some of the features of the Ornstein Uhlenbeck (OU) process, i.e. a Gaussian mean reverting process. However, unlike the OU process, our data is presumably non-Gaussian. Using the relationship between stochastic differential equations and discrete time processes, we are building a reliable model using tools from stochastic analysis including the Malliavin Calculus, and objects such as AR(1) process and the OU process, rather than relying upon the so-called Gumbel noise, which is an object used largely in Machine Learning.

Bengt Arnetz
Department of Family Medicine
College of Human Medicine
Michigan State University
USA

Forecasting Patient No-show and long-term Health Trajectory using Bayesian Analysis

Background: The healthcare cost explosion and quality deficiencies represent major challenges to the United States' financial, social, and population health. By some estimates, 30% of healthcare resources are wasted. Another challenge is to develop a better understanding of the long-term trajectory of patients with chronic diseases, e.g., hypertension, diabetes, and metabolic syndrome. Clinical questions that beg better answers include: Which patients are most likely to have the worst health trajectory, and therefore require intensive medical interventions? Which patients will manage fine with only minimal interventions? What are the optimal mixture of various treatment strategies, e.g., pharmaceutical, environmental, and/or behavioral for specific patient groups? Another challenge to healthcare systems is patients that for various reasons do not show-up for their medical appointments. This results in underutilization of sparse resources and, possibly, worse patient health outcomes.

Approach: Bayesian analysis that incorporates close cooperation between statisticians and clinicians represents a promising new approach to address such challenges. It does not suffice to merely apply machine learning and artificial intelligence. By involving clinicians, specific models can be developed and tested in cooperation with the statisticians. Through an interactive, self-corrective, and reiterative model developing process, the end-result will assist clinicians and healthcare systems to apply an evidence-based strategy to improve patient outcomes, processes, and resource utilization.

A team of Bayesian statisticians, health systems researchers, and clinicians are in the process of applying this interactive and reiterative process to determine whether chronic disease patients, whom use around 80% of the healthcare resources, can be managed more efficiently. Furthermore, can clinics, especially those serving underserved and vulnerable patients, better predict no-shows and implement policies that facilitate patients' ability to attend medical appointments?

Conclusion: Healthcare costs represent almost 20% of the United States gross national product, the highest in the world. However, based on outcomes, the United States are not even ranked among the top 10. By forging a closer collaboration with key stakeholders, it is likely that we will achieve a marked improvement in patient outcomes, at the same time as the pressure on an already strained healthcare system is decreased.

Judy Arnetz, Sukhesh Sudan, Bengt Arnetz
Department of Family Medicine
Michigan State University

Taps Maiti, Frederi Viens, Léo Neufcourt
Department of Statistics and Probability
Michigan State University
USA

Workplace Bullying among Nurses and Patient Outcomes – An Exploratory Bayesian Analysis

Background: The health and safety of hospital patients is largely dependent on the skills, safety and well-being of the nurses who care for them. Workplace bullying is a commonly recognized aspect of the nursing culture that presents a significant challenge to nurse safety. Bullying entails repeated negative behavior directed towards a particular individual with the intention of offending, abusing, or intimidating them. Nurse bullying has been inversely associated with nurses' physical and psychological health, job satisfaction, work productivity, and job turnover. However, the possible impact of bullying on patient outcomes has not been explored. The aim of this study was to examine the association between self-reported nurse bullying and documented patient quality indicators in a single hospital.

Methods: An anonymous survey on workplace bullying was administered in 2017 via email to all registered nurses (n=1780) in a regional hospital in Michigan. A total of 432 nurses (24.3% response rate) from 37 units responded. The proportion of nurses that reported being bullied and/or witnessing bullying, respectively, in the last six months was calculated at the unit level. Unit-level de-identified nursing-sensitive patient adverse event data from the last two quarters from the National Database of Nursing Quality Indicators (NDNQI®) were used as dependent variables: (1) hospital-acquired pressure ulcers; (2) patient falls; (3) central line associated blood stream infections (CLABSI); (4) catheter associated urinary tract infections (CAUTI); and (5) ventilator associated events (VAEs). A collaborative team of Family Medicine researchers and statisticians utilized the subject-area knowledge of the Family Medicine team to identify unit-level priors, including for nurse qualification (highest nursing degree completed), nurse staffing (ratio of the number of nurse days to patient days), and failure to report work unit due to the sensitivity of the questionnaire subject. Bayesian methodology was used to investigate whether self-reported and witnessing bullying, respectively, predicted patient adverse events after taking nurse qualification and staffing into account.

Results: 139 nurses (36.9%) reported having personally experienced bullying and 191 (51.5%) reported having witnessed someone else being bullied in the past 6 months. Participants' personal experiences of bullying were directly related to CLABSI (95% credible interval, CI [0.017, 0.212] but not to the other patient adverse events. Witnessing bullying was not significantly related to any patient adverse events at a 95% credibility level. However, both forms of bullying approached significance (posterior probability >0.81) in association with VAEs. Nurse qualification was positively and significantly associated with patient falls and unassisted falls at the 10% level (posterior probability >0.9). Nurse staffing associated positively with CLABSI CI [-0.086, 1.3] and CAUTI CI [0.14, 1.53] for those who experienced bullying and those that witnessed bullying CI [CLABSI -0.106, 1.45]; CI [CAUTI 0.14, 1.53].

Conclusions: Results suggest that efforts to reduce bullying among hospital nurses could result in corresponding moderate but likely significant reductions in rates of patient adverse events, specifically CLABSI, CAUTI, and VAEs.

Selin Aviyente
Department of Electrical and Computer Engineering
Michigan State University
USA

Discriminative Dictionary Learning for Tensor Data

Dictionary learning methods aim to learn atoms that can best represent a signal class. In recent years, these methods have been extended to learn discriminative dictionaries such that the learned atoms are specific to each class enabling better classification accuracy. With the increase of high dimensional and multi-aspect data, there is a growing need to extend dictionary learning algorithms to tensor type data. In this talk, we propose an efficient, separable and orthogonal dictionary structure for learning class-specific dictionaries for tensor objects. The proposed cost function tries to minimize the representation error as well as within-class scatter while putting a sparsity constraint on the learned representation. The algorithm is applied to different tensor object classification tasks with extensive evaluations of the effect of sparsity, discriminability and reconstruction error on classification accuracy.

Richard T. Baillie
Department of Economics
Michigan State University
USA

The Heterogeneous Autoregressive Model; the Role of Long Memory in Realized Volatility

The presence of long memory in Realized Volatility (RV) of asset prices is a widespread stylized fact. The origins of RV have been attributed to jumps, structural breaks, nonlinearities, or pure long memory. The alternative economic explanations are extensions of the Heterogeneous Autoregressive (HAR) model with jumps and good volatility. This paper assesses the contribution of these rival explanations through the use of fractional long memory models combined with extended HAR models and random coefficient extended HAR models. We find evidence that the statistical modeling of long memory is generally important, in addition to more structural model explanations.

Asish Banik
Department of Statistics and Probability
Michigan State University
USA

Bayesian classification of Alzheimer's disease stages from longitudinal volumetric MRI data

The primary objective of this article is to build a classification method using longitudinal volumetric magnetic imaging (MRI) data from five regions of interest (ROIs) (hippocampus (H), entorhinal cortex (EC), middle temporal cortex (MTC), fusiform gyrus (FG) and whole brain (WB)). A functional data analysis method is used to handle the longitudinal measurement of ROIs and later the functional coefficients are used in the classification models. We propose a Polya-gamma augmentation method to classify normal controls and diseased patients based on the functional MRI measurements. We get a fast posterior sampling by avoiding slow and complicated Metropolis-Hastings algorithm. Our main motivation is to determine the important ROIs which have the highest separating power for classifying our dichotomous response. We compared the sensitivity, specificity and accuracy of classification based on single ROIs and also with various combinations of them. We obtained sensitivity over 85\% and specificity around 90\% for most of the combinations. Addition of few baseline Mental state exam scores in the model improve our results. The combination in which all five important ROIs and baseline patients' scores are included provides the best result among all other combinations.

