1. **Session title**: High-dimensional Tensor Data Analysis
   Organizer: Xin "Henry" Zhang (Florida State U)
   Chair: Xin "Henry" Zhang (Florida State U)
   Time: June 4th, 9:00-10:30am
   Location: VEC 404/405

   Speech 1: Multilayer Tensor Factorization with Applications to Recommender Systems
   Speaker: Xuan Bi (Yale)
   abstract: Recommender systems have been widely adopted by electronic commerce and entertainment industries for individualized prediction and recommendation, which benefit consumers and improve business intelligence. In this talk, we propose an innovative method, namely the recommendation engine of multilayers (REM), for tensor recommender systems. The proposed method utilizes the structure of a tensor response to integrate information from multiple modes, and creates an additional layer of nested latent factors to accommodate between-subjects dependency. One major advantage is that the proposed method is able to address the "cold-start" issue in the absence of information from new customers, new products or new contexts. Specifically, it provides more effective recommendations through sub-group information. To achieve scalable computation, we develop a new algorithm for the proposed method, which incorporates a maximum block improvement strategy into the cyclic blockwise-coordinate-descent algorithm. In theory, we investigate algorithmic properties for convergence from an arbitrary initial point and local convergence, along with the asymptotic consistency of estimated pa- rameters. Finally, the proposed method is applied in simulations and IRI marketing data with 116 million observations of product sales. Numerical studies demonstrate that the proposed method outperforms existing competitors in the literature.

   Speech 2: Dynamic Tensor Clustering
   Speaker: Will Wei Sun (University of Miami)
   abstract: Dynamic tensor data are becoming prevalent in numerous applications. Existing tensor clustering methods either fail to account for the dynamic nature of the data, or are inapplicable to a general-order tensor. Also there is often a gap between statistical guarantee and computational efficiency for existing tensor clustering solutions. In this talk, I will introduce a new dynamic tensor clustering method, which takes into account both sparsity and fusion structures, and enjoys strong statistical guarantees as well as high computational efficiency. The efficacy of our approach will be illustrated via two real applications: brain dynamic functional connectivity analysis, and advertisement clustering for market segmentation. This is a joint work with Lexin Li.

   Speech 3: Covariate-adjusted tensor classification in high-dimensions
   Speaker: Qing Mai (Florida State U)
   abstract: In contemporary scientific research, it is of great interest to predict a categorical response based on a high-dimensional tensor and additional covariates. We introduce the CATCH model (in short for Covariate-Adjusted Tensor Classification in High-dimensions), that efficiently integrates the covariates and the tensor to predict the categorical outcome and jointly explains the relationships among the covariates, the

tensor predictor, and the categorical response. To tackle the new computational and statistical challenges arising from the intimidating tensor dimensions, we propose a group penalized approach and an efficient algorithm. Theoretical results confirm that our method achieves variable selection consistency and optimal prediction, even when the tensor dimension is much larger than the sample size. The superior performance of our method over existing methods is demonstrated in extensive simulation studies, a colorimetric sensor array data, and two neuroimaging studies.

2. **Session title**: High-dimensional inference: assumption-lean or assumption-laden?
Organizer: Ryan Tibshirani (CMU)
Chair: Jelena Bradic (UCSD)
Time: June 4th, 9:00-10:30am
Location: VEC 902/903

Speech 1: Inferential goals, targets, and principles in high-dimensional regression
Speaker: Todd Kuffner (Washington U)
Abstract: This talk will focus more on theory than methodology. We will analyze the inferential goals, the targets of inference, and the various principles employed to justify prominent methodologies. A new perspective, motivated by philosophy of science, will be presented on how to discern between competing inferential procedures for high-dimensional regression.

