Plenary Talks

Bayesian Tensor Regression for Neuroimaging Data

Montserrat Fuentes, Virginia Commonwealth University

Imaging data with thousands of spatially-correlated data points are common in many fields. In Neurosciences, magnetic resonance imaging (MRI) is a primary modality for studying brain structure and activity. Modeling spatial dependence of MRI data at different scales is one of the main challenges of contemporary neuroimaging, and it could allow for accurate testing for significance in neural activity. The high dimensionality of this type of data (millions of voxels) presents modeling challenges and serious computational constraints. Methods that account for spatial correlation often require very cumbersome matrix evaluations which are prohibitive for data of this size, and thus current methods typically reduce dimensionality by modeling covariance among regions of interest, coarser or larger spatial units, rather than among voxels. However, ignoring spatial dependence at different scales could drastically reduce our ability to detect important activation patterns in the brain and hence produce misleading results. To overcome these problems, we introduce a novel Bayesian Tensor approach, treating the brain image as response and having a vector of predictors. Our method provides estimates of the parameters of interest using a generalized sparsity principle. This method is implemented using a fully Bayesian approach to characterize different sources of uncertainty. We demonstrate posterior consistency and develop a computational efficient algorithm. The effectiveness of our approach is illustrated through simulation studies and the analysis of the effects of cocaine addiction on the brain structure. We implement this method to identify the effects of demographic information and cocaine addiction on the functioning of the brain.

Goodness-of-fit Testing of Error Distribution in Linear Measurement Error Models

Hira L Koul, Michigan State University

We shall discuss the two types of goodness-of-fit tests for fitting a response error distribution in linear regression models when covariates are observed with additive measurement errors. We consider the Berkson measurement errors and the classical errors-in-variables models. The two types of tests considered are based on certain residual empirical processes and on direct and deconvoluted regression error density estimates. In the former case we first establish the asymptotic uniform linearity (AUL) of the residual empirical process in both models. The AUL results in turn are used to establish the asymptotic distribution free property of a Khmaladze type transformation of the residual empirical process in each model. The second class of tests are based on integrated square distances between a density estimate and its expectation under the null.
hypothesis. The two types of density estimates are considered. One is kernel type density estimate while the other is the deconvoluted density estimate, both based on direct residuals. All of these tests are asymptotically distribution free. But the tests based on residual empirical process have the classical $n^{1/2}$-consistency rate and while those based on density estimates have relatively much slower consistency rates.

Adaptive Designs for Confirmatory Trials

► Cyrus Mehta, Cytel Inc.

Adaptive designs have been proposed as a means to increase the efficiency of randomized clinical trials, potentially benefitting trial participants and future patients while reducing costs and enhancing the likelihood of finding a true benefit, if one exists of the therapy being studied. These designs can modify the future course of an on-going trial, by applying prospectively specified decision rules to the unblinded interim data, without undermining the validity of the final analysis. Typical design changes implemented in a confirmatory setting are sample size re-assessment, dropping of non-performing treatment arms in a multi-arm study, and changes in the inclusion/exclusion criteria with special reference to biomarker based population enrichment. I will present an example of each such design type, drawn from our consulting experience at Cytel. Type-1 error control, operational and regulatory considerations will be discussed.

Bahadur Lecture

Inference with High-Dimensional Parameters of Interest and Many Controls

► Alexandre Belloni, Duke University

In this work we propose and analyze procedures to construct confidence regions for $p$ (infinite dimensional) parameters of interest after model selection for general moment condition models where $p$ is potentially larger than the sample size $n$. This allows us to cover settings with functional response data where each of the $p > n$ parameters of interest is a function. The procedure is based on the construction of generalized score functions which are new moment functions with an additional orthogonality condition. The proposed uniform confidence bands for all parameters relies on uniform central limit theorems for high-dimensional vectors (and not on Donsker arguments as we allow for $p > n$). The construction of the bands are based on a multiplier bootstrap procedure which is computationally efficient as it only involves resampling the estimated score functions (and does not require resolving the high-dimensional optimization problems). We formally apply the general theory to inference on regression coefficient process in the distribution regression model with a logistic link, where two implementations are analyzed in detail. Simulations and an application to real data are provided to help illustrate the applicability of the results.

Jayanta K. Ghosh (JKG) Memorial Session

Statistical issues in Time Domain Astronomy

► Jogesh Babu, The Pennsylvania State University

Majority of data from gravitational wave detectors, exoplanet surveys, forthcoming data from Large Synoptic Survey Telescope and other projects surveying large portions of the visible sky is in the form of irregularly spaced time series. Brief review of statistical issues in time-domain astronomy will be presented.

Challenges in Dealing with Multiplicities

► James Berger, Duke University

JKG was fascinated by, and contributed enormously, to the interplay of Bayesian and frequentist analysis. One of the many areas in which he did this was that of multiple testing. We briefly review some of his work in the area, and then move on to discuss some of the remaining mysteries. In the frequentist approach
to handling of multiplicities, one develops a procedure that, e.g., controls for the false positive rate. In the Bayesian approach, one ignores certain multiplicities and deals with others only through the assignment of prior probabilities. These apparently conflict with frequentist analysis, so the challenge is to find common ground. We begin by reviewing two situations in which common ground is thought to be not possible: sequential (or interim) analysis and sequential endpoint testing, where the multiplicities arise through multiple looks at the data and through having multiple decision endpoints, respectively. While the first is a resolvable BFF problem, the second does not appear to be so. We then turn to two multiple testing scenarios where agreement is possible, testing mutually exclusive hypotheses and simultaneous multiple testing.

Fast nonparametric estimation of a smooth mixing density

► Ryan Martin, North Carolina State University

In a mixture model setting, estimation of the mixing distribution is a challenging problem. Typical estimators, such as the nonparametric MLE, are discrete, but what can be done if the mixing distribution is known to have a smooth density? In the first half of this talk, I will describe the “predictive recursion” (PR) algorithm designed for fast nonparametric estimation of a smooth mixing density, present its theoretical convergence properties, and briefly discuss extensions, applications, and open questions. In some recent work, certain insights about PR’s general structure have proved useful for understanding seemingly unrelated algorithms and for designing new ones. In the second half of this talk, I will present some of these new details. (The PR algorithm was the topic of my PhD dissertation under Professor JK Ghosh’s guidance, so he played a significant role, directly or indirectly, in all of the work presented here.)

Special Invited Sessions

Bayesian method for causal inference in spatially-correlated multivariate time series

► Subhashis Ghoshal, North Carolina State University

Measuring the causal impact of an advertising campaign on sales is important for advertising companies. The growing uncertainty of prediction with time lag often obscures potential signals in a time series model preventing detection of causal impact. We propose a novel Bayesian method to infer causality which allows us to detect even weaker impacts. We compare two posterior distributions of a latent variable—one obtained by using the observed data from the test stores and the other one obtained by using the data from their counterfactual potential outcomes. The counterfactual potential outcomes are obtained from the data of synthetic controls given by a sparse linear combination of sale figures at many control stores over the causal period. We use a multivariate structural time series model to capture the spatial correlation between test stores. Sparsity is imposed on the precision matrix through a G-Wishart prior. Stationarity is imposed on the local linear trend of the model to prevent the prediction intervals from being explosive. A two-stage algorithm is proposed to estimate the parameters of the model. In Stage 1, a modified EMVS algorithm is applied to select control stores. In Stage 2, an MCMC algorithm is used to obtain the samples of the rest parameters. We present extensive simulation results to show the effectiveness of the proposed method. The new method is applied to measure the causal effect of an advertising campaign for a consumer product sold at stores of a large national retail chain.

Recent Progression in Machine Learning and Precision Medicine

► Michael Kosorok, University of North Carolina at Chapel Hill

Precision medicine, the paradigm of improving clinical care through data driven approaches to tailoring treatment to the individual, is an important area of statistical and biomedical research. Individualized treatment rules (ITR’s) formalize precision medicine as mappings from the space of patient covariates to the set of available treatments or, equivalently, as mappings which identify covariate-defined subgroups for which different treatments should be applied. ITR’s are thus an important tool to improve patient outcomes through utilizing biomarkers to target treatment. Machine learning has become an increasingly utilized and evolving method-
ology for ITR discovery, and we discuss recent progress in this area and present examples in type I diabetes and bipolar disorder.

**A Statistical Analysis of Crowdsourced Mobile Data**

**Soumendranath Lahiri**, North Carolina State University

Big data is ubiquitous in many areas of applications. While abundance of data provides opportunities for better inference, it also comes with unique challenges, including the issues of volume and veracity of the data that make the task of information extraction very difficult. In this talk, we will consider a crowd-sourced data set collected through an App on mobile devices that gives low quality, high volume, unstructured, noisy measurements on several variables including the ambient temperature. We introduce a way to associate veracity scores to a data value that says how reliable and useful the data value is. We then use this measure of veracity to develop statistical methods for the low quality data and show that it provides significant improvements over naive applications of the standard prediction methodology. We also explore ways of augmenting information available sources using the crowd-sourced data.

**Analysis of Stratified Clinical Trials: Binary and Time-to-Event Endpoints**

**Devan V. Mehrotra**, Merck Research Laboratories

Stratified randomization is commonly used in clinical trials when the endpoint of interest (infection cure, cancer progression, etc.) is expected to be influenced by one or more known prognostic factors (gender, age group, ethnicity, biomarker status, etc.). To help promote better statistical practice in such settings, I will discuss two topics with a focus on binary and time-to-event endpoints: (i) estimation and inference for the “overall” (i.e., averaged across all patients) treatment effect, and (ii) assessment of whether the treatment effect is consistent across strata. I will contrast different statistical approaches for addressing (i) and (ii) using examples from the infectious disease, oncology and vaccine therapeutic areas. I will also present simulation results to support key recommendations for the design and analysis of stratified clinical trials.

**An overview of VAR models and their extensions under high-dimensional scaling**

**George Michailidis**, University of Florida

Vector Autoregressive Models (VAR) are widely used in applied economics and finance, as well as in functional genomics and neuroscience. In this talk we provide an overview of recent work on estimating large scale VAR models under structured sparsity constraints. Further, we consider extensions of the standard VAR model that include exogenous variables, as well as latent factors. The models are illustrated on a number of data sets addressing different macroeconomics questions.

**Revisiting the Genomewide Threshold of $5 \times 10^{-8}$ in 2018**

**Bhramar Mukherjee**, University of Michigan

During the past two years, there has been much discussion and debate around the perverse use of the P-value threshold of 0.05 to declare statistical significance for single null hypothesis testing. A recent recommendation by many eminent statisticians is to redefine statistical significance at $p < 0.005$ [Benjamin et al, *Nature Human Behaviour*, 2017]. This new threshold is motivated by the use of Bayes Factors and true control of false positive report probability. In genomewide association studies, a much smaller threshold of $5 \times 10^{-8}$ has been used with notable success in yielding reproducible results while testing millions of genetic variants. I will first discuss the historic rationale for using this threshold. We will then investigate whether this threshold that was proposed about a decade ago needs to be revisited with the current genomewide data we have in terms of the newer sequencing platforms, imputation strategies, testing rare versus common variants, the existing knowledge we have gathered regarding true association signals, or for controlling other metrics associated with multiple hypotheses testing beyond the family wise error rate. I will discuss notions of
Bayesian error rates for multiple testing and use connections between the Bayes Factor and the Frequentist Factor (the ratio of power and Type 1 error) for declaring new discoveries. Empirical studies using data from the Global Lipids Consortium will be used to evaluate if we applied various thresholds/decision rules in 2008 or 2009, how many of the most recent GWAS results (in 2013) would we detect and what would be our “true” false discovery rate.

Science and Statistics: Analyzing Data from 16S rRNA Microbiome Experiments

Glen A Satten, Centers for Disease Control and Prevention

Data from a microbiome study is typically a $n \times p$ table of counts ($n = \text{sample size}, \ p = \text{number of taxa}$). Tempted by superficial similarities to RNA-Seq data, a number of potentially inappropriate methods have been applied to microbiome data to find ‘differentially expressed taxa.’ Ecologists have taken a different approach based on calculating a matrix of pairwise distances between samples, followed by ordination (plotting samples in 2 or 3 dimension using principal components); this community-based approach often finds striking differences between meaningful groups (e.g., cases vs controls). However, methods for testing hypotheses about the observed patterns or finding the taxa that contribute most are not well developed. Further, microbiome studies can have all the complexities of observational studies such as confounding covariates and clustered observations. To provide a single analysis path that includes distance-based ordination, global hypothesis tests of any effect of the microbiome, and hypotheses tests of the effects of individual taxa, and control for confounding covariates and complex study designs, we have developed the Linear Decomposition Model (LDM), a distance-based model that combines community-level and taxon-level analyses to assesses the importance of explanatory variables based on the proportion of variability in the data that they explain. Any distance matrix can be used. The variance explained is easily calculated, so that its significance can be established using permutation. Using simulation, we show that our approach can have more power to detect overall association than existing methods. If time allows, we may also discuss analyses of multi-omics data and analyses of hierarchical hypotheses implied by the phylogenetic relationships between taxa.

Statistical challenges in obstetrics: developing quantitative support for obstetrical care

Rajeshwari Sundaram, National Institutes of Health

The area of obstetrics presents exciting analytical challenges for statisticians. In this talk, we will focus on the problem of labor management. The talk will focus on whether the current American College of Obstetrics and Gynecology (ACOG) guidelines for clinical management of labor progression, mostly based on historical or expert consensus are supported by data and whether they can/need to be personalized to individual woman. Defining labor progression started with the landmark work of Friedman (1955), who developed so-called “graphicostatistical” analytical methods to analyze such data. This formed the basis of obstetrical practices for multiple decades. With the changing landscape of obstetrical population, i.e., advancing age of mother at first child birth and rising obesity rates and various advancements in medical field, the need for revisiting the patterns of labor progression considering these issues to better guide clinical practices has become critical. Our approach to assessing labor progression is by viewing it as an example of multivariate survival data: multistage models. We will discuss various aspects of clinical guidelines, including (i) how long does a woman take to traverse through first stage of labor till she gets fully dilated (ii) how long can a woman be safely allowed to push without considerably increasing her risk for maternal morbidities or neonatal morbidities (iii) can one predict when a woman’s labor will go on arrest, a leading cause for cesarean delivery. We will provide some insight into these issues based on methods involving panel count data with unobserved start time and informative examination times, joint modeling of multiple survival times subject to various type of incompleteness, as well as stochastic differential equations. These methods will be used to assess the ACOG guidelines mentioned above using the seminal studies designed for setting the standards in this field, the Collaborative Perinatal Project and the Consortium of Safe Labor.
Fréchet Analysis of Variance for Random Objects

► Paromita Dubey, University of California, Davis

Co-authors: Hans-Georg Mueller

Fréchet mean and variance provide a way of obtaining mean and variance for metric space valued random variables and can be used for statistical analysis of data objects that lie in abstract spaces devoid of algebraic structure and operations. Examples of such data include covariance matrices, graph Laplacians of networks and univariate probability distribution functions. We derive a central limit theorem for Frechet variance under mild regularity conditions, utilizing empirical process theory, and also provide a consistent estimator of the asymptotic variance. These results lead to a test for comparing $k$ populations of metric space valued data objects in terms of Fréchet means and variances. We examine the finite sample performance of this novel inference procedure through simulation studies for several special cases that include probability distributions and graph Laplacians, which leads to a test for comparing populations of networks. The proposed methodology has good finite sample performance in simulations for different kinds of random objects. We illustrate the proposed methods with data on mortality profiles of various countries and resting state Functional Magnetic Resonance Imaging data.

Generalized Sparse Additive Models

► Asad Haris, University of Washington

Co-authors: Noah Simon, Ali Shojaie

We present a unified framework for estimation of generalized additive models in high-dimensional regression problems. The framework defines a large class of penalized regression estimators, encompassing many existing methods. An efficient computational algorithm for this class is presented that easily scales to thousands of observations and features. Additionally, we prove minimax optimal convergence bounds on these estimators under a weak compatibility condition. In addition, we characterize the rate of convergence when this compatibility condition is not met. Finally, we also show that the optimal penalty parameters for our structure and sparsity penalties are linked, allowing cross-validation to be conducted over only a single tuning parameter. We complement this with empirical studies comparing some exiting methods.

Variable selection using pseudo-variables

► Wenhao Hu, North Carolina State University

Co-authors: Eric Laber, Leonard Stefanski

Penalized regression has become a standard tool for model building across a wide range of application domains. Common practice is to tune the amount of penalization to tradeoff bias and variance or to optimize some other measure of performance of the estimated model. An advantage of such automated model-building procedures is that their operating characteristics are well-defined, i.e., completely data-driven, and thereby they can be systematically studied. However, in many applications it is desirable to incorporate domain knowledge into the model building process; one way to do this is to characterize each model along the solution path of a penalized regression estimator in terms of an operating characteristic that is meaningful within a domain context and then to allow domain experts to choose from among these models using these operating characteristics as well as other factors not available to the estimation algorithm. We derive an estimator of the false selection rate for each model along the solution path using a novel variable addition method. The proposed estimator applies to both fixed and random designs and allows for $p \gg n$. The proposed estimator can be used to estimate a model with a pre-specified false selection rate or can be overlaid on the solution path to facilitate interactive model exploration. We characterize the asymptotic behavior of the proposed estimator in the case of a linear model under a fixed design; however, simulation experiments show that the proposed estimator provides consistently more accurate estimates of the false selection rate than competing methods.
across a wide range of models.

Spectral estimation for high-dimensional linear processes

► Jamshid Namdari, University of California Davis

Co-authors: Debashis Paul and Alexander Aue

We propose a novel estimation procedure for a class of high dimensional linear time series by means of estimating the joint eigenvalue distribution of the coefficient matrices of the process. The process being considered is of the form $X_t = \sum_{\ell=0}^{\infty} A_{\ell} Z_{t-\ell}$ where the innovations $\{Z_t\}$ are i.i.d., $p$-dimensional random vectors with zero mean, and the coefficient matrices $\{A_{\ell}\}$ and $\text{Var}(Z_t)$ are diagonalizable in a common orthonormal basis. The proposed estimators rely on the asymptotic behavior of weighted integrals of the sample periodogram, where throughout we work under the asymptotic regime where $p, n \to \infty$ such that $p/n \to c \in (0, \infty)$. Under this setting, we first establish a Marchenko-Pastur type limiting distribution for the aforementioned weighted sample periodograms. This result is expressed in terms of the convergence of the Stieltjes transforms of the respective empirical spectral distributions. We utilize this result to develop a class of estimators, by minimizing an $L^\kappa$ discrepancy measure (for $\kappa > 0$) between the empirical and limiting Stieltjes transforms, of the joint spectral distribution of the coefficient matrices of the linear process, by assuming that the latter is a discrete mixture of point masses. Furthermore, we prove consistency of the estimator corresponding to the $L^2$ discrepancy measure. We illustrate the methodology through extensive simulations and through an analysis of stock prices from the S&P 500 series.

Stationary Subspace Analysis of Nonstationary Processes

► Raanju Sundararajan, Texas A&M University

Co-authors: Mohsen Pourahmadi

Stationary subspace analysis (SSA) is a recent technique for finding linear transformations of nonstationary multivariate processes which are stationary in the limited sense that the first two moments or means and covariances are time-invariant. The key optimization problem is that of finding a matrix minimizing the Kullback-Leibler divergence between Gaussian distributions measuring the non-constancy of the means and covariances across several segments. We present a frequency domain alternative to SSA for general multivariate second-order nonstationary processes. Using the asymptotic uncorrelatedness of the discrete Fourier transform of a stationary time series, a measure of departure from stationarity is introduced and minimized to find the stationary subspace. The dimension of the subspace, the key parameter, is estimated using a sequential testing procedure and its asymptotic properties are studied. We illustrate the broader applicability and better performance of the frequency domain method in comparison to time domain SSA methods through simulation examples. We apply our method to filter out noise in EEG brain signals from an economic choice task experiment. This improves prediction performance and more importantly reduces the number of trials needed from individuals in neuroeconomic experiments thereby aligning with the principle of simple and controlled designs in experimental and behavioral economics.

Student Paper Competition: Applications

Temporal Exponential-Family Random Graph Models with Time-Evolving Latent Block Structure for Dynamic Networks

► Amal Agarwal, Pennsylvania State University

Co-authors: Kevin Lee, Lingzhou Xue

Model-based clustering of dynamic networks has emerged as an essential research topic in statistical network analysis. It is critical to effectively and efficiently model the time-evolving latent block structure of dynamic networks in practice. However, the focus of most existing methods is on the static or dynamically invariant block structure. We present a principled statistical clustering of large-scale dynamic networks through the dy-
Dynamic exponential-family random graph models with a hidden Markov structure. The hidden Markov structure is used to infer the time-evolving block structure of dynamic networks. We prove the identification conditions for both network parameters and transition matrix in our proposed model-based clustering. We propose an effective model selection criterion based on the integrated classification likelihood to choosing an appropriate number of clusters. We develop a variational expectation-maximization algorithm to solve the approximate maximum likelihood estimate. The numerical performance of our proposed method is demonstrated in simulation studies and real data applications to dynamic international trade networks and dynamic email networks of a large institute.

Unique Entity Estimation with Application to the Syrian Conflict

Beidi Chen, Rice University
Co-authors: Anshumali Shrivastava, and Rebecca C. Steorts

Entity resolution identifies and removes duplicate entities in large, noisy databases and has grown in both usage and new developments as a result of increased data availability. Nevertheless, entity resolution has tradeoffs regarding assumptions of the data generation process, error rates, and computational scalability that make it a difficult task for real applications. In this paper, we focus on a related problem of unique entity estimation, which is the task of estimating the unique number of entities and associated standard errors in a data set with duplicate entities. Unique entity estimation shares many fundamental challenges of entity resolution, namely, that the computational cost of all-to-all entity comparisons is intractable for large databases. To circumvent this computational barrier, we propose an efficient (near-linear time) estimation algorithm based on locality sensitive hashing. Our estimator, under realistic assumptions, is unbiased and has provably low variance compared to existing random sampling based approaches. In addition, we empirically show its superiority over the state-of-the-art estimators on three real applications. The motivation for our work is to derive an accurate estimate of the documented, identifiable deaths in the ongoing Syrian conflict. Our methodology, when applied to the Syrian data set, provides an estimate of $191,874 \pm 1772$ documented, identifiable deaths, which is very close to the Human Rights Data Analysis Group (HRDAG) estimate of 191,369. Our work provides an example of challenges and efforts involved in solving a real, noisy challenging problem where modeling assumptions may not hold.

The Automatic Shape-constrained Nonparametric Regression

Zhikun Gao, The George Washington University
Co-authors: Judy Wang, Yanlin Tang, Guangying Wu, Jeff Lin

Like human beings, many animals produce sounds for communication and social interactions. The vocalizations of mice have the characteristics of songs, consisting of syllables of different types determined by the frequency modulations and structure variations. To characterize the impact of social environments and genotypes on vocalizations, it is important to identify the patterns of syllables based on the shapes of frequency contours. Using existing hypothesis testing methods to determine the shapes would require testing various null and alternative hypotheses for each curve, and is impractical for vocalization studies where the interest is on a large number of frequency contours. To overcome this challenge, we propose a new penalization-based method, which provides function estimation and automatic shape identification simultaneously. The method estimates the functional curve through quadratic B-spline approximation, and captures the shape feature by penalizing the positive and negative parts of the first two derivatives of the spline function in a group manner. Under some regularity conditions, we show that the proposed method can identify the correct shape with probability approaching one, and the resulting nonparametric estimator can achieve the optimal convergence rate. Simulation shows the proposed method gives more stable curve estimation than the unconstrained B-spline estimator, and it is competitive to the shape-constrained estimator assuming prior knowledge of the functional shape. The proposed method is applied to the motivating vocalization study to examine the effect of Mecp2 gene on the vocalizations of mice during courtship.

Online Sequential Monitoring Of Disease Incidence Rates With An Application To The Florida Influenza-Like Illness Data

Kai Yang, University of Florida
Online sequential monitoring of the incidence rates of chronic or infectious diseases is critically important for public health and stability of our society. Governments around the world have invested a great amount of resource in building global, national and regional disease reporting and surveillance systems. In these systems, conventional control charts, such as the cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) charts, are usually included for disease surveillance purpose. However, these charts require many assumptions on the observed data, including the ones of independent and identically normally distributed data when no disease outbreaks are present. These assumptions can hardly be valid in practice, making the results from the conventional control charts unreliable. Motivated by an application to monitor the Florida influenza-like illness data, we develop a new sequential monitoring approach, which can accommodate the dynamic nature of the disease incidence rates, spatio-temporal data correlation, and non-normality. It is shown that the new method is much more reliable to use in practice than the commonly used conventional charts for sequential monitoring of disease incidence rates.

Revisiting the proton-radius problem using constrained Gaussian processes

Shuang Zhou, Florida State University
Co-authors: Pablo Giulani, Debdeep Pati, Anirban Bhattacharya, Jorge Piekarewicz

The proton radius puzzle is an unanswered problem in physics relating to the size of the proton. Historically the proton radius was measured via two independent methods, which converged to a value of about 0.8768 femtometers (1 fm = 10^{-15} m). This value was challenged by a 2010 experiment utilizing a third method, called the muonic lamb shift experiment which produced a radius about 5% smaller than this. The discrepancy is explained in the current literature either by changing the laws of physics or suspecting that the original data collected from the electron scattering experiment were erroneous. Although new datasets with high precision measurements confirm that the radius might actually be closer to 0.84 fm, the discrepancy stemming from the original dataset remains unresolved, and is a topic of ongoing research. In this article, we approach this problem from a novel nonparametric Bayesian function estimation perspective, with physical constraints explicitly accounted for in the estimation procedure. Our analysis of the electron-form factor measurements versus potential transfer values data confirms the value obtained from the new datasets (0.84 fm) as the radius. Incorporating the physical constraints substantially reduces the uncertainty and 95% credible intervals obtained from our method do not contain the previous value of 0.8768 fm.

Poster Session

Tests For Association Using Cutpoint Based Categorization of Prognostic Variables

Saptarshi Chatterjee, Northern Illinois University
Co-authors: Shrbanit Chowdhury, Sanjib Basu

Cutoff detection in prognostic variables is an important area of research in clinical data analysis. Medical practitioners often prefer to categorize predictor variables in order to interpret the association with the outcome in a clinically meaningful way. Minimum $p$-value or maximally selected test statistic is commonly used for optimally categorizing a predictor variable, however, this approach can strikingly inflate the type I error. Several adjustments have been proposed in the literature, but they are often difficult or conservative and remain less used in practice. We propose omnibus cut-point based tests for assessing the association between an outcome and a predictor. These tests are based on the theory of permutation tests and are readily implementable. We show that these tests maintain their level and have strong power as blackbox tests in detecting both linear and nonlinear associations, even with outlier-prone and heavy-tailed error distributions. We further examine the level and power of these tests for binary outcome.

Using $U$-statistics to compare grouped marginal distributions of right censored event and waiting times in observational studies
Yichen Chen, University of Florida
Co-authors: Somnath Datta

$U$-statistics cover a wide class of non-parametric test statistics for comparing distributions of outcome in two or more groups. In epidemiology, the primary objective of an observational study is to compare disease outcome in two or more groups. The existence of confounding covariates blurs the causal relationship between group membership and disease outcome. We consider applying $U$-statistics to analyze observational studies with right censored event and transition times (state entry or exist times). We propose confounding and censoring adjusted (CCA) $U$-statistics motivated by the adjustment of marginal distribution function in each group. The summands of classical $U$-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 a multistate system, two types of the adjustments are proposed corresponding to two different selection methods. The dependence of censoring time on state entry time, group membership and/or observed covariates are considered. Difference normalization approaches are compared. 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. A data analysis strategy is suggested based on simulation results. From simulation results, the CCA $U$-statistics have the overall smallest mean square error when censoring time depends on group membership. If censoring time does not depend on group membership, $U$-statistics only account for confounders are also reasonable and suggested. In the multistate system, complete and conditional normalization factor are recommended for the two types of the adjustments, specifically. Our procedure is applied to a real data set on recovery process following bone marrow transplant for acute leukemia. The pairwise values of CCA U-statistics within three risk groups are 0.708, 0.364 and 0.275, which result in CCA Kruskal-Wallis test statistic equals to 28.147, $p < 0.0001$. CCA $U$-statistics can be used for analyzing observational studies with observed confounding covariates and right-censored outcomes. The only relationship requires to be determined before applying CCA $U$-statistics is between censoring time and group membership. Log-rank test is a simple and useful tool for this determination.