Andrew Bender
Department of Epidemiology and Bio-Statistics
Michigan State University
USA

Critical Perspectives in Statistical Analysis of Neuroimaging Data

Neuroimaging research is inherently interdisciplinary, but divergent perspectives between disciplines can complicate the success of a collaborative project. Statistical modeling and analysis of neuroimaging data can strongly benefit from neuroscientists' perspectives on 1) brain organization, appreciating the brain as a complex organ composed of multiple, integrated subsystems; and 2) limitations and assumptions associated with different types of neuroimaging data, including design and processing of neuroimaging study data and potential sources of systematic error and bias. For example, among neuroimaging investigators, analytic perspectives widely range from strictly confirmatory, theoretically-guided approaches to exploratory or 'hypothesis-free' analyses, and each has its own advantages and downsides. Whereas confirmatory approaches may lend to easier interpretation or translation, they may under-identify novel effects. In contrast, exploratory approaches may provide new insights into neural organization, but have greater potential for generating spurious findings. This presentation will 1) address goals and values in neuroimaging data, including common points of practical divergence, 2) describe specific challenges in statistical analysis of structural and functional neuroimaging data, 3) identify current needs for new directions in statistical analysis, and, 4) propose alternatives to common approaches that may circumvent some of these issues.

Shrijita Bhattacharya
Department of Statistics and Probability
Michigan State University
USA

Data-adaptive trimming of the Hill estimator and detection of outliers in the extremes of heavy-tailed data

We introduce a trimmed version of the Hill estimator for the index of a heavy-tailed distribution, which is robust to perturbations in the extreme order statistics. In the ideal Pareto setting, the estimator is essentially finite-sample efficient among all unbiased estimators with a given strict upper break-down point. For general heavy-tailed models, we establish the asymptotic normality of the estimator under second order regular variation conditions and show it is minimax rate-optimal in the Hall class of distributions. We also develop an automatic, data-driven method for the choice of the trimming parameter, which yields a new type of robust estimator that can adapt to the unknown level of contamination in the extremes. This adaptive robustness property makes our estimator particularly appealing and superior to other robust estimators in the setting where the extremes of the data are contaminated. As an important application of the data-driven selection of the trimming parameters, we obtain a methodology for the principled identification of extreme outliers in heavy tailed data. Indeed, the method has been shown to correctly identify the number of outliers in the previously explored Condroz data set.

Md Al Masum Bhuiyan
Computational Science Program
The University of Texas at El Paso
USA

Marcinkiewicz's strong law of large numbers for nonlinear expectations

The sub-linear expectation space is a nonlinear expectation space having advantages of modeling the uncertainty of probability and distribution. In the sub-linear expectation space, we use capacity and sub-linear expectation to replace probability and expectation of classical probability theory. In this paper, the method of selecting subsequence is used to prove Marcinkiewicz's strong law of large numbers under sub-linear expectation space. This result is a natural extension of the classical Marcinkiewicz's strong law of large numbers to the case where the expectation is nonlinear. In addition, this paper also gives a theorem about convergence of a random series.

Guorong Dai
Texas A&M University
USA

Efficient Estimators for Expectations in Conditional Mean Models with Responses Missing at Random

We consider regression models in which only the mean response given the covariates and the regression function is modeled parametrically. This model is useful when restrictive assumptions on the structure of the random errors cannot be justified. We propose estimators for expectations of the joint distribution of response and covariates when responses are possibly missing, with the missingness explained by the covariates. Our estimator is a non-parametric estimator involving a Nadaraya-Watson type estimator for conditional expectations, improved by an additive correction term that takes into account the non-linear regression structure. We prove that the estimator is asymptotically efficient in the Hajek and Le Cam sense. Simulations and an example using real data confirm the optimality of our approach.

Somnath Datta
Department of Biostatistics
University of Florida
USA

Adjustments of Multi-Sample U-Statistics to Right Censored Data and Confounding Covariates

We consider U-statistics that can be used for comparing distribution of outcomes in two groups. We propose adjustments to the classical U-statistics for mediating potential bias come from right-censoring of the outcomes and presence of confounding covariates. These newly proposed U-statistics are appropriate when, in addition to right censored outcome, some fixed covariates are observed and associated with both group membership and the outcome in an observational study. The summands of U-statistics are re-weighted and normalized based on a combination of inverse probability of censoring weights and propensity score based weights. Censoring time may depend on the group membership or some observed time-dependent covariates, which may result in censoring mechanisms of varying degrees of complexity. In total, four censoring mechanisms are considered for the two group comparison. Simulation results are used to illustrate the impact of confounding covariates and right-censoring on the performance of the newly proposed U-statistics under different censoring mechanisms. We also demonstrate that large sample inferences for the adjusted U-statistics are valid using jackknife variance estimator. Comparisons of more than two groups are also considered from certain pairwise group comparisons.

Richard Davis
Department of Statistics
Columbia University
USA

Extreme Value Analysis Without the Largest Values: What Can Be

In this paper we are concerned with the analysis of heavy-tailed data when a portion of the extreme values is unavailable. This research was motivated by an analysis of the degree distributions in a large social network. The degree distributions of such networks tend to have power law behavior in the tails. We focus on the Hill estimator, which plays a starring role in heavy-tailed modeling. The Hill estimator for this data exhibited a smooth and increasing "sample path" as a function of the number of upper order statistics used in constructing the estimator. This behavior became more apparent as we artificially removed more of the upper order statistics. Building on this observation we introduce a new version of the Hill estimator. It is a function of the proportion theta of the upper order statistics used in the estimation, but also depends on the proportion delta of unavailable extremes values. We establish functional convergence of the normalized Hill estimator to a Gaussian process. An estimation procedure is developed based on the limit theory to estimate the number of missing extremes and extreme value parameters including the tail index and the bias of Hill's estimate. We illustrate how this approach works in both simulations and real data examples. This is joint work with Jingjing Zou and Gennady Samorodnitsky.

Soukaina Douissi
University Cadi Ayyad Marrakech
Morocco

Parameter estimation for general Gaussian processes with discrete observations.

In this talk we give a general framework to study parameter estimation problems for general Gaussian sequences. We use tools from analysis on Wiener space. No assumption of stationarity is required. The only assumptions made on the sequence are the existence of an asymptotic variance, that a least-squares-type estimator for this variance parameter has a bias and a variance which can be controlled, and that the sequence's covariance function, which may exhibit long memory, has a no-worse memory than that of fractional Brownian motion with Hurst parameter not greater than 3/4. The applications we give concern the estimation of the asymptotic variance for various fractional-noise-driven Ornstein-Uhlenbeck processes. This is joint work with Khalifa Es-Sebaiy (Kuwait University, Kuwait) and Frederi Viens (Michigan State University, USA).

Omar De la Cruz Cabrera
Department of Mathematical Sciences
Kent State University
USA

Penalized estimating equations for a stochastic model for compositional data

We study several approaches for generating penalized estimating equations for the parameters of a stochastic process that models the evolution in continuous time of compositional data (i.e., vectors of positive numbers that add up to 1). The parameters to be estimated specify the Dirichlet distribution that is the invariant distribution of the process, as well as a time scale parameter. Regularization by penalization becomes especially important when the number of categories (and thus the number of parameters) is large, as is the case in many applications. This is joint work with Oana Mocioalca and Yicheng Su.

Phillip M Duxbury
Dean, College of Natural Science
Department of Physics and Astronomy
Michigan State University
USA

Three vignettes at the interface of statistics, algorithms and statistical physics: Random fields in Ising systems; Image reconstruction and beam emittance

I will talk about two problems that my group works on; random field Ising problems and Image reconstruction; where active physics/math collaborations have revealed new methods for solving hard physics problems. I will also talk about the emittance of many particle beams which also deals with issues that are familiar to statisticians, but where modern statistics results have not yet been applied.