Speech 2: Should We Model X in High-Dimensional Inference?
Speaker: Lucas Janson (Harvard)
Abstract: For answering questions about the relationship between a response variable Y and a set of explanatory variables X, most statistical methods focus their assumptions on the conditional distribution of Y given X (or Y | X for short). I will describe some benefits of shifting those assumptions from the conditional distribution Y | X to the joint distribution of X, especially for high-dimensional data. First, modeling X can lead to assumptions that are more realistic and verifiable. Second, there are substantial methodological payoffs in terms of much greater flexibility in the tools an analyst can bring to bear on their data while also being guaranteed exact (non-asymptotic) inference. I will briefly mention some of my recent and ongoing work on methods for high-dimensional inference that model X instead of Y, as well as some challenges and interesting directions for the future.

Speech 3: Towards a Better Understanding of ``High-Dimensional'' Linear Least Squares Regression
Speaker: Andreas Buja (U Penn)
Abstract: High-dimensional regression is conventionally interpreted as an optimization problem of, most commonly, a penalized or constrained LS criterion. Instead, we propose to reinterpret high-dimensional regression as follows: Data is first explored either in a principled way (e.g., lasso or best subset selection) or in an exploratory/unprincipled way to select a manageable set of variables; subsequently the reduced data are subjected to

linear regression. The final set of variables is often much smaller than the sample size and the total number of initial variables. We will treat the combination of both steps as forming high-dimensional linear regression. A first question we consider is to ask what the nature of the OLS estimator is if regressors have been subselected by some variable selection procedure. We answer this question in full generality by proving a deterministic uniform-in-model result about linear regression, and this provides an interpretation of what is being estimated irrespective of the data-dependent variable selection procedure.A second question we consider is how to perform statistical inference using the OLS estimator obtained from a variable selection procedure. This problem is exactly the problem of valid Post-Selection Inference (PoSI). This talk will focus on an approach to PoSI based on an asymptotic linear representation and a high-dimensional central limit theorem.All our results are proved without assuming any probability models, and they allow for non-identically distributed random vectors. In addition, they apply equally to independent and functionally dependent data. Finally, our results do not require any sparsity assumptions. Joint work with the Wharton Linear Models Group including Lawrence
Brown, Edward George and Linda Zhao. Some of this talk is based on https://arxiv.org/abs/1802.05801.

3. **Session title**: Modern nonparametric statistics
Organizer: Richard Samworth (Cambridge)
Chair: Zhengling Qi (UNC)
Time: June 4th, 9:00-10:30am
Location: VEC 1202/1203

Speech 1: Regimes of label-noise determine the benefits of Active Learning
Speaker: Samory Kpotufe (Princeton)
Abstract: In active learning (for classification tasks), the learner has the ability to request the labels of carefully chosen points over space. Intuitively, this might speedup the learning process — in terms of the number of labels required to achieve a fixed error— over the usual passive setting where the learner accesses i.i.d labeled data. Unfortunately, despite significant progress on the subject, the benefits of active over passive learning remain largely unclear: for example, in the usual PAC setting with VC classes, label requirements in active learning are of the same order as in passive learning outside of strong assumptions on label noise. However, a clearer picture of the benefits of active learning emerges under refined parameterization of label noise - this is considered e.g. in work by Hanneke and by Koltchinskii, however under the strong assumption of bounded 'disagreement-coefficient'.In this talk, we aim to gain a better picture of the benefits of active learning over passive learning. In particular, we will consider parametrizations of label noise that help capture a continuum from easy to hard classification problems, and elicit a clearer picture of the benefits of active learning along this continuum. Such parametrizations draw on intuition from the so-called `cluster assumption' in ML, and more generally on so-called 'margin conditions' common in both ML and Statistics. Our results reveal interesting phase transitions (in label requirements) driven by the interaction between noise parameters, marginal distribution, and data dimension. In particular, we manage to address a previous conjecture about the existence of some such

transitions. Furthermore, our algorithmic strategies are adaptive, i.e., require no a priori knowledge of distributional parameters, yet are rate-optimal.The talk is based on recent collaborations with S. Ben-David, R. Urner, A. Locatelli, and A. Carpentier.