A new class of geometrically anisotropic spatial models for random fields on regular lattices

Fan Dai, Iowa State University
Co-authors: Somak Datta

Geometric anisotropy arises when the variogram of a spatial random field varies with direction. We propose a new class of geometrically anisotropic models for intrinsic random fields on regular lattices. The proposed class arises from fractional Laplacian differencing on the lattice and contains second or higher order anisotropic intrinsic autoregressions as special cases. Furthermore, with diminishing lattice spacing, these models approximate certain continuum anisotropic Matern class of models. The variogram functions of the proposed anisotropic models are not known analytically and are expressed in terms of multidimensional integrals. Extending the numerical methods of Dutta and Mondal (2016, *Environmetrics*), we demonstrate the computations of theoretical variogram values and perform exploratory analyses using data on ocean chlorophyll concentrations obtained from satellite measurements by MODIS-Aqua project of NASA.

Discovering Common Change-Point Patterns in Functional Connectivity Across Population

Mengyu Dai, Florida State University
Co-authors: Zhengwu Zhang, Anuj Srivastava

This paper studies change-points in human brain functional connectivity (FC) and seek patterns that are common across multiple subjects under identical external stimulus. FC, represented mathematically as a covariance or a correlation matrix, relates to the similarity of fMRI responses across different brain regions when a brain is simply resting or performing a task under an external stimulus. While the dynamical nature of FC is well accepted, this paper develops a formal statistical test for finding change-points in time series associated
with FC observed over time. It represents instantaneous connectivity by a symmetric positive-definite matrix, and uses a Riemannian metric on this space to develop a graphical method for detecting change-points in a time series of such matrices. It also provides a graphical representation of estimated FC for stationary subintervals in between detected change-points. Furthermore, it uses a temporal alignment of the test statistic, viewed as a real-valued function over time, to remove intersubject variability and to discover common change-point patterns across subjects. This method is illustrated using the Human Connectome Project (HCP) database for multiple subjects and tasks. This paper finds a consistent pattern of change-points in FC in task-related fMRI across subjects.

Adaptive Inference with a Multidimensional Multiscale Statistic

Pratyay Datta, Columbia University

Co-authors: Bodhisattva Sen

We observe a stochastic process $Y$ on $[0, 1]^d$ ($d \geq 1$) satisfying $dY(t) = n^{1/2}f(t)dt + dW(t)$, $t \in [0, 1]^d$, where $n \geq 1$ is a given scale parameter (‘sample size’), $W$ is a standard Brownian sheet on $[0, 1]^d$ and $f \in L_1([0, 1]^d)$ is the unknown function of interest. We propose a multivariate multiscale statistic in this setup and prove its almost sure finiteness; extending a previous work that proposed the analogous statistic for $d = 1$. We use the proposed multiscale statistic to construct optimal tests for testing $f = 0$ versus (i) appropriate Hölder classes of functions, and (ii) alternatives of the form $f = \mu_n 1_{B_n}$, where $B_n$ is a rectangle in $[0, 1]^d$ with sides parallel to the coordinate axes and $\mu_n \in R$, $\mu_n$ and $B_n$ unknown. Using our proposed multiscale statistic we also construct confidence bands for $f$ with guaranteed finite-sample coverage probability, assuming $f$ is shape-constrained; e.g., $f$ is multivariate isotonic/convex. The constructed confidence bands are shown to be adaptive and optimal (in an appropriate sense) with respect to the smoothness of the underlying function $f$. Moreover, these bands are adaptive to the unknown intrinsic dimension (as opposed to the ambient dimension $d$) of $f$. Our numerical studies corroborate our main findings, and illustrate the usefulness of the proposed approach.

Bayesian approaches for modeling functional connectivity in neuroimaging

Naira Ghosal, University of Illinois at Chicago

Co-authors: Sanjib Basu

Functional connectivity considers temporal dependence of activation patterns in functionally linked and anatomically separated brain regions which are in continuous connection with each other. Functional connectivity can be measured by considering co-activation of brain regions in resting-state functional magnetic resonance imaging (fMRI). The objective of this research is to develop statistical methodologies for exploring differential functional connectivities in subjects with Autism Spectrum Disorder (ASD) in the context of the controversy between contrasting under-connectivity and over-connectivity theories for ASD subjects. Autism Brain Imaging Data Exchange (ABIDE) is an international collaboration among medical centers across the globe and is the largest repository of fMRI data for autism. We consider functional connectivity among 84 regions in the Brodmann areas of the brain. Two important features here are the different levels of connectivities among different brain regions, including connected and non-connected regions and the differential connectivities between normal and ASD subjects. We consider different models for these two features. The connected and non-connected regions can be modeled by a mixture model with stochastic constrains to ensure identifiability. We consider a nonparametric Bayesian approach to model differential connectivities between normal and ASD subjects that allows for flexibility in modeling and analysis. We further develop a Conditional auto-regressive model based on an intuitive neighborhood structure among the region pairs $(r, s)$ where the neighbors are defined by shared vertices. Our time-dependent model for functional connectivity jointly considers the time sequence of fMRI measurements and the modularity structure of brain regions. In extensive numerical studies, our nonparametric models reported superior sensitivity and specificity compared to the parametric counterparts under both correctly and incorrectly specified data generating models. We applied the developed methodologies to analyze functional connectivities in ABIDE data collected at two different centers and found heterogeneities in differential functional connectivities inferred by different models and in-
Interestingly, from same models across different centers. There are however brain regions whose differential
connectivities between autism and normal subjects are identified in the common intersection of results from
the different models and centers.

Hypothesis Testing in Functional Linear Concurrent Regression

▶ Rahul Ghosal, North Carolina State University

Co-authors: Arnab Maity

Functional linear concurrent regression model arises when the response and covariates are both function of
same argument $t$ (e.g. time or any continuous index) and the dependence of the response $y(t)$ on the co-
variates $x_j(t)$ is assumed to be linear. More specifically, the predictor variables $x_j$ influence the response
variable $y(t)$ only through its value at $t$. Like in usual linear regression often the primary interest of a study
is to find out whether a covariate is truly significant or not i.e. to test for association between a predictor
variable and the response. Building point wise confidence interval of the estimated regression function does
not answer the question of overall significance of the covariate. Thus there is a need for developing testing
methods to find out significant predictors in functional data setting. In this paper we have proposed a new
testing method for testing the null hypothesis of no effect of a covariate on response in the context of Func-
tional Linear Concurrent Regression. We have established an equivalent random effects formulation of our
functional model under which our testing problem reduces to testing for zero variance component for random
effects. For this purpose we have used a one sided score test approach which is an extension of the classi-
cal score test. We have given theoretical justification why asymptotically our testing procedure has the right
levels under null under standard assumptions. Using numerical simulations we have shown that our testing
method has the desired type 1 error rates and also that our proposed testing procedure has higher power
than the bootstrapped F test currently existing in literature. Our model and testing procedure is shown to give
good performances even when the data is sparsely observed and the covariate is contaminated with noise.
We also illustrate our method by applying to two real data applications: the gait data, and dietary calcium
absorption study data.

Pseudo-likelihood Based Consistent Variable Selection for High Dimensional Bayesian VAR
Models

▶ Satyajit Ghosh, University of Florida

Co-authors: Dr. Khitij Khare and Dr. George Michailidis

Vector autoregressive (VAR) models aim to capture linear temporal interdependencies among multiple time
series. They have been widely used in macro and financial econometrics and more recently have found novel
applications in functional genomics and neuroscience. These applications have also accentuated the need to
investigate the behavior of the VAR model in a high dimensional regime, which will provide novel insights into
the role of temporal dependence for regularized estimates of the models parameters. However, hardly any-
thing is known regarding posterior model selection consistency for Bayesian VAR models in such regimes. In
this work we develop a pseudo-likelihood based Bayesian approach to variable selection in high dimensional
VAR models by considering hierarchical normal priors on the autoregressive coefficients as well as on the
model space. We show the posterior ratio and strong selection consistency of the proposed method in the
sense that the posterior probability of the true model converges to one even when the dimension $p$ of the VAR
system grows nearly exponentially with the sample size $n$. Moreover posterior model selection holds without
imposing any sparsity or diagonal structure assumption on the error covariance matrix $\Sigma$. As long as the max-
imum eigenvalue of $\Sigma$ remains bounded above by a constant we can recover the true model as $n \to \infty$. And
most importantly the strong selection consistency does not require any restriction on the maximum number of
edges. To the best of our knowledge, these results are the first of their kind for high-dimensional multivariate
models exhibiting temporal dependence and as a by-product of these results, we establish strong selection
consistency for the high-dimensional linear regression model with serially correlated errors.
Title: Predicting Survival Times for Neuroblastoma Patients Using RNA-Seq Expression Profiles

► Tyler Grimes, University of Florida

Co-authors: Susmita Datta, Somnath Datta

Background: Neuroblastoma is the most common tumor of early childhood and is notorious for its high variability in clinical presentation. Accurate prognosis has remained a challenge for many patients. In this study, expression profiles from RNA-sequencing are used to predict survival times directly. Several models are investigated using various annotation levels of expression profiles (genes, transcripts, and introns), and an ensemble predictor is proposed as a heuristic for combining these different profiles. Results: The use of RNA-Seq data is shown to improve accuracy in comparison to using clinical data alone for predicting overall survival and event-free survival times. Furthermore, clinically high-risk patients can be subclassified based on their predicted overall survival times. In this effort, the best performing model was the elastic net using both transcripts and introns together. This model separated patients into two groups with 2-year overall survival rates of $0.32 \pm 0.10$ ($n = 24$) versus $0.84 \pm 0.05$ ($n = 66$), $p < 0.001$. The ensemble approach gave similar results, with groups $0.36 \pm 0.10$ ($n = 22$) versus $0.83 \pm 0.05$ ($n = 68$), $p = 0.009$. This suggests that the ensemble is able to effectively combine the individual RNA-Seq datasets. Conclusions: Using predicted survival times based on RNA-Seq data can provide improved prognosis by subclassifying clinically high-risk neuroblastoma patients.

Optimal Gaussian Approximation For Multiple Time Series

► Sayar Karmakar, University of Chicago

Co-authors: Wei Biao Wu

We obtain an optimal bound for Gaussian approximation of a large class of vector-valued random processes. Our results substantially generalize earlier ones which assume independence and/or stationarity. Based on the decay rate of functional dependence measure, we quantify the error bound of the Gaussian approximation based on the sample size $n$ and the moment condition. Under the assumption of $p$th finite moment, with $p > 2$, this can range from the worst $n^{1/2}$ to the optimal $n^{1/p}$ rate.

Feature selection and multi-class classification using sparse envelope model

► Minji Lee, University of Florida

Co-authors: Zhihua Su

We propose a new method for multi-class classification and feature selection under the sparse envelope framework. The sparse envelope model introduced by Su et al. (2016) can conduct variable selection on the responses in a multivariate regression model and achieve the efficiency gains provided by the standard envelope model. Compared to the original context, the proposed method is based on a one-way multivariate analysis of variance which enables it to perform a feature selection naturally. In the feature selection, even though the part of features is not significant, non-selected features should not be removed to improve efficiency of significant features. Simulation studies show that our method chooses true features in a stable manner and has lower misclassification rates than other multi-class classification methods. Consistency and the oracle property of the proposed model are established and the asymptotic distribution of the estimator is obtained.

Statistical Downscaling with Spatial Misalignment: Application to Wildland Fire PM$_{2.5}$ Emissions Forecasting

► Suman Majumder, North Carolina State University

Co-authors: Yawen Guan, Ana G. Rappold, Brian J. Reich
Fine particulate matter PM$_{2.5}$ has been documented to have adverse long term health effects on human and wildfires send out tons of such particles into the air affecting the nearby areas. Forecasters use numerical models to predict PM$_{2.5}$ concentrations in different areas for the next 24 to 72 hours to warn the public of impending health risk. Statistical methods are needed to calibrate numerical model forecast using monitor data and statistical challenges such as spatial misalignment and potential model bias often come up. Typical model calibration techniques do not allow correction of errors due to misalignment of geographic locations. We propose a spatiotemporal downscaling methodology that, using image registration techniques, identifies the spatial misalignments and accounts for and corrects the bias produced by such warping. Our model is fitted in the Bayesian framework to provide uncertainty of space-time warping function as well as the forecasts. Two sources of data: a short term forecasts of PM$_{2.5}$ concentration from a deterministic model and spatially sparse monitor data is used in our model. We apply this method to simulated datasets as well as on a real dataset to demonstrate the utility of the proposed method and its applicability as a real-time forecast method. This is a joint work with Dr. Yawen Guan and Dr. Brian J. Reich of North Carolina State University and Ana G. Rappold of United States Environmental Protection Agency.

**A Geometric Variational Approach to Bayesian Inference**

▶ Abhijoy Saha, The Ohio State University

Co-authors: Karthik Bharath, Sebastian Kurtek

We propose a novel Riemannian geometric framework for variational inference in Bayesian models based on the nonparametric Fisher-Rao metric on the manifold of probability density functions. Under the square-root density representation, the manifold can be identified with the positive orthant of the unit Hilbert sphere, and the Fisher-Rao metric reduces to the standard $L_2$ metric. Exploiting such a Riemannian structure, we formulate the task of approximating the posterior distribution as a variational problem on the hypersphere based on the alpha-divergence. This provides a tighter lower bound on the marginal distribution when compared to, and a corresponding upper bound unavailable with, approaches based on the Kullback-Leibler divergence. We propose a novel gradient-based algorithm for the variational problem based on Frechet derivative operators motivated by the geometry of the sphere, and examine its properties. Through simulations and real-data applications, we demonstrate the utility of the proposed geometric framework and algorithm on several Bayesian models.

**A Test for Isotropy on a Sphere using Spherical Harmonic Functions**

▶ Indranil Sahoo, North Carolina State University

Co-authors: Joseph Guinness, Brian J. Reich

Analysis of geostatistical data is often based on the assumption that the spatial random field is isotropic. This assumption, if erroneous, can adversely affect model predictions and statistical inference. Nowadays many applications consider data over the entire globe and hence it is necessary to check the assumption of isotropy on a sphere. In this paper, a test for spatial isotropy on a sphere is proposed. The data are first projected onto the set of spherical harmonic functions. Under isotropy, the spherical harmonic coefficients are uncorrelated whereas they are correlated if the underlying fields are not isotropic. This motivates a test based on the sample correlation matrix of the spherical harmonic coefficients. In particular, we use the largest eigenvalue of the sample correlation matrix as the test statistic. Extensive simulations are conducted to assess the Type I errors of the test under different scenarios. We show how temporal correlation affects the test and provide a method for handling temporal correlation. We also gauge the power of the test as we move away from isotropy. The method is applied to the near-surface air temperature data which is part of the HadCM3 model output. Although we do not expect global temperature fields to be isotropic, we propose several anisotropic models with increasing complexity, each of which has an isotropic process as model component and we apply the test to the isotropic component in a sequence of such models as a method of determining how well the models capture the anisotropy in the fields.

**Central limit theorems on random fields for indicated sampling designs**

▶ Leshun Xu, The University of Auckland
Co-authors:

In many real situations, a modelled population is driven by the auxiliary information from other populations, and at the same time the sample is affected by some complex factors with unequal selection probabilities. In this paper, we use random fields to model superpopulations. We introduce a function to describe the relationship between the modelled population and the outside information, and an indicator random field to express the sampling strategy. With some certain assumptions, such as the assumption on the independent indicators given the information from the other population, we consider central limit theorems on strong mixing random fields for the indicated sampling design.

Invited Talks

Microbiome as a causal mediator

► Alexander V. Alekseyenko, Medical University of South Carolina

Session: IP33

Common research designs of microbiome studies do not allow to attribute and measure direct causal role of the microbiota on the observed outcomes. To do that, additional experiments are typically necessary. In many instances however, it is possible to quantify the indirect effect of interventions through the microbiome in a statistical mediation analysis framework. The main challenge for such modeling is the fact that microbiome measurements are highly multivariate, under-sampled, compositional and over-dispersed. Several alternatives exist for estimating the mediation effect of individual taxa via multiple regressions with appropriate penalties. We propose an omnibus method for causal mediation analysis with microbiome data that is designed to measure the total multivariate mediated effect of all measured microbial communities. Our method relies on distance-based energy statistics and builds on existing hypothesis testing-based frameworks for causal mediation analysis.

Design and analysis of trials with expected delayed treatment effect

► Keaven M Anderson, Merck Research Laboratories

Session: IP42

The potential for delayed treatment care warrants special consideration for the design and analysis of clinical trials with a time-to-event endpoint. Doing analyses too soon and planning analyses based on event counts only are concepts that deserve scrutiny. Early analyses should focus more on patient safety rather than a traditional futility approach. Testing and estimation with alternatives to Cox regression and logrank procedures may improve power. Combination tests may make a design more robust to various alternate hypotheses, with and without proportional hazards. Getting longer-term follow-up at some appropriate level should be ensured prior to final analysis. Practical guidance reflecting the above will be provided.

High-dimensional nonparametric Bayes models for doubly robust causal inference

► Joseph Antonelli, Harvard University

Session: IP24

We introduce causal inference methods for high-dimensional scenarios that utilize nonparametric Bayesian methods to alleviate modeling assumptions, and sparsity inducing priors to reduce the dimension of the covariate space. By leveraging the Bayesian framework, we can acquire the posterior distribution of any estimator that can be defined as a function of the treatment and outcome model (e.g. standard doubly robust estimator). We illustrate how the proposed framework allows us to extend many existing estimators to highly nonlinear
or high-dimensional scenarios while maintaining desirable small sample properties. We account for all of the uncertainty inherent in the data by combining our posterior samples with an efficient resampling procedure that will only slightly increase computation time, while giving credible intervals that obtain good frequentist properties in finite samples and achieve nominal coverage rates. We show via simulation the ability of the proposed estimator to flexibly estimate causal effects in high-dimensions. Finally, we apply our proposed procedure to estimate the effect of environmental exposures on health outcomes.

Enhancement of Visual Analytics for Efficacy Data
► Vipin Arora, Eli Lilly and Company

Session: IP48

Efficacy data present many challenges with regard to analysis and interpretation. Outcomes have high variability in measurements and are multidimensional and interrelated in nature. It is well recognized that visual analytics present a useful alternative to tabular outputs for exploring data and present a great opportunity to enhance evaluation of drug efficacy. Graphs can play a big role in facilitating communication of results with regulators, investigators, DMC, and other stakeholders and help convey multiple pieces of information concisely and more effectively than tables. The possibility of tapping into all subgroups (identified or yet to be identified) is almost impossible when using the traditional tools (Tables and listings). With the availability of recent tools (Spotfire, Tableau, R etc.) there are significant opportunities that have become available to explore the data and provide quick responses to stakeholders how the treatment performed in subgroups or a set of subgroups and the associated efficacy assessments. In this discussion, we will consider the role of graphical methods and possible enhancements to these graphs in efficacy and show how to get maximum gain from using visual analytics taking into account some considerations that must be borne in mind for effective visualization. We will show via some examples how visual analytics play a critical role in the ongoing evaluation of a drug during study conduct or later.

Applications of Geometry to Network Inference
► Dena M. Asta, Ohio State University

Session: IP29

Graphs are often formed by sampling points from some manifold and connecting edges with some probability dependent on the manifold distances between points. There exists a plethora of research showing how the geometry of the manifold constrains properties of the sampled graphs. We survey some such research, and motivated as such, we present some recent work relating the geometry of the manifold with properties of the random graphs and their estimators.

Structural breaks in functional time series
► Alexander Aue, University of California, Davis

Session: IP6

This talk explores methods for the analysis of structural breaks in functional time series data, where structural breaks might be in the mean function, the covariance operator or both. The proposed methods can be fully functional (for the mean break case) or based on dimension reduction techniques (for the covariance break case). Theoretical results will be presented and illustrated with simulations and an application to annual temperature profiles.

SCnorm: A quantile-regression based approach for normalization of single-cell RNA-seq data
► Rhonda L. Bacher, University of Florida
Session: IP59

Single cell RNA-sequencing (scRNA-seq) is a promising tool that facilitates study of the transcriptome at the resolution of a single cell. However, along with the many advantages of scRNA-seq come technical artifacts not observed in bulk RNA-seq studies. The normalization methods traditionally used in bulk RNA-seq were not designed to accommodate these features and, consequently, applying them to the single-cell setting results in artifacts that bias downstream analyses. To address this, we developed SCnorm to enable efficient and accurate scRNA-seq normalization. Simulation and case study results demonstrate that SCnorm provides for increased accuracy in fold-change estimation as well as improvements in downstream inference.

Radiomics promise and challenges in Oncology.

▶ Yoganand Balagurunathan, Moffitt Cancer Center

Session: IP45

Cancer Imaging has been traditionally used as a non-invasive diagnostic tool in both identification and tracking disease progression. Conventional radiological science has used these images to provide heuristic observation by a trained expert. Improvement in scanner detector technologies has allowed obtaining high resolution imaging close to sub-millimeter range. Recently, quantification of regions on the cancer imaging has evolved as a disciple ('Radiomics') to characterize tumor, whose greatest utility has been to relate the characteristics to patient outcome and disease diagnosis 1, 2. The radiomics metrics has shown acceptable level of reproducibility in lung tumors and some of these stable metrics are used as a surrogate marker for progression of disease. In our recent exploration of lung screening CT images of with a goal to develop image based malignancy predictors. We built a cohort which was randomly into two train and test parts. Lung nodules were identified by a radiologist, and on the delineated regions we extracted 219 quantitative image features describing size, shape and texture of the nodules. Using features in the size & shape categories, the average area under the receiver operator characteristics (AUROC) was 0.79, while using non-size based features (location, co-occurrence, run length, pixel histogram and texture: laws & wavelets) the AUROC was in the range of 0.83. Although the medical imaging has shown great promise, the underlying technology is optimized for conventional radiology. The vast availability of medical images as part of clinical workflow has certainly opened up avenues to improve diagnosis and patient care. The interinstitutional variability has been diminished with the increased discordance between patient scans. Standardization of computer generated radiomic descriptors has been a challenge. The field had shown promise with abundant potential to grow, especially with recent advancement in learning methods.

Spatial modeling of the American community survey

▶ Soutir Bandyopadhyay, Colorado School of Mines

Session: IP38

The American Community Survey (ACS) publishes county-level estimates for hundreds of demographic and economic variables, with the statistical uncertainty only being quantified from the standpoint of sampling error. We propose a nonstationary spatial model (the lattice-kriging model) for ‘flow’ estimates (i.e., representing aggregate activity over a region) that accounts for the sampling error while also utilizing the latent structure of the population process. Our goals are two-fold: ?first, to offer estimates of all counties (including ‘missing’ counties, for which the ACS omits to publish estimates due to the sampling variability being too high) along with a variance estimate that combines both the population dynamics and the sampling mechanism; second, to offer estimates of custom regions that correspond to the needs of users at the state and local level. These custom regions can overlap county boundaries, and can have nontrivial homotopy (e.g., be discus-shaped).

A spatial Wishart process model for diffusion tensor imaging data from cocaine addiction studies

▶ Dipankar Bandyopadhyay, Virginia Commonwealth University
In this talk, we propose a Bayesian spatial process model to capture spatial dependencies between proximal positive definite matrices (summary of DTI measures corresponding to each voxel), assuming the matrices to be marginally distributed as Wishart. An adaptation of the Nearest Neighbor Gaussian Process facilitates the corresponding Markov chain Monte Carlo computing framework. Simulation studies and applications to a real dataset on cocaine addiction illustrates the advantages of our method over available standard techniques in detecting brain hotspots.

Variable Selection in the Context of Spatially Dependent Covariates in High Dimensional Data

► Anjishnu Banerjee, Medical College of Wisconsin

About 90% of all malignant tumors diagnosed in the United State are solid tumors. Improving diagnostic accuracy is essential for preventing unnecessary procedures. Current imaging techniques lack the ability to accurately diagnose prostate cancer grade, specially in peripheral and high fluid density regions. Multiparametric MRI is gaining acceptance as the standard of care for prostate imaging. We discuss novel statistical methods for variable selection and feature determination in the presence of high spatial correlation by predicting underlying histological information based on prior radiographic and pathologic knowledge. This will give radiologists the ability to accurately pinpoint the location and grade of prostate cancer nodules, thereby improving quality of life the patients. In this proposal, we propose “ensemble” selection methods, which combine a set of within group and between spatial cluster selection tools. We develop a Bayesian nonparametric weighting scheme, which besides borrowing strength across classifiers, also helps mitigate issues due to minor misalignment between MR images and the histology. This possible gains are with real life MR images from patients with cancers of the prostate and glioblastoma.

Steady states of queueing networks

► Sayan Banerjee, University of North Carolina, Chapel Hill

Consider a system of $N$ parallel single-server queues with unit-exponential service time distribution and a single dispatcher where tasks arrive as a Poisson process of rate $\lambda(N)$. When a task arrives, the dispatcher assigns it to one of the servers according to the Join-the-Shortest Queue (JSQ) policy. Eschenfeldt and Gamarnik (2015) established that in the Halfin-Whitt regime where $(N - \lambda(N))/\sqrt{N} \to \beta > 0$ as $N \to \infty$, appropriately scaled occupancy measure of the system under the JSQ policy converges weakly on any finite time interval to a certain non-elliptic diffusion process as $N \to \infty$. Recently, it was further established by Braverman (2018) that the convergence result extends to the steady state as well, i.e., stationary occupancy measure of the system converges weakly to the steady state of the diffusion process as $N \to \infty$, proving the interchange of limits result. In this talk, I will talk about analyzing the detailed behavior of the steady state of this diffusion process using tools from renewal theory. We will investigate the tails and bulk behavior of the steady state distribution and sample path fluctuations of the diffusion process.


► Saonli Basu, University of Minnesota

The development of a complex disease is an intricate interplay of genetic and environmental factors. ‘Heritability’ is defined as the proportion of total trait variance due to genetic factors within a given population. Studies with monozygotic (MZ) and dizygotic (DZ) twins allow us to estimate heritability by fitting an ‘ACE’ model which estimates the proportion of trait variance explained by additive genetic (A), common shared environment (C), and unique non-shared environmental (E) latent effects, thus helping us better understand
disease risk and etiology. In this paper, we develop a flexible generalized estimating equations framework ('GEE2') for fitting twin ACE models that requires minimal distributional assumptions; rather only the first two moments need to be correctly specified. In addition, we prove that two commonly used methods for estimating heritability, the normal linear mixed effects model (LMM) and Falconer's method, can both be fit within this unified GEE2 framework, which additionally provides robust standard errors. Given non-normal data, we show that the GEE2 models attain significantly better coverage of the true heritability compared to the traditional LMM and Falconer's methods. Finally, we demonstrate that Falconer's method can consistently estimate heritability even when the total variance differs between MZ and DZ twins; whereas the LMM will produce biased estimates in such settings.

Bayesian Approaches to Analysis of Adverse Events

► Sanjib Basu, University of Illinois at Chicago

Session: IP15

Adverse events or side effects from treatment are of serious concern for patients on one side and pharmaceutical industries on the other side. We investigate the methodological issues in analysis of sparse adverse events and show that the parameterization and the choice of priors play crucial roles in the statistical analysis. For improper priors, we establish conditions for posterior propriety; these conditions are related to the concepts of complete and quasi-complete separation in logistic regression. In extensive simulation studies, we find that our proposed Bayesian estimates, in particular, the semiparametric model based estimates, perform significantly better than the continuity corrected and other estimates proposed in the literature. We illustrate the proposed methods in two examples of meta analysis. This is joint work with Arpita Chatterjee from Georgia Southern University.

PK/PD Data Extrapolation Models for Improved Pediatric Efficacy and Toxicity Estimation

► Cynthia Basu, Pfizer Inc.

Session: IP62

In most drug development settings, the regulatory approval process is accompanied by extensive studies performed to understand the drug’s pharmacokinetic (PK) and pharmacodynamic (PD) properties. In this talk, we attempt to utilize the rich PK/PD data to inform the borrowing of information from adults during the pediatric drug development. In pediatric settings, it is especially crucial that we are parsimonious with the patients recruited for experimentation. We illustrate our approaches in the context of clinical trials in pediatric and adult patients where we model both the efficacy and safety endpoint. We use population PK/PD modeling to quantitatively assess the similarity between adults and children, and use this information in various hierarchical Bayesian adult borrowing rules whose statistical properties can then be evaluated.