Philip Ernst
Department of Statistics
Rice University
USA

Optimal Real-time Detection of a Drifting Brownian Coordinate and Statistical Inference for Paths of Stochastic Processes

Consider a three-dimensional Brownian motion whose two coordinate processes are standard Brownian motions with zero drift, and the third (unknown) coordinate process is a standard Brownian motion with a non-zero drift. Given that only the position of the three-dimensional Brownian motion X is being observed, the problem is to detect, as soon as possible and with minimal probabilities of the wrong terminal decisions, which coordinate process has the non-zero drift. We solve this problem in the Bayesian formulation under any prior probabilities of the non-zero drift being in any of the three coordinates when the passage of time is penalized linearly.

Time permitting, we will then briefly discuss how we recently resolved a longstanding open statistical problem. The problem, formulated by the British statistician Udny Yule in 1926, is to mathematically prove Yule's 1926 empirical finding of "nonsense correlation". The solution of this problem has prompted our recent investigation into tests of independence for paths of stochastic processes.

Jianqing Fan
Department of Operations Research and Financial Engineering
Princeton University
USA

Farming in Data Rich Environment

Correlated and heavy-tailed data arise frequently in a wide range of scientific and engineering problems: from genomics, medical imaging to neuroscience and finance. This talk introduces Factor-Adjusted Robust Multiple testing (FARM-test) and Factor-Adjusted Robust Model selection (FARM-select). The former is introduced to control the false discovery proportion for large-scale simultaneous inference when variables are highly correlated, and the latter deals with variable selection problems when covariates are highly correlated. We demonstrate that robust factor adjustments are extremely important in both improving the power of the tests and controlling FDP. We identify general conditions under which the proposed method produces a consistent estimate of the FDP. We also prove that factor adjustments significantly reduce the conditions needed for selection consistency. The results will be illustrated by numerical experiments. This is joint work with Yuan Ke, Qiang Sun, Wenxin Zhou, and Kaizheng Wang.

Ahmad Flaih
University of Al-Qadisiyah
Iraq

Bayesian Regression with Errors from ESDIW Distribution

The Epsilon Skew Double Inverted Weibull (ESDIW) distribution is introduced as generalization to the Inverted Weibull distribution. The ESDIW density is another lifetime model that can be used in reliability. Our goal is to make inference on using Bayesian techniques for deriving the posterior density function of the regression coefficient vector when the errors are from ESDIW density. The non-normalized joint posterior density function of the parameters and the Fisher information matrix of the sample from ESDIW probability density function have been estimated.

Richard Furnstahl
Ohio State University
USA

Bayesian statistics for effective field theories

The physics of the atomic nucleus gives rise to a tower of emergent phenomena at widely varying energy scales. A method called effective field theory (EFT) turns the challenge of dealing with these disparate scales into an advantage by using their ratios as expansion parameters. But despite the promised systematic nature of this approach, a robust framework to include theoretical uncertainties in EFT predictions of experiment has been lacking. This has changed with the first applications of Bayesian statistics to EFTs. Truncation errors for the EFT expansion of nuclear force predictions in continuous domains (functions of energy and scattering angle) are well accounted for by a discrepancy model using Gaussian processes (GPs). Posteriors for the GP parameters give novel physical insight into the nature of EFT expansions, such as their breakdown scales. Model checking diagnostics are proving to be useful not only to validate credibility intervals but as tools for physics discovery. The first successes motivate future applications using Bayesian model selection and model averaging.

Yuan Gan
The Chinese University of Hong Kong
Hong Kong

Multiple-regime Self-excited Vector Threshold Autoregressive Models with Multivariate Threshold Variables

This talk proposes a multivariate extension of the well-known threshold autoregressive (TAR) model in nonlinear time series literature. We consider k dimensional multiple-regime self-excited vector threshold autoregressive models with multivariate threshold variables, where the regime switches are governed by the lag d series. Specifically, the regimes are the subsets partitioned by unknown threshold hyperplanes in the k dimensional space, and the time series follows different models when the lag d series falls into different partitioning subsets. One challenge in estimating such models arises from the great complexity of regimes under high dimensional settings. We formulate the task of model estimation into a minimization problem based on minimum description length (MDL) principle and develop a genetic algorithm that can estimate the number of threshold hyperplanes, the parametric equations of threshold hyperplanes and vector autoregressive (VAR) model parameters in each regime simultaneously. The consistency of such estimators is established theoretically with an illustration by numerical simulations.

Joseph Gardiner
Director, Division of Biostatistics
Michigan State University
USA

Competing Risks Analyses: Overview of Regression Models

In competing risks analyses, the time to a terminal event such as death is analyzed together with its cause. Death by one cause precludes occurrence of death by any other cause, because an individual can die only once. The cumulative incidence function CIF(j, t) is the probability of death by time t from cause j. Cause-specific hazard functions are the analogs of the hazard function when only a single cause is present. By incorporating explanatory variables in cause-specific hazard functions, provides an approach to accessing their impact on the CIF and on overall survival. We discuss methods for estimation of the CIF from event times and their associated causes, allowing for right censoring when the event and its cause are not observed. When covariates are present, a semi-parametric approach similar to Cox regression models the cause-specific hazards. The Fine-Gray model defines a sub-distribution hazard function that has an expanded risk set comprised of individuals at risk of the event by any cause at t, together with those who died before t from any cause other than the cause j of interest. Finally, with additional assumptions a full parametric analysis is also feasible. We illustrate the application of these methods with an empirical data set.

Pei Geng
Mathematics
Illinois State University
USA

Regression model checking with Berkson measurement errors in covariates

A minimum distance regression model checking approach is proposed when covariates are observed with Berkson measurement errors. When the measurement error density is unknown, it is assumed that validation data is available. This assumption makes it possible to estimate the calibrated regression function consistently. The proposed tests are based on a class of minimized integrated square distances between a nonparametric estimate of the calibrated regression function and the parametric null model being fitted. The asymptotic normality of these tests under the null hypothesis and the consistency against certain alternatives are established. A simulation study shows desirable performance of a member of the proposed class of estimators and tests.

Samiran Ghosh
Center for Molecular Medicine and Genetics
Wayne State University
USA

Non-Inferiority Design in Comparative Effectiveness Research: Should We be Bayesian for a While?

Randomized controlled trials (RCT's) are an indispensable source of information about efficacy of treatments in almost any disease area. With the availability of multiple treatment options, comparative effectiveness research (CER) is gaining importance for better and informed health care decisions. However, design and analysis of effectiveness trial is much more complex than the efficacy trial. The effect of including an active comparator arm/s in a RCT is immense. This gives rise to superiority and non-inferiority trials. The non-inferiority (NI) RCT design plays a fundamental role in CER, which will be also focus of this talk. In the past decade many statistical methods have been developed, though largely in the Frequentist setup. However, availability of historical placebo-controlled trial is useful, and if integrated in the current NI trial design, it can provide better precision for CER. This may reduce sample size burden and improves statistical power significantly in current trials. Bayesian paradigm provides a natural path to integrate historical as well as current trial data via sequential learning in the NI setup. In this talk, we will discuss both fraction margin and fixed margin based Bayesian approach for three-arm NI trial. We will also discuss some interesting open problems related to CER using NI trial.

Liudas Giraitis
School of Economics and Finance
Queen Mary, University of London

Inference on Time Series with Changing Mean and Variance

The paper develops point estimation and asymptotic theory with respect to a semiparametric model for time series with moving mean and unconditional heteroscedasticity. These two features are modelled nonparametrically, whereas autocorrelations are described by a short memory stationary parametric time series model. We first study the usual least squares estimate of the coefficient of the first-order autoregressive model based on constant but unknown mean and variance. Allowing for both the latter to vary over time in a general way we establish its probability limit and a central limit theorem for a suitably normed and centered statistic, giving explicit bias and variance formulae. As expected mean variation is the main source of inconsistency and heteroscedasticity the main source of inefficiency, though we discuss circumstances in which the estimate is consistent for, and asymptotically normal about, the autoregressive coefficient, albeit inefficient. We then consider standard implicitly-defined Whittle estimates of a more general class of short memory parametric time series model, under otherwise more restrictive conditions. When the mean is correctly assumed to be constant, estimates that ignore the heteroscedasticity are again found to be asymptotically normal but inefficient. Allowing a slowly time-varying mean we resort to trimming out of low frequencies to achieve the same outcome. Returning to finite order autoregression, nonparametric estimates of the varying mean and variance are given asymptotic justification, and forecasting formulae developed. Finite sample properties are studied by a small Monte Carlo simulation, and an empirical example is also included. This is joint work with V. Dalla and P.M. Robinson.