Speech 2: Sampling design and stochastic gradient descent for relational data
Speaker: Peter Orbanz (Columbia)
Abstract: State-of-the art learning procedures for relational data typically involve several steps that randomly subsample a data set: (1) During data acquisition from the underlying population. (2) During data splitting or cross validation. (3) During training, if learning involves stochastic gradient descent. There are many natural ways to subsample relational data, and in practice, it is not the exception but the rule that different sampling schemes are used in the different steps. That raises a number of problems: If the sampling schemes do not cohere in a suitable sense, the meaning of prediction becomes ambiguous, error estimates are biased, etc. I will discuss what conditions are required to avoid such problems, and describe a new method for learning from relational data that incorporates the sampling scheme as an explicit model design choice.

Joint work with Victor Veitch, Wenda Zhou, Morgane Austern and David Blei.

Speech 3: Statistical Properties of Maximum Mean Discrepancy with Gaussian Kernels
Speaker: Tong Li(Columbia)
Abstract: Despite the popularity of reproducing kernel based techniques for nonparametric hypothesis testing, the choice of kernel in these approaches is usually ad hoc and how to do so in a more principled way remains one of the most critical challenges in practice. To overcome this difficulty, we provide here justifications for one of the most common and successful choices, Gaussian kernels with a flexible shape parameter. More specifically, we study the statistical properties of maximum mean discrepancy (MMD) based testing procedures with Gaussian kernels. We show that they arise naturally when maximizing MMD over a general class of radial basis function kernels. Moreover, we show that when the underlying distributions are sufficiently smooth, MMD with Gaussian kernels gives rise to a test adaptive over different levels of smoothness in that it attains the minimax optimal detection rates, up to a logarithmic factor, for any given smoothness index.

4. **Session title**: New methods for directed acyclic Gaussian graph and adaptive data analysis
Organizer: Yichao Wu (UIC)
Chair: Weibin Mo (UNC)
Time: June 4th, 9:00-10:30am
Location: VEC 1302/1303

Speech 1: Sublinear-Time Adaptive Data AnalysisSpeaker: Lev Reyzin (UIC)
Abstract: The topic of this talk lies in the burgeoning area of adaptive data analysis, where the goal is to design mechanisms that can give statistically valid answers to adaptively generated queries. In this talk, I will discuss sublinear-time mechanisms for

answering adaptive queries into datasets. These mechanisms provide a polynomial speed-up per query over previous approaches, without increasing the total amount of data needed. I will also discuss a new method for achieving statistically-meaningful responses even when the mechanism is only allowed to see a constant number of samples from the data per query. Finally, I will show how these techniques also yield improved bounds for adaptively optimizing convex and strongly convex functions over a dataset. This is based on work joint with Benjamin Fish and Benjamin I. P. Rubinstein.

Speech 2: Reconstruction of a directed acyclic Gaussian graph for observational and interventional data
Speaker: Xiaotong Shen (U Minnesota)
Abstract: Directed acyclic graphs are widely used to describe, among interacting units, causal relations. Causal relations are estimated by reconstructing a directed acyclic graph's structure, presenting a great challenge when the unknown total ordering of a DAG needs to be estimated.In such a situation, it remains unclear if a graph's structure is reconstructable in the absence of an identifiable likelihood with regard to graphs, and in facing super-exponentially many candidate graphs in the number of nodes. In this talk, I will introduce a global approach to process observational data and interventional data, to identify all estimable causal directions and estimate model parameters. This approach uses constrained maximum likelihood with nonconvex constraints reinforcing the non-loop equirement to yield an estimated directed acyclic graph, where super-exponentially many constraints characterize the major challenge. Computational issues will be discussed in addition to some theoretical aspects. This work is joint with Y. Yuan, W. Pan, Z. Wang and S. Peng..

Speech 3: Convex clustering over an undirected graph
Speaker: Yunzhang Zhu (OSU)
Abstract: Cluster analysis is a fundamental problem in statistics. It aims at categorizing the observations into different groups, called clusters, such that observations in the same cluster tend to be more similar to each other than those from different clusters. In this talk, I will introduce a new optimization-based clustering method called convex clustering over (weighted) undirected graph. The choice of both the graph and its weights is crucial to clustering performance as well as the algorithm's computational efficiency. Specifically, we consider two types of graphs: a minimum spanning tree and a so-called K-means bipartite graph. Computationally, both graphs make the associated optimization problems easier to solve compared to that with a complete graph; and statistically, both lead to better clustering results. Further numerical comparisons with the K-means clustering algorithm and a density-based algorithm demonstrate the superior performance of the proposed algorithms.