Small Area Estimation for Conservation Effects Assessment Project

► Emily J. Berg, Iowa State University

Session: IP60

The Conservation Effects Assessment Project (CEAP) is a complex survey that measures several types of erosion on cropland and pastureland in the United States. One challenge in implementing small area estimation procedures using CEAP data is that distributions are skewed and contaminated with zeros. We develop an empirical Bayes small area predictor based on a zero-inflated lognormal model. We evaluate the properties of the predictor and bootstrap mean squared error estimator through simulation. We apply the zero-inflated lognormal model to obtain small area estimates of erosion using CEAP data.

The Graphical Horseshoe Estimator for Inverse Covariance Matrices

19
Anindya Bhadra, Purdue University

Session: IP20

We develop a new estimator of the inverse covariance matrix for high-dimensional multivariate normal data using the horseshoe prior. The proposed graphical horseshoe estimator has attractive properties compared to other popular estimators, such as the graphical lasso and the graphical smoothly clipped absolute deviation. The most prominent benefit is that when the true inverse covariance matrix is sparse, the graphical horseshoe provides estimates with small information divergence from the sampling model. The posterior mean under the graphical horseshoe prior can also be almost unbiased under certain conditions. In addition to these theoretical results, we also provide a full Gibbs sampler for implementing our estimator. MATLAB code is available for download from github at http://github.com/liyf1988/GHS. The graphical horseshoe estimator compares favorably to existing techniques in simulations and in a human gene network data analysis.

Bayesian Inference for the Relationship between two categorical Variables with covariates for clustered data

Dilli Bhatta, University of South Carolina Upstate

Session: IP60

Analysis of categorical data, presented in a two-way contingency table, is a well known statistical problem in many real life applications. With such data, we make an inference about the possible association between two cross-classified categorical variables. Usually, the data presented in a two-way table are obtained from a simple random sampling (SRS), where the responses of the individuals are assumed to be independent of each other. In this paper, we consider cluster sample (type of complex sample survey data) and provide a test of independence between two categorical variables by incorporating PSU and SSU related covariates. It is likely that these covariates can be associated with two cross-classified categorical variables and thus may have impact on their association. We incorporate covariates via a multi-level logistic regression model and propose a two-stage Bayesian procedure in order to make an inference about association. We apply our methodology to the TIMSS data.

Asymptotics of autocovariance matrices

Monika Bhattacharjee, University of Florida

Session: IP19

We consider the high-dimensional moving average (MA) and autoregressive (AR) processes. Our goal is to explore the asymptotics for eigenvalues of the sample autocovariance matrices. This asymptotics help in the estimation of unknown order of the high-dimensional MA and AR processes. Our results also provide tests of different hypotheses on coefficient matrices.

Scalable computation with Bayesian shrinkage priors

Anirban Bhattacharya, Texas A&M University

Session: IP9

Gaussian scale mixture priors are frequently employed in Bayesian analysis of high-dimensional models, and several members of this family have optimal risk properties when the truth is sparse. While optimization-based algorithms for the extremely popular Lasso and elastic net procedures can scale to dimension in the hundreds of thousands, corresponding Bayesian methods that use Markov chain Monte Carlo (MCMC) for computation are limited to problems at least an order of magnitude smaller. This is due to high computational cost per step of the associated Markov kernel and growth of the variance of time-averaging estimators as a function of dimension. Here we propose an MCMC algorithm for computation in these models that combines block updating and approximations of the Markov kernel to directly combat both of these factors. Our algorithms...
rithm gives orders of magnitude speedup over the best existing alternatives in high-dimensional applications.

Safety Assessment and Monitoring in Clinical Drug Development - An Overview

▶ Amit Bhattacharyya, ACI Clinical

Session: IP48

Patient Safety is paramount for any pharmaceutical drug research and development (R&D). Society expects drugs to be fully safe and effective in treating diseases. Almost all pharmaceutical drugs have some inherent risks which the manufacturers, regulators, physicians and often patients do understand the need for assessing the benefit of the drugs in treating serious diseases, compared to the risk that they carry for the patients. Several established guidelines exist from the regulators worldwide to assess and monitor appropriate safety for any new investigational research for new drugs. Mechanisms such as multi-disciplinary committees - internal to the drug manufacturers, and most importantly independent external monitoring and assessment committees are essential for identifying, adjudicating and monitoring safety issues during trials and provide recommendations for mitigating the risks to the patients. This introductory talk will set the scene discussing the landscape of safety assessment in clinical trials and the monitoring of patients’ safety through various committees, to help the audience understand the importance of the research on drug safety.

Pre-term birth, inter-pregnancy interval and parity: A joint modeling approach

▶ Joe Bible, Clemson University

Session: IP15

In any study the task of modeling the association between candidate factors and risk associated with an outcome is complicated by the observation process. Often, it is implicitly assumed that the candidate factors and outcome are not related in a substantive way with the observation process. Methodologies have been developed to address settings where the time to event (informative time to event), or when the number of events experienced or observed (informative cluster size) is related to the outcome. However, at this time literature which addresses settings where both may be a concern appears to be lacking. We propose a shared random parameter model, which accommodates heterogeneity in the outcome, time to event and cluster size portions of the model. We provide a simulation study which demonstrates that when these relationships exist and are related to the observation process that naive modeling can lead to severely biased inference. We provide the Consecutive Pregnancy Study data as a motivation example and propose a score test to identify significant dependencies between gap times between pregnancies and the number of pregnancies as well as subject specific propensity towards pre-term birth.

Bayesian Models for High-Dimensional Non-Gaussian Dependent Data

▶ Jonathan R. Bradley, Florida State University

Session: IP1

A Bayesian approach is introduced for analyzing high-dimensional dependent data that are distributed according to a member from the exponential family of distributions. This problem requires extensive methodological advancements, as jointly modeling high-dimensional dependent data leads to the so called ‘big n problem’. The computational complexity of this problem is further exacerbated by allowing for non-Gaussian data models. Thus, we develop new computationally efficient distribution theory for this setting. In particular, we introduce a class of conjugate multivariate distributions for the exponential family. We discuss several theoretical results regarding conditional distributions, an asymptotic relationship with the multivariate normal distribution, parameter models, and full-conditional distributions for a Gibbs sampler. We demonstrate the modeling framework through several examples, including an analysis of a large environmental dataset.

Automated Scalable Bayesian Inference via Data Summarization
The use of Bayesian models in large-scale data settings is attractive because of the rich hierarchical relationships, uncertainty quantification, and prior specification they provide. Standard Bayesian inference algorithms are often computationally expensive, however, making their direct application to large datasets difficult or infeasible. We leverage the insight that data often exhibits redundancies to instead obtain a weighted subset of the data (called a “coreset”) that is much smaller than the original dataset. We can then use this small coreset in any number of existing posterior inference algorithms without modification. We provide theoretical guarantees on the size and approximation quality of the coreset. The proposed approach also permits efficient construction of the coreset in both streaming and parallel settings.

Single cell analyses reveal distinct “species” of cancer cells within tumor ecosystems

Therapeutic approaches to cancer aim to eliminate tumors. But, tumors are heterogeneous ecosystems populated by normal and cancer cells. Cancer cells vary in their resource uptake, modes of metabolism (glycolytic and acid producing versus more efficient oxidative phosphorylation), motility and susceptibility to the immune system. Tumors can be heterogeneous in blood flow, oxygen levels, nutrient availabilities, and exposure to the immune system. Much work has focused on genetic variability among cancer cells. Less work has been done on identifying cell types by their morphology or phenotypic characteristics, and linking these characteristics to environmental heterogeneity. The edge versus the interior of a tumor provides a source of heterogeneity. The edge should be more resource rich but more exposed to the immune system than the interior. In examining histologies of breast cancer patients, the cancer cells on the tumor’s edge were more glycolytic (helpful as an immune defense through profligate with resources) and had higher cell turnover rates than those in the interior. At smaller scales within the tumor, distance to vasculature should matter. In breast cancer tumors estrogen positive (cells requiring estrogen for growth and proliferation) and estrogen negative cancer cells partition space. ER+ cells aggregate nearer blood vessels, and ER- farther away. Furthermore, women scored as ER+ showed greater blood vasculature than those scored ER-. The former is considered a better patient prognosis than the latter. So, how does one treat a community of cancer cells? The tumors of men with castrate resistant metastatic prostate cancer may have up to three species of cancer cells. Two can be successfully treated with androgen deprivation and/or targeted therapies. The third is essentially untreatable resulting in progression and death. In a recent clinical trial, we showed how maintaining a balance of competing cancer cell types, via adaptive therapy, significantly extends progression free survival.

Model-Based Standardization Using an Outcome Model with Random Effects

Model-based standardization uses a statistical outcome model or exposure model to estimate an unconfounded population-average effect. With it, one can compare groups using a distribution of confounders identical in each group to that of a standard population. We develop an approach based on an outcome model, in which the mean of the outcome is modeled conditional on the exposure and the confounders. In our approach, there is a confounder that clusters the observations into a very large number of categories. We treat the parameters for the clusters as random effects. We use a between-within model to account for the association of the random effects not only with the exposure but also with the cluster population sizes. We review alternative approaches presented in the literature, and we compare the outcome-modeling approach to recently proposed exposure-modeling approaches incorporating random effects. To illustrate, we use 2014 Truven Health MarketScan Research Data to compare proportions of acute respiratory tract infection diagnoses with an antibiotic prescription for emergency department versus outpatient visits, adjusting for confounding by unmeasured patient level variables. We also present results of a simulation study.
Multivariate semiparametric spatial methods for imaging data

- **Guanqun Cao**, Auburn University

Session: IP56

Univariate semiparametric methods are often used in modeling nonlinear age trajectories for imaging data, which may result in efficiency loss and lower power for identifying important age-related effects that exist in the data. As observed in multiple neuroimaging studies, age trajectories show similar nonlinear patterns for the left and right corresponding regions and for the different parts of a big organ such as the corpus callosum. To incorporate the spatial similarity information without assuming spatial smoothness, we propose a multivariate semiparametric regression model with a spatial similarity penalty, which constrains the variation of the age trajectories among similar regions. The proposed method is applicable to both cross-sectional and longitudinal region-level imaging data. We show the asymptotic rates for the bias and covariance functions of the proposed estimator and its asymptotic normality. Our simulation studies demonstrate that by borrowing information from similar regions, the proposed spatial similarity method improves the efficiency remarkably. We apply the proposed method to two neuroimaging data examples. The results reveal that accounting for the spatial similarity leads to more accurate estimators and better functional clustering results for visualizing brain atrophy pattern.

Assessing agreement with uncommon data

- **Josep Carrasco**, University of Barcelona

Session: IP8

The assessment of agreement usually consists of the estimation of an index or coefficient that summarizes the amount of closeness among paired data. The index and its inferential properties are closely related to the nature of the data. For example, when data is quantitative it is commonly assumed that the outcome (conditioned to covariates) is continuous, normally distributed and the relation between the outcome and the covariates is linear. However, in occasions such assumptions do not meet. Here, I will introduce three examples were the common assumptions for quantitative data cannot be accepted. Specifically, the case examples will deal with spatially correlated data, asymmetric or extreme data and dependent count data. Dome strategies to analyze such data will be introduced and the total deviation index will be estimated to assess the degree of agreement.


- **Stefano Castruccio**, University of Notre Dame

Session: IP32

Facing increasing domestic energy consumption from population growth and industrialization, Saudi Arabia is aiming to reduce its reliance on fossil-fuels and broaden its energy mix by expanding its investment in renewable energy sources, including wind energy. A preliminary task in the development of wind energy infrastructure is the assessment of the wind energy potential, a key aspect of which is the characterization of its spatio-temporal behavior. In this study we examine the impact of internal climate variability on seasonal wind power density variability using the large Ensemble (LENS) developed at the National Center for Atmospheric Research. A spatio-temporal model for daily wind speed is proposed, with a multi-resolution skew-t distribution designed to capture the spatial patterns of higher order moments. The model assumes a low-dimensional latent structure that captures the regional dependence, has a closed form for the EM inference steps, and allows affordable inference for millions of data points. The statistical model acts as a fast stochastic approximation (a stochastic generator) of the original climate model, and is used to generate synthetic data that adequately reproduce the LENS dataset, thus allowing for fast uncertainty quantification of wind power density.
Scalable Bayesian support vector machines with feature selection

Sounak Chakraborty, University of Missouri-Columbia

Session: IP49

In this paper, we consider Bayesian analysis of binary and multiclass support vector machines with feature selection. We consider the fully supervised support vector machine problem and place Gaussian spike and slab priors. We propose a computationally scalable Gibbs sampling algorithm, which has linear computational complexity. We also consider Bayesian semi-supervised learning and propose a novel Bayesian approach for variable selection with scalable Gibbs algorithm. Our proposed novel Gibbs sampler called “Skinny Gibbs” which is much more scalable to high dimensional problems, both in memory and in computational efficiency. It can also avoid large matrix computations needed in standard Gibbs sampling algorithms. In terms of computational complexity for our Skinny Gibbs, it grows only linearly in the number of predictors. Efficiency of our method for supervised and semi-supervised SVM models are demonstrated based on several simulation studies and data analysis.

Adaptive Bayesian Modeling for Data Assimilation and Uncertainty Quantification

Avishek Chakraborty, University of Arkansas

Session: IP4

Deterministic computer models (or simulators) are common in many scientific disciplines to explore the dynamics of physical events. However, running these computer models are usually resource and time expensive and they may not adequately capture the real-world system. An emulator, a stochastic surrogate for the computer code, is used within a hierarchical framework to combine realizations from the computer model and field observations for estimation of parameters governing the system and prediction at new input combinations. In this talk, we are going to look at emulators with an over-specified set of functions where number of such functions, and their inclusion probabilities, are treated as unknown quantities. This approach is found to have smaller predictive uncertainty and computational efficiency than the standard Gaussian process approach to emulation and calibration. We shall consider two choices for such functions - splines and polynomial chaos expansions. We will also look at associated inverse problems where we aim to retrieve the unknown experimental input from the observed output. We will use an example from Astrophysics to illustrate the methods we discuss.

Adaptive Risk Bounds in Total Variation Denoising

Sabyasachi Chatterjee, University of Illinois at Urbana Champaign

Session: IP23

Total Variation Denoising (TVD) for matrices is a very successful and widely used technique in Image processing. However, nuanced statistical risk bounds for the TVD estimator are sparsely available. We study the constrained form of the TVD estimator for matrices and obtain risk bounds from a worst case and an adaptive perspective. Firstly, we show the minimax rate optimality of this estimator for matrices with total variation bounded by some number $V > 0$. We then show the rate of convergence can be faster than the minimax rate when the true image is piecewise constant on rectangles and the correct tuning parameter is chosen. This fact gives statistical backing to the folklore that the TVD estimator preserves sharp edges in an image. Lastly, we explore the existence of estimators which are completely tuning parameter free yet can still achieve the minimax rate in this problem.

Evaluating the Quality of Climate Models

Snigdhansu Chatterjee, University of Minnesota
Climate models produce output over decades or longer at high spatial and temporal resolution. Starting values, boundary conditions, greenhouse gas emissions, and so forth make the climate model an uncertain representation of the climate system. A standard paradigm for assessing the quality of climate model simulations is to compare what these models produce for past and present time periods, to observations of the past and present. Many of these comparisons are based on simple summary statistics called metrics. We propose an alternative: evaluation of competing climate models through probabilities derived from tests of the hypothesis that climate-model-simulated and observed time sequences share common climate-scale signals. The probabilities are based on the behavior of summary statistics of climate model output and observational data, over ensembles of pseudo-realizations. We compare monthly sequences of CMIP5 model output of average global near-surface temperature anomalies to similar sequences obtained from the well-known HadCRUT4 data set, as an illustration. This talk includes joint work with several students and colleagues.

A Bayesian Three-arm Non-Inferiority Trial with Negative Binomial Endpoint

Arpita Chatterjee, Georgia Southern University

Noninferiority clinical trials have gained immense popularity within the last decades. Such trials are designed to demonstrate that a new experimental drug is not unacceptably worse than an active control by more than a pre-specified small margin. Three-arm non-inferiority trials have been widely acknowledged as the Gold Standard because they can simultaneously establish both non-inferiority and the assay sensitivity. Bayesian approach to assess Non-inferiority under continuous end-point has been studied in recent past. However, such models were never investigated using in the presence of count data. One obvious recommendation will be the choice of Poisson distribution to model the primary end-point, which is often affected by over-dispersion. Instead, a Poisson-Gamma mixture, which ultimately results in Negative Binomial Distribution, can be a suitable alternative to model count data. In this paper, we propose a Bayesian hierarchical model to perform simultaneous testing of Non-Inferiority and Assay Sensitivity in a three-arm trial for negative binomially distributed endpoint. Also, we examined the effect of power prior (Ibrahim and Chen, 2000) under the proposed model setup through a simulation study. Finally, the performance of the proposed model is evaluated based on simulated dataset under varying scenarios.

Covariance based Moment Equations for Improved Variance Component Estimation

Sanjay Chaudhuri, National University of Singapore

ANOVA-type estimators of variance components for nested error regression models are always constructed based on moment equations related to residual variance. We consider moment equations associated with covariance and construct improved ANOVA-type estimators. These estimators are seen to be consistent, asymptotically unbiased and have better performances than traditional estimators of variance components for almost all kinds of sample allocations. Their improved performance is demonstrated analytically as well as through detailed simulation studies and applications to real data sets.

Potential drug repurposing using a simple omics integration approach for melanoma patients

Yian (Ann) Chen, Moffitt Cancer Center

Much progress has been made in developing therapeutic strategies based on mutation information in melanoma patients. After accounting for major known driver mutations, such as BRAF, NF1 and NRAS, there are still approximately 25% of melanoma patients without clear driver mutations. In addition, although a few im-
munotherapies, such as PD1 and CTLA4, have shown great promise to improve patient outcomes, after treatment failure, very limited treatment options remained available. Methods: We developed an integrated approach for potential drug repurposing and treatment prioritization by investigating the association between patients gene expression and mutation status of target genes from each candidate drug with patients clinical outcome. We applied our proposed approach to analyze the whole exome sequencing and RNAseq data from two melanoma cohorts: 459 TCGA patients and 135 BMS patients as a proof of principle study. The association between mutation, gene expression from 18,143 genes, and OS was evaluated using two Cox proportional hazard models, adjusting for age and prior treatment information. A Wilcoxon test was performed to assess potential functional implication of mutations on RNA expression using an eQTL analysis. Fisher’s Product method was used to synthesize the results from 3 analyses for each candidate drug over the two cohorts. False discovery rate (FDR) was used for treatment ranking. A total of 5,835 candidate treatments and their targets were included in this study: 5,787 drugs from DsigDB, 38 HDAC inhibitors, and 10 immunotherapies. Results: The immunotherapy, PDL1 (FDR= \(4.92 \times 10^{-10}\)), FDA approved for melanoma treatment, was ranked the second. It serves as a positive control for our simple integrated approach. Another immunotherapy, LAG3 (FDR= \(3.2 \times 10^{-8}\)), which is currently clinical trials was also among the top 10 most significant treatments. Uramustine (FDR= \(3.4 \times 10^{-8}\)), a chemotherapy drug used in lymphatic malignancies, was ranked fifth, was also promising. Some of the top ranked drugs were not previously used for cancer treatment but potential mechanisms and involved signaling pathways have been recently studied and reported. Our proposed simple approach and preliminary results are very promising. We are currently extending this to prioritizing combination treatments.

Identifying disease-associated copy number variations by a doubly penalized regression model

▶ Yichen Cheng, Georgia State University

Session: IP63

Copy number variation (CNV) of DNA plays an important role in the development of many diseases. However, due to the irregularity and sparsity of the CNVs, studying the association between CNVs and a disease outcome or a trait can be challenging. Up to now, not many methods have been proposed in the genetic literature for this problem. Most of the current researchers reply on an ad hoc two-stage procedure by first identifying CNVs in each individual genome and then performing an association test using these identified CNVs. This potentially leads to information loss and as a result a lower power to identify disease associated CNVs. In this paper, we describe a new method that combines the two steps into a single coherent model to identify the common CNV across patients that are associated with certain diseases. We use a double penalty model to capture CNVs’ association with both the intensities and the disease trait. We validate its performance in simulated datasets and a data example on platinum resistance and CNV in ovarian cancer genome.

Integrated Statistical Inference on Genomic Handles of Traits (InSIGHT)

▶ Cheng Cheng, St. Jude Children’s Research Hospital

Session: IP22

The genomic biochemical process underlying a complex trait often involves a large number and several types of molecules derived from the genome. The advancement of biotechnology has enabled researchers to measure genomic and epigenomic elements (DNA sequence variations, methylation levels, microRNA expressions), as well as gene (and protein) expressions, on genome-wide scale and often in relatively large cohorts of subjects. The technological advancement provides opportunities to gain more comprehensive understanding of the genomic biological process underlying complex traits, yet raises a challenge for statistical inference in terms of how information from different types of molecules can be fused to enhance the power of discovering genomic regions (especially genes) responsible for the trait variations. In this talk we present one component of InSIGHT – the integrated inference of all measured cis elements around genes for their associations with a quantitative trait. We will discuss a novel procedure called truncated aggregation of P values (TAP) test, with a description of the methodology, a real data example, and if time permits, some simulation
Generalized Scale-change Models for Recurrent Event Processes Under Informative Censoring

► Sy Han Chiou, The University of Texas at Dallas

Session: IP58

Two major challenges arise in regression analyses of recurrent event data: firstly, regression models may not fit the data adequately and therefore not characterize covariate effects properly under the applied models; secondly, the censoring time remains informative about the risk of experiencing recurrent events after accounting for covariates. We tackle both challenges by a general class of semiparametric scale-change models that allows a scale-change covariate effect and a multiplicative covariate effect. The model is flexible and nests a number of existing models, including the popular proportional rates model, the accelerated mean model, and the accelerated rate model. Moreover, it accommodates informative censoring through a subject-level latent frailty whose distribution is left unspecified. A robust approach is proposed to estimate the model parameters, which does not need parametric assumption on the distribution of the frailty and the recurrent event process. The asymptotic properties of the resulting estimator are established, with the asymptotic variance estimated from a novel resampling approach. The structure of the model permits model selection among the submodels via hypothesis testing of model parameters. Numerical studies show that the proposed estimator and the model selection procedure perform well under both noninformative and informative censoring scenarios. The methods are applied to data from two transplant cohorts to study the risk of infections after transplantation.

Understanding ICH E17 and Multi-Regional Clinical Trials (MRCTs)

► Michael W. Chiu, Cytel

Session: IP3

Globalization of drug development has increased the use multi-regional clinical trials (MRCTs) for regulatory submission in ICH regions and non-ICH regions. Regulatory authorities often face challenges in evaluating data from MRCT for drug approval. These challenges could be due to operating and scientific difficulties or sometimes conflicting requirements, there are no harmonized guidelines until recently International Conference for Harmonization (ICH) has developed and issued a new E17 guideline for MRCTs. The main objective of this guideline is to provide common principles in planning and designing MRCTs and to minimize potential conflicting advice provided by regulatory agencies on the trial design. With proper execution it is anticipated, MRCT may facilitate more efficient drug development and improve the chance of market approval in different regional regulatory authorities, thus reducing time lag of drug launch and allowing earlier access to innovative therapies for patients. This session will provide an overview of the guideline and introduce key statistical considerations, planning strategy and potential impact.

Computing conditional density of eigenvalues in high-dimension

► Yunjin Choi, National University of Singapore

Session: IP23

We propose a method for evaluating conditional density of eigenvalues of a Wishart matrix in high-dimension. Evaluating the density of eigenvalues involve multi-dimensional integration, while multi-dimensional integration can be computationally challenging especially in high-dimensional setting. Johnstone (2001) addressed this issue by utilizing approximation of a random matrix kernel and proposed a method for evaluating the marginal distribution of the largest eigenvalue of a Wishart matrix. We extend this approach and propose a method for evaluating the conditional distribution of any k-th eigenvalue with its preceding eigenvalues conditioned. The proposed method can be used for testing the significance of the principal components.
Assessing Agreement with Functional Data

▶ Pankaj Choudhary, University of Texas at Dallas

Session: IP8

We discuss a statistical methodology for modeling and analysis of paired functional data arising in method comparison studies. The observed data consist of repeated measurements of a continuous variable made using two methods of measurement on a sample of subjects. The data are viewed as two smooth curves per subject that are observed with noise at a common set of discrete time points which may vary from subject to subject. The methodology uses functional principal components analysis within the framework of mixed-effects models to represent each curve in terms of a small number of principal components. The dependence in the two curves of a subject is modeled using common principal component scores for the curves. Penalized splines are used for estimation of mean and covariance functions. Bootstrapping is employed to obtain estimates of bias and covariance matrix of model parameter estimates. These in turn are used to compute relevant confidence intervals for parameters and functions thereof such as measures of similarity and agreement between the measurement methods. The performance of the methodology is evaluated using simulation. Its application is illustrated using a core body temperature dataset from the literature.

Non-inferiority Testing in Three-arm Trials for Binary Outcome with Application in Depression Study

▶ Shrabanti Chowdhuri, Icahn School of Medicine at Mount Sinai

Session: IP3

In comparative effectiveness research (CER) one of the main goals is to study comparative benefits of competing interventions. Non-inferiority (NI) trials are of great importance for CER when one cannot guarantee superiority of one active treatment over the other. Such trials aim to demonstrate that an experimental treatment is non-inferior to an existing clinically proven active comparator by not more than a pre-specified margin. The inclusion of a placebo arm giving rise to the three-arm trial is prudent, if ethically reasonable, since it requires less stringent assumptions as compared to the two-arm placebo-free NI trial. We consider fraction margin approach where NI margin is formulated as a pre-determined negative fraction of the unknown effect size of reference drug in current trial. While risk difference (RD) is the most common and well explored functional form for testing efficacy (or effectiveness), however, recent FDA guideline suggested measures such as relative risk (RR), odds ratio (OR), among others, on the basis of which NI test can be performed for binary outcome. Bayesian paradigm provides a natural path to integrate historical and current trials’ as well as uses patients’ and clinicians’ opinions as prior information which may greatly reduce the cost burden on current trial in terms of effective sample size. We propose novel Frequentist approach incorporating the condition of assay sensitivity (AS) as well as Bayesian procedures for testing NI in the three-arm trial with binary end points considering RD, RR and OR as the functions of interest. In addition, we discuss sample size calculation which could be readily used while designing such trials in practice. To illustrate our proposed approach we analyze one clinical trial dataset from mental health depression study where the primary end point was the change in HAMD-17 score measured at the end of six weeks from baseline.

Evaluating the statistical properties of Bayesian basket trial designs

▶ Kristen Cunanan, Memorial Sloan Kettering Cancer Center

Session: IP39

Recent oncology trials are dominated with the development of targeted therapies. In certain cancer populations, these alternative treatment options have initially been shown to provide more effective and durable tumor responses with more favorable safety profiles, as compared to conventional approaches such as chemotherapy. But, how can we identify these populations early on? In addressing this, creative yet complex oncology clinical trial designs have emerged. One such class of designs has been termed “basket trials”, whereby treatment allocation is biomarker-driven rather than disease-driven. In the simple basket trial setting, patients are enrolled into a trial if they have the required genomic mutation and then placed into baskets based
on their anatomical diseases. The mutations are often infrequent and result in small sample sizes in each basket, especially for rare cancers. Consequently, implementing independent designs for each basket in parallel is not always feasible and often, not an optimal design for the setting. Depending on previous regulatory approval in other disease indications, investigators may be inclined to expect broad efficacy across all baskets at the onset of a trial; however, since these trials involve multiple disease populations, investigators can expect some heterogeneity in responses across baskets. A Bayesian adaptive design is an appealing approach for a basket trial to capitalize on the expected correlated efficacies between baskets, while screening out futile baskets. This strategy can potentially improve power and trial efficiency, as compared to independent designs. Adaptive basket trial designs using Bayesian hierarchical modeling and Bayesian mixture modeling have been proposed with varying degrees of model complexity. In our work, we investigate the impact of model complexity for different hierarchical and mixture models on the operating characteristics of the design for multiple clinical settings. In preliminary work, we found as the number of baskets and the heterogeneity between baskets increases, some methods perform poorly: resulting in large false positive rates in declaring the drug is effective in futile baskets. In this talk we present our findings from this investigation of potential gains of such complexities and delineate when they are needed.