David Han
Department of Management Science & Statistics
University of Texas at San Antonio
USA

Bayesian Design Optimization of a Non-specific Sensor System for Calibration of Analyte Responses

In the current and future generation of products, the nature of field reliability data is changing rapidly and dramatically. With modern sensor technology, innovative data analytics is emerging in reliability and quality technology. In order to reduce unexpected failures of a system in service, it is crucial to assess its condition/health accurately in real time. Using an array of sensors with well calibrated but different tuning curves, it is possible to appreciate a wide range of stimuli. The objective of this research is to adopt Bayesian framework to develop a statistically sound estimation method given sensor responses by elucidating the uncertain nature of environment-dependent stimuli through a choice of prior. Using decision-theoretic approach, the design optimization of a sensory system is also explored via maximization of the expected utility. Furthermore, to characterize the fundamental analytical capability of a measurement system, general-purpose selectivity is defined in an information-theoretic manner.

Yue Hao
Department of Physics and Astronomy, NSCL/FRIB
Michigan State University
USA

Preliminary Application of Bayesian Inference in Accelerator Commissioning

In this talk we will report the preliminary application of the Bayesian Inference of the unknown parameters of accelerator model using the Facility for Rare Isotope Beams (FRIB) commissioning data. The inference result not only indicates the value of the unknown parameter, but also the confidence of adopting the value. The Bayesian approach provides an alternative method to understand the difference between the accelerator model and the hardware and may help achieving ultimate beam parameters of FRIB.

Tailen Hsing
Department of Statistics
University of Michigan
USA

Model and Inference of Local Stationarity

Stationarity is a common assumption in spatial statistics. The justification is often that stationarity is a reasonable approximation to the true state of dependence, if we focus on spatial data "locally." In this talk, we first review various known approaches for modeling nonstationary spatial data. We then examine the notion of local stationarity in more detail. To illustrate, we focus on the multi-fractional Brownian motion, for which a thorough analysis could be conducted assuming data are observed on a regular grid. A theoretical lower bound for the minimax risk of this inference problem is established for a wide class of smooth Hurst functions. We also propose a new nonparametric estimator and show that it is rate optimal. Implementation issues of the estimator including how to overcome the presence of a nuisance parameter and choose the tuning parameter from data will be considered. Finally, extensions to more general settings that relate to Matheron's intrinsic random functions will be briefly discussed.

Tomoyuki Ichiba
Department of Statistics & Applied Probability
Center for Financial Mathematics and Actuarial Research, and Mathematics
University of California Santa Barbara
USA

Detecting Mean-Field in Diffusions on a Large Network

We shall consider a detection problem in a large system of stochastic network in continuous time. We describe the system by an infinite-dimensional, nonlinear stochastic differential equation of McKean-Vlasov type. We determine a dichotomy of presence or absence of mean-field interaction, and discuss the problem of detecting its presence from the observation of a single component process. We shall also discuss its discrete-time analogue and corresponding estimation problems.

Umar Islambekov
Mathematical Sciences
University of Texas at Dallas
USA

Unsupervised Space-Time Clustering using Persistent Homology

We present a new clustering algorithm for space-time data using the method of persistent homology. This method, having its roots in algebraic topology, is a popular tool in topological data analysis and it is used to extract topological information from data. A notable aspect of persistent homology consists in analyzing data at multiple resolutions which allows to distinguish true features from noise based on the extent of their persistence. We evaluate the performance of our algorithm on synthetic data and compare it to other well-known clustering algorithms such as K-means, hierarchical clustering and DBSCAN. We illustrate its application in the context of a case study of water quality in the Chesapeake Bay.

Mark Iwen
Department of Mathematics
Department of Computational Mathematics, Science and Engineering
Michigan State University
USA

Sublinear-Time Algorithms for Approximating Functions of Many Variables

The development of sublinear-time compressive sensing methods for signals which are sparse in Bounded Orthonormal Product Bases (BOPBs) will be discussed. These new methods are obtained from CoSaMP by replacing its usual support identification procedure with a new faster one inspired by fast Sparse Fourier Transform (SFT) techniques. The resulting sub-linearized CoSaMP method allows for the rapid approximation of BOPB-sparse functions of many variables which are too hideously high-dimensional to be learned by other means. Both numeric and theoretical recovery guarantees will be presented. This is joint work with Bosu Choi and Felix Krahmer.

Richard A. Johnson
Department of Statistics
University of Wisconsin
USA

Some Comments on the Izawa Bivariate Gamma Distribution

With an application to microarray data in mind, we consider the bivariate gamma distribution due to Izawa (1965). However, inference procedures are still lacking, perhaps because the density function contains a modified Bessel function of the first kind of order k and is not particularly tractable. Looking ahead to the microarray application, we focus our attention on the model which has equal shape parameters. We verify the conditions for the MLE to be asymptotically normal and also determine which elements of the Fisher information matrix can be calculated in closed form and which require numerical integration. The numerical details are non-trivial. Simulation studies illuminate the properties of maximum likelihood estimators. We also establish an asymptotic test for independence in this non-regular case.

In the context of microarray data, Izawa's bivariate gamma model produces two-dimensional patterns for gene expressions that are similar to those in many applications. We entertain a hierarchical model where latent mean expression levels are first generated from the Izawa bivariate gamma distribution. Then, conditional on the mean levels, independent and identically distributed gamma variables produce the gene expression observations. Our hierarchical model implies a mean-variance relationship observed in many earlier studies.

For gene-specific inference, we take an empirical Bayes approach to estimate the mode of a posterior distribution which we obtain in closed form. We conclude with an application to a well-studied data set.

This is joint work with Sang-Hoon Cho.

Jongyun Jung
Department of Mathematics and Statistics
Minnesota State University, Mankato
USA

Understanding Cryptocurrency Price Formation from Time Series of Local Blockchain Graph Features

Over the last couple of years, digital cryptocurrencies and the Blockchain technology that forms their basis have witnessed a flood of attention. With the emergence and rapid adoption of Blockchain and the associated cryptocurrencies understanding the network dynamics behind Blockchain technologies has emerged as an important research direction. Unlike other financial networks such as stock and currency trading, blockchains have the entire time series of interaction graph accessible to the public. This facilitates a thorough analysis of the network data with a time series approach. A natural question to ask is whether the network dynamics of a cryptocurrency impact its price in dollars. We show that on the one hand, time series of standard global graph features such as degree distribution are not enough to capture the network dynamics that impact the underlying cryptocurrency price. In contrast, multiple time series of persistent topological homologies can explain higher level interactions among nodes in Blockchain graphs and can be used to build more accurate price prediction models.

Jana Jureckova
Charles University and The Czech Academy of Sciences
Czech Republic

Testing the Tail Index in AR Model and Average AR Quantile

The talk is partially based on the joint results with Hira L. Koul. We consider the linear autoregressive model of order p with possibly heavy tailed innovations. The AR quantiles and the dual AR rank scores were introduced by Koul and Saleh in 1995 following the regression quantiles and rank scores of the linear regression model. Specifically, the averaged AR alpha-quantile follows the tail behavior of the innovations, it is monotone in alpha, and it enables inference on distribution properties of the innovations. Being solution to a specific linear programming problem, under every number of observations, it is a linear combination of a finite number (p) of observations corresponding to the optimal base. The extreme averaged AR quantile is of a particular interest; besides being a linear combination of p basic observations, it can be expressed by means of (rank) R-estimate of the AR coefficients. A frequent problem of interest is test of hypothesis on the tail index of the innovations. An asymptotic nonparametric test, based on the empirical process of maxima of segments of the time series, was constructed by Koul and coauthors in 2009. Another test can be likely based also on the averaged AR quantiles.