5. **Session title**: Statistical Inference in Clustering Problems
Organizer: Jacob Bien (Cornell)
Chair: Jacob Bien (Cornell)
Time: June 4th, 9:00-10:30am
Location: VEC 1402

Speech 1: Inference for variable clustering under correlation-like similarities

Speaker: Max G'Sell (CMU)
Abstract: 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 weakening the required assumptions and generalizing the measure of correlation, and the computational issues that arise in this pursuit.

Speech 2: Large scale cluster analysis via L1 fusion penalization
Speaker: Gourab Mukherjee (USC)
Abstract: We study the large sample behavior of a convex clustering framework, which minimizes the sample within cluster sum of squares under an L_1 fusion constraint on the cluster centroids. We establish that the sample procedure consistently estimates its population analog. We derive the corresponding rates of convergence and develop a novel methodology for feature screening in the clustering of massive datasets. We demonstrate empirically the applicability of our method to cluster analysis of big datasets arising in single-cell gene expression studies.

Speech 3: Density Tree and Density Ranking in Singular Measures
Speaker: Yen-Chi Chen (UW)
Abstract: A density tree (also known as a cluster tree of a probability density function) is a tool in topological data analysis that uses a tree structure to represent the shape of a density function. Even if the density function is multivariate, a density tree can always be displayed on a two-dimensional plane, making it an ideal tool for visualizing the shape of a multivariate dataset. However, in complex datasets such as GPS data, the underlying distribution function is singular so the usual density function and density tree no longer exist. To analyze this type of data and generalize the density tree, we introduce the concept of density ranking and ranking tree (also called an α-tree). We then show that one can consistently estimate the density ranking and the ranking tree using a kernel density estimator. Based on the density ranking, we introduce several geometric and topological summary curves for analyzing GPS datasets.

6. **Session title**: Statistical methods of integrating -omics data
Organizer: Wei, Ying (Columbia U)
Chair: X. Song (Mount Sinai)
Time: June 4th, 9:00-10:30am
Location: VEC 1403

Speech 1: A Statistical Framework for Leveraging Information across Multiple Traits in Genetic Studies
Speaker: Gen Li (Columbia)
Abstract: In genetic studies, pleiotropy occurs when a genetic variant affects multiple traits simultaneously. The true effect sizes for different traits usually have significant correlations. Most existing genome-wide association studies only focus on one trait at a

time and fail to leverage the relationships between different traits. Motivated by the multi-tissue expression quantitative trait loci (eQTL) analysis in the Genotype-Tissue Expression (GTEx) project, we develop a two-stage method to address this limitation. It effectively borrows strength across multiple traits to identify genetic variants regulating a target trait. The method is based on summary statistics and allows genotype data to partially overlap between traits. We apply the proposed method to the GTEx data and identify more eQTLs with potential functionality.

Speech 2: A new method to study the change of miRNA–mRNA interactions due to environmental exposures
Speaker: Pei Wang (Icahn School of Medicine at Mount Sinai)
Abstract: ntegrative approaches characterizing the interactions among different types of biological molecules have been demonstrated to be useful for revealing informative biological mechanisms. One such example is the interaction between microRNA (miRNA) and messenger RNA (mRNA), whose deregulation may be sensitive to environmental insult leading to altered phenotypes. In this work, we introduce a new network approach—integrative Joint Random Forest (iJRF), which characterizes the regulatory system between miRNAs and mRNAs using a network model. iJRF is designed to work under the high-dimension low-sample-size regime, and can borrow information across different treatment conditions to achieve more accurate network inference. It also effectively takes into account prior information of miRNA–mRNA regulatory relationships from existing databases. We then apply iJRF to data from an animal experiment designed to investigate the effect of low-dose environmental chemical exposure on normal mammary gland development. We detected a few important miRNAs that regulated a large number of mRNAs in the control group but not in the exposed groups, suggesting the disruption of miRNA activity due to chemical exposure. Effects of chemical exposure on two affected miRNAs were further validated using breast cancer human cell lines.