ARMA Cholesky factor models for the covariance matrix

► Michael Daniels, University of Florida

Session: IP17

In longitudinal studies, serial dependence of repeated outcomes must be taken into account to make correct inferences on covariate effects. As such, care must be taken in modeling the covariance matrix. However, estimation of the covariance matrix is challenging because there are many parameters in the matrix and the estimated covariance matrix should be positive definite. To overcome these limitations, two Cholesky decomposition approaches have been proposed: modified Cholesky decomposition for autoregressive (AR) structure and moving average Cholesky decomposition for moving average (MA) structure, respectively. Unfortunately, the correlations of repeated outcomes are often not captured parsimoniously using either approach separately. In this paper, we propose a class of flexible, nonstationary, heteroscedastic models that exploits the structure allowed by combining the AR and MA modeling of the covariance matrix that we denote as ARMACD. We analyze a recent lung cancer study to illustrate the power of our proposed methods.

Horseshoe Regularization for Feature Subset Selection

► Jyotishka Datta, University of Arkansas

Session: IP14

Feature subset selection arises in many high-dimensional applications of statistics, such as compressed sensing and genomics. The \( \ell_0 \) penalty is ideal for this task, the caveat being it requires the NP-hard combinatorial evaluation of all models. A recent area of considerable interest is to develop efficient algorithms to fit models with a non-convex \( \ell_\gamma \) penalty for \( \gamma \in (0, 1) \), which results in sparser models than the convex \( \ell_1 \) or lasso penalty, but is harder to fit. We propose an alternative, termed the horseshoe regularization penalty for feature subset selection, and demonstrate its theoretical and computational advantages. The distinguishing feature from existing non-convex optimization approaches is a full probabilistic representation of the penalty as the negative of the logarithm of a suitable prior, which in turn enables an efficient expectation-maximization algorithm for optimization and MCMC for uncertainty quantification. In synthetic and real data, the resulting algorithm provides better statistical performance, and the computation requires a fraction of time of state of the art non-convex solvers.

Robust Regression Analysis of Temporal Data under Censoring

► Somnath Datta, University of Florida

Session: IP44
We consider regression analysis of temporal data when the temporal correlation is modeled by an autoregressive process. Robust R estimators of regression and autoregressive parameters are obtained. Our estimators are valid under censoring caused by detection limits. Efficient computation of the estimators is discussed. Theoretical and simulation studies of the estimators are presented. We analyze a real data set on air pollution using our methodology.

Robustness and semiparametric estimators in causal inference

► Xavier de Luna, Umea University

Session: IP24

Semiparametric inference for causal parameters (effect of a treatment say) is based on models where some nuisance components of the data generating mechanism (DGM) are not of interest and therefore not restricted. Sometimes, though, some of the nuisance components (e.g., working model for treatment assignment mechanism) are specified through parametric models. Semiparametric inference is not necessarily robust to model misspecification of such nuisance components. Moreover, one may also be interested in robustness to situations where some observations are not generated by the DGM of interest but comes from some nuisance distribution. Classical semiparametric inference is not robust to such contamination, and a single observation may have an unbounded influence on semiparametric estimators, such as inverse probability weighting, outcome regression and double robust estimators. Moreover, because such estimators typically overweight some observations, the effect of contamination may be devastating. In this talk, we introduce semiparametric estimators of causal effects, which are robust to contamination in the sense that their influence function is bounded to contamination of a single observation. We give asymptotic properties and study finite sample behavior through simulations. In particular, we contrast the consequences of model misspecification and contamination for different class of estimators. One interesting aspect is that the auxiliary models used to adjust for confounding are also useful to protect against contamination. We also illustrate through a case study that both adjustment to observed confounding (using auxiliary models) and protection against contamination of the sample are achieved through weighting schemes, and that these weights can be contrasted to gain further insights.

Vaccine efficacy trial design for emerging infectious disease threats

► Natalie E. Dean, University of Florida

Session: IP62

The 2013-2016 West African Ebola epidemic was an unprecedented public health emergency. The large scale of the outbreak provided a rare opportunity to assess the field efficacy of Ebola vaccines, but the phase III trials planned faced challenges in achieving study power. A trial in Guinea successfully demonstrated high efficacy of the rVSV Ebola vaccine using an innovative ring vaccination design that adaptively followed the epidemic as cases developed. For ethical reasons, this trial used a novel delayed vaccination comparator. We describe a statistical framework for predicting bias and power for the per protocol analysis of a vaccine efficacy trial with or without a delayed vaccination arm. We focus on outbreak settings where incidence may quickly wane, negatively impacting power. We further discuss adaptive trial designs for pathogens causing outbreaks with unpredictable timing and duration. We advise prospectively planning to combine efficacy data across multiple outbreaks. We provide recommendations for the design of Lassa fever and plague vaccine trials. This work is motivated by the World Health Organization’s R&D Blueprint for Action to Prevent Epidemics.

Time Depending Shape Regression with Application to Bioshape Analysis

► Yifang Deng, Florida State University

Session: IP57
A random object \( X \) is a random point on a complete metric space \((\mathcal{M}, \rho)\). Important examples of objects are shapes of configurations extracted from digital images, or from medical imaging outputs. For such data, the associated objects considered are points on the so called Kendall shape spaces, on affine shape spaces or on projective shape spaces. Other examples of object spaces are axial spaces, or spaces of directions. The afore mentioned object spaces have a structure of compact spaces, therefore if we consider an object space \( \mathcal{M} \) provided with a “chord” distance \( \rho_j \) associated to an embedding \( j : \mathcal{M} \rightarrow \mathbb{R}^N \), \( \rho_j(p_1, p_2) = ||j(p_1) - j(p_2)|| \), the statistical analysis performed relative to \( \rho_j \) distance is termed extrinsic data analysis. The expected square distance from the random object \( X \) to an arbitrary point \( p \) defines the so called Fréchet function associated with \( X \): \( F_j(p) = \mathbb{E} ||j(X) - j(p)||^2 \). The set of maximizers of the Fréchet function, is called the extrinsic antimean set. In case the extrinsic antimean set has one point only, that point is called extrinsic antimean of \( X \) and is labeled \( \alpha E_j(X) \). Given a pair of random objects \((Y, X)\) on a product of object spaces \( \mathcal{N} \times \mathcal{M} \), and an embedding \( j \) of \( \mathcal{M} \) as above, the antiregression function \( f_j : \mathcal{N} \rightarrow \mathcal{M} \) is defined by the conditional extrinsic antimean formula \( f_j(y) = \alpha E_j(X|Y = y) \). Here we give necessary and sufficient conditions that insure that the antiregression function is well defined, and an example of extrinsic antiregression in shape analysis of a clamshells species found in the Florida panhandle, where the predictor is age and the response is 3D projective shape.

Treatment Effects on Ordinal Outcomes: Causal Estimands and Sharp Bounds

**Peng Ding**, University of California, Berkeley

Session: IP41

Under the potential outcomes framework, we can define causal effects as comparisons between the potential outcomes under treatment and control. However, unfortunately, the average causal effect, often the parameter of interest, is difficult to interpret for ordinal outcomes. To address this challenge, we propose to use two causal parameters, which are defined as the probabilities that the treatment is beneficial and strictly beneficial for the experimental units. However, although well-defined for any outcomes and of particular interest for ordinal outcomes, the two aforementioned parameters depend on the association between the potential outcomes, and are therefore not identifiable from the observed data without additional assumptions. We present the sharp bounds of the aforementioned causal parameters for ordinal outcomes, under fixed marginal distributions of the potential outcomes. We also extend the discussion to the relative treatment effect, which has been widely used in biostatistics.

Likelihood ratio tests and confidence intervals based on the shape constraint of concavity

**Charles R Doss**, University of Minnesota

Session: IP5

We consider estimation and inference for a log-concave density and for a concave regression function. These problems have some similarities because they both rely on an underlying shape constraint of concavity. Forming confidence intervals or hypothesis tests in nonparametric settings is often challenging. We propose using likelihood ratio statistics to form hypothesis tests (which can be inverted to form confidence intervals). We consider tests or intervals for the location of the mode of the concave function and for the value of the concave function. The statistics we propose are tuning parameter free, a rarity in nonparametric settings. We demonstrate that the likelihood ratio statistics are asymptotically pivotal (satisfy the so-called Wilks phenomenon). Thus, they have universal critical values not depending on any unknown parameters, allowing the tests or intervals to be computed in practice.

Bayesian activation region detection in fMRI studies

**Somak Dutta**, Iowa State University

Session: IP38

Functional Magnetic Resonance Imaging (fMRI) has been widely adopted as a non-invasive technique for
understanding spatial localization of neural associates of human cognitive and motor functions. In this talk we propose a whole-brain 3D analysis for identifying regions with high neural activities associated with external stimuli. We follow a hierarchical Bayesian route with a three states 3D Potts model on the latent activation classes and a spike-and-two-slabs prior on the activation signals. We incorporate the prior information on activation proportion in the Potts prior by suitably selecting the hyper parameters. Keeping in mind the critical slowdown of Gibbs sampler in simulating from Potts model, we adopt the Swendsen-Wang algorithm for sampling from the Potts prior and partial decoupling algorithm for sampling from the posterior distribution of the activation class indicators. Via extensive simulation studies we also demonstrate the robustness of our method near the critical temperature of the Potts prior model. We illustrate our methodology on a sport imagination experiment and an audio-visual experiment.

On Perfect Classification and Clustering for Gaussian Processes

▶ **Subhajit Dutta**, IIT Kanpur

Session: IP34

We first study the problem of discriminating Gaussian processes by analyzing the behavior of the underlying probability measures in an infinite-dimensional space. Motivated by singularity of a class of Gaussian measures, we propose a joint transformation and investigate its theoretical properties. In a classification problem, this transformation induces complete separation and the misclassification probability of a simple component-wise classifier used on this transformed data asymptotically converges to zero (i.e., perfect classification). In finite samples, the empirical classifier is constructed and related properties are stated. In the second part of the talk, we shall discuss the problem of clustering for Gaussian processes. Using this transformation, we develop a clustering algorithm along with a procedure for consistently estimating the number of clusters.

Individualized Treatment Recommendation (ITR) for Survival Outcomes

▶ **Haoda Fu**, Eli Lilly and Company

Session: IP36

ITR is a method to recommend treatment based on individual patient characteristics to maximize clinical benefit. During the past a few years, we have developed and published methods on this topic with various applications including comprehensive search algorithms, tree methods, benefit risk algorithm, multiple treatment & multiple ordinal treatment algorithms. In this talk, we propose a new ITR method to handle survival outcomes for multiple treatments. This new model enjoy the following practical and theoretical features. Instead of fitting the data, our method directly search the optimal treatment policy which improves the efficiency. To adjust censoring, we propose a doubly robust estimator. Our method only requires either censoring model or survival model is correct, but not both. When both are correct, our method enjoys better efficiency. Our method handles multiple treatments with intuitive geometry explanations. Our method is Fisher's consistent even under either censoring model or survival model misspecification (but not both). This method has potential applications in multiple therapeutic areas. One direct impact for Diabetes business unit is that how we can leverage Lilly Diabetes’ broad treatment options to reduce or delay diabetes comorbidities such as CV event, diabetes related retinopathy, nephropathy, or neuropathy.

Inference for variable clustering under correlation-like similarities

▶ **Maxwell G'Sell**, Carnegie Mellon University

Session: IP12

Clustering is often applied to detect dependence structure among the variables in large data sets. However, it is typically difficult to determine the appropriate amount of clustering to carry out in a given application. We will take a selective inference approach to testing of hierarchical clustering of variables based on measures of their correlation. We will see that this yields reasonable goodness-of-fit stopping rules for selecting the number of clusters. We will consider extending weakening the required assumptions and generalizing the measure of correlation, and the computational issues that arise in this pursuit.
A Bayesian Nonparametric Model for Predicting Pregnancy Outcomes Using Longitudinal Profiles

▶ Jeremy Gaskins, University of Louisville

Session: IP20

Across several medical fields, developing an approach for disease classification is an important challenge. The usual procedure is to fit a model for the longitudinal response in the healthy population, a different model for the longitudinal response in the disease population, and then apply the Bayes’ theorem to obtain disease probabilities given the responses. Unfortunately, when substantial heterogeneity exists within each population, this type of Bayes classification may perform poorly. In this paper, we develop a new approach by fitting a Bayesian nonparametric model for the joint outcome of disease status and longitudinal response, and then use the clustering induced by the Dirichlet process in our model to increase the flexibility of the method, allowing for multiple subpopulations of healthy, diseased, and possibly mixed membership. In addition, we introduce an MCMC sampling scheme that facilitates the assessment of the inference and prediction capabilities of our model. Finally, we demonstrate the method by predicting pregnancy outcomes using longitudinal profiles on the $\beta$–HCG hormone levels in a sample of Chilean women being treated with assisted reproductive therapy.

Robust Bayesian Model Averaging

▶ Joyee Ghosh, The University of Iowa

Session: IP14

The majority of Bayesian variable selection methods/algorithms for linear regression have focused on normal errors, which is a venerable problem in its own right, when the number of variables exceeds the sample size. Since estimates obtained under the normality assumption can be sensitive to outliers, robustifying the error distribution may be of interest, especially in high dimensions, when standard model diagnostics do not work well. The Bayesian variable selection approach can handle an unknown degree of sparsity by placing a prior on the inclusion probability of variables. In this work, we develop Bayesian models that allow additional flexibility by letting the likelihood depend on an unknown degree of tail heaviness. We compare and contrast the results with those obtained under a traditional model with normal errors.

Bivariate Beta and Kumaraswamy Models developed using the Arnold-NG bivariate beta distribution

▶ Indranil Ghosh, University of North Carolina, Wilmington

Session: IP50

In this paper we explore some mechanisms for constructing bivariate and multivariate beta and Kumaraswamy distributions. Specifically, we focus our attention on the Arnold-Ng (2011) eight parameter bivariate beta model. Several models in the literature are identified as special cases of this distribution including the Jones-Okin-Liu-Libby-Novick bivariate beta model, and certain Kotz and Nadarajah bivariate beta models among others. The utility of such models in constructing bivariate Kumaraswamy models is investigated. Structural properties of such derived models are studied. Parameter estimation for the models is also discussed. For illustrative purposes, a real life data set is considered to exhibit the applicability of these models in comparison with rival bivariate beta and Kumaraswamy models.

Range of the transient “frog” random walk model on $\mathbb{Z}$

▶ Arka Ghosh, Iowa State University

Session: IP47
We observe the frog model, an infinite system of interacting random walks, on $\mathbb{Z}$ (integers) with an asymmetric underlying random walk. For certain initial frog distributions, we construct an explicit formula for the moments of the leftmost visited site, as well as their asymptotic scaling limits as the drift of the underlying random walk vanishes. We also provide conditions in which the lower bound can be scaled to converge in probability to the degenerate distribution at 1 as the drift vanishes.

Two manifestations of rigidity in point processes: forbidden regions and maximal degeneracy

► Subhroshekhar Ghosh, National University of Singapore

Session: IP47

A point process is said to be “rigid” if its local observables are completely determined (as deterministic functions) of the point configuration outside a local neighborhood. For example, it has been shown in earlier work that, in the Ginibre ensemble (a.k.a. the 2D Coulomb gas at inverse temperature $\beta = 2$), the point configuration outside any bounded domain determines, almost surely, the number of points in the domain. In this talk, we will explore two recent manifestations of such rigidity phenomena. For the zeros of the planar Gaussian analytic function, we show that outside every large “hole”, there is a “forbidden region” which contains a vanishing density of points. This should be seen in contrast with the corresponding situation for classically understood models (e.g. random matrix ensembles), where no such effects are known to occur. In the second part of the talk, we will consider “stealthy” hyperuniform systems, which are systems whose structure function (i.e., the Fourier transform of the two-point correlation) vanishes near the origin. We show that such systems exhibit “maximal degeneracy”, namely the points outside a bounded domain determine, almost surely, the entire point configuration inside the domain. En route, we establish a conjecture of Zhang, Stillinger and Torquato that such systems have (deterministically) bounded holes. Based on joint works with Joel Lebowitz and Alon Nishry.

Adaptive elastic net for group testing

► Karl Gregory, University of South Carolina

Session: IP35

For disease screening, group (pooled) testing can be a cost-saving alternative to one-at-a-time testing, with savings being realized from assaying pooled biospecimen (e.g. urine, blood, saliva) from multiple individuals at once. It is often of interest to relate individuals’ true disease statuses to covariate information with a binary regression model. Several authors have developed regression methods for group testing data, which is challenging because individuals’ true statuses are never observed; due to imperfect testing, all testing outcomes (on pools and individuals) are subject to misclassification. To further complicate matters, individuals may be involved in several testing outcomes. For the analysis of such data, we provide a methodology which generalizes and extends the aforementioned regression techniques and which incorporates regularization. Specifically, for model fitting and variable selection, we propose an adaptive elastic net estimator under the logistic regression model which can be used to analyze data arising from any group testing strategy. We provide an efficient algorithm for computing the estimator along with guidance on tuning parameter selection. Moreover, we establish the asymptotic properties of the proposed estimator in the group testing setting and show that it possesses oracle properties. We evaluate the finite-sample performance of the estimator under data-based tuning parameter selection through Monte Carlo studies and illustrate the methodology in an analysis of chlamydia data collected by the State Hygienic Laboratory in Iowa City.

Sparse Penalized Quantile Regression: Method, Theory, and Algorithm

► Yuwen Gu, University of Connecticut

Session: IP52

Sparse penalized quantile regression is a useful tool for variable selection, robust estimation, and het-
eroscedasticity detection in high-dimensional data analysis. We discuss the variable selection and estimation properties of the lasso and folded concave penalized quantile regression via non-asymptotic arguments. We also consider consistent parameter tuning therein. The computational issue of the sparse penalized quantile regression has not yet been fully resolved in the literature, due to non-smoothness of the quantile regression loss function. We introduce fast alternating direction method of multipliers (ADMM) algorithms for computing the sparse penalized quantile regression. Numerical examples demonstrate the competitive performance of our algorithm: it significantly outperforms several other fast solvers for high-dimensional penalized quantile regression.

Scalable Bayesian methods for high-dimensional inverse problems

► Nilabja Guha, University of Massachusetts at Lowell

Session: IP31

We discuss inverse problems arising in porous media flow, where the quantity of interest can be an unknown high contrast high dimensional spatial parameter. Estimating such parameters based on observed data and underlying PDE model constitutes an inverse problem. We propose scalable methods for inverse problem and uncertainty quantification for such a class of problems.

Bayesian Nonparametric Differential Analysis for Dependent Multigroup Data with Application to DNA Methylation Analyses in Cancer

► Subharup Guha, University of Florida

Session: IP31

Cancer `omics datasets are characterized by widely varying sizes and scales, measurement variables, and correlation structures. An overarching scientific goal in cancer research is to develop general statistical techniques that can cleanly sift the signal from the noise in identifying genomic signatures of the disease across a set of experimental or biological conditions. We propose BayesDiff, a nonparametric Bayesian approach based on a novel class of first order mixture models, called the Sticky Poisson-Dirichlet process or multicuisine restaurant franchise. The BayesDiff methodology flexibly utilizes information from all the measurements and adaptively accommodates any serial dependence in the data, accounting for the inter-probe distances, to perform simultaneous inferences on the variables. The technique is applied to analyze the motivating DNA methylation colorectal cancer dataset, which displays both serial correlations and complex interaction patterns. In simulation studies, we demonstrate the effectiveness of the BayesDiff procedure relative to existing techniques for differential DNA methylation. Returning to the motivating dataset, we detect the genomic signature for multiple subtypes of colorectal cancer. The analysis results support and complement known features of DNA methylation as well as gene association with colorectal cancer.

A Bayesian Nonparametric Spiked Process Prior for Dynamic Model Selection

► Michele Guindani, University of California, Irvine

Session: IP23

In many applications, investigators consider processes that vary in space and time, with the goal of identifying temporally persistent and spatially localized departures of those processes from a baseline or “normal” state. In this talk, I will discuss how to cast the identification problem in the context of a Bayesian nonparametric model selection approach for the analysis of spatio-temporal data, which takes into account the non-exchangeable nature of measurements collected over time and space. I will outline the model and discuss its performances by means of an application to a disease surveillance problem, for detecting outbreaks of pneumonia and influenza mortality in the continental United States.

Distributional Independent Component Analysis for Diverse Neuroimaging Modalities

► Ying Guo, Emory University
Session: IP21

Recent advances in neuroimaging technologies such as MRI have provided opportunities to acquire brain images of different modalities for studying human brain organization from both functional as well as structural perspectives. Analysis of images derived from various modalities involves some common goals such as dimension reduction, denoising, and feature extraction. However, since these modalities have vastly different data characteristics, current analysis is usually performed using distinct analytical tools that are only suitable for a specific imaging modality. In this talk, we present a novel Distributional Independent Component Analysis (DICA) that could provide a general method for decomposing data derived from diverse neuroimaging modalities including functional MRI (fMRI) and diffusion tensor imaging (DTI). Unlike classical ICA which separates observed data as a mixture of independent components, the proposed DICA represents a fundamentally different approach that aims to perform ICA on the distribution level. The connection and distinction between the classical ICA and DICA will be discussed. The DICA aims to provide a unified framework for extracting features across imaging modalities that have different scales, representations and variability. We perform statistical inference on the model parameters under Bayesian framework and develop a posterior computation algorithm based on Markov chain Monte Carlo (MCMC) method. Theoretical properties related to DICA and its estimation are established. We evaluate the performance of the DICA as compared with classical ICA methods through extensive simulation studies. For real-world applications, we apply DICA to a multimodal neuroimaging study to investigate functional networks as well as white matter structural networks in the brain.

Estimating and Interpreting Treatment Effect in Oncology Trials with Immune Therapies

Susan Halabi, Duke University

Session: IP42

It is standard in a phase III trial to use one measure, the hazard ratio, to estimate the treatment effect for time-to-event endpoints. The hazard ratio is a reasonable metric when the proportional hazards assumption holds. In trials with immunotherapies a delayed treatment effect leads to non-proportional hazards (NPH). In this talk I will bring up some of the statistical challenges in summarizing clinical benefit and interpreting data from trials with NPH. Several potential metrics are available that estimate the treatment effect in the setting of NPH, including weighted hazard ratios, piecewise hazard ratios, and restricted mean survival time (RMST). I will discuss the advantages and disadvantages of the above mentioned measures. In addition, real life examples will be used to demonstrate that one metric may be inadequate to estimate clinical benefit.

Generative Adversarial Networks for Data Augmentation: Promise and Pitfalls

Lawrence O. Hall, University of South Florida

Session: IP45

Deep learning has been shown to be a powerful approach for building predictive models from images. There are many examples, especially in medical domains, where the number of labeled data examples is not large enough to effectively leverage deep learning models. This has led to efforts to use images with noise, that are rotated, that are flipped to augment training sets. Augmentation has been shown to enable more accurate classifiers to be built. We discuss using generative adversarial networks to generate images that may be used for augmentation. These deep networks attempt to create images that are indistinguishable from some population of images. Of course, if the images are exact copies augmentation is unlikely to help. So, we discuss ways that you may generate useful images for augmentation as well as pitfalls with examples.

Propensity score calipers and the overlap condition

Ben Hansen, University of Michigan

Session: IP24
Propensity scores (Rosenbaum and Rubin, 1983) are used widely to address measured confounding in quasi-experiments. They also arise in connection with the antecedent question of whether non-equivalent treatment and control groups are suitable for comparison at all, with or without covariate adjustments. “Common support”, the assumption that propensity scores are bounded away from 1, is so named because it means that the support of the treatment group in covariate spaces is contained within that of the control group. This is less simple to check than is often supposed: even if treatment and control groups’ values of the true propensity score overlap, when arranged in order of estimated propensity scores they may appear not to. The naive method of discarding those members of the treatment group whose estimated propensity scores fall above all the controls, and those members of the control group whose propensity scores fall below all those estimated within the treatment group, is needlessly wasteful of sample size. It is possible to address common support by restricting the range of the estimated propensity score within which comparisons are permitted, but this requires careful determination of the tolerance enforced for matching discrepancies; available heuristics and guidelines attend only to some of the issues that must be considered. I present a new formula for determining caliper widths for matching based on propensity scores, and other constructed index functions. The method is compatible with conventional means of propensity score estimation, although it relaxes some of the more tenuous of the conventional assumptions, in particular permitting the dimension of the parameter to grow with $n$.

Using Wavelets to Discover Relationships Among Tree-Ring Records

Megan B. Heyman, Rose-Hulman Institute of Technology

Session: IP61

Tree-rings provide a proxy measurement for yearly climactic information such as precipitation or temperature. This is because, as trees grow, their trunk diameter increases, and in a typical year a tree-ring is produced. The width of this ring reflects growing conditions during the year, and when standardized, a wider ring indicates better growing conditions. Tree-ring records exist for thousands of years in many locations across the earth. Dendrologists use these records to understand past climate. Although tree-ring records are related over location and time, many explorations of the relationships among these records tend to focus on either single element. In this talk, we present a novel way to utilize the discrete wavelet transformation in modeling spatial and temporal tree-ring record relationships simultaneously. This model will be demonstrated on subsets of records from the international tree-ring data bank (ITRDB). We find significant relationships with respect to distance but also differing over various time scales.

Asymptotically Stable Drift and Minorization for Markov Chains with Application to Albert and Chib’s Algorithm

James P. Hobert, University of Florida

Session: IP26

Recent work suggests that, while often useful for establishing convergence rates when $n$ and $p$ are fixed, techniques based on drift and minorization conditions may not be versatile enough to handle the more delicate task of convergence complexity analysis. In this talk, general methods of sharpening drift and minorization conditions are introduced, and an application of these ideas to Albert and Chib’s (1995, JASA) data augmentation algorithm for the Bayesian probit model is described.

Mining new information in medical imaging

Scott S. Hsieh, University of California, Los Angeles

Session: IP67

Modern medical imaging modalities have enabled precise capture of the anatomical characteristics of disease. In addition to qualitative impressions from radiological images, such as the presence or absence of
a tumor, in many cases it is possible to extract a quantitative biomarker that can help diagnose a pathology or signal its likely prognosis. We will provide examples of biomarkers already accepted for use today, such as RA950 for emphysema or coronary artery calcium for cardiovascular disease, and continue by exploring emerging biomarkers that are still in the process of validation. These include "radiomics" features that may be connected to the underlying genetic profile and features that arise from convolutional neural networks which may be interpretable using an expert-in-the-loop framework. Improvements in biomarkers require understanding the underlying disease process and careful statistical analysis but have the potential to transform patient management.

**Analyzing Matched Sets of Microbiome Data**

**Yijuan Hu**, Emory University

Session: IP16

Matched data arise frequently in microbiome studies. For example, the gut microbiome data may be collected pre- and post-treatment from a set of individuals, or the vaginal microbiome samples may be collected longitudinally in each trimester of a pregnancy. We recently developed the Linear Decomposition Model (LDM), a powerful and flexible method that integrates the global testing of any microbiome effect and the detection of differentially abundant OTUs. Here we extend the LDM for analyzing matched sets of microbiome data. In this analysis, the microbiome characterizing the set is treated as a ‘nuisance parameter’, allowing all effort to focus on the (common) differences within sets. We compare the power of the matched analysis with that of the standard (unmatched) analysis using the LDM as well as other existing methods.

**Real Time PM$_{2.5}$ Mapping and Anomaly Detection from AirBoxes in Taiwan**

**Guowen Huang**, National Tsing Hua University

Session: IP13

Fine particulate matter (PM$_{2.5}$) has gained increasing attention due to its adverse health effects to human. In Taiwan, it was conventionally monitored by large environmental monitoring stations of the Environmental Protection Administration. However, only a small number of 77 monitoring stations are currently established. Recently, a project using a large number of small devices, called AirBoxes, was launched in March 2016 to monitor PM$_{2.5}$ concentrations. Although thousands of AirBoxes have been deployed across Taiwan to give a broader coverage, they are mostly located in big cities, and their measurements are less accurate. In this paper, we propose to use a spatial statistics technique, called kriging, which gives PM$_{2.5}$ level with its standard error at any location in Taiwan by incorporating spatial dependence structure of AirBox data in a statistically optimal way. This provides a smoothly varied real-time PM$_{2.5}$ concentration map and its associated standard error map. In addition, we develop a novel spatio-temporal control chart that monitors anomalous measurements by utilizing neighboring AirBox information. Our method automatically adaptive to different neighboring structures at different AirBox locations without the need to specify a neighborhood range. The proposed method has abilities to detect potential emission sources, malfunctioned AirBoxes, and AirBoxes that are wrongly put indoors.