Rejaul Karim, Taps Maiti
Department of Statistics and Probability
Michigan State University
USA

Rongrong Wang
Department of Computational Mathematics, Science and Engineering
Michigan State University
USA

Analysis of support tensor machines in high dimensional data

One of the most popular classification algorithms is the support vector machines (SVM) [Vapnik, 2013]. The classifier is sparse function of training data since it aims to minimize Vapnik dimensions. Analogously, Support Tensor Classifier (STM) [Tao et al., 2007] can be constructed for tensor inputs. Algorithmic properties of low rank representation of support tensor is studied. The method is tested on real MRI data.

Kun Ho Kim
College of Economics and Finance
Hanyang University
South Korea

Forward Premium Anomaly: A New Insight through Time-varying Parameter Approach

This paper employs uniform confidence bounds to investigate the forward premium anomaly, where spot currency returns are generally found to be negatively correlated with lagged interest rate differentials, or the forward premium. This violates uncovered interest rate parity which implies a positive coefficient of unity. The results here indicate remarkable variation in the time periods where the anomaly occurs and also considerable similarity in co-movements across currencies. This paper also investigates reasons for the failure of uncovered interest parity and finds that the standard fundamentals associated with the monetary model and also variables associated with time dependent risk premium contain a lot of predictive power in explaining the extent and degree of the anomaly. This is a joint work with Richard T. Baillie.

Soumendra N. Lahiri
Statistics Department
North Carolina State University
USA

On limit horizons in high dimensional inference

We consider a common situation arising in many high dimensional statistical inference problems where the dimension d diverges with the sample size n and the statistic of interest is given by a function of component-wise summary statistics. The limit distribution of the statistic of interest is often influenced by an intricate interplay of underlying dependence structure of the component-wise summary statistics. Here, we introduce a new concept, called limit horizon (L.H.) that gives the boundary of the growth rate of d as a function of n where the natural approach to deriving the limit law by iterated limits works. Further, for d growing at a faster rate beyond the L.H., the natural approach breaks down. We investigate the L.H. in some specific high dimensional problems.

Peide Li
Department of Statistics and Probability
Michigan State University
USA

Tensor base fMRI data classification

Functional magnetic resonance imaging studies human brain activities, which can be used to identify human physical status. In this paper, we use tensor to represent fMRI image data to preserve its spatial-temporal information, and select features from tensor with different methods. We further apply tensor classification methods to fMRI data and compare the classification accuracy as well as computation cost. It turns out that tensor discriminant analysis has a better performance and low computation cost. Some issues about tensor classification problems are discussed at the end of paper.

Yi Li
Biostatistics Department
University of Michigan
USA

Multiclass Linear Discriminant Analysis with Ultrahigh-Dimensional Features

Within the framework of Fisher's discriminant analysis, we propose a multiclass classification method, which embeds variable screening for ultrahigh-dimensional predictors. Leveraging inter-feature correlations, we show that the proposed linear classifier recovers informative features with probability tending to one and can asymptotically achieve a zero misclassification rate. We evaluate the finite sample performance of the method via extensive simulations and use this method to classify post-transplantation rejection types based on patients' gene expressions.

Jinghang Lin
Department of Statistics and Probability
Michigan State University
USA

Marcinkiewicz's strong law of large numbers for nonlinear expectations

The sub-linear expectation space is a nonlinear expectation space having advantages of modeling the uncertainty of probability and distribution. In the sub-linear expectation space, we use capacity and sub-linear expectation to replace probability and expectation of classical probability theory. In this paper, the method of selecting subsequence is used to prove Marcinkiewicz's strong law of large numbers under sub-linear expectation space. This result is a natural extension of the classical Marcinkiewicz's strong law of large numbers to the case where the expectation is nonlinear. In addition, this paper also gives a theorem about convergence of a random series.

Anna Little
Department of Computational Mathematics Science and Engineering
Michigan State University
USA

Path-Based Spectral Clustering: Guarantees, Robustness to Outliers, and Fast Algorithms

This talk will discuss new performance guarantees for robust path-based spectral clustering and an efficient approximation algorithm for the longest leg path distance (LLPD) metric, which is based on a sequence of multiscale adjacency graphs. LLPD-based clustering is informative for highly elongated and irregularly shaped clusters, and we prove finite-sample guarantees on its performance when random samples are drawn from multiple intrinsically low-dimensional clusters in high-dimensional space, in the presence of a large number of high-dimensional outliers. More specifically, we derive a condition under which the Laplacian eigengap statistic correctly determines the number of clusters for a large class of data sets, and prove guarantees on the number of points mislabeled by our method. Our methods are quite general and provide performance guarantees for spectral clustering with any ultrametric. We also propose a fast algorithm for implementing path-based spectral clustering which has complexity quasilinear in the number of data points.

Regina Liu
Department of Statistics
Rutgers University
USA

Prediction with Confidence – a General Framework for Predictive Inference

We propose a general framework for prediction in which a prediction is in the form of a distribution function called "predictive distribution function." This predictive distribution function is well suited to prescribing the notion of confidence under the frequentist interpretation, and it can provide meaningful answers for prediction-related questions. A general approach under this framework is formulated and illustrated using the so-called confidence distributions (CDs). This CD-based prediction approach inherits many desirable properties of CD, including its capacity to serve as a common platform for directly connecting the existing procedures of predictive inference in Bayesian, fiducial, and frequentist paradigms. We discuss the theory underlying the CD-based predictive distribution and related efficiency and optimality issues. We also propose a simple yet broadly applicable Monte-Carlo algorithm for implementing the proposed approach. This concrete algorithm together with the proposed definition and associate theoretical development provide a comprehensive statistical inference framework for prediction. Finally, the approach is demonstrated by simulation studies and a real project on predicting the volume of application submissions to a government agency. The latter shows the applicability of the proposed approach to dependent data settings.

This is joint work with Jieli Shen, Minge Xie, and Goldman Sachs.

Wenjian Liu
Mathematics and Computer Science
City University of New York
USA

DNA Reconstruction of K80 Evolution Mode

BACKGROUND: Determining the reconstruction threshold of broadcast models on d-ary regular tree, as the interdisciplinary subject, has attracted more and more attention from probabilists, statistical physicists, biologists, etc. OBJECTIVE: Consider a (2q)-state symmetric transition matrix as the noisy communication channel on each edge of a regular d-ary tree. Suppose there are two categories, one of which contains exactly q states, and 3 transition probabilities: remain in the same state, mutate to other state but remain in the same category, and mutate to the other category. Consider all the symbols received at the vertices of the n-th generation. Does this leaves-configuration contain a non-vanishing information on the letter transmitted by the root, as n goes to infinity? By means of a refined analysis of moment recursion on a weighted version of the magnetization, concentration investigation, and large degree asymptotics, we construct a nonlinear second-order dynamical system and show that the Kesten–Stigum reconstruction bound is not tight when q>=4. On the other side, when q=2, that is, Kimura Model of DNA evolution, the interactions of nodes on tree become weaker as d increases. This allows us to utilize the Gaussian approximation. Therefore, we explore stability of the fixed point of Gaussian approximation function in order to verify the tightness of Kesten-Stigum reconstruction bound.

APPLICATION: The reconstruction problem is concerned essentially with a tradeoff between noise and duplication in a tree communication network; phylogenetic reconstruction is a major task of systematic biology; reconstruction thresholds on trees are believed to determine the dynamic phase transitions in many constraint satisfaction problems including random K-SAT and random colorings on random graphs; the reconstruction threshold is also believed to play an important role in the efficiency of the Glauber dynamics on trees and random graphs.