Speech 3: smFARM: sparse multivariate Factor Analysis Regression Model in integrative genomics analysis
Speaker: Peter Song(Umich)
Abstract:
The multivariate regression model is a useful tool to explore complex associations between multiple response variables (e.g. gene expressions) and multiple predictors (e.g. SNPs). When the multiple responses are correlated, ignoring such dependency will impair statistical power in the data analysis. Motivated by an integrative genomic data, we propose a new methodology – sparse multivariate factor analysis regression model (smFARM), in which the covariance of the response variables is modeled by a factor analysis model with latent factors. This proposed method not only allows us to address the challenge that the number of genetic predictors is larger than the sample size, but also to adjust for unobserved genetic and/or non-genetic factors that potentially conceal the underlying real response-predictor associations. The proposed smFARM is implemented efficiently by utilizing the strength of the EM algorithm and the group-wise coordinate descend algorithm. In addition, the identified latent factors are explained by the means of gene enrichment analysis. The proposed methodology is evaluated and compared to the

existing methods through extensive simulation studies. We apply smFARM in an integrative genomics analysis of a breast cancer dataset on the relationship between DNA copy numbers and gene expression arrays to derive genetic regulatory patterns relevant to breast cancer.

7. **Session title**: Recent Advances in Statistical Learning
Organizer: Ming Yuan (Columbia)
Chair: Dong Xia (UCSD)
Time: June 4th, 11:00am-12:30pm
Location: VEC 404

Speech 1: Geometry of Optimization Landscapes and Implicit Regularization of Optimization Algorithms.
Speaker: Jason Lee (USC)
Abstract:

We first study the problem of learning a Gaussian input two-layer ReLU network with positive output layer and and the symmetric matrix completion problem. Despite the non-convexity of both problems, we prove that every local minimizer is a global minimizer. Since gradient descent converges to local minimizers, this shows that simple gradient-based methods can find the global optimum of these non-convex problems.

In the second part, we analyze the implicit regularization effects of various optimization algorithms. In particular we prove that for least squares with mirror descent, the algorithm converges to the closest solution in terms of the bregman divergence. For linearly separable classification problems, we prove that the steepest descent with respect to a norm solves SVM with respect to the same norm. For over-parametrized non-convex problems such as matrix sensing or neural net with quadratic activation, we prove that gradient descent converges to the minimum nuclear norm solution, which allows for both meaningful optimization and generalization guarantees.

This is joint work with Rong Ge, Suriya Gunasekar, Tengyu Ma, Mor Shpigel, Daniel Soudry, and Nati Srebro.

Speech 2: M-estimation with the Trimmed L1 penalty
Speaker: Aurelie Lozano (IBM)
Abstract:

We study high-dimensional M-estimators with the trimmed L1 penalty. While standard L1 penalty incurs bias (shrinkage), trimmed L1 leaves the h largest entries penalty-free. This family of estimators include the Trimmed Lasso for sparse linear regression and its counterpart for sparse graphical model estimation. The trimmed L1 penalty is non-convex, but unlike other non-convex regularizers such as SCAD and MCP, it is not amenable and therefore prior analyzes cannot be applied.

We characterize the support recovery of the estimates as a function of the trimming parameter h. Under certain conditions, we show that for any local optimum, (i) if the trimming parameter h is smaller than the true support size, all zero entries of the true parameter vector are successfully estimated as zero, and (ii) if h is larger than the true support size, the non-relevant parameters of the local optimum have smaller absolute values than relevant parameters and hence relevant parameters are not penalized. We then bound the L2 error of any local optimum. These bounds are asymptotically comparable to those for non-convex amenable penalties such as SCAD or MCP, but enjoy better constants. We specialize our main results to linear regression and graphical model estimation.