**Meta-analytic and Integrative Framework for Sparse k-means to Identify Disease Subtypes**

**Zhiguang Huo**, University of Florida

Session: IP6

Disease phenotyping by omics data has become a popular approach that potentially can lead to better personalized treatment. Identifying disease subtypes via unsupervised machine learning is the first step towards this goal. With the accumulation of massive high-throughput omics data sets, omics data integration is essential to improve statistical power and reproducibility. In this talk, two extensions from sparse $K$-means method will be introduced. The first extension is towards a meta-analytic framework to identify novel disease subtypes...
when expression profiles of multiple cohorts are available. The lasso regularization and meta-analysis identify a unique set of gene features for subtype characterization. An additional pattern matching reward function guarantees consistent subtype signatures across studies. The second extension is towards integrating multi-level omics datasets with the guidance of prior biological knowledge via sparse overlapping group lasso. An algorithm using alternating direction method of multiplier (ADMM) will be applied for fast optimization. Simulation and real applications in breast cancer and leukemia will show the superior clustering accuracy, feature selection, functional annotation and computing efficiency.

A machine learning approach to prognostic and predictive covariate identification for subgroup analysis

**David A. James**, Novartis

Session: IP36

We illustrate the use of machine learning techniques to identify prognostic and predictive features to characterize potential subgroups of patients with higher risk for various endpoints and for discriminating differential treatment effects among patients in morbidity and mortality (M&M) trials. Exploratory analysis techniques include the use of “cognostics” (Tukey 1965, Wilkinson, Anand, and Grossman 2005, Guha et al 2010) to rank order large number of visual displays to efficiently explore relatively large number of candidate features; various recursive partitioning trees, including survival trees (Breiman et al 1984, LeBlanc and Crowley 1992, Therneau and Atkins 1997), random survival trees (Breiman 2001, Ishwaran et al 2008), and model-based partitioning trees (Seibold, Zeiles, and Hothorn 2014). We conclude with lessons learned from these in-depth analyzes and highlight their strengths and weaknesses vis-a-vis traditional statistical techniques, such as Cox proportional hazard regression, logistic regressions, generalized additive models, etc.

Component-wise Markov chain Monte Carlo

**Galin Jones**, University of Minnesota

Session: IP26

It is common practice in Markov chain Monte Carlo to update the simulation one variable (or sub-block of variables) at a time, rather than conduct a single full-dimensional update. When it is possible to draw from each full-conditional distribution associated with the target this is just a Gibbs sampler. Often at least one of the Gibbs updates is replaced with a Metropolis-Hastings step, yielding a Metropolis-Hastings-within-Gibbs algorithm. Strategies for combining component-wise updates include composition, random sequence and random scans. While these strategies can ease MCMC implementation and produce superior empirical performance compared to full-dimensional updates, the theoretical convergence properties of the associated Markov chains have received limited attention. We pay particular attention to the connections between the convergence rates of the various component-wise strategies. This is important since it ensures the existence of tools that an MCMC practitioner can use to be as confident in the simulation results as if they were based on independent and identically distributed samples. We illustrate our results in two examples.

Global multivariate point pattern models for rain type occurrence

**Mikyoung Jun**, Texas A&M University

Session: IP32

We seek statistical methods to study the occurrence of multiple rain types observed by satellite on a global scale. The main scientific interests are to relate rainfall occurrence with various atmospheric state variables and to study the dependence between the occurrences of multiple types of rainfall (e.g. short-lived and intense versus long-lived and weak; the heights of the rain clouds are also considered). Commonly in point process model literature, the spatial domain is assumed to be a small, and thus planar domain. We consider the log-Gaussian Cox Process (LGCP) models on the surface of a sphere and take advantage of cross-covariance models for spatial processes on a global scale to model the stochastic intensity function of the
LGCP models. We present analysis results for rainfall observations from the TRMM satellite and atmospheric state variables from MERRA-2 reanalysis data over the tropical Eastern and Western Pacific Ocean, as well as over the entire tropical and subtropical ocean regions. Statistical inference is done through Monte Carlo likelihood approximation for LGCP models. We employ covariance approximation to deal with massive data.

**Novel response adaptive allocation in randomized factorial designs: a case study**

► John Kairalla, University of Florida

Session: IP65

Response adaptive randomization uses observed treatment outcomes from preceding participants to change allocation probabilities. Traditionally, the strategy can fulfill the ethical desire to increase the likelihood of giving an individual the best-known treatment at the time of randomization. In a multi-arm clinical trial setting with ordered testing priorities, novel response adaptive allocation methods may allow for more flexibility and efficiency with respect to information allocation decisions made during study accrual. We will review two such novel response adaptive allocation designs recently funded by the NIH that are currently in early accrual phases. Both are multi-stage $2 \times 2$ factorial designs with fixed total sample size. In one, studying biopsychosocial influence on shoulder pain, the primary hypothesis is tested at both the interim and final stages. The other, studying augmented cognitive training in older adults, involves interim testing for a secondary hypothesis to go along with allocation decisions. Study operating characteristics will be extensively explored, summarized, and compared to alternatives, with recommendations for improvements to the designs given.

**Scalar-on-Image Regression via the Soft-Thresholded Gaussian Process**

► Jian Kang, University of Michigan

Session: IP63

This work concerns spatial variable selection for scalar-on-image regression. We propose a new class of Bayesian nonparametric models and develop an efficient posterior computational algorithm. The proposed soft-thresholded Gaussian process provides large prior support over the class of piecewise-smooth, sparse, and continuous spatially-varying regression coefficient functions. In addition, under some mild regularity conditions the soft-thresholded Gaussian process prior leads to the posterior consistency for parameter estimation and variable selection for scalar-on-image regression, even when the number of predictors is larger than the sample size. The proposed method is compared to alternatives via simulation and applied to an electroencephalography study of alcoholism.

**Errors-in-Variables Jump Regression Using Unsupervised Learning**

► Yicheng Kang, Bentley University

Session: IP27

Errors-in-variables regression is widely used in econometric and financial models. The statistical analysis becomes challenging when the regression function is discontinuous and the distribution of measurement error is unknown. In this talk, we propose a novel jump-preserving curve estimation method. A major feature of our method is that it can remove the noise effectively while preserving the jumps well, without requiring much prior knowledge about the measurement error distribution. The jump-preserving property is achieved mainly by unsupervised learning. We show that the proposed curve estimator is statistically consistent, and it performs favorably, in comparison with an existing jump-preserving estimator. Finally, we demonstrate our method by an application to a health insurance tax policy study in Australia.

**A Bayesian approach for joint estimation of multiple networks**

► Kshitij Khare, University of Florida

Session: IP17
In this paper, we develop a novel Bayesian approach for joint estimation of multiple graphical models. This problem arises in many applications, such as understanding co-expression networks from high-dimensional Omics data obtained from different biological conditions or disease subtypes. We pursue a pseudo-likelihood based approach which provides robustness and computational efficiency. We illustrate the efficacy of our approach using simulated and real datasets. This is joint work with George Michailidis and Peyman Jalali.

Bayesian longitudinal causal inference in the analysis of public health impact of air pollution

► Chanmin Kim, Boston University

Session: IP37

Pollutant emissions from coal-burning power plants have been deemed to adversely impact ambient air quality and public health conditions. Despite the noticeable reduction in emissions and the improvement of air quality since the Clean Air Act (CAA) became the law, the public-health benefits from changes in emissions have not been widely evaluated yet. In terms of the chain of accountability (HEI Accountability Working Group, 2003), the link between pollutant emissions from the power plants (PM$_{2.5}$) and public health conditions (cardiovascular and respiratory diseases) with counting for changes in ambient air quality (PM$_{2.5}$), we provide the first assessment of the longitudinal effect of specific pollutant emission (SO$_2$) on public health outcome that is mediated through changes in the ambient air quality. It is of particular interest to examine the extent to which the effect that is mediated through changes in local ambient air quality differs from year to year. In this paper, we propose a Bayesian approach to estimate time-varying mediation effects in the presence of mediators and responses measured every year. We replace the commonly invoked sequential ignorability assumption with a new set of assumptions which are sufficient to estimate the natural indirect and direct effects in this setting. Also, a Bayesian updating method is used to minimize modeling assumptions.

A Model for Large Multivariate Datasets

► William Kleiber, University of Colorado at Boulder

Session: IP32

Multivariate spatial modeling is a rapidly growing field, but most extant models are infeasible for use with large spatial datasets. In this talk we introduce a highly flexible, interpretable and scalable multiresolution approach to multivariate spatial modeling. Relying on compactly supported basis functions and Gaussian Markov random field specifications for coefficients results in efficient and scalable calculation routines for likelihood evaluations and co-kriging. We analytically show that special parameterizations approximate popular existing models. Moreover, the specification allows for control over the spectral coherence between component fields. The model is illustrated through Monte Carlo studies and on a complex large bivariate observational minimum and maximum temperature dataset over the western United States.

Modeling the Cholesky factors of Covariances in Multivariate Longitudinal Data

► Priya Kohli, Connecticut College

Session: IP17

Modeling the covariance matrix of multivariate longitudinal data is more challenging as compared to its univariate counterpart due to the presence of correlations among multiple responses. The high-dimensionality problem is also exacerbated as the number of parameters grows quadratically with both the number of variables and repeated time points. Graphical tools to visualize dependence patterns may also involve a multitude of graphs. Therefore, few, if any, graphical methods have been explored for multivariate longitudinal data. The modified Cholesky block decomposition reduces the task of covariance modeling into parsimonious modeling of its two matrix factors: the regression coefficient matrices and the innovation covariance matrices. These parameters are statistically interpretable, however ensuring positive-definiteness of several (innovation) covariance matrices presents itself as a new challenge. In this work, we overcome some of the challenges by modeling the innovation covariance matrices using a slight modification of the linear covariance models.
and writing them as linear combinations of known positive-definite basis matrices with unknown non-negative scalar coefficients. A novelty of this approach is that positive-definiteness is guaranteed by construction; it removes a drawback of Anderson's model and hence makes linear covariance models more realistic and viable in practice. Another major contribution of this work is the extension of regressograms for visualization of the dynamic association patterns present in the multivariate longitudinal studies. Multivariate longitudinal studies are quite common in clinical trials and biological research. Studying the joint evolution and dependence patterns among health outcomes of patients over time is of interest in making recommendations for the treatment and follow-up of patients to guide the development of health promoting interventions. We illustrate the applicability of the proposed approach in such studies.

Rank-Based Procedures in Factorial Designs: Hypotheses about Nonparametric Treatment Effects

► Frank Konietschke, University of Texas at Dallas

Session: IP58

Existing tests for factorial designs in the nonparametric case are based on hypotheses formulated in terms of distribution functions. Typical null hypotheses, however, are formulated in terms of some parameters or effect measures, particularly in heteroscedastic settings. In this talk we extend this idea to nonparametric models by introducing a novel nonparametric ANOVA-type-statistic based on ranks which is suitable for testing hypotheses formulated in meaningful nonparametric treatment effects in general factorial designs. This is achieved by a careful in-depth study of the common distribution of rank-based estimators for the treatment effects. Since the statistic is asymptotically not a pivotal quantity we propose different approximation techniques, discuss their theoretic properties and compare them in extensive simulations together with two additional Wald-type tests. An extension of the presented idea to general repeated measures designs is briefly outlined. The proposed rank-based procedures maintain the pre-assigned type-I error rate quite accurately, also in unbalanced and heteroscedastic models. A real data example illustrates the application of the proposed methods.

Scalable Bayesian Variable Selection for Structured High Dimensional Data

► Suprateek Kundu, Emory University

Session: IP14

Variable selection for structured covariates lying on an underlying known graph is a problem motivated by practical applications, and has been a topic of increasing interest. However, most of the existing methods may not be scalable to high dimensional settings involving tens of thousands of variables lying on known pathways such as the case in genomics studies. We propose an adaptive Bayesian shrinkage approach which incorporates prior network information by smoothing the shrinkage parameters for connected variables in the graph, so that the corresponding coefficients have a similar degree of shrinkage. We fit our model via a computationally efficient expectation maximization algorithm which scalable to high dimensional settings. Theoretical properties for fixed as well as increasing dimensions are established, even when the number of variables increases faster than the sample size. We demonstrate the advantages of our approach in terms of variable selection, prediction, and computational scalability via a simulation study, and apply the method to a cancer genomics study.

Joint Modeling of Multiple Skewed Longitudinal Processes with Excess of Zeroes and Time-to-event Using a Shared Parameter Model

► Subrata Kundu, George Washington University

Session: IP18

In this talk, we proposes a joint modeling approach to study the association between the two longitudinal processes, the covariates and the time to event. A mixed effects model with two correlated random effects
was proposed to model the longitudinal processes and a discrete hazard model with a log-log link to model the time to event. This model was used to analyze the Oxford Conception Study and to observe the patterns in menstrual cycle lengths and their association with time to pregnancy.

**Composite responder rate estimation under non-ignorable missingness**

▶ Madan G Kundu, AbbVie Inc.

Session: IP53

Longitudinal changes in a population of interest are often heterogeneous and may be influenced by missing data that are potential source of bias in clinical trials including the trials evaluating (composite) responder rate based on one or more measurements at a particular visit. Composite response criteria are useful when therapeutic benefit is determined by simultaneous improvement on more than one characteristics and are in use in many therapeutic area including irritable bowel syndrome, Myelofibrosis and rheumatoid arthritis. The drug regulatory agencies often recommend to adopt missing-not-at-random (MNAR) mechanism for handling missing data. Existing likelihood based MNAR missing data methods (e.g., selection model or pattern mixture model) make some distributional assumption such as multi-variate normal across visits and the resulting conclusion can be very sensitive to this assumption. Therefore, we propose an intuitive and straightforward strategy to estimate responder rate under MNAR missingness without making any distributional assumption. By not using distributional assumption our method avoids bias due to mis-specification of distribution across visits. Our method falls in the class of inverse-probability-weighted (IPW) estimation. However, unlike IPW approach, the proposed approach proceeds by estimating the responder rates separately among completers and dropped-out patients rather than computing weights for individual completers. The proposed method arguably is more intuitive and requires less technical sophistication compared to existing approaches and also less computation intensive. The comparative performance of proposed estimator was evaluated through simulation and the method was applied to clinical trial data comparing composite responder rates of two treatments.

**Radiologic Image-based Statistical Shape Analysis of Brain Tumors**

▶ Sebastian A. Kurtek, The Ohio State University

Session: IP13

We propose a curve-based Riemannian-geometric approach for general shape-based statistical analyses of tumors obtained from radiologic images. A key component of the framework is a suitable metric that (1) enables comparisons of tumor shapes, (2) provides tools for computing descriptive statistics and implementing principal component analysis on the space of tumor shapes, and (3) allows for a rich class of continuous deformations of a tumor shape. The utility of the framework is illustrated through specific statistical tasks on a dataset of radiologic images of patients diagnosed with glioblastoma multiforme, a malignant brain tumor with poor prognosis. In particular, our analysis discovers two patient clusters with very different survival, subtype and genomic characteristics. Furthermore, it is demonstrated that adding tumor shape information into survival models containing clinical and genomic variables results in a significant increase in predictive power.

**Bayesian Low-Rank Graph Regression Models for Mapping Human Connectome Data**

▶ Eunjee Lee, University of Michigan

Session: IP67

We propose a Bayesian low-rank graphic regression model (BLGRM) framework for the regression analysis of matrix response data across subjects. This work is motivated by detailed comparisons of functional and structural connectivity data across subjects, groups, and time and relating neural connections to particular behavioral measures. The BLGRM can be regarded as a novel integration of principal component analysis, tensor decomposition, and regression models. In BLGRM, we find a common low-dimensional subspace for efficiently representing all matrix responses. Based on such low-dimensional representation, we can easily
quantify the effects of various predictors of interest, such as age and diagnosis, and then perform regression analysis in the common subspace, leading to both substantial dimension reduction and much better prediction. We adapt a parameter expansion approach to our graphic regression model to address weak identifiability and high posterior dependence among parameters in our decomposition model. Posterior computation proceeds via an efficient Markov chain Monte Carlo algorithm. A simulation study is performed to evaluate the finite sample performance of BLGRM and to compare it with several competing approaches. We apply BLGRM to the resting state functional magnetic resonance imaging data set obtained from the Alzheimer’s Disease Neuroimaging Initiative.

Nonparametric graphical models: Theories and Applications

► Kuang-Yao Lee, Temple University

Session: IP58

With the advance of high-throughput technologies, massive and complex data are routinely collected and these data need to be processed and analyzed differently from conventional data. In this presentation I will discuss a nascent concept for analyzing statistical networks—additive conditional independence (ACI)—a three-way statistical relation that shares many similarities with conditional independence. However, its nonparametric characterization does not involve multivariate kernel, which enjoys the flexibility of nonparametric estimators but avoids the curse of dimensionality in high-dimensional settings. We facilitate the implementation of ACI via a case study on nonparametric graphical models, and describe a general framework for adopting ACI to a broader scope. Additionally, to emphasize the increasing impact of ACI we also introduce several recent developments under various statistical settings. We investigate the properties of the proposed estimators through both theoretical and simulation analyses. The usefulness of our procedures is also demonstrated through an application to gene regulatory network (GRN) inference using a DREAM Challenge dataset. This is joint work with Bing Li (Penn State), Hongyu Zhao (Yale), Lexin Li (UC Berkeley).

High dimensional general linear hypothesis with non-linear spectral shrinkage

► Haoran Li, University of California, Davis

Session: IP19

In recent years, extensive work has been done for hypothesis testing on mean vectors in high dimensional regimes, especially for two-sample problems. In this paper, we are interested in general linear hypotheses under multivariate linear regression model, which includes many existing hypotheses about mean vectors as its special cases. We propose a family of rotation-equivariant tests with a class of analytic functions as spectral shrinkage. Asymptotic normality is derived under finite moments conditions. The asymptotic local power of the test is studied under various probabilistic prior models of local alternatives. As an application of the theory, we construct a test with ridge-regularization and suggest possible extensions to higher-order approximation. A simulation study is carried out to examine the numerical performance of the test in MANOVA set-up and compare with other MANOVA tests in the literature.

Integrative multi-view reduced-rank regression: bridging group-sparse and low-rank models.

► Gen Li, Columbia University

Session: IP36

In this talk, I will introduce a novel integrated reduced-rank framework for multivariate regression. Predictors are multi-view data, which naturally form different groups. Each predictor group has its unique low-rank coefficient matrix. The framework flexibly captures the relationship between multivariate responses and predictors, and subsumes many existing methods such as reduced rank regression and group lasso as special cases. We develop an efficient alternating direction method of multipliers (ADMM) algorithm for model fitting, and ex-
exploit a majorization approach to deal with binary responses or missing values in responses. We demonstrate the efficacy of the proposed methods with simulation studies and a real application to the Longitudinal Study of Aging.

Conditional Networks: A New Framework for Integrative Analyses

▶ Yi Li, University of Michigan

Session: IP63

Understanding how genes interplay with each other and how their regulations are associated with other high-dimensional genomic markers may uncover the underlying mechanism of disease progression processes. Graphical models have commonly been used in simultaneously learning the response network and the associations between the response, but these models assume a homogeneous population and ignore the heterogeneity between individual-level networks. In this talk, we propose a conditional graphical model with functional precision parameters. We propose a Fisher scoring matching approach for variable selection and network recovery. We show that the proposed method can consistently select important predictors and recover the response network structure. The proposed method is computationally inexpensive and can be directly applied to analyzing “omic” scaled networks and DNA data, such as the cancer genome atlas (TCGA) data, to study cancer-triggering biological pathways.

Standardization of spatial Gaussian mixture models and background adjustment of PET images in brain oncology

▶ Meng Li, Rice University

Session: IP21

In brain oncology, it is routine to evaluate the progress or remission of the disease based on the differences between a pre-treatment and a post-treatment Positron Emission Tomography (PET) scan. Background adjustment is necessary to reduce confounding by tissue-dependent changes not related to the disease. When modeling the voxel intensities for the two scans as a bivariate Gaussian mixture, background adjustment translates into standardizing the mixture at each voxel, while tumor lesions present themselves as outliers to be detected. In this talk, we address the question of how to standardize the mixture to a standard multivariate normal distribution, so that the outliers (i.e., tumor lesions) can be detected using a statistical test. To address standardization in spatially heterogeneous image data, we propose a spatial and robust multivariate expectation-maximization (EM) algorithm, where prior class membership probabilities are provided by transformation of spatial probability template maps and the estimation of the class mean and covariances are robust to outliers. The proposed methods are illustrated through simulations and real data applications.

Sampling error in meta-analysis with small sample sizes

▶ Lifeng Lin, Florida State University

Session: IP1

Meta-analyses frequently include studies with small sample sizes. Researchers usually account for sampling error in meta-analyses by modeling the observed study-specific effect sizes with their estimated within-study variances, but these variances are traditionally treated as the truths, so the sampling error in the variances is ignored. However, this sampling error may be considerable when sample sizes are small. We illustrate a type of bias in meta-analysis results completely due to sampling error, which has received relatively less attention compared with publication bias or selective reporting. This bias is basically caused by the intrinsic association between the observed effect sizes and their within-study variances. We conducted extensive simulation studies to investigate this bias. Sampling error did not cause bias when the effect size is mean difference, but standardized mean difference, odds ratio, risk ratio, and risk difference suffered from this bias to different extents. Also, surprisingly, Hedges’ g, which is a bias-corrected estimate of standardized mean difference within studies, might lead to larger bias than Cohen’s d in meta-analyses. Cautions are needed to perform
Bayesian Regression Tree Ensembles that Adapt to Smoothness and Sparsity

▶ Antonio R. Linero, Florida State University

Session: IP46

Ensembles of decision trees are a useful tool for obtaining flexible estimates of regression functions. Examples of these methods include gradient boosted decision trees, random forests, and Bayesian CART. Two potential shortcomings of tree ensembles are their lack of smoothness and vulnerability to the curse of dimensionality. We show that these issues can be overcome by instead considering sparsity inducing soft decision trees in which the decisions are treated as probabilistic. We implement this in the context of the Bayesian additive regression trees framework, and illustrate its promising performance through testing on benchmark datasets. We provide strong theoretical support for our methodology by showing that the posterior distribution concentrates at the minimax rate (up-to a logarithmic factor) for sparse functions and functions with additive structures in the high-dimensional regime where the dimensionality of the covariate space is allowed to grow near exponentially in the sample size. Our method also adapts to the unknown smoothness and sparsity levels, and can be implemented by making minimal modifications to existing BART algorithms.

Bootstrapping Spectral Statistics in High Dimensions

▶ Miles Lopes, University of California, Davis

Session: IP19

Spectral statistics play a central role in many multivariate testing problems. It is therefore of interest to approximate the distribution of functions of the eigenvalues of sample covariance matrices. Although bootstrap methods are an established approach to approximating the laws of spectral statistics in low-dimensional problems, these methods are relatively unexplored in the high-dimensional setting. The aim of this paper is to focus on linear spectral statistics (LSS) as a class of "prototype statistics" for developing a new bootstrap method in the high-dimensional setting. In essence, the method originates from the parametric bootstrap, and is motivated by the notion that, in high dimensions, it is difficult to obtain a non-parametric approximation to the full data-generating distribution. From a practical standpoint, the method is easy to use, and allows the user to circumvent the difficulties of complex asymptotic formulas for LSS. In addition to proving the consistency of the proposed method, we provide encouraging empirical results in a variety of settings. Lastly, and perhaps most interestingly, we show through simulations that the method can be applied successfully to statistics outside the class of LSS, such as the largest sample eigenvalue and others.

Multiple Changepoint Detection in Climate Time Series

▶ Robert B. Lund, Clemson University

Session: IP61

This talk presents methods to estimate the number of changepoint time(s) and their locations in time-ordered data sequences when prior information is known about some of the changepoint times. A Bayesian version of a penalized likelihood objective function is developed from minimum description length (MDL) information theory principles. Optimizing the objective function yields estimates of the changepoint number(s) and location(s). Our MDL penalty depends on where the changepoint(s) lie, but not solely on the total number of changepoints (such as classical AIC and BIC penalties). The techniques allow for autocorrelation in the observations and mean shifts at each changepoint time. This scenario arises in climate time series where a "metadata" record exists documenting some, but not necessarily all, of station move times and instrumentation changes. Applications to climate time series are presented throughout.

On order determination using augmentation predictor

▶ Wei Luo, Baruch College
In many statistical dimension reduction problems, including principal component analysis, canonical correlation analysis, independent component analysis, and sufficient dimension reduction, etc., it is often of interest to determine the rank of a matrix parameter based on a consistent matrix estimator. In this paper, we propose a method called the augmentation estimator for this purpose, with the aid of an augmentation predictor that is artificially generated and merged with the original predictor. Similar to the ladle estimator (Luo and Li, 2016), the augmentation estimator uses information from both the eigenvalues and the eigenvectors of the matrix estimator. Compared with the existing order-determination methods, it is easy to implement, computationally efficient, consistent under general conditions, and applicable in high-dimensional cases. Its effectiveness is supported by simulation studies. The way we employ the augmentation predictor is novel, which may inspire independent research interest.

Aympirical Method: A new paradigm for statistical analysis for samples

Ping Ma, University of Georgia

Traditional statistical theory and methods are developed for small and mild size samples. In particular, statistical model fitting and inference are conducted in the small and mild size samples to get empirical results. Asymptotic theory is established to extrapolate the performance of the empirical results to large samples. However, this traditional coherent statistical analysis paradigm falls apart in large samples. The key challenge is that many traditional statistical methods are computational too expensive to get meaningful empirical results. New statistical paradigm is in urgent need for the statistical analysis in large samples. In this talk, I will present an asympirical (asymptotic + empirical) method, which is designed by the principle that theory informs practice. I will present it in the context of smoothing spline ANOVA models. Simulation and real data analysis will be used to demonstrate the performance of the new paradigm.

Bias Reduction in Logistic Regression with Missing Responses when the Missing Data Mechanism is Nonignorable

Arnab K. Maity, Texas A&M University

In logistic regression with nonignorable missing responses, Ibrahim and Lipsitz (1996) proposed a method for estimating regression parameters. It is known that the regression estimates obtained by using this method are biased when the sample size is small. Also, another complexity arises when the iterative estimation process encounters separation in estimating regression coefficients. In this article we propose a method to improve the estimation of regression coefficients. In our likelihood based method, we penalize the likelihood by multiplying it by a non-informative Jeffreys prior as a penalty term. The proposed method reduces bias and is able to handle the issue of separation. Simulation results show substantial bias reduction for the proposed method as compared to the existing method. Analyses using real world data also support the simulation findings. An R package called brlrmr is developed implementing the proposed method and the Ibrahim and Lipsitz method.

Ultrahigh-dimensional Robust and Efficient Sparse Regression using Non-Concave Penalized Density Power Divergence

Subhabrata Majumdar, University of Florida

We propose a sparse regression method based on the non-concave penalized density power divergence loss function which is robust against infinitesimal contamination in very high dimensionality. Present methods of sparse and robust regression are based on $\ell_1$-penalization, and their theoretical properties are not well-investigated. In contrast, we use a general class of folded concave penalties that ensure sparse recovery and...
consistent estimation of regression coefficients. We propose an alternating algorithm based on the Concave-Convex procedure to obtain our estimate, and demonstrate its robustness properties using influence function analysis. Under some conditions on the fixed design matrix and penalty function, we prove that this estimator possesses large-sample oracle properties in an ultrahigh-dimensional regime. The performance and effectiveness of our proposed method for parameter estimation and prediction compared to state-of-the-art are demonstrated through simulation studies.

Bayesian variational approaches for inverse problems

➤ Bani Mallick, Texas A&M University

Session: IP4

We consider a Bayesian approach to inverse problems with complex error structure. A hierarchical Bayesian model is developed in this inverse problem setup. The Bayesian approach contains a natural mechanism for regularization in the form of a prior distribution. Different regularized prior distribution has been used to strongly induce sparseness. We propose a variational type algorithm by minimizing the Kullback-Liebler divergence between the true posterior and a separable approximated one. The proposed method is illustrated on several two dimensional linear and nonlinear inverse problems.