Zhixin Lun
Department of Math
Oakland University
USA

Regression Imputation for Skewed Multivariate Data using Copula Transformation

Missing data is a common phenomenon in various data analyses. Imputation is viewed as a flexible method for handling missing-data problems since it efficiently uses all the available information in the rest of the data. Effectively, we predict the missing values from observed data and therefore, the performance of prediction plays a critical role in the imputation methodology. Most of the methodology available for missing data imputation revolves around data, which are assumed to be distributed as multivariate normal, and thus, when the data are skewed, these methods may not be very effective. To deal with data which may have non-normal distribution, we introduce an approach based on Copula transformation which was recently introduced by Bahuguna and Khattree (2018, A Generic All Purpose Transformation for Multivariate Modeling through Copulas, Preprint). We demonstrate that under mild assumptions, the copula transformation can be successfully used to impute the skewed multivariate data. In this talk, we confine to regression methods for imputation under the assumption of missing at random (MCAR) and through extensive simulations with various probability densities and different correlation structures, study and compare the performance of our approach and the one when multivariate normality is (incorrectly) assumed. Based on the simulations, we demonstrate that this new approach performs considerably better for the imputation of missing values in terms of smaller average sum squares of residuals. Further, percent of times our approach gives smaller sum of squares of residual is almost always and considerably more than 50 percent.

Selma Meradji
LaPS Laboratory
Badji Mokhtar University Annaba
Algeria

Stochastic differential equations for eigenvalues and eigenvectors of a G-Wishart process with drift

The aim of this paper is to give a system of G-SDEs for the eigenvalues and the eigenvectors of the G-Wishart process, defined from a G-Brownian motion matrix as in the classical case. Since we have not necessarily the independence between the entries of the G-Brownian motion matrix, we assume in our model that their quadratic co-variations are zero. An intermediate result, which states that the eigenvalues never collide, was also obtained. This extends Bru's results obtained for the classical Wishart process (1989).

Uschi (Ursula U, Mueller)
Department of Statistics
Texas A&M University
USA

Residual-based inference for semiparametric models

In this talk, I will review some joint research with Hira Koul, Anton Schick and Wolfgang Wefelmeyer on estimating the error distribution in nonparametric and semiparametric regression, with emphasis on regression models with independent errors and covariates. We will identify various regression models where the residual-based empirical distribution function allows a simple uniform expansion, which, in particular, characterizes an efficient estimator of the error distribution function. The expansion also provides the basis for constructing goodness-of-fit tests, for example, distribution free martingale-transform tests about the form of the error distribution. I will further explain how the results can be adapted to missing data scenarios. The derivation uses the "transfer principle" for obtaining limiting distributions of complete case statistics (for general missing data models) from corresponding results in the complete data model. To conclude, I will present some related research on estimating the error distribution function in single-index regression.

Léo Neufcourt
Department of Statistics and Probability
Michigan State University
USA

Bayesian approach to model-based extrapolation of nuclear observables

Quantifying the mass, or nuclear binding energy, of atomic nuclei is fundamental for understanding the origin of elements in the universe. The astrophysical processes responsible for the nucleosynthesis in stars often take place far from the valley of stability, where experimental masses are not known. Taking advantage of the information contained in mass model residuals where the experimental information exists, we utilize Bayesian machine learning techniques to provide the missing nuclear information using extreme extrapolations of theoretical predictions. Our methodology is developed on the two-neutron separation energies S_{2n} of even-even nuclei, where we consider 10 global models. Quantified emulators of S_{2n} residuals are constructed using Bayesian Gaussian processes and Bayesian neural networks of which we assess respectively predictive power and honesty of credibility intervals with rms deviation and empirical coverage probability. We consider the AME2003 dataset as our training dataset with a testing dataset at its external boundary composed of all ulterior measurements. While both statistical models reduce the rms deviation from experiment significantly, GP offers a better and much more stable performance. After statistical corrections all models display similar rms deviations on the testing dataset. This methodology is applied to the one- and two-neutron separations energies of the nuclei in the region of heavy calcium isotopes, which forms the frontier of experimental and theoretical nuclear structure research. The recent discovery of the extremely neutron-rich nuclei around ^{60}Ca and the experimental determination of masses for $^{55-57}$Ca provide unique information about the binding energy surface in this region. To assess the impact of these recent discoveries on the nuclear landscape, we compute the posterior probability for nuclides between Si and Ti to be bound to neutron emission. We find that extrapolations for drip-line locations are consistent across the global mass models used, in spite of significant variations between their raw predictions. In particular we predict that ^{68}Ca has an average posterior probability $p_{ex} \approx 76\%$ to be bound to two-neutron emission while ^{70}Ca is a threshold system with $p_{ex} \approx 57\%$. The nucleus ^{61}Ca is expected to decay by emitting a neutron ($p_{ex} \approx 46\%$).

Severien Nkurunziza
Department of Mathematics and Statistics
The University of Windsor
Canada

Robust inference in generalized Ornstein-Uhlenbeck processes with multiple change-points

In this talk, we present improved inference methods in generalized Ornstein-Uhlenbeck processes with multiple unknown change-points when the drift parameter satisfies uncertain constraint. A Salient feature of this investigation consists in the fact that the number of change-points and the locations of the change-points are unknown. We generalize some recent findings in five ways. First, our inference method incorporates the uncertain prior knowledge. Second, we derive the unrestricted estimator (UE) and the restricted estimator (RE) and we derive their asymptotic properties. Third, we derive a test for testing the hypothesized restriction and we derive its asymptotic power. Fourth, we propose a class of shrinkage estimators (SEs) which includes as special cases the UE, RE, and classical SEs. Fifth, we study the relative risk dominance of the proposed estimators, and we establish that SEs dominate the UE. The novelty of the established results consists in the fact that the dimensions of the proposed estimators are random. Because of that, the asymptotic power of the proposed test and the asymptotic risk analysis do not follow from the results in statistical literature

Michael Perlmutter
Department of Mathematics
Michigan State University
USA

Gabor Scattering Moments for Sparse Signals and Poisson Processes

We present a unified machine learning model for sparse signal analysis in both the deterministic and the statistical setting. Like the wavelet scattering transform introduced by S. Mallat, our construction is a mathematical model of Convolutional Neural Networks and is naturally invariant to translations and reflections. Our model replaces wavelets with Gabor type measurements and decouples the roles of scale and frequency. In the deterministic setting, this will allow us to establish a compressive-sensing type result where, under mild assumptions, we can completely recover a sparse signal (up to translations and reflections) with a sufficient number of measurements. In the statistical setting, we will assume that our sparse signal can be modeled as compound Poisson noise and show that our measurements allow us to estimate the Poisson arrival rate λ as well as the first and second moments of the arrival values $A_i.$

Scott Pratt
Department of Physics and Astronomy, NSCL/FRIB
Michigan State University
USA

Big Data Meets Big Models

Increasingly, sophisticated numerically intensive simulations are confronting large-scale heterogeneous data sets. Often, the goal is to determine model parameters from the model/data comparison, and to quantitatively, and rigorously, express the constraint. In this talk we show how model emulators were employed to analyze data from the nuclear collisions of relativistic heavy ions, with the aim of determining fundamental properties of the quark-gluon plasma.

Sudesh Pundir
Preventive Medicine (Health and Biomedical Informatics)
Northwestern University
USA

Reducing the Batch and Platform Effects in Trancriptome Data Analysis

High throughput technologies, such as microarrays and massive parallel sequencing, have brought new challenges for gene expression or transcriptome data analysis. While numerous datasets for both disease and normal tissues (or cells) are publicly available, traditional statistical methods usually fail in integrating and reproducing the results performed across different datasets. Two major issues; platform (microarray and sequencing) and batch differences; if not accounted during the statistical analysis might lead to misleading results. In addition, the high dimensionality, complexity and sparsity of data pose additional problems for data integration. We will present an overview of the current statistical methods, comparative evaluation of those methods on two The Cancer Genome Atlas transcriptome datasets, and discuss the need for development of new statistical methods.