Finally, we develop a fast provably convergent optimization algorithm for the trimmed regularizer problem. The algorithm has the same rate of convergence as difference of convex (DC)-based approaches, but is faster in practice and finds better objective values than recently proposed algorithms for DC optimization. Empirical results further demonstrate the value of L1 trimming.

Speech 3: Methods of network comparison
Speaker: Sofia Olhede (UCL)
Abstract: The topology of any complex system is key to understanding its structure and function. Fundamentally, algebraic topology guarantees that any system represented by a network can be understood through its closed paths. The length of each path provides a notion of scale, which is vitally important in characterizing dominant modes of system behavior. Here, by combining topology with scale, we prove the existence of universal features which reveal the dominant scales of any network. We use these features to compare several canonical network types in the context of a social media discussion which evolves through the sharing of rumors, leaks and other news. Our analysis enables for the first time a universal understanding of the balance between loops and tree-like structure across network scales, and an assessment of how this balance interacts with the spreading of information online. Crucially, our results allow networks to be quantified and compared in a purely model-free way that is theoretically sound, fully automated, and inherently scalable.

This work is joint with Patrick Wolfe.

8. **Session title**: Supervised and unsupervised learning of complex data
Organizer: Junhui Wang (Citi U of HK)
Chair: Junhui Wang (Citi U of HK)
Time: June 4th, 11:00am-12:30pm
Location: VEC 405

Speech 1: Systems of partially linear models with gradient boosting
Speaker: Yongzhao Shao (NYU)
Abstract: We develop systems partially linear models with gradient boosting for prediction in multicenter studies or regression-based clustering in large scale data. Simultaneous variable selection and effect estimation are achieved using LASSO type penalty functions and ADMM. Simulation studies and real data examples are used to illustrate effectiveness of the proposed methods.

Speech 2: Supervised Dimensionality Reduction for Exponential Family Data
Speaker: Yoonkyung Lee (OSU)
Abstract: Supervised dimensionality reduction techniques, such as partial least squares and supervised principal components, are powerful tools for making predictions with a large number of variables. The implicit squared error terms in the objectives, however, make it less attractive to non-Gaussian data, either in the covariates or the responses. Drawing on a connection between partial least squares and the Gaussian distribution, we show how partial least squares can be extended to other members of the exponential family - similar to the generalized linear model - for both the covariates and the responses. Unlike previous attempts, our extension gives latent variables which are easily interpretable as linear functions of the data and is computationally efficient. In particular, it does not require additional optimization for the scores of new observations and therefore predictions can be made in real time. This is joint work with Andrew Landgraf at Battelle Memorial Institute.

Speech 3: Transform-based unsupervised point registration and unseeded low-rank graph matching
Speaker: Yuan Zhang (OSU)
Abstract:

Unsupervised estimation of the correspondence between two point sets has long been an attractive topic to CS and EE researchers. In this paper, we focus on the vanilla form of the problem: matching two point sets that are identical over a linear transformation. The problem is well-studied and many classical algorithms exist, yet, many of them suffer one or several of the following shortcomings: slow computation on large data sets, limited applicable distribution families and lack of theoretical analysis. Arguably, the bottleneck of computation lies in the need of many methods to evaluate n^2 many pairwise similarity or distance measures in each iteration. Also, few results exist in bounding the error of some specific point matching algorithm, where dependence might be a main obstacle.

In this paper, we propose a novel method using Laplace transformation to directly match the underlying distributions of the two point sets. Our method is fast because it avoids the n^2 many pairwise evaluations in iterations. On the theory side, we propose a new error bound on the Wasserstein distance between two distributions in terms of the integrated difference between their Laplace transforms. Based on this, we can establish

consistency of our method. Our method is also distinct by its versatility in handling a wide range of distribution families, while most existing methods typically require the data generating distributions to be continuous.

We then show how our method applies to the problem of match up nodes in two low-rank networks such that the aligned networks "look similar". Numerical comparisons illustrate our method's significant advantages in both speed and accuracy over existing methods.