A Modeling Approach for Predicting Disease Status Using Functional Data in the Absence of a Gold Standard

➤ Amita Manatunga, Emory University

Session: IP13

Data from clinical studies involving risk predictions provide a wealth opportunities for statistical research in particular to the development of prediction models and their evaluations. I will introduce a special clinical decision making problem in nuclear medicine, present its statistical challenges and discuss some potential solutions. In our study, two consecutive curves over time are observed per subject and in some cases, the second curve for some subjects are not observed. There is no gold standard for determining disease status, instead, the ratings for disease status from multiple experts are available. We consider a latent class modeling approach for predicting disease status of a subject based on observed functional data and its ratings from multiple experts. I will present our work including the modeling procedure, prediction models consisting of several prediction schemes, and their evaluation via simulation studies. I will demonstrate the practicality of our method and will show that proposed modeling procedure reasonably captures the patterns of observed curves and provide sensible clinical interpretations. I will conclude with a brief discussion of future work.

Robust Methods in Small Area Estimation

➤ Abhyuday Mandal, University of Georgia

Session: IP39

Modern societies have an ever-increasing appetite for reliable and up to date data to make informed decisions in both public and private sectors alike. While censuses, usually conducted once in a decade, provide reliable information about the population across various geography and demography, such information quickly get outdated each passing year after a census. To obtain a current picture of a population under study, suitable surveys are conducted to collect data from only a fraction of the population. Due to budget constraints, these surveys are inherently limited in size. While information gained from such surveys may be adequate for the entire population, the same data is often inadequately small when it is sliced and diced across geographic and demographic sub-populations. These sub-populations are termed small areas. National Statistical Offices around the world have been mandated for many years to produce reliable small area statistics for many important variables such as population, income, unemployment, health outcomes, etc. Statistical summaries based on traditional direct estimates, computed using only sample data from individual small areas, are usu-
ally very unreliable. In small area estimation, by borrowing strength from other data sources, appropriate statistical methodologies have been developed to improve on the traditional direct estimates. In this talk, we propose new alternatives to some popular models in small area estimation. Model-based small area estimates are developed by shrinking direct estimates to suitable regression synthetic estimates, generated from the regression model for small area population means. Our new models are based on finite mixture of normal distributions. We implement our models using a hierarchical Bayesian approach. This talk is based on collaboration with Gauri Sankar Datta and Adrijio Chakraborty.

A New Test for Two-Sample Location Problem Based on Empirical Distribution Functions

► Sunil Mathur, Texas A&M University-Corpus Christi

Session: IP11

Ranked-set sampling (RSS) is used to get the better representation of the population for the variable of interest and RSS has been shown to result in improved statistical inference as compared to simple random sampling (SRS). A few methods have been developed based on RSS by modifying the existing statistical tests. In this paper, we propose a new test based on empirical distribution functions for two-sample location problem for ranked set samples. The test statistic is constructed based on divergence between the empirical distribution functions obtained from two samples. We show that the proposed test is distribution-free. The quantiles of the proposed test under null distribution are provided. We empirically show that the sampling distribution of the proposed test in the RSS case is stochastically smaller than the SRS case. Monte Carlo power simulations study shows that the proposed test performs better than its competitor. Application of the proposed test is presented using real public health big data.

Computational Discrete Mathematical Optimization meets Structured Sparsity

► Rahul Mazumder, MIT Sloan School of Management

Session: IP34

Nonconvex problems arise frequently in modern applied statistics and machine learning, posing outstanding challenges from a computational and statistical viewpoint. Continuous especially convex optimization, has played a key role in our computational understanding of (relaxations or approximations of) these problems. However, some other well-grounded techniques in mathematical optimization (for example, mixed integer optimization) have not been explored to their fullest potential. When the underlying statistical problem becomes difficult, simple convex relaxations and/or greedy methods have shortcomings. Fortunately, many of these can be ameliorated by using estimators that can be posed as solutions to structured discrete optimization problems. To this end, I will demonstrate how techniques in modern computational mathematical optimization (especially, discrete optimization) can be used to address the canonical problem of best-subset selection and cousins. I will describe how recent algorithms based on local combinatorial optimization can lead to high quality solutions in times comparable to (or even faster than) the fastest algorithms based on \( L_1 \)-regularization. I will also discuss the relatively less understood low Signal to Noise ratio regime, where usual subset selection performs unfavorably from a statistical viewpoint; and propose simple alternatives that rely on nonconvex optimization. If time permits, I will outline problems arising in the context robust statistics (least median squares/least trimmed squares), low-rank factor analysis and nonparametric function estimation where, these techniques seem to be promising.

Fitting stochastic epidemic models to incidence time series and gene genealogies

► Vladimir N. Minin, University of California, Irvine

Session: IP20

Stochastic epidemic models describe how infectious diseases spread through a population of interest. These models are constructed by first assigning individuals to compartments (e.g., susceptible, infectious, and recovered) and then defining a stochastic process that governs the evolution of sizes of these compartments through time. Here, we propose a new strategy for fitting these models to data, which turns out to be a
challenging task. The main difficulty is that even the most vigilant infectious disease surveillance programs offer only noisy snapshots of the number of infected individuals in the population. We present a Bayesian data augmentation strategy that makes statistical inference with stochastic epidemic models computationally tractable. Besides standard incidence data, our approach can also handle more exotic data types, such as genealogies/phylogenies of infectious disease agent genetic sequences collected during outbreak monitoring. We present results of using our new approach to fit stochastic epidemic models to data from outbreaks of influenza and Ebola viruses.

Assessing the distribution of discrete survival time in presence of recall error

► Sedigheh Mirzaei, National Institutes of Health

Session: IP30

Time-to-pregnancy (TTP), or the number of months non-contracepting couples require to become pregnant is a measure of human fecundity. It is an example of discrete survival time. Prospectively measured TTP is the gold standard, though few studies exist worldwide that provide such data prompting researchers to rely on retrospective data. We propose a multistage model that utilizes a woman’s retrospectively-reported TTP as well as her certainty level about it to estimate the TTP distribution. Our proposed model utilizes a discrete survival function that accounts for random heterogeneity arising from between women TTP data as well as a multinomial regression model for certainty level attached to recall that accounts for ‘memory fading with time’, i.e., depends on time since pregnancy in estimating the TTP distribution. Other novel features of the model include attention to whether the pregnancy was (un)planned as well as providing approach to predict survival function for women without a reported TTP. Our model allows for the consideration of covariate association for each of the underlying factors of (un)planned pregnancy, measure of certainty and TTP distribution. We use Monte Carlo simulations to assess the finite sample performance for the proposed estimators. We illustrate our proposed method by analyzing the Upstate KIDS Study.

Bayesian Non-parametric mixtures on directional data: Consistency and applications

► Riten Mitra, University of Louisville

Session: IP55

We provide a detailed proof of strong and weak posterior consistency in models using Matrix Langevin likelihoods and a Dirichlet Process prior. In the course of proof, we highlight the importance of the properties of the base measure in establishing sufficient conditions. We draw some comparisons across similar models developed in literature and mention possible generalizations to other mixing measures and other directed data densities. Next, we discuss briefly a posterior sampling scheme that combines a generic slice sampling method with posterior conditionals developed for the parametric version. We conclude with implications of such Bayesian models in flexible modeling of covariance matrices.

Information-theoretic observation selection strategies for linear Bayesian inverse problems

► Jayanth Jagalur Mohan, Massachusetts Institute of Technology

Session: IP4

Many inverse problems involve large data sets; yet these data are seldom equally informative, and may in fact contain redundancies. Moreover, practical constraints on storage, communication, and computational costs may limit the number of observations that one wishes to employ. We introduce information theoretic criteria and strategies for selecting subsets of the data that yield accurate approximations of the inverse solution. This goal can also be understood in terms of optimal experimental design. Our strategies exploit the structure of inverse problems in the Bayesian statistical setting leading to greedy algorithms based on modular bounds and can be efficiently implemented using low-rank approximations.
Penalized concomitant regression for microbiome data

Christian Lorenz Mueller, Simons Foundation

Session: IP33

Targeted amplicon sequencing data, such as 16S rRNA and ITS sequence data, are compositional in nature. Using these data for regression tasks is thus challenging due to the constant sum constraint. In addition, typical microbiome data are overdispersed and zero-inflated. To alleviate the challenges associated with these data, we present novel regression models for microbiome data where both the regression vector and scales are estimated concomitantly. The presented model estimation tasks admit convex optimization formulations that can be solved efficiently using proximal algorithms. We show improved prediction performance compared to state-of-the-art methods both on synthetic and real microbiome data, ranging from host-associated to environmental amplicon data.

A computationally efficient change-point based SPC chart to detect arbitrary distributional change

Partha Sarathi Mukherjee, Boise State University

Session: IP40

In the current era of computers, statistical monitoring of various types of observations is an important research area. In problems such as monitoring the quality of industrial products, health variables like blood glucose, climatological variables including temperature, precipitation, etc., we are often interested in detecting a change in the process distribution in general, not just in mean or variance. This paper proposes an efficient statistical process control (SPC) chart for Phase II monitoring of univariate continuous processes. Unlike most SPC charts in the literature, it neither assumes any in-control probability distribution nor requires any in-control Phase I data, and it aims to detect arbitrary distributional change. This chart uses a computationally efficient method to find the possible change-point which can be applied to similar problems. Moreover, the proposed chart uses the p-value based data pruning approach to further increase the efficiency. Another major contribution of this paper is that the chart combines the strengths of two different tests of hypotheses, which also has a potentially broad application. Numerical simulations, a climate data analysis, and a blood glucose monitoring data analysis show that it can be used in various process monitoring problems when the nature of distributional change is unknown.

Targeted Random Projection for Prediction from high-dimensional features

Minerva Mukhopadhyay, Duke University

Session: IP49

We consider the problem of computationally-efficient prediction from high-dimensional and highly correlated predictors in challenging settings where accurate variable selection is effectively impossible. Direct application of penalization or Bayesian methods implemented with Markov chain Monte Carlo can be computationally daunting and unstable. Hence, some type of dimensionality reduction prior to statistical analysis is in order. Common solutions include application of screening algorithms to reduce the regressors, or dimension reduction using projections of the design matrix. The former approach can be highly sensitive to threshold choice in finite samples, while the later can have poor performance in very high-dimensional settings. We propose a TArgeted Random Projection (TARP) approach that combines positive aspects of both the strategies to boost performance. In particular, we propose to use information from independent screening to order the inclusion probabilities of the features in the projection matrix used for dimension reduction, leading to data-informed sparsity. Theoretical results on the predictive accuracy of TARP is discussed in detail along with the rate of computational complexity. Simulated data examples, and real data applications are given to illustrate gains relative to a variety of competitors.
Statistical Considerations in the Sizing and Analysis of Immune-Oncology Studies

Pralay Mukhopadhyay, AstraZeneca

Session: IP42

The development of immune-oncology (IO) agents poses some unique challenges that may not be fully addressed using traditional statistical approaches. One well known problem is the anticipation of a delayed separation of Kaplan-Meier curves and the need for sample size considerations assuming non-proportional hazards (NPH). We discuss the impact of a delayed treatment effect and implications on sample size and study duration during the design stage. We provide an approach of computing sample size based on the overall hazard ratio (HR) as a weighted average of two piecewise HRs (before and after the treatment lag) using the method proposed by Schemper. Hypothesis testing with log-rank test are likely to be underpowered to detect treatment differences under NPH. We explore power of the study using the Fleming-Harrington (FH) class of weighted log-rank test (WLRT) and a new combination test based on the FH WLRT. We use simulations and case studies from two oncology trials to evaluate the pros and cons of the different testing approaches.

Measurement error correction methods to improve parameter estimates relating 2D body composition data with physical activity status

Anarina L. Murillo, University of Alabama at Birmingham

Session: IP10

Many methods are available to assess body composition. However, body fat estimates may be prone to inherent and unavoidable measurement error, which, in turn, could lead to biased estimates of the association between body fat and health outcomes. This study evaluates the performance of existing methods to correct model parameters to reduce potential biases due to measurement errors when evaluating the effects of body fat on the probability of being physically active. Cross-sectional data from 588 non-Hispanic Black and White adults aged 19-80 years were analyzed. Dual-energy X-ray absorptiometry (DXA) scans were considered the reference measure for estimating body fat percentage. A 2D photographic-based method used to estimate body fat percentage was assumed to be prone to measurement error. Three methods were applied: regression calibration (RC), simulation extrapolation (SIMEX), and multiple imputation (MI). Results showed that unadjusted body fat had upward biases of 30%. MI-corrected values yielded a 9% downward bias, RC-corrected values had a 13% upward bias, and SIMEX-corrected values had an 91% downward bias. In conclusion, this work introduced three statistical approaches to reduce potential estimation biases due to measurement errors and illustrated its value using real data. MI performed the best in comparison to RC and SIMEX methods to reduce biases due to measurement error in models. Thus, measurement error methods can improve the reliability of analyses relating body fat measures to health outcomes.

Meta analytic approaches and multiple treatments assessment in clinical trials

Saman Muthukumarana, University of Manitoba

Session: IP64

In clinical trials, Meta-analysis can be used for combining or contrasting the results from multiple studies. We develop a Bayesian approach for meta-analysis and extend the approach for binary data in the presence of excessive zeros by defining a modified unconditional odds ratio which accounts for excessive zeros. We then discuss a network meta-analysis approach which is also commonly used to incorporate direct and indirect evidence in comparing treatments. We compare the sources of inconsistency identified by our approach and existing loop-based contrast based methods using real and simulated datasets. Results from the data analyses, simulation studies, and the log-pseudo marginal likelihood (LPML) model selection procedure indicate that the proposed models perform better than conventional alternative methods.

Bayesian analysis of numerous multinomial counts for small areas with sub-areas

Balgobin Nandram, Worcester Polytechnic Institute
A four-stage hierarchical Bayesian model is used to accommodate heterogeneity for multinomial cell counts from numerous sub-areas within small areas. The first stage accommodates the sub-area multinomial counts. The second stage has a Dirichlet distribution for the cell probabilities associated with the sub-areas. The third stage has another Dirichlet distribution for the area probabilities. This is a bit awkward because the two Dirichlet distributions are not conjugate, thereby leading to difficulties with computations. The fourth stage has the prior for the hyper-parameters. We compare two computational methods, an approximate method and an exact method, to fit the model. The approximation method provides fast computation, and the exact method, which provides a check on the approximate method, will have prohibitive computational time for some data sets with numerous sub-areas. We apply our method to the Nepal Living Standards Survey that has sparse counts of household members within wards for three health status groups.

Bayesian approaches for high-dimensional quantile models
▶ Naveen Naidu Narisetty, University of Illinois at Urbana-Champaign

Quantile regression provides a more comprehensive relationship between a response and covariates of interest compared to mean regression and is especially advantageous for censored data. In this talk, I will first discuss a new Bayesian approach for censored quantile regression that can handle high dimensional covariates. Our approach uses continuous spike and slab priors with sample size dependent parameters to induce adaptive shrinkage and sparsity. A scalable Gibbs sampling algorithm for posterior computation will be presented, which has desired theoretical properties. Our theoretical results deal with the challenges raised by nonconvexity of the objective function involved.

Agreement between raters’ ordinal classifications in population-based studies
▶ Kerrie P Nelson, Boston University

Ordinal categorical scales are commonly used in screening and diagnostic tests to classify a patient’s disease status. However, discrepancies commonly observed between raters’ classifications in screening tests have motivated large-scale studies of agreement involving multiple raters to be conducted. In this talk we describe a flexible model-based approach and measure of agreement based upon the class of generalized linear mixed models to model the agreement between many raters in a unified comprehensive manner. We apply the methods to a recent large-scale cancer agreement study.

Applications of Gamma Degradation Model
▶ Hon Keung Tony Ng, Southern Methodist University

The gamma degradation model has been used to characterize the evolution of degradation measurements. In this talk, I will first provide an introduction to the gamma degradation model. Then, I will discuss an application of the gamma degradation model in biopharmaceutical statistics. The gamma degradation model is used for assessing the similarity of two drug dissolution profiles and its merits are discussed. The performances of the proposed methods are compared with a multivariate test procedure via Monte Carlo simulation studies. All the methods are illustrated with a numerical example.

Likelihood-based Classification for Fire Debris Analysis
▶ Liqiang Ni, University of Central Florida
A simple method is introduced for assessing the evidentiary value of fire debris samples. The method relies on models built by random draws from a database of ignitable liquid and substrate pyrolysis samples. The likelihood ratios are estimated by direct calculation with the assumption of multivariate between-object Gaussian kernel density distributions and rely on the mean and covariance structure resulting from each random draw. Bootstrap cross-validation is used to characterize the models, which are subsequently tested on laboratory burn samples.

**Scalable Bayesian Nonparametric Clustering and Classification with Application to Medical Records Data**

► Yang Ni, UT Austin

We develop a scalable multi-step Monte Carlo algorithm for inference under a large class of nonparametric Bayesian models for clustering and classification. Each step is “embarrassingly parallel” and can be implemented using the same Markov chain Monte Carlo sampler designed for “blocked” samples. The simplicity and generality of our approach makes a wide range of Bayesian nonparametric methods applicable to large datasets. Specifically, we apply the approach to inference under a product partition model with regression on covariates. We show results for inference with two motivating data sets: a large set of electronic health records (EHR) and a bank telemarketing dataset. We find interesting clusters and favorable classification performance relative to other widely used competing classifiers.

**A convolved subsampling approach to the block bootstrap**

► Dan Nordman, Iowa State University

The block bootstrap and subsampling are different resampling approaches for dependent data, both aiming to approximate sampling distributions by resampling data blocks. However, we may use a structural relationship to subsampling in order to characterize the block bootstrap in a new and general manner. In the important case of sample means, the block bootstrap distribution of a sample mean equals the k-fold self-convolution of a subsampling distribution. From the perspective of convolved subsampling, simple conditions are then possible for showing that a convolved subsampling (or block bootstrap) estimator validly approximates a normal target limit. Consequently, the block bootstrap for means can be established more easily, and under much weaker assumptions, than previously considered in many dependence settings. Beyond sample means, convolved subsampling does not necessarily match the block bootstrap, but instead provides an alternative hybrid-type of resampling estimator. Under minimal dependence conditions, we also justify such convolved subsampling for general statistics having normal limits.

**A Birth-death Markov chain Monte Carlo approach to modeling the ecological structure of metagenomic data**

► Jack O’Brien, Bowdoin College

Mixture models are a common approach to understanding the ecological implications of metagenomic data: they can be relatively easily interpreted, form a framework for regression, and have tractable solutions even for demandingly large data. A hard problem in this models is determining the number of components to use. In this talk, I show how a birth-death Markov chain Monte Carlo approach can be used to estimate the Bayes factors between components and present an efficient implementation that runs practically even for large datasets. Similar solutions may be extensible to regression approaches and other models.
Flexible treatment selection rules based on high dimensional data

► Todd Ogden, Columbia University

We propose a projection pursuit regression approach specifically designed to model interactions between a treatment indicator and a large number of covariates. The approach extends the modified covariate approach of Tian et al. (2014) by utilizing nonparametrically defined treatment-specific link functions. Equipped with an appropriate regularization procedure, the method is effective in selecting predictors that interact with the treatment indicator (i.e., treatment effect modifiers). The proposed method provides a valuable approach for developing a treatment decision rule based on patients’ data measured at baseline.

Estimating Gestational Age from Maternal Anthropometry

► Ana Ortega-Villa, National Institutes of Health

Determining the date of conception is important for estimating gestational age and monitoring whether the fetus and mother are on track in their development and pregnancy. In high resource countries, several methods using ultrasound have been proposed to assess gestational age at birth, however, these methods are not easily accessible for low-resource populations. We develop a shared random-parameter model for estimating the date of conception by using longitudinal assessment of multiple maternal anthropometry and cross-sectional neonatal anthropometry. In addition, we develop methodology that incorporates additional maternal information from the last menstrual period, a measure of gestational age collected with sizable measurement error. The proposed methodology is evaluated using simulation studies under a training-test set paradigm to examine the robustness of the methods to model misspecification. We illustrate this new methodology with data from the National Institute of Child Health and Development Fetal Growth Studies.

Topological Data Analysis on Object Spaces

► Robert L. Paige, Missouri University of Science & Technology

In recent years computational topology has found increasing importance in the exploratory analyses and data mining of complex data sets which are often made of high dimensional point clouds. A particularly notable example is Topological Data Analysis (TDA), a fairly new and rapidly developing area of Statistics, which uses computationally feasible persistent homology techniques to find structure in point cloud data. In this paper we elucidate the subtle topological aspects of TDA which bear upon a statistician’s understanding of what TDA is really doing and what it’s results really mean. In particular, we make clear that it is not the shape of a data set but in fact the homotopy type of the data which TDA explores. Also, we explain that the type of data structure TDA describes is more general than that of a manifold (such as in manifold learning) even more general than a CW complex structure. In fact, Topological Object Data Analysis (TODA), defined as TDA on a stratified object space, is more inclusive that classical TDA. Finally, we consider applications of TDA on the leaf data, which leads to increased statistical power when one goes from a using projective shape technique, as has been considered previously, to homotopy type data techniques of TODA via a persistent landscape analysis.

Inference for Multi-stage Destructive Cure Rate Model

► Suvra Pal, University of Texas at Arlington

In this talk, I will present a generalization of the destructive cure rate models by assuming that each com-
peting cause undergoes a destructive process for more than one time. The states of the causes, after each destructive process, are assumed to form a sequence of Markov dependent Bernoulli trials. To find the maximum likelihood estimates of the model parameters, I will discuss the steps of the expectation maximization algorithm. I will also present the results of an extensive simulation study to demonstrate the performance of the model and the proposed estimation method.

A Bayesian Framework for Modeling Data on the Stiefel Manifold

**Subhadip Pal**, University of Louisville

Session: IP55

A Bayesian framework for the Matrix Langevin distribution on the Stiefel manifold is presented. The model exploits a particular parametrization of the Matrix Langevin distribution, various aspects of which are elaborated on. A general, and novel, family of conjugate priors, and an efficient Markov chain monte carlo (MCMC) sampling scheme for the corresponding posteriors is then developed for the posterior inference. Theoretical properties of the prior and posterior distributions are explored in detail. A novel procedure for efficient computation of a class of the hypergeometric function of a matrix argument is developed to facilitate the posterior inference.

Regularized Aggregation of Statistical Parametric Maps

**Cheolwoo Park**, University of Georgia

Session: IP45

Combining statistical parametric maps (SPM) from individual subjects is the goal in some types of group-level analyses of functional magnetic resonance imaging (fMRI) data. Brain maps are usually combined using a simple average across subjects, making them susceptible to subjects with outlying values. Furthermore, t tests are prone to false positives and false negatives when outlying values are observed. We propose a regularized unsupervised aggregation method for SPMs to find an optimal weight for aggregation, which aids in detecting and mitigating the effect of outlying subjects. We also present a bootstrap-based weighted t test using the optimal weights to construct an activation map robust to outlying subjects. We validate the performance of the proposed aggregation method and test using simulated and real data examples. Results show that the regularized aggregation approach can effectively detect outlying subjects, lower their weights, and produce robust SPMs.

Tests to compare competing risks.

**Sujata Patil**, Memorial Sloan Kettering Cancer Center

Session: IP25

Competing risks occur when a patient experiencing one type of event that changes the probability of developing a second type of event. For example, it may be of interest to examine whether the incidence of death due to disease is different in two subgroups of patients, where death due to other causes is defined as a competing risk. Tests to compare the cumulative risk of one event in subgroups of patients have been developed and are commonly used. However, another hypothesis of interest is in testing whether the cumulative distributions of two competing events are different. In this talk, I will discuss previously developed methods on this problem and also share our work on a new solution.

Locating targets via wireless sensor networks

**Rohit K. Patra**, University of Florida

Session: IP5
Wireless sensor networks (WSNs) serve as key technological infrastructure for monitoring diverse systems across space and time. Examples of their widespread applications include: precision agriculture, surveillance, animal behavior, drone tracking, and emergent disaster response and recovery. A WSN consists of hundreds or thousands of identical sensors at fixed locations where each individual sensor observes the surrounding at fixed time intervals. In this work we estimate the location of a (signal emitting) target under the assumption that magnitude of signal detected at the sensor is a strictly decreasing function of the distance between the sensor and the signal emitting target. We propose an automated $\sqrt{n}$-consistent estimator of the location the target under under only the monotonicity assumption. Our estimator is tuning parameter free. We show that our estimator has a Gaussian limit distribution and construct asymptotic confidence region for the location target. This is a joint work with George Michailidis and Moulinath Banerjee.

Key aspects of high dimensional Object Data analysis

► Vic Patrangenaru, Florida State University

Session: IP46

Sample spaces of interest in Statistics are essentially infinitely dimensional. A first step in analyzing certain aspects from a sample, consists in measuring the relevant objects from it, like extracting features from images of observed individuals. Representing such objects as points on a metric space, with a sufficiently smooth high dimensional structure to allow derivation of estimators for location parameters, that lead to pivotal statistics, is the second objective. While the difficulty in such selection consists mainly in an appropriate object matching, the silver bullet of Object Data analysis, is dimension reduction to key variables relevant to certain case studies. Another useful tool in high dimensional Object Data analysis is the technique of energy statistics, that helps differentiate high dimensional distributions at a reasonable computational cost.

Estimation and clustering in a Dynamic Stochastic Block Model

► Marianna Pensky, University of Central Florida

Session: IP29

We consider a Dynamic Stochastic Block Model (DSBM) under the assumptions that the connection probabilities, as functions of time, are smooth and that only few nodes can switch their class memberships between two consecutive time points. The objective is estimation of the tensor of connection probabilities and clustering of the nodes. In particular, in the context of the DSBM, we derive a penalized least squares estimator of the tensor of connection probabilities and show that it satisfies an oracle inequality and also attains minimax lower bounds for the risk. For the purpose of clustering, we estimate the edge probability tensor by a kernel-type procedure and extract the group memberships of the nodes by spectral clustering. The procedure is computationally viable, adaptive to the unknown smoothness of the functional connection probabilities, to the rate of membership switching and to the unknown number of clusters. In addition, it is accompanied by non-asymptotic guarantees for the precision of estimation and clustering.

Fréchet estimation of dynamic covariance matrices, with application to regional myelination in the developing brain

► Alexander Petersen, University of California Santa Barbara

Session: IP56

Assessing brain development for small infants is important for determining how the human brain grows during the early period of life when the rate of brain growth is at its peak. The development of MRI techniques has enabled the quantification of brain development. A key quantity that can be extracted from MRI measurements is the level of myelination, where myelin acts as an insulator around nerve fibers and its deployment makes nerve pulse propagation more efficient. The co-variation of myelin deployment across different brain regions provides insights into the co-development of brain regions and can be assessed as correlation matrix that
varies with age. Typically, available data for each child are very sparse, due to the cost and logistic difficulties of arranging MRI brain scans for infants. We showcase here a method where data per subject are limited to measurements taken at only one random age, so that one has cross-sectional data available, while aiming at the time-varying dynamics. This situation is encountered more generally in cross-sectional studies where one observes $p$-dimensional vectors at one random time point per subject and is interested in the $p \times p$ correlation matrix function over the time domain. The challenge is that at each observation time one observes only a $p$-vector of measurements but not a covariance or correlation matrix. For such very sparse data, we develop a Fréchet estimation method. Given a metric on the space of covariance matrices, the proposed method generates a matrix function where at each time the matrix is a non-negative definite covariance matrix, for which we demonstrate consistency properties. We discuss how this approach can be applied to myelin data in the developing brain and what insights can be gained.

Integration of Methylation, Expression, and Outcome Data Motivates a New Clinical Trial in Pediatric Acute Myeloid Leukemia

► Stanley B. Pounds, St. Jude Children's Research Hospital

Session: IP22

Recent advances in biotechnology have empowered researchers to obtain multiple forms of molecular data from each of a series of tissue samples. Researchers now have the exciting opportunity and challenge of integrating multiple forms of molecular data in order to better understand biological mechanisms. We developed CC-PROMISE as a novel method that combines canonical correlation (CC) analysis and projection onto the most interesting statistical evidence (PROMISE) to integrate methylation, RNA expression, and clinical outcomes data. The CC-PROMISE method found that increased expression of the methyltransferase gene DNMT3B associated with greater total methylation across the entire genome, the RNA expression levels of thousands of genes, and worse clinical outcomes in a pediatric acute myeloid leukemia (pAML) data set. These associations were subsequently confirmed in a separate pAML cohort and motivated the administration of demethylating agents prior to standard combination chemotherapy to reduce genome-wide methylation burden to improve outcomes in a recently opened clinical trial (https://clinicaltrials.gov/show/NCT03164057). This experience indicates that using classical statistical methods in ways that optimize power to detect biologically meaningful associations can find results that may be rapidly translated into new therapeutic strategies.