Lianfen Qian
Department of Mathematical Sciences
Florida Atlantic University
USA

Estimating parameters of a frailty semi-competing model with measurement errors in covariates

In lifetime data analysis, it is common to observe multiple endpoints of risks. In this paper, we consider a shared frailty semi-competing model with measurement errors in covariates for cluster data with two semi-competing risks. Under the assumptions of shared Gamma frailty within each cluster and Weibull baseline hazards, we propose a corrected maximum likelihood estimation for covariate effects and Bayes estimation for the frailties. We derive the theoretical formulas for EM algorithm which is utilized for numerical optimization. To evaluate the finite sample performance of this method, we conduct the simulation studies which show that the proposed method works better than the Bayes estimation with MCMC algorithm. Moreover, the proposed method is robust to model mis-specification in terms of with or without measurement errors. For illustration purpose, we apply the proposed method to the monoclonal gammopathy of undetermined significance data. The results show that age is significant for all three baseline hazards, while the size of the monoclonal protein spike at diagnosis is significant only for the hazard from healthy to plasma cell malignancy. This is joint work with Caiya Zhang and Xiaolu Gu.

Annie Qu
Department of Statistics
University of Illinois at Urbana-Champaign
USA

Multilayer Tensor Factorization With Applications to Recommender Systems

Recommender systems have been widely adopted by electronic commerce and entertainment industries for individualized prediction and recommendation, which benefit consumers and improve business intelligence. In this article, we propose an innovative method, namely the recommendation engine of multilayers (REM), for tensor recommender systems. The proposed method utilizes the structure of a tensor response to integrate information from multiple modes, and creates an additional layer of nested latent factors to accommodate between-subjects dependency. One major advantage is that the proposed method is able to address the "cold-start" issue in the absence of information from new customers, new products or new contexts. Specifically, it provides more effective recommendations through sub-group information. To achieve scalable computation, we develop a new algorithm for the proposed method, which incorporates a maximum block improvement strategy into the cyclic block-wise-coordinate-descent algorithm. In theory, we investigate both algorithmic properties for global and local convergence, along with the asymptotic consistency of estimated parameters. Finally, the proposed method is applied in simulations and IRI marketing data with 116 million observations of product sales. Numerical studies demonstrate that the proposed method outperforms existing competitors in the literature. This is joint work with Xuan Bi and Xiaotong Shen.

Mark Reimers
Neuroscience Program
Michigan State University
USA

Statistical issues arising in emerging modalities of dynamic brain imaging

As big data in genomics has stimulated many new statistical endeavors, the emergence of optical imaging methods, generating terabytes of dynamic high-resolution brain activity data, is leading to many new statistical problems. There are the usual problems of an emerging technology, to do with noise reduction and artifact removal. However, these rich data pose many more scientifically interesting statistical problems addressing theories of brain dynamics, and bridging dynamical systems ideas and high-dimensional statistical inference. This talk will focus on some of these more scientifically interesting problems, and discuss current attempts to address them.

Farzad Sabzikar
Department of Statistics
Iowa State University
USA

Asymptotic Theory for Near Integrated Processes Driven by Tempered Linear Processes

In this talk, we establish asymptotic theory for near-integrated random processes and associated regressions including the score function in more general settings where the errors are tempered linear processes. Tempered processes are stationary time series that have a semi-long memory property in the sense that the autocovariogram of the process resembles that of a long memory model for moderate lags but eventually diminishes exponentially fast according to the presence of a decay factor governed by a tempering parameter. When the tempering parameter is sample size dependent, the resulting class of processes admits a wide range of behavior that includes both long memory, semi-long memory, and short memory processes. The limit results relate to tempered fractional processes that include tempered fractional Brownian motion and tempered fractional diffusion process of the second kind.

Alexander Sakhanenko
Novosibirsk State University
Sobolev Institute of Mathematics
Novosibirsk, Russia

On accuracy of approximation in Koul's Theorem for weighted empirical processes

Using coupling we obtain estimates for the distribution of the uniform distance between a given weighted empirical process and an accompanying gaussian process, with the same mean and covariance function. For such weighted empirical processes in the book, Hira L. Koul “Weighted Empirical Processes in Dynamic Nonlinear Models” was only established weak convergence of distributions. However, our estimates do not require the continuity of the limiting Gaussian process. They are also valid for the uniform norm. This is a joint work with O.A. Sukhovershina.

Lyudmila Sakhanenko
Department of Statistics & Probability
Michigan State University
USA

Integral Curve Estimation for High Angular Resolution Diffusion Imaging

High Angular Resolution Diffusion Imaging is a vivo brain imaging technique that allows to understand axonal anatomy. However, the images have a notoriously high level of noise. We model the uncertainty in images via a super-tensor model, where the components of a diffusion super-tensor are the slopes in a system of regression equations that are measured by a MRI scanner. Then we study how the uncertainty propagates from the tensor field to a vector field to the integral curves, which serve as the models for axonal fibers. We construct the estimators of the fibers, show their asymptotical normality, and develop computationally fast tractography approach. As a result, brain images are enhanced via confidence tubes enveloping the estimated fibers that quantify and picture the uncertainty present in the data. The location of fibers and their connectivity are important to neuroscientists, since aging and some diseases such as Alzheimer's change how the brain regions connect to each other.

Anton Schick
Department of Mathematical Sciences
Binghamton University
USA

Estimation of the error distribution function in a varying coefficient regression model

This talk discusses estimation of the error distribution function in a varying coefficient regression model. Three estimators are introduced and their asymptotic properties described by uniform stochastic expansions.

The first estimator is a residual-based empirical distribution function utilizing an under-smoothed local quadratic smoother of the coefficient function. The second estimator exploits the fact that the error distribution has mean zero. It improves on the first estimator, but is not yet efficient. An efficient estimator is obtained by adding a stochastic correction term to the second estimator.

Ralf Schmälzle
Department of Communication
Michigan State University
USA

Statistical applications in neuroscience: A view from Neurocognitive Communication at MSU

Neurocognitive Communication, a new focus area of Michigan State's College of Communication Arts and Sciences, examines the neural underpinnings of human communication and social interaction. The goal of this talk is to inspire statisticians to see application potential in this field and to stimulate discussion about similarities and differences between classical "cognitive neuroscience" statistics, which has been developed over the past decades. I will showcase work that measures common brain dynamics between people (e.g. during movie-watching or during dyadic interaction), work that links neural- and social-level variables (e.g. SES, social network metrics), and work that uses brain activity to predict outcomes (e.g. predict which messages will be persuasive or get shared on social media). Finally, I will briefly touch on the field's emerging computational infrastructure, reproducibility, and big-data/data-sharing initiatives.

Peter Schmidt
Department of Economics
Michigan State University
USA

A New Family of Copulas, with Application to Estimation of a Production Frontier System

In this talk we propose a new family of copulas for which the copula arguments are uncorrelated but dependent. Specifically, if w_1 and w_2 are the uniform random variables in the copula, they are uncorrelated, but w_1 is correlated with |w_2 - ½|. We show how this family of copulas can be applied to the error structure in an econometric production frontier model. We also generalize the family of copulas to three or more dimensions, and we give an empirical application. This is joint work with Christine Amsler and Artem B. Prokhorov.

Leila Setayeshgar
Department of Mathematics & Computer Science
Providence College
USA

Large Deviations for a Class of Stochastic Semi-linear Partial Differential Equations

Standard approaches to large deviations analysis for stochastic partial differential equations (SPDEs) are often based on approximations. These approximations are mostly technical and often onerous to carry out. In 2008, Budhiraja, Dupuis and Maroulas, employed the weak convergence approach and showed that these approximations can be avoided for many infinite dimensional models. Large deviations analysis for such systems instead relied on demonstrating existence, uniqueness and tightness properties of certain perturbations of the original process. In this talk, we use the weak convergence approach, and establish the large deviation principle for the law of the solutions to a class of semi-linear SPDEs. Our family of semi-linear SPDEs contains, as special cases, both the stochastic Burgers' equation, and the stochastic reaction-diffusion equation.

Anuj Srivastava
Department of Statistics
Florida State University
USA

Shapes Analysis of Functional Data

Functional data has a growing presence in all branches of science and engineering, partly due to tremendous advances made in data collection and storage technologies. Such data is mostly analyzed using the classical Hilbert structure of square-integrable function spaces, but that setup ignores the shapes of these functions. Shape implies the ordering and the heights of peaks and valleys but is flexible on their exact locations. To focus on shapes of functions, we have introduced Elastic functional data analysis that allows time warpings of functions in order to register functional data, i.e. match their peaks and valleys. This, in turn, requires elastic Riemannian metrics that enable comparisons and testing of shape data modulo warping group action. I will present some statistical procedures resulting from their framework, including estimation of shape-constrained densities, ANOVA on shape space of curves, shape estimation and analysis of large biomolecules, and shape analysis of brain anatomical structures.

Donatas Surgailis
Institute of Mathematics and Informatics
Vilnius University
Lithuania

Testing for Long Memory in Random-coefficient AR(1) Panel

It is well-known since Granger (1980) that random-coefficient AR(1) process can have long memory covariance, if the tail distribution function of the random coefficient regularly varies at the unit root with exponent $\beta \in (1,2)$. This talk discusses statistical inference for N x T panel consisting of N independent RCAR(1) series, each of length T, as N and T jointly increase possibly at a different rate. In the first part, we discuss the asymptotic distribution of the sample mean and the sample variance for the above panel and show that these distributions crucially depend on the mutual increase rate of N, T, with the critical rate $T \sim N^{\beta}$ separating different limit regimes. The second part deals with nonparametric estimation of β for the same panel. The estimator $\tilde{\beta}_N$ is constructed as a version of the tail index estimator of Goldie and Smith (1987) applied to the serial correlation coefficients of each of N rows. The asymptotic normality of $\tilde{\beta}_N$ is obtained under certain conditions on N, T, β and some other quantities of our statistical model. Based on this result, we construct a statistical test to test the null hypothesis $\beta \geq 2$ against the alternative $\beta < 2$, or that the RCAR(1) panel data exhibit long memory. A simulation study illustrates finite-sample performance of the introduced estimator and testing procedure.

This is joint work with Remigijus Leipus, Anne Philippe, and Vytaute Pilipauskaite

Frederi Viens
Department of Statistics & Probability
Michigan State University
USA

Parameter estimation in long-memory and other Gaussian processes

We consider the class of all stationary Gaussian processes. When the spectral density is parametrically explicit, we defined a Generalized Method of Moments estimator that satisfies consistency and asymptotic normality, using the Breuer-Major theorem which applies to long-memory processes. This result is applied to the joint estimation of the three parameters of a stationary fractional Ornstein-Uhlenbeck (fOU) process driven for all Hurst parameters. For general processes observed at fixed discrete times, no matter what the memory length, we use state-of-the-art Malliavin calculus tools to prove Berry-Esseen-type and other speeds of convergence in total variation, for estimators based on power variations. This is joint work with Luis Barboza (U. Costa Rica), Khalifa es-Sebaiy (U. Kuwait), and Soukaina Douissi (U. Cadi Ayyad, Morocco).

Rongrong Wang
Department of Computational Mathematics, Science and Engineering
Department of Mathematics
Michigan State University
USA

A simple nonlinear dimension reduction technique for high dimension data visualization

Over the past two decades, many nonlinear dimension reduction techniques are developed to address some of the limitations of linear dimension reduction techniques. However, every nonlinear DR technique has their own limitation. For example, LLE is known as best in preserving local structures, but is very unstable to outliers and sensitive to the number of k nearest neighbors. tSNE is excellent in data clustering, but cannot preserve the geometry of the high dimensional data. Isomap is good at preserving the geodesic distances but suffers from topological instability. In this talk, we propose a simple algorithm that has the advantages of both LLE and tSNE, i.e., it preserves both the locally linear structure and the clusters in the dataset.

Ali Zare
Department of Computer Science
Michigan State University
USA

Tensor Based Feature Extraction Methods for MRI Data Classification

In this work, the performance of various dimensionlity reduction and feature extraction techniques applied to MRI data has been investigated. The performance assessment is done based on binary classification results as well as computational aspects of a method, such as run time and memory usage. These methods include Multilinear Principal Component Analysis (MPCA) and Matrix Product State (MPS). Other methods such as random subsampling, random projection and random FFT have also been tried for classification. In some cases, certain combinations of these methods have also been used, e.g., MPCA/MPS applied after random subsampling of the data. The main method used in this work for classification is 1-Nearest Neighbor (1-NN) applied to features, which can also follow the Linear Discriminant Analysis (LDA) projection of the features. The Alzheimer's disease Neuroimaging Initiative (ADNI) data have been used in the experiments, and results have been reported for various levels of dimension reduction, subsampling ratios, etc., for different data sizes.

Caiya Zhang
Department of Statistics
Zhejiang University City College
China

Asymptotic Properties of the QMLE in a Log-linear RealGARCH Model with Gaussian Errors

To incorporate the realized volatility in stock return, Hansen et al. (2012) proposed a RealGARCH model and conjectured some theoretical properties about the quasi-maximum likelihood estimation (QMLE) for parameters in a log-linear RealGARCH model without rigorous proof. Under Gaussian errors, we derives the detailed proof of the theoretical results including consistency and asymptotic normality of the QMLE, hence it solves the conjectures in Hansen et al. (2012).

Hongjuan Zhou
School of Mathematical and Statistical Sciences
Arizona State University
USA

Parameter estimation for fractional Ornstein-Uhlenbeck processes of general Hurst parameter

I will talk about several statistical estimators for the drift and volatility parameters of an Ornstein-Uhlenbeck process driven by fractional Brownian motion, whose observations can be made either continuously or at discrete time instants. Power variations are used to estimate the volatility parameter. The almost sure convergence of the estimators and the corresponding central limit theorems are obtained for all the Hurst parameter range $H \in (0,1)$. The least squares estimator is used for the drift parameter. A central limit theorem is proved when the Hurst parameter $H \in (0, \frac{3}{4}]$ and a noncentral limit theorem is proved for $H \in (\frac{3}{4}, 1)$. This is a joint work with Yaozhong Hu and David Nualart.

Quan Zhou
Department of Statistics
Rice University
USA

When is it best to follow the leader?

We study a classical continuous-time optimal scanning problem with N boxes. A missing object is hidden in one box according to a given prior probability distribution. The goal is to find out which box contains the object. When we search a box, we observe a signal evolving as a Brownian motion with a constant drift if the box contains the object, or as a Brownian motion with drift zero otherwise. This problem was first studied by Posner and Rumsey (1966), who conjectured that the optimal search strategy is to always observe the box with the largest posterior. However, we prove that this conjecture is not true in general and provide counterexamples for some specific prior distributions. But it remains open whether this strategy is optimal for the uniform prior distribution. Joint work with Philip Ernst and L.C.G. Rogers.

David C. Zhu
Department of Radiology
Michigan State University
USA

Statistical Challenges for Clinical Trials with Neuroimaging

Clinical trials with neuroimaging often collect data from various techniques of MRI (magnetic resonance imaging), various techniques of PET (positron emission tomography), genetics, cerebrospinal fluid and blood biomarkers, cognitive tests and clinical measurements. Ideally, all data are used to define the condition of a brain. However, these data are complex with multiple dimensions and multiple image modalities, and thus pose a true big data challenge. Researchers often focus on one data type or incorporate two or three modalities together to investigate the brain and how the brain function and structure change over time. In this talk, I will provide an overview of the data collected from the large multi-site ADNI (Alzheimer's Disease Neuroimaging Initiative, http://adni.loni.usc.edu) project and the multi-site rrAD (Risk Reduction for Alzheimer's Disease, http://www.rradtrial.org) project. The ADNI project collects brain data from older subjects (normal, mild cognitive impairment (MCI) and Alzheimer's disease (AD)) over multiple time points with intervals of months and years to observe the brain modification. The rrAD project is an intervention study to understand whether aerobic exercise and intensive medical management of blood pressure and cholesterol can reduce the risk of AD in older adults who are at risk of AD. From an imaging researcher perspective, I will also introduce a method of incorporating image data collected from various modalities, along with genetics, cognitive tests and clinical data. Nevertheless, vigorous statistical integration of big data remains a challenge.