On the use of cross validation for molecular classification

► Li-Xuan Qin, Memorial Sloan Kettering Cancer Center

Session: IP66

Reproducibility of scientific experimentation has become a major concern, due to the perception that many published biomedical studies cannot be replicated. In this talk we draw attention to the connection between inflated over-optimistic findings and the use of cross-validation for error estimation in molecular classification studies, in the presence of confounding handling effects in the data. We demonstrate this important yet overlooked complication of cross validation using a unique pair of datasets on the same set of tumor samples. One dataset was collected with uniform handling to prevent handling effects; the other dataset was collected without uniform handling and exhibited handling effects. The paired datasets were used to estimate the biological effects of the samples and the handling effects of the arrays in the latter dataset, which were then used to simulate data using virtual re-hybridization following various array-to-sample assignment schemes. Our study showed that (1) cross-validation tended to under-estimate the error rate when the data possessed confounding handling effects, (2) depending on the relative amount of handling effects, normalization may further worsen the under-estimation of the error rate, (3) balanced assignment of arrays to comparison groups allowed cross-validation to provide an unbiased error estimate. Our study demonstrates the benefits of balanced array assignment for reproducible molecular classification and calls for caution on the routine use of data normalization and cross-validation in such analysis.

58
Image Comparison for Online Image Monitoring

► Peihua Qiu, University of Florida

Session: IP27

In manufacturing industries, images are commonly used for quality control purposes. In such applications, if the quality of the products is good, then their images should be all similar to the image of a good-quality product. Therefore, comparison of images is a fundamental task in image-based quality control. This problem, however, is complicated in the sense that 1) observed images often contain noise, and 2) the related images need to be geometrically matched up first because images of different products could be geometrically mismatched due to the fact that the relative positions between a camera and different products are often not exactly the same. The first issue requires a statistical method that can remove noise, and the second issue is related to the so-called image registration problem in the image processing literature. In this talk, we present an effective method for detecting difference between two images of products, and our proposed method can accommodate both noise and geometric mismatch mentioned above. Theoretical results and numerical examples show that it can work effectively in applications. This is a joint research with Dr. Long Feng.

Precision Therapeutic Biomarker Identification with Application to Drug Screening for Cancer

► J. Sunil Rao, University of Miami

Session: IP11

Cancer cell lines have frequently been used to link drug sensitivity and resistance with genomic profiles. To capture genomic complexity in cancer, the Genomics in Drug Sensitivity in Cancer (GDSC) Project screened over 600 human tumor cell lines with over 130 drugs ranging from known chemotherapeutic agents to experimental compounds. Questions of interest include: i) can cancer-specific therapeutic biomarkers be detected, ii) can drug resistance patterns be identified along with predictive strategies to circumvent resistance using alternate drugs, iii) can biomarkers of drug synergies be predicted? To tackle these questions, following statistical challenges still exist: i) biomarkers cluster among the cell lines; ii) clusters can overlap (e.g. a cell line may belong to multiple clusters); iii) drugs should be modeled jointly. We introduce a multivariate regression model with a latent overlapping cluster indicator variable to address above issues. A generalized finite mixture of multivariate regression (FMMR) model in connection with the new model and a new EM algorithm for fitting are proposed. Re-analysis of the dataset sheds new light on the therapeutic inter-relationships between cancers as well existing and novel drug behaviors for the treatment and management of cancer.

Graph-based regularization for variable selection with highly-dependent covariates

► Garvesh Raskutti, University of Wisconsin-Madison

Session: IP12

Theoretical guarantees for sparse high-dimensional linear regression and inverse problems typically assume that explanatory variables are only mildly correlated. However, in modern applications ranging from functional MRI to genome-wide association studies, we observe highly correlated explanatory variables and associated design matrices that do not satisfy theoretical properties such as RIP or the restricted eigenvalue condition. Here I present an approach that involves penalization using graph-based regularizers induced by the covariance matrix of the variables. Graph-based similarity measures have led to successful regularization in the cases of the fused LASSO, edge LASSO, and graph trend filtering; however, using graph-based regularizers to develop theoretical guarantees for highly correlated covariates has not been previously examined. We show how our proposed graph-based regularization scheme interacts with correlated design matrices to yield mean-squared error guarantees for a broad range of covariance graph structures and correlation strengths which in many cases are optimal. The proposed approach outperforms other methods for highly correlated design in a variety of experiments on simulated and real fMRI data.
Debashree Ray, Johns Hopkins University

Session: IP39

Genome-wide association studies (GWAS) for complex diseases have focused primarily on single trait analyses for disease status and disease-related quantitative traits. For example, GWAS on risk factors for coronary artery disease analyze genetic associations of plasma lipids such as total cholesterol, LDL-cholesterol, HDL-cholesterol, and triglycerides separately. However, traits are often correlated and a joint analysis may yield increased statistical power for association over multiple univariate analyses. Recently several multivariate methods have been proposed which require individual-level data. Here, we develop metaUSAT, a novel unified association test of a single genetic variant with multiple traits that uses only summary statistics from existing GWAS. This novel method does not require individual-level data and can test genetic associations of binary and/or continuous traits. One can also use metaUSAT to analyze a single trait over multiple studies, appropriately accounting for overlapping samples, if any. metaUSAT provides an approximate asymptotic p-value for association and is computationally efficient for implementation at a genome-wide level. Simulation experiments show that metaUSAT maintains proper type-I error at low error levels. It has similar and sometimes greater power to detect association across a wide array of scenarios compared to existing methods, which are usually powerful for some specific association scenarios only. When applied to plasma lipids summary data from the METSIM and the T2D-GENES studies (with some individuals that are common to both), metaUSAT detected genome-wide significant loci beyond the ones identified by univariate analyses. Evidence from larger studies suggest that the variants additionally detected by metaUSAT are, indeed, associated with lipid levels in humans. In summary, metaUSAT can provide novel insights into the genetic architecture of a common disease or traits. Our software (in R) is available at https://github.com/RayDebashree/metaUSAT

Gregory Rice, University of Waterloo

Session: IP6

Most methods for analyzing functional time series rely on the estimation of lagged autocovariance operators. As in univariate time series analysis, testing whether such operators are zero is an important diagnostic step that is well understood when the data or residuals form a strong white noise. When functional data are constructed from dense records of asset prices or returns allowing for conditional heteroscedasticity is often more realistic. Applying standard inferential procedures for the autocovariance to such data often leads to the erroneous conclusion that the data exhibit significant autocorrelation. We develop methods for performing inference for lagged autocovariance operators of functional time series that are valid under general conditional heteroscedasticity conditions. These include a portmanteau test to assess the cumulative significance of empirical autocovariance operators, and methods for obtaining confidence bands for a functional version of the autocorrelation.

Davide Risso, Weill Cornell Medicine

Session: IP51

Dropout events in single-cell RNA sequencing (scRNA-seq) cause many transcripts to go undetected and induce an excess of zero read counts, leading to power issues in differential expression (DE) analysis. We present a general and flexible zero-inflated negative binomial model (ZINB-WaVE), which leads to low-dimensional representations of the data that account for zero inflation (dropouts), over-dispersion, and the count nature of the data. The model can also be used to compute observational weights to unlock bulk RNA-seq DE pipelines for zero-inflated data, boosting performance for scRNA-seq.
Use of Co-data in Early Phase Oncology Trials

Satrajit Roychoudhury, Pfizer Inc

Session: IP25

Approval of 21st century cure act has increased the value of real-world evidence (RWE) in the development of medical products. This framework allows the use of co-data, which comprises of all relevant (historical and concurrent) trial-external data, for the inference of the parameter in the actual trial. While the use of co-data in clinical trials is attractive, it is also ambitious. For example, avoiding undue weight of co-data (relative to actual trial data) is important. This can often be achieved by plausible assumptions about between-trial heterogeneity and allowance for nonexchangeability across trial parameters. A robust meta-analytic combined approach is proposed to combine data from different platform with potential heterogeneity. Two examples from early phase Oncology trials will be discussed. The first one will be a randomized Basket design in Oncology with multiple-tumor types with historical control information. The second example will include a phase I dose escalation trial for combination product with multiple regimens. Performance of the proposed model will be explored by a real life example and simulation in both examples.

Placebo Response as a Latent Characteristic: Application to Analysis of Sequential Parallel Comparison Design Studies

Denis Rybin, Pfizer Inc

Session: IP62

In clinical trials, high placebo response can affect the inference about efficacy of the studied treatment. It is important to have a robust way to classify trial subjects with respect to their response to placebo. Simple, criterion-based classification may lead to classification error and bias the inference. The uncertainty about placebo response characteristic has to be factored into the treatment effect estimation. This presentation describes a novel approach that views the placebo response as a latent characteristic and the study sample as an unlabeled mixture of ‘placebo responders’ and ‘placebo non-responders’. The likelihood-based methodology is utilized to estimate the treatment effect corrected for placebo response as defined within Sequential Parallel Comparison Design (SPCD). The performance of the novel analysis methodology is examined on example of an SPCD study.

A Divide and Conquer Strategy for High Dimensional Bayesian Factor Models

Gautam Sabnis, University of Michigan

Session: IP9

We propose a distributed computing framework, based on a divide and conquer strategy and hierarchical modeling, to accelerate posterior inference for high-dimensional Bayesian factor models. Our approach distributes the task of high-dimensional covariance matrix estimation to multiple cores, solves each subproblem separately via a latent factor model, and then combines these estimates to produce a global estimate of the covariance matrix. Existing divide and conquer methods focus exclusively on dividing the total number of observations $n$ into subsamples while keeping the dimension $p$ fixed. Our approach is novel in this regard: it includes all of the $n$ samples in each subproblem and, instead, splits the dimension $p$ into smaller subsets for each subproblem. The subproblems themselves can be challenging to solve when $p$ is large due to the dependencies across dimensions. To circumvent this issue, we specify a novel hierarchical structure on the latent factors that allows for flexible dependencies across dimensions, while still maintaining computational efficiency. Our approach is readily parallelizable and is shown to have computational efficiency of several orders of magnitude in comparison to fitting a full factor model. We report the performance of our method in synthetic examples.

Statistical Methods in Blinded Review of Safety Data
The FDA Draft Guidance for IND Safety Reporting Final Rule requires expedited reporting whenever aggregate analysis indicates events occur more frequently in the investigational drug than in a concurrent or historic control group and requires use of all available data. Suspected adverse reactions should be evaluated on an ongoing basis. The Guidance requires Sponsors to periodically review accumulating safety data, integrated across multiple studies, both completed and ongoing, provide a quantitative framework for measuring the evidence of an association for anticipated events or a clinically important increase for expected events and making a judgment about “reasonable possibility” for IND safety reporting. The FDA’s preferred approach calls for a Safety Assessment Committee (SAC) that regularly performs unblinded comparisons across treatment groups to detect numerical imbalances with appropriate steps taken to maintain overall study blinding. An alternative approach would be to only unblind if the overall rate for all treatment groups of a specific event is substantially higher than a predicted rate. This requires sponsors to pre-specify predicted rates of anticipated events and expected events and guidelines for determining when an observed rate has exceeded the predicted rate. Companies have implemented, or are in the process implementing the guidance using different novel statistical methods and along with logistical considerations requiring cross-functional collaboration. This discussion will review and compare some of the statistical methods that have been proposed in the context of blinded safety monitoring in detecting adverse event risk based on inference on blinded safety data.

MCMC methods for Bayesian semi-supervised learning

**Daniel Sanz-Alonso**, Brown University

Semi-supervised learning concerns the inference of labels on inputs by the use of two types of data: unlabelled data consisting only of inputs, and labelled data consisting of pairs of inputs and labels. In many applications, unlabelled data abounds, but labels are limited. The aim of this talk is to present new theoretical results on the Bayesian approach to semi-supervised learning that address precisely this regime of interest: fixed size of labeled data sets, large size of unlabelled data. I will first establish a statistical consistency result, and then I will propose an MCMC method whose rate of convergence is independent of the size of the unlabelled data set.

Bayesian Semiparametric Modeling of High-Dimensional Longitudinal Metabolomics Data

**Abhra Sarkar**, The University of Texas at Austin

Our research is motivated by an application to the EarlyBird cohort study aiming to explore how anthropometrics, clinical and metabolic processes are associated with obesity and glucose control during childhood. There is interest in inferring the relationship between dynamically changing and high-dimensional metabolites and a longitudinal response. Important aspects of the analysis include the selection of the important set of metabolites and accommodation of missing data in both response and covariate values. With this motivation, we propose a flexible but parsimonious Bayesian semiparametric joint model for the outcome and the covariate generating processes, making novel use of nonparametric mean processes, latent factor models and different classes of continuous shrinkage priors. The proposed approach efficiently addresses the daunting dimensionality challenges, greatly simplifies the imputation tasks, and also automates the selection of the important predictors. Implementation via efficient Markov chain Monte Carlo algorithm appropriately accounts for uncertainty in various aspects of the analysis. Simulation experiments illustrate the efficacy of the proposed methodology. Application to the EarlyBird cohort study illustrates its practical utility to enable statistical integration of different molecular processes involved in glucose production and metabolism.

Estimating disparity measures for count data
Melanocytic nevi are among the important known risk factors for melanoma, increase with age during childhood and appear to be related in part to genetics and sun exposure. A study of nevi in children demonstrated considerable disparity in nevus counts, with some children having no nevi and other having moderate to large counts. We used Bayesian hierarchical models to estimate various measures of disparity and examined the role of genetics on disparity. This talk will present these results and the properties of various measures examined.

Getting involved with biostatistical controversies: Some Examples

Jonathan Shuster, University of Florida
Session: IP43

In this talk, I review some examples of statistical controversies that can help shape your statistical practice. This is an interactive session, as the controversy will be introduced and discussed.

Multiple testing with covariates

Bodhisattva Sen, Columbia University
Session: IP5

We consider the problem of multiple testing when additional covariate information is available on each of the hypothesis tests. We propose a model for such data and develop likelihood based methods for estimating the unknown parameters. The theoretical properties of the proposed estimators are studied. We illustrate the practical efficacy of our methodology in applications in neuroscience, astronomy and genomics.

A Bayesian Mixture Model for Clustering on the Stiefel Manifold

Subhajit Sengupta, NorthShore University HealthSystem
Session: IP55

Deciphering latent structure in data is one of the fundamental challenges that the statistical and machine learning community has been grappling with in recent years. Particularly in the Big-Data driven problems, there is a potential scope to improve the statistical inference techniques if the geometry and topology of the sample space are also taken into consideration. Modeling directional data is one of the major sub-fields in modern statistics. Beyond those fashioned for simpler non-Euclidean spaces like the circle or the sphere, there is a pressing need for methodology development for general sample spaces such as the Stiefel or the Grassmann manifold to support modern applications, increasingly seen in the fields of computer vision, medical image analysis, astronomy, and biology, to name but a few. Mixture modeling techniques on general analytic manifolds have not been thoroughly explored so far and there has been only a limited amount of work on special manifolds like the Stiefel and the Grassmann manifold. In this work, using the framework discussed in we present a Bayesian inference technique of a mixture model on the Stiefel manifold that remains computationally tractable even with large data sizes. With ever-growing computational power, we argue that it is now feasible to apply Bayesian methods to real-world large and directional data. Here we have used one of the popular exponential family distributions on the Stiefel manifold, known as the Matrix Langevin (ML) or the Von-Mises Fisher matrix distribution. An efficient Markov chain Monte Carlo (MCMC) sampling scheme for the posterior distribution is developed for the mixture model. Posterior consistency for the mixture model is also discussed. In order to identify the optimum number of the cluster, we use Deviance Information Criteria for Bayesian model selection (DIC). We performed numerous experiments with several score functions for DIC apart from the standard definition of DIC. In order evaluating clustering method, we carry out extensive simulation experiments. It is important to measure the assignment of each data point to the appropriate
cluster. Note that, even if the number of clusters is right, the performance of the clustering method could be low because of incorrect cluster assignments. In order to measure clustering efficiency, we calculate several external cluster evaluation metrics like purity, normalized mutual information (NMI), rand index (RI), adjusted rand index (ARI), Jaccard index (JI) and F-measure. Finally, we apply our framework in two real-world example. A large-scale diffusion tensor imaging (DTI) dataset is analyzed to demonstrate the computational tractability of the approach. We also use the clustering framework in Near Earth Comet dataset which is one of the most commonly used datasets in this area.

Efficient community detection via network subsampling

► Srijan Sengupta, Virginia Tech
Session: IP29

Statistical network analysis is a fast growing research area with diverse applications spanning several scientific disciplines. The community structure observed in networks has been of particular interest in the statistics literature, along with the closely related task of community detection, i.e., discovering similar groups of nodes. Popular community detection methods include clustering-based methods like Spectral clustering, and modularity maximization methods like Extreme points. Such methods have excellent statistical properties but are often computationally expensive making them prohibitively costly particularly for large networks. We propose SONNET, a general subsampling strategy based on overlapping sub-networks, that is computationally scalable and can substantially speed up a remarkable range of community detection methods with little loss of statistical accuracy. We demonstrate the advantages of our method using spectral clustering, heterogeneous spectral clustering, and extreme points, on both simulated networks and benchmark network datasets like political blogs and DBLP.

A new approach to dimension reduction for multivariate time series

► Xiaofeng Shao, University of Illinois at Urbana-Champaign
Session: IP35

In this talk, we introduce a new methodology to reduce the number of parameters in multivariate time series modeling. Our method is motivated from the consideration of optimal prediction and focuses on the reduction of the effective dimension in conditional mean of time series given the past information. In particular, we seek a contemporaneous linear transformation such that the transformed time series has two parts with one part being conditionally mean independent of the past information. Our dimension reduction procedure is based on eigen-decomposition of the so-called cumulative martingale difference divergence matrix, which encodes the number and form of linear combinations that are conditional mean independent of the past. Interestingly, there is a factor model representation for our dimension reduction framework and our method can be further extended to reduce the dimension of volatility matrix. We provide a simple way of estimating the number of factors and factor loading space, and obtain some theoretical results about the estimators. The finite sample performance is examined via simulations and data analysis in comparison with some existing methods.

Annotation Regression for Genome-Wide Association Studies

► Sunyoung Shin, University of Texas at Dallas
Session: IP28

Although genome-wide association studies (GWAS) have been successful at identifying many disease-associated genetic variants, these studies are hampered by two obstacles. First, despite ever-increasing sample sizes, these studies are still underpowered for variants with weak effect sizes. Second, and more importantly, a large percentage of identified variants reside in non-coding regions, making them difficult to interpret. In this talk, I will propose a general regression framework utilizing functional annotation data in approaching the challenges. The annotation regression framework for GWAS (ARoG) is based on finite mixture of linear re-
gression models where GWAS association measures are viewed as responses and functional annotations as predictors. This mixture framework addresses heterogeneity of effects of genetic variants by grouping them into clusters and high dimensionality of the functional annotations by enabling annotation selection within each cluster. The framework will be illustrated with computational experiments and analyses of schizophrenia data from Psychiatric Genomics Consortium.

**PCA-based Two Group Comparison in Pain Response using Arterial Spin Labeling (ASL) Imaging**

▶ Haochang Shou, University of Pennsylvania

Session: IP67

The biophysiological mechanisms of pain and how human respond is not well understood. In addition, self-rated pain scores are often prone to subjective biases. Hence there is a need to identify neuroimaging biomarkers that are involved in pain. We were motivated by a study of real pain model where the samples were repeatedly scanned post 3rd molar tooth extractions, before and after taking pain medication or placebo. Arterial Spin Labeling (ASL) functional MRI were collected over 5-min blocks and tracked the changes of blood flow in individual brain voxels. A principal component analysis (PCA)-based two group comparison was conducted between responders and non-responders on the longitudinal cerebral blood flow (CBF) maps and their associations with changes of self-reported pain scores. Regions such as amygdala that is involved with emotions and perceptions were identified as differentially activated between groups. In addition, connectivity across regions between the two groups were also tested using distance-based ANOVA.

**Domain adaptation: Models and applications in Neuroscience and Neuroimaging**

▶ Vikas Singh, University of Wisconsin-Madison

Session: IP56

Analysis of imaging, functional, clinical and other demographic data acquired at multiple sites via pooling leads to a number of issues, even within simple models such as linear regression. This talk will cover some of our recent results covering new theoretical results and publicly available toolboxes implementing the algorithms that can facilitate pooling and analysis of neuroscience and neuroimaging datasets. The talk will describe recent results, settled and open questions and the types of scientific questions we are attempting to answer with our scientific collaborators.

**Analysis of Matched Case-Control Study with a Misclassified Exposure**

▶ Samiran Sinha, Texas A&M University

Session: IP53

Matched case-control studies are used for finding the disease-exposure association after controlling the effect of important confounding variables. It is a known fact that the estimators of the association (regression) parameters are biased when the exposure is misclassified, and case-control and matched case-control study is of no exception. Any bias correction method relies on a validation data that contain the true exposure and the misclassified exposure value. In contrast, here we propose a consistent method of correcting such bias in the analysis of a matched case-control data when an instrumental variable is measured for each subject of the study. The significance of this approach is that it works without any validation data that often are not available where measurement of the true exposure is impossible or too costly. The asymptotic property of the method is theoretically justified. The operation characteristics of the proposed method are assessed via simulation studies. The method is illustrated by applying the method to a real dataset.

**Single Index Model with Increasing Link in Mental Health Studies**
Debajyoti Sinha, Florida State University

Session: IP25

We propose both frequentist and Bayesian approaches to fit a novel monotone single index model using Bernstein polynomial basis to model the link function. The monotonicity of the unknown link function of the regression gives us an useful index that can guide psychiatrists and patients regarding relative importance of different covariates to control adolescent depression. We use an iterative profile likelihood algorithm to estimate the index vector and the unknown link function. To facilitate the algorithm, we use linearization of the objective function using first order Taylor expansion of the link function. For Bayesian analysis, we implement a novel and efficient Metropolis Hastings (MH) step to generate from the conditional posterior distribution of the index parameter using a proposal density with same mode of the target density. These methodologies are illustrated via simulation studies and the analysis of a mental-health study of dysphoria scores of adolescent girls.

PERFect: permutation filtration of microbiome data

Ekaterina Smirnova, University of Montana

Session: IP16

Microbiota composition, which is associated with a number of diseases including obesity and bacterial vaginosis, requires preprocessing steps that take into account sparsity of counts and large number of taxa. Filtering is defined as removing taxa that are present in a small number of samples and have small counts in the samples where they are observed. Currently, there is no consensus on filtering standards and quality assessment. This can adversely affect downstream analyses and reproducibility of results. We introduce PERFect (https://github.com/katiasmirn/PERFect), a novel filtering approach designed to address two unsolved problems in microbiome data processing: (i) define and quantify loss due to filtering, and (ii) introduce and evaluate a permutation test for filtering loss to provide a measure of excessive filtering. Methods are assessed on mock data, where the true taxa compositions are known, and are applied to a vaginal microbiome data set. PERFect correctly removes contaminant taxa, quantifies filtering loss, and provides a uniform data-driven filtering criteria for real microbiome data sets.

Statistical inference for one-shot device testing data

Hon Yiu So, University of Waterloo

Session: IP64

One shot devices are products that can only be used once. Typical one-shot devices include air-bags, fire-extinguisher, fireworks etc.. The observations from those devices are either success and failure at the time of test/use. So, there is usually a huge loss of information and, hence, the estimation of life characteristics becomes a difficult problem. The estimation problem in this case has been discussed by a number of authors, mostly in a parametric setting. In this talk, I will focus on the following aspects of one-shot devices test data. First, I will discuss the Bayesian estimation and a semi-parametric estimation method for simple one-shot devices. Due to the fact that most one-shot devices contain many components and that failure of any one of them may lead to the failure of the device, a competing risk model will be discussed next in a one-shot device testing context. In the second section, I will discuss the maximum likelihood estimation of model parameters using EM algorithm, and also the Bayesian estimation as well as the semi-parametric estimation for such a competing risk scenario. Finally, I will conclude the presentation by mentioning some open problems.

Improving Efficiency Through Augmentation in the Semiparametric Accelerated Failure Time Model With Missing Covariates

Jon Steingrimsson, Brown University

Session: IP30
In failure time studies where the event of interest is rare large cohorts are often needed in order to get reliable information about the effect of covariates on the failure time. Collecting covariate information on the whole cohort can be expensive and a case-cohort design is a cost reducing design where covariate information is partitioned into stage one covariate information collected on everyone and stage two covariate information collected on all failures but only on a subsample of the non-failures. In this talk we consider a class of augmented estimating equations for analyzing the semiparametric accelerated failure time model under a case-cohort study, focusing on improving efficiency over the simple but inefficient Horvitz-Thompson type estimators. We summarize asymptotic properties of the class of estimators, identify the most efficient estimator within that class, and give guidance for calculating the augmented estimator in practice. Finite sample performance is evaluated via simulations as well as by analyzing data arising from the Wilms’ tumor studies. This is joint work with Professor Rob Strawderman (University of Rochester).

**Entity Resolution with Societal Impacts in Statistical Machine Learning**

**Rebecca Steorts**, Duke University

Session: IP43

Very often information about social entities is scattered across multiple databases. Combining that information into one database can result in enormous benefits for analysis, resulting in richer and more reliable conclusions. Among the types of questions that have been, and can be, addressed by combining information include: How accurate are census enumerations for minority groups? How many of the elderly are at high risk for sepsis in different parts of the country? How many people were victims of war crimes in recent conflicts in Syria? In most practical applications, however, analysts cannot simply link records across databases based on unique identifiers, such as social security numbers, either because they are not a part of some databases or are not available due to privacy concerns. In such cases, analysts need to use methods from statistical and computational science known as entity resolution (record linkage or de-duplication) to proceed with analysis. Entity resolution is not only a crucial task for social science and industrial applications, but is a challenging statistical and computational problem itself. In this talk, we describe the past and present challenges with entity resolution, with applications to the Syrian conflict but also official statistics, and the food and music industry. This work, which is a joint collaboration with researchers at Rice University and the Human Rights Data Analysis Group (HRDAG) touches on the interdisciplinary research that is crucial to problems with societal impacts that are at the forefront of both national and international news.

**Envelope-based sparse partial least squares**

**Zhihua Su**, University of Florida

Session: IP7

Sparse partial least squares (SPLS, Chun and Keles, 2010) is a widely used method that performs dimension reduction and variable selection simultaneously in linear regression. Despite its popularity in applied sciences, its theoretical properties are largely unknown. In this paper, we use a connection between envelope models and partial least squares (PLS) to construct an envelope-based SPLS estimator and establish its consistency, oracle property and asymptotic normality. The large-sample scenario and high-dimensional scenario are both considered. We also develop the envelope-based SPLS estimators under the context of generalized linear models, and discuss its theoretical properties including consistency, oracle property and asymptotic distribution. Numerical experiments and examples show that the envelope-based SPLS estimator has better variable selection and prediction performance over the SPLS estimator.

**Convex Tensor Co-clustering with Applications to Online Advertising**

**Will Wei Sun**, University of Miami

Session: IP2
Tensor as a multi-dimensional generalization of matrix has received increasing attention in industry due to its success in modeling data with complex structures. One typical circumstance is in online advertising, where the users' click behavior on different ads from multiple publisher platforms forms a user-ad-publisher tensor. Our goal is to simultaneously group users, ads, and publishers for better targeted advertising. In literature, prevalent clustering methods mainly focus on vector or matrix-variate data and are not applicable to general-order tensors. And there is a gap between statistical guarantees and computational efficiency for existing tensor clustering solutions due to the nature of their non-convex formulations. In this work, we bridge this gap by developing a provable convex formulation of tensor co-clustering. Our convex co-clustering (CoCo) estimator enjoys stability guarantees and is both computationally and storage efficient. We further establish a non-asymptotic error bound for the CoCo estimator, which reveals a surprising "blessing of dimensionality" phenomenon that does not exist in vector or matrix-variate cluster analysis. To demonstrate the potential business impact of our method, we conduct convex clustering on the user-ad-publisher tensor data obtained from a major online company. Our clustering results provide interesting insights in understanding users' click behavior.

**ROC-Guided Classification and Survival Trees**

▶ **Yifei Sun**, Columbia University

Session: IP30

Tree-based methods are popular statistical tools for creating simple and interpretable prediction rules. In this article, we introduce a unified framework for risk prediction trees with both binary and survival time outcomes. The algorithms for growing and pruning trees are guided by decision-theoretic criteria based on Neyman-Pearson Lemma. Using the idea of randomized tests, we develop generalized ROC curves to evaluate the performance of discrete-valued risk scores resulting from tree models. We establish the optimality of the target functions with respect to the ROC curves and related concordance measures. The proposed ROC-guided algorithms aim to maximize the concordance of the tree classifier, as opposed to the common approaches that maximize the between-node heterogeneity or within-node homogeneity in the tree building procedure. The methods are applied to an HIV clinical trial for illustration.

**Estimating bidirectional mediation effects to understand the reversible relationship between obesity and diabetes.**

▶ **Rajesh Talluri**, University of Mississippi Medical Center

Session: IP54

Obesity and type 2 diabetes are major public health issues with known interdependence. Genetic variants have been associated with obesity, type 2 diabetes, or both; thus, we hypothesize that some single nucleotide polymorphisms associated with both conditions may be mediated through obesity to affect type 2 diabetes or vice versa. We propose a framework for bidirectional mediation analyses. Simulations show that this approach accurately estimates the parameters, whether the mediation is unidirectional or bidirectional. In many scenarios, when the mediator is regressed on the initial variable and the outcome is regressed on the mediator and the initial variable, the resulting residuals are correlated because of other unmeasured covariates not in the model. We show that the proposed model provides accurate estimates in this scenario, too. We applied the proposed approach to investigate the relationship between type 2 diabetes and obesity using genetic data from the Multi-Ethnic Study of Atherosclerosis cohort. Specifically, we used body mass index as a measure for obesity and fasting glucose as a measure for type 2 diabetes. We evaluated the top 6 SNPs associated with both body mass index and fasting glucose. Two SNPs (rs3752355 and rs6087982) had indirect effects on body mass index mediated through fasting glucose (0.2677; 95% confidence interval (CI) [0.0007, 0.6548] and 0.3301; 95% CI [0.0881, 0.8544], respectively). The remaining four SNPs (rs7969190, rs4869710, rs10201400 and rs12421620) directly affect body mass index and fasting glucose diabetes and obesity without mediating effects.

**Subgroup analysis in regression using mixture of finite mixture models**

▶ **Aixin Tan**, University of Iowa
In many regression problems, there exist latent subgroups such that subjects in different subgroups differ in how the response relates to the explanatory variables, say, through group-specific intercepts and slopes. We use a Bayesian model called the mixture of finite mixtures (MFM) to identify such subgroups, where the number of subgroups itself is a random variable. This natural Bayesian solution to handle unknown number of subgroups was not commonly implemented due to computational difficulties. An alternative is the Dirichlet Process Mixture Model (DPMM), which is relatively easy to compute, but tends to over-estimate the number of groups. We develop efficient algorithms for MFMs in regression setups that closely resembles that for DPMMs, based on the results in Miller and Harrison (2017). We make comparisons of the MFM solution to other methods in numerical studies and real data analysis.

**SIMEX approach to estimation of the sparse conditional functional quantile regression with measurement error**

► Carmen D. Tekwe, Texas A&M University

Quantile regression is a semiparametric approach used for modelling associations between variables. It is most helpful when the covariates have a complex relationship with the location, scale, and shape of the outcome distribution. Despite its robustness to distributional assumptions and outliers in the outcome, regression quantiles may be biased in the presence of measurement error in the covariates. While studies have investigated the case of scalar-valued covariates, the impact of function-valued covariates contaminated with error has not yet been examined. In this paper, we present an instrumental variable approach for consistently estimating linear quantile regression models that include a function-valued covariate measured with error. A two-stage approach to estimation is proposed. In the first stage, an instrumental variable is used to identify the model and the covariance matrix of the measurement error is estimated. In the second stage, the simulation extrapolation (SIMEX) approach for measurement error correction is used to simulate additional measurement error with increasing variance which is added to the observed measure for the function valued covariate. After adding the additional error to the observed function-valued covariate, the standard quantile check function is minimized. Standard errors are estimated by means of point-wise nonparametric bootstrap. We present a simulation study to assess the robustness of the proposed estimator in the presence of measurement errors. The proposed methods are applied to the NHANES database to study obesity and physical activity among U.S. adults with a history of cancer. In our analysis, we estimate body mass index (BMI) percentiles conditional on objective measures of physical activity intensity levels. Measures of daily physical activity intensity represent the functional data observed with error, while measures of abdominal adiposity are used as the instrumental variable.

**Robust PCA by Manifold Optimization**

► Zhang Teng, University of Central Florida

Robust PCA is a widely used statistical procedure to recover a underlying low-rank matrix with grossly corrupted observations. This work considers the problem of robust PCA as a nonconvex optimization problem on the manifold of low-rank matrices, and proposes two algorithms (for two versions of retractions) based on manifold optimization. It is shown that, with a proper designed initialization, the proposed algorithms are guaranteed to converge to the underlying low-rank matrix linearly. Compared with a previous work based on the Burer-Monterio decomposition of low-rank matrices, the proposed algorithms reduce the dependence on the conditional number of the underlying low-rank matrix theoretically. Simulations and real data examples confirm the competitive performance of our method.

**Dynamic Functional Connectivity in Movie-Watching fMRI Data**
Recent neuroimaging studies suggest the temporal evolution of brain connectivity where interactions between distinct brain regions change over time, in both resting state and response to tasks or stimuli. Characterizing and inferring different types of changing connectivity patterns from brain signals for a large number of nodes are classical statistical challenges in modelling non-stationarity and high-dimensionality in time series data. In this talk, I will consider a problem of analyzing task-based dynamic functional connectivity from fMRI data recorded from subjects watching a same movie. Identifying dynamic connectivity during naturalistic conditions presented with mixture of unconstrained, multimodal stimuli, like movie watching, is far more challenging than the conventional settings with carefully controlled stimuli typically of a particular type. We present an approach based on Markov-switching vector autoregressive VAR (SVAR) models to capture time-evolving directed functional connectivity that alternate between a set of re-occurring latent brain states. Within a state-space formulation, change-points of connectivity regimes and the dependencies between regions in each regime can be estimated efficiently using Kalman filtering and expectation-maximization algorithm. We also consider some extensions to account for high-dimensional connectivity based on the factor analysis, and to identify common states that are shared across multiple subjects. Finally, the estimated time courses of dynamic connectivity states will be correlated to the meanings of sequences of different audio-visual stimuli, including the speech, language and movement gestures annotated from the movie.

Estimation of a parametric function of a two-parameter exponential distribution under double censoring

Yogesh M Tripathi, Indian Institute of Technology, Patna

We consider estimation of a parametric function \( k\sigma \) using doubly censored samples from a two-parameter exponential with \( \sigma \) being location and scale parameters respectively, \( k \) is a known constant. The inadmissibility of the minimum risk equivariant estimator (MRE) is established under an arbitrary strictly convex loss function. We also develop a new class of estimators improving upon the MRE under the quadratic loss function. Studied classes of estimators are shown to contain some known results from literature.

Bayesian Models of High-Dimensional Count Data

Marina Vannucci, Rice University

Many of the real applications prevalent in the modern data science involve heterogeneous and mixed data (e.g. count, binary, continuous, skewed continuous, among other data types). In this talk we will consider hierarchical Bayesian models for high-dimensional count data that incorporate variable selection. Zero-inflation, skewness, and overdispersion all cause difficulties when modeling count data. In this talk I will first look at a Bayesian Dirichlet-Multinomial regression model which uses spike-and-slab priors for the selection of significant associations between a set of available covariates and taxa from a microbiome abundance table. I will then describe a negative binomial mixture regression model for the analysis of sequence counts and methylation data. In addition to feature selection, models include priors that capture structural dependencies among the variables.

Markov Chain Monte Carlo in High-Dimensional Bayesian Regression

Dootika Vats, University of Warwick

Inference in Bayesian regression is often made using Markov chain Monte Carlo (MCMC) methods. For MCMC to perform well, its rate of convergence to the target distribution must be relatively fast. In particular,
uniform or geometric rates of convergence are of specific interest since they allow for output analysis of the MCMC samples. We review some popular high-dimensional Bayesian regression models and discuss the effects of the priors on the MCMC sampling procedure. In particular, we discuss how posterior consistency for Bayesian models connects to the sampling quality of the Markov chains.

Progressively Type-II Censored Competing Risks Data from the Linear Exponential Distribution

► William D. Volterman, Syracuse University

Session: IP64

Across different types of lifetime studies, whether it be in the medical or engineering sciences, the possibility of competing causes of failures needs to be addressed. Typically referred to as competing risks, in this paper we consider progressively type-II censored competing risks data when the lifetimes are assumed to come from a linear exponential distribution. We develop likelihood inference and demonstrate the performance of the estimators via an extensive Monte Carlo simulation study. We also provide an illustrative example using a small data set.

Unraveling bacterial fingerprints of city subways from microbiome 16S gene profiles

► Alejandro R. Walker, University of Florida

Session: IP11

As part of the 2017 CAMDA MetaSUB Inter-City Challenge, next generation sequencing (NGS) data was generated from swipe samples collected from subway stations in Boston, New York City hereafter New York, and Sacramento. Operational Taxonomic Unit (OTU) from NGS data was generated in open reference mode with QIIME, then normalized and analyzed. Principal component analysis (PCA) showed clear clustering of the samples for the three cities, with a substantial proportion of the variance explained by the first three components. We ran two different classifiers and results were robust for error rate (<6%) and accuracy (>95%). The analysis of variance (ANOVA) demonstrated that overall, bacterial composition across the three cities is significantly different. A similar conclusion was reached using a novel bootstrap based test using diversity indices. Last but not least, a co-abundance association network analyses for the taxonomic levels “order”, “family”, and “genus” found different patterns of bacterial networks for the three cities. Association based network analysis using the bacterial fingerprint profiles emphasized similarities between the closest cities sharing common bacterial composition. This work advocates a data analysis pipeline, which could be followed in order to get biological insight from this data. However, the biological conclusions from this analysis is just an early indication out of a pilot microbiome data provided through CAMDA challenge and will be subject to change as we get more complete data sets in the future.

On the stability of extrinsic sample means and antimeans

► Yunfan Wang, Florida State University

Session: IP57

For random objects on compact metric spaces, Patrangenaru and Ellingson (2015), introduced a new population parameter, the extrinsic antimean. Necessary and sufficient conditions for the existence of the extrinsic antimean were given in Patrangenaru et al. (2016), in terms of the antifocal set of the population. The asymptotic behavior of the extrinsic sample mean was studied in Patrangenaru et al(2016a). Inspite of consistency of the sample means and antimeans, computational challenges, leading to unstable sample means or antimeans, arise when the mean of the embedded data in the numerical ambient space is too close to the focal, respectively nonfocal set of the population. As computational of stability of sample means and antimeans, we consider means of Kendall planar shapes data for a group of midface landmarks in children from X-rays, from Bookstein (1997).
Probabilistic Dimensionality Reduction via Structure Learning

▶ Li Wang, University of Texas at Arlington

Session: IP2

We propose an alternative probabilistic dimensionality reduction framework that can naturally integrate the generative model and the locality information of data. Based on this framework, we present a new model, which is able to learn a set of embedding points in a low-dimensional space by retaining the inherent structure from high-dimensional data. The objective function of this new model can be equivalently interpreted as two coupled learning problems, i.e., structure learning and the learning of projection matrix. Inspired by this interesting interpretation, we propose another model, which finds a set of embedding points that can directly form an explicit graph structure. We proved that the model by learning explicit graphs generalizes the reversed graph embedding method, but leads to a natural interpretation from Bayesian perspective. This can greatly facilitate data visualization and scientific discovery in downstream analysis. Extensive experiments are performed that demonstrate that the proposed framework is able to retain the inherent structure of datasets and achieve competitive quantitative results in terms of various performance evaluation criteria.

Supervised $t$-SNE for Data Visualization and Classification

▶ Xinlei Wang, Southern Methodist University

Session: IP66

In areas such as biology, medicine and health care, high-dimensional data are often observed. In practice, such data are likely to lie within a lower dimensional manifold. Many dimension reduction methods have been developed with applications in data clustering, visualization, and regression/classification, among others. An unsupervised dimension reduction is known as a data preprocessing technique that can achieve easy visualization and remove redundant information for downstream analysis. If the primary task is to predict some variable of interest using the reduced data, we may wish to incorporate the outcome information of labeled data into the dimension reduction process. However, such a supervised dimension reduction task can be challenging since it is unclear how the labeled information should be used. We propose a novel supervised method called supervised $t$-distributed stochastic neighbor embedding (St-SNE) that achieves the dimension reduction by preserving similarities between data points in both covariate and outcome spaces. A major advantage of St-SNE is its robustness to the underlying data-generating mechanism due to its nonparametric setup. We show that the proposed method provides better prediction power than existing supervised methods and serves as a competitive visualization tool.

A Bayesian Design for Phase I Cancer Therapeutic Vaccine Trials

▶ Chenguang Wang, Johns Hopkins University

Session: IP37

Phase I clinical trials are the first step in drug development to test a new drug or drug combination on humans. Typical designs of Phase I trials use toxicity as the primary endpoint and aim to find the maximum tolerable dosage. However, these designs are poorly applicable for the development of cancer therapeutic vaccines because the expected safety concerns for these vaccines are not as much as cytotoxic agents. The primary objectives of a cancer therapeutic vaccine phase I trial thus often include determining whether the vaccine shows biologic activity and the minimum dose necessary to achieve a full immune or even clinical response. In this paper, we propose a new Bayesian phase I trial design that allows simultaneous evaluation of safety and immunogenicity outcomes. We demonstrate the proposed clinical trial design by both a numeric study and a therapeutic human papillomavirus vaccine trial.

Estimation of mediating effect in a mediation model with a censored mediator

▶ Jian Wang, The University of Texas MD Anderson Cancer Center
A mediation model assesses the direct and indirect effects of an initial variable on an outcome by including one mediator. In practice, instead of the complete data, investigators can observe censored data. The current approaches for a mediation model with censored variables focus on the censored outcome variable. However, the mediation model can include a censored mediator. We proposed to use the accelerated failure time model and a multiple imputation approach to analyze the mediation model when the mediator is a censored variable. We further proposed a measure of the indirect effect for the mediation model with a censored mediator. Simulations were conducted to investigate the performance of the proposed approach and compared it to that of the existing approaches. The proposed approach accurately estimated the indirect effects and percentages of the total effects mediated; while the existing approaches provided biased estimations. The existing and proposed approaches were applied to investigate the indirect effect of the age at menopause on the association between genetic variants and type-2 diabetes risk.

Challenges in Deep Learning for Lung Cancer Detection

▶ Xiaofeng Wang, Cleveland Clinic

Session: IP27

Lung cancer screening programs using low-dose chest computed tomography (LDCT) imaging are currently being implemented in the U.S. One of the major challenges arising from the implementation of LDCT screening and the widespread use of diagnostic CT imaging is the enormous number of CT images that must be analyzed by radiologists. However, even for highly trained radiologists, malignant lung nodules can be difficult to separate from benign nodules. Deep learning with convolutional neural networks (CNN) has been successfully applied to medical imaging applications. We will give an overview of the current stage on the deep-learning-based CAD systems that automatically detect pulmonary nodules on CT scans. We will discuss the current challenges and opportunities in this line of research.

Two-stage Adaptive Enrichment Design for Testing an Active Factor

▶ Samuel S. Wu, University of Florida

Session: IP65

We propose a two-stage clinical trial method to identify an active factor, where the first stage trial provides preliminary estimates for designing a more powerful second stage trial. Using a weighted Fisher's combination to combine p-values from both stages, simulations indicate that our proposed test generally improves upon a single-stage trial of equivalent size. The two-stage active effect test is very robust compared to a traditional main effect test and achieves big power improvement when interaction exists.

A Diagnostic Procedure For High-Dimensional Data Streams Via Missed Discovery Rate Control

▶ Dongdong Xiang, East China Normal University

Session: IP40

In monitoring complex systems involving high-dimensional data streams, apart from quick real-time detection of abnormal changes of system performance, accurate and efficient diagnosis of responsible streams has become increasingly important in many data-rich statistical process control (SPC) applications. Diagnostic procedures in the literature, designed for low/moderate dimensional multivariate process, may suffer from missing too much important information on the out-of-control (OC) streams with high signal-to-noise ratio (SNR) or wasting too many resources for finding useless in-control (IC) streams with low SNR. In addition, these procedures do not treat the streams differently according to their severities (e.g., the importance, shift sizes and intensities). In this article, we formulate the diagnosis problem of high-dimensional data streams as a multiple testing problem and provide a computationally fast and easy-to-implement diagnostic procedure to control the weighted missed discovery rate (wMDR) at some satisfactory level. The proposed procedure
overcomes the limitations of traditional diagnostic procedures by controlling the wMDR and minimizing the expected number of true negatives (ETN) as well. We show theoretically that the proposed procedure is asymptotically valid and optimal in a certain sense. We compare empirically using simulations the performance of the proposed procedure with existing methods over a range of distributions with various correlation structures, data dimensionality and signal sparsity. We apply the proposed procedure to analyze data from a semi-conductor manufacturing process.

**Autoregressive models for matrix-valued time series**

*Han Xiao*, Rutgers University

Session: IP44

In finance, economics and many other fields, observations in a matrix form are often observed over time. For example, several key economic indicators are reported in different countries every quarter. Various financial characteristics of many companies are reported over time. Import-export figures among a group of countries can also be structured in a matrix form. Although it is natural to turn the matrix observations into a long vector then use standard vector time series models, it is often the case that the columns and rows of a matrix represent different sets of information that are closely interplayed. We propose a novel matrix autoregressive model that maintains and utilizes the matrix structure to achieve greater dimensional reduction as well as easier interpretable results. The model can be further simplified by a set of reduced rank assumptions. Estimation procedure and its theoretical properties are investigated and demonstrated with simulated and real examples.

**High-dimensional nonparametric density estimation via symmetry and shape constraints**

*Min Xu*, University of Pennsylvania

Session: IP12

We tackle the problem of high-dimensional nonparametric density estimation by taking the class of log-concave densities on \( \mathbb{R}^p \) and incorporating within it a symmetry assumption. We say that a log-concave density is \( K \)-symmetrical for a convex body \( K \) if all the level sets of the density are scaled versions of \( K \). This extends the notion of spherical symmetry and leads naturally to a two-stage estimation algorithm where we first estimate the location and the contour \( K \) of the density and then estimate the slope profile of the density. We prove that the risk (in Hellinger loss) of this estimator can be bounded by a sum of two terms the first of which accounts for the errors in estimating the location and \( K \) and the second of which is an adaptive rate that is independent of \( p \) and identical to the rate of estimating a univariate log-concave density. This immediately yields a bound on the rate of convergence for estimating an elliptically-symmetric log-concave density. Our future work focuses on developing estimators for the contour when \( K \) is a general convex body.

**Causal inference with GWAS-based Mendelian randomization**

*Hongyan Xu*, Augusta University

Session: IP54

Mendelian randomization (MR) is an effective approach for causal analysis of risk factors for complex diseases. It is critical to use appropriate genetic variants as instrumental variables for valid and efficient Mendelian randomization analysis using data from genomewide association studies. We developed an approach to systematically select single nucleotide polymorphisms (SNPs) to investigate the causality of blood lipids on coronary heart disease (CAD) and type 2 diabetes mellitus (T2DM). Using the 338 SNPs selected for multivariate MR analysis of CAD, we found that low density lipoprotein cholesterol and triglycerides increased risk for CAD (\( \beta = 0.424 \) and 0.200; \( p = 6.23E-22 \) and 1.95E-07, respectively), high density lipoprotein cholesterol (HDL-c) had protective effect against CAD (\( \beta = -0.286; \ p = 2.21E-13 \)). With 309 SNPs for similar analysis of T2DM, HDL-c was found to decrease risk for diabetes (\( \beta = -0.1677, \ p =5.06E-07 \)). MR-Egger analysis gave
consistent results. T2DM was significantly associated with increased risk for CAD ($\beta = 0.11499, p = 2.93E-10$) with MR analysis. Our results provided evidence of casual effect of blood lipids on CAD and T2DM and T2D on CAD.

Performance Assessment of High-dimensional Variable Identification

▶ Yuhong Yang, University of Minnesota

Session: IP46

One important task in statistical/machine learning is to identify the most important variables that influence the response of interest. Examples include the identification of biomarkers of a disease in medical research and choice of behavioral characteristic variables of customers for effective advertisement. Since such variable selection process is ubiquitous in data analysis, reproducibility of the statistical results demands a serious evaluation of reliability of the employed model selection method, no matter what label it may have in terms of good properties. Instability measures have been proposed for evaluating model selection uncertainty. However, low instability does not necessarily indicate that the selected model is trustworthy, since low instability can also arise when a certain method tends to select an overly parsimonious model. $F-$ and $G-$ measures have become increasingly popular for assessing variable selection performance in theoretical studies and simulation results in machine learning and statistics. However, they are not computable in practice. In this work, we propose an estimation method for $F-$ and $G-$ measures and prove their desirable properties of uniform consistency. This gives the data analyst a valuable tool to compare different variable selection methods based on the data at hand. Extensive simulations are conducted to show the very good finite sample performance of our approach. We further demonstrate the application of our methods using several micro-array gene expression data sets. The work is joint with Yanjia Yu and Yi Yang.

Statistical Adjustment for Reporting Bias in Surveillance Data of Infectious Diseases

▶ Yang Yang, University of Florida

Session: IP59

Reporting bias is common in the surveillance of infectious diseases, often a result of efforts to identify the most vulnerable subpopulation or to balance between public health impact and cost. In this talk, we focus on two kinds of biases motivated by real surveillance data. The first is the Chinese influenza outbreak investigation data where outbreaks in institutional settings (e.g., schools and workplaces) were reported and investigated only if there were 10 or more cases of influenza-like illness. For inference adjusted for such a selection bias, we condition the likelihood for time series of symptom onsets on that for the final size of cases exceeding the lower limit of detection. The second example is the Zika surveillance data from Colombia where female cases in reproductive age were much more actively sought than other demographic groups. We use a Bayesian framework to estimate the basic reproductive number of Zika and the effects of age and gender on transmissibility, while correcting for reporting bias. The performance of the proposed methods is evaluated in simulated epidemics.

On Quality Control of Mixed Quality Variables: Some Thoughts and Ideas

▶ Arthur B. Yeh, Bowling Green State University

Session: IP40

In the last two decades, the applications of quality control tools, especially control charts, have expanded to non-manufacturing sectors, such as service, healthcare and retail industries. New challenges naturally arise from new applications, which, in turn, spur creations of new methodologies or innovative retooling of existing methodologies. In this exposition, we focus on one such challenge where multiple correlated mixed quality variables need to be monitored simultaneously. Here, by “mixed quality variables”, we mean that the multivariate quality characteristic to be monitored contains a collection of continuous, count and categorical
variables. In such a scenario, resorting to developing the multivariate joint distribution can be quite challenging, especially when the dimensionality is large. A more practical alternative, we propose, is simultaneous hypothesis testing procedure. We discuss how Holm’s step-down testing procedure can be implemented such that it controls the overall family error-rate when the process is in-control, and has reasonable detecting power when the process is out-of-control. An added bonus is its ability to identify, to a certain extent, which of the variables are out of control immediately after the chart signals.

Dissecting tumor microenvironment using single-cell RNA-seq data

► Xiaoqing Yu, H. Lee Moffitt Cancer Center & Research Institute

Session: IP51

A deep understanding of the tumor-infiltrating lymphocytes is critical to the identification of effective biomarkers and development of novel immunotherapy. The newly emerged single-cell sequencing technologies provide us such an opportunity for better understanding of the tumor-infiltrating immune cells in the complex tumor microenvironment. In this study, we analyzed in depth a pooled single cell RNA-seq dataset from peripheral blood, tumor and adjacent normal tissues, and derived cell type-specific gene expression signatures for major tumor-infiltrating lymphocytes. A modified t-SNE mapping method was performed on the combined data to identify cluster of similar cells, followed by a kernel-based clustering method to further separate the T cell subtypes. Based on the cell-specific signatures derived from single cell RNA-seq, we further developed a deconvolution algorithm to decipher the immune cell compositions in bulk tumor sequencing. Results showed that our method was able to better separate T cell subpopulation and provide more informative immune outcomes from tumor samples.

Incorporating Information from Nearby Loci for Detection of Imprinting and Maternal Effects Based on Partial Likelihood

► Fangyuan Zhang, Texas Tech University

Session: IP28

Numerous statistical methods have been developed to explore genomic imprinting and maternal effects by identifying parent-of-origin patterns in complex human diseases. However, because most of these methods only use available locus-specific genotype data, it is sometimes impossible for them to infer the parental origin of a variant allele, especially when there are missing genotypes. In this paper, we propose a two-step approach, LIMEhap, to improve upon a recent partial likelihood inference method. The first step is to infer the distribution of the missing genotypes by using the information from nearby loci. This is then followed by the second step in which the partial likelihood method is applied to the inferred data. To substantiate the validity of the proposed procedure, we simulated data in a genomic region of gene GPX1. The results show that by borrowing genetic information from nearby loci, the power of the proposed method can be close to that with complete genotype data at the locus of interest. This illustrates that the use of nearby marker in linkage disequilibrium can help resolve parental origin ambiguity. We further studied the robustness of LIMEhap to violation of Hardy-Weinberg Equilibrium (HWE), as inference on the genotype distribution is made under the assumption of HWE. To illustrate its practical utility, LIMEhap was applied to autism study data.

Theoretical studies in the big data era for spatial statistics

► Hao Zhang, Purdue University

Session: IP38

One of the areas where big data are collected is in spatial statistics. There have been good developments in computational methods to handle the big data. In this talk, I will make the case for the need in more theoretical studies that can help and guide the development in computational methods. Algorithms and prediction have been the focus in many of the studies and have produced new methodologies such as deep learning, while statisticians have not provided breakthrough theories for these big and new data. One has to wonder how statistics deals with the changing landscape. In this talk, I will try to shed light as to where statistics is going
and show some clear trend in statistics. I will also provide some advices for young researchers to remain productive and carry out impactful research.

**Bayes Factors For Survival Model Selection**

**Haiming Zhou**, Northern Illinois University

Session: IP3

Choosing an appropriate survival model is crucial for efficient design of clinical trials with time-to-event endpoints. In this talk I will present a super model that includes six commonly-used interpretable semiparametric survival models (e.g., PH, AFT, PO) as special cases. Approximate Bayes factors are developed for testing and choosing among these six models. Baseline survival is modeled with a novel transformed Bernstein polynomial prior. All manner of censored survival times are simultaneously accommodated including uncensored, interval censored, current-status, left and right censored, and mixtures of these. A new R function which calls efficient compiled C++ in the R package `spBayesSurv` is developed for fitting the super model and all Bayes factors. The methodology is broadly illustrated with simulations and real data applications.

**A Powerful Bayesian Test for Equality of Means in High Dimensions**

**Roger Zoh**, Texas A&M University School of Public health

Session: IP10

We develop a Bayes factor based testing procedure for comparing two population means in high dimensional settings. In ‘large-p-small-n’ settings, Bayes factors based on proper priors require eliciting a large and complex $p \times p$ covariance matrix, whereas Bayes factors based on Jeffrey’s prior suffer the same impediment as the classical Hotelling $T^2$ test statistic as they involve inversion of ill-formed sample covariance matrices. To circumvent this limitation, we propose that the Bayes factor be based on lower dimensional random projections of the high dimensional data vectors. We choose the prior under the alternative to maximize the power of the test for a fixed threshold level, yielding a restricted most powerful Bayesian test (RMPBT). The final test statistic is based on the ensemble of Bayes factors corresponding to multiple replications of randomly projected data. We show that the test is unbiased and, under mild conditions, is also locally consistent. We demonstrate the efficacy of the approach through simulated and real data examples.

**A Model-Based Conditional Power Assessment for Decision Making in Randomized Controlled Trial Studies**

**Baiming Zou**, University of Florida

Session: IP65

Conditional power based on summary statistic by comparing outcomes (such as the sample mean) directly between two groups is a convenient tool for decision making in randomized controlled trial studies. In this paper, we extend the traditional summary statistic-based conditional power with a general model-based assessment strategy, where the test statistic is based on a regression model. Asymptotic relationships between parameter estimates based on the observed interim data and final unobserved data are established, from which we develop an analytic model-based conditional power assessment for both Gaussian and non-Gaussian data. The model-based strategy is not only flexible in handling baseline covariates and more powerful in detecting the treatment effects compared with the conventional method, but also more robust in controlling the overall type I error under certain missing data mechanisms. The performance of the proposed method is evaluated by extensive simulation studies and illustrated with an application to a clinical study.