9. **Session title**: **Advances in estimation and prediction for understanding complex disorders**
Organizer: Heping Zhang (Yale)
Chair: Naveen Narisetty(UIUC)
Time: June 4th, 11:00am-12:30pm
Location: VEC 902

Speech 1: Uncertainty Quantification of Treatment Regime in Precision Medicine by Confidence Distributions
Speaker: Min-ge Xie (Rutgers)
Abstract: Personalized decision rule in precision medicine can be viewed as a "discrete parameter", for which theoretical development for statistical inference is lacking. In this talk, we propose a new way to quantify the estimation uncertainty in a personalized decision based on recent developments of confidence distribution (CD). Specifically, in a parametric regression model setup, suppose the decision for treatment versus control for an individual x_a is determined by a linear decision rule $D_a = I(x_a\beta > x_a\gamma)$, where β and γ are unknown regression coefficients in models for potential outcomes of treatment and control, respectively. The data-driven decision hat-D_a relies on the estimates of b and g, which in turn introduces uncertainty on the decision. In this work, we propose to find a CD for $\eta_a = x_a\beta - x_a\gamma$ and compute a "confidence measure" of the decision $\{Da = 1\} = \{ha > 0\}$. This measure has a value between 0 and 1, and provides a frequency-based assessment on how reliable our decision is. For example, if the confidence measure of the decision $\{Da = 1\}$ is 63%, then we know that, out of 100 patients who are the same as patient xa, 63 will benefit to have the treatment and 38 will be better off to be in the control group. Numerical study suggests that this new measurement is inline with classical assessments (such as sensitivity, specificity, etc.), but different from the classical assessments, this measurement can be directly computed from the observed data without the need to know the truth of $\{Da = 1\}$ or $\{Da = 0\}$. Utility of this new measure will also be demonstrated in an application of an adaptive-design clinical trial. (Joint work with Yilei Zhan and Sijian Wang)Personalized decision rule in precision medicine can be viewed as a "discrete parameter", for which theoretical development for statistical inference is lacking. In this talk, we propose a new way to quantify the estimation uncertainty in a personalized decision based on recent developments of confidence distribution (CD). Specifically, in a parametric regression model setup, suppose the decision for treatment versus control for an individual $x_a$is determined by a linear decision rule $D_a = I(x_a\beta > x_a\gamma)$, where β and γ are unknown regression coefficients in models for potential outcomes of treatment and control, respectively. The data-driven decision hat-D_a relies on the estimates of andg, which in

turn introduces uncertainty on the decision. In this work, we propose to find a CD for $\eta_{\alpha}=\xi_a\beta-\xi_a\gamma$ and compute a "confidence measure" of the decision $D_a = 1\} = \{\eta_a > 0\}$. This measure has a value between 0 and 1, and provides a frequency-based assessment on how reliable our decision is. For example, if the confidence measure of the decision $D_a = 1\}$ is 63%, then we know that, out of 100 patients who are the same as patient x_a, 63 will benefit to have the treatment and 38 will be better off to be in the control group. Numerical study suggests that this new measurement is inline with classical assessments (such as sensitivity, specificity, etc.), but different from the classical assessments, this measurement can be directly computed from the observed data without the need to know the truth of $D_a = 1\}$ or $D_a = 0\}$. Utility of this new measure will also be demonstrated in an application of an adaptive-design clinical trial. (Joint work with Yilei Zhan and Sijian Wang)

Speech 2: Semiparametric Estimation in the Secondary Analysis of Case-Control Studies
Speaker: Yanyuan Ma (Penn State)
Abstract:

We study the regression relationship among covariates in case-control data, an area known as the secondary analysis of case-control studies. The context is such that only the form of the regression mean is specified, so that we allow an arbitrary regression error distribution, which can depend on the covariates and thus can be heteroscedastic. Under mild regularity conditions we establish the theoretical identifiability of such models. Previous work in this context has either (a) specified a fully parametric distribution for the regression errors, (b) specified a homoscedastic distribution for the regression errors, (c) has specified the rate of disease in the population (we refer this as true population), or (d) has made a rare disease approximation. We construct a class of semiparametric estimation procedures that rely on none of these. The estimators differ from the usual semiparametric ones in that they draw conclusions about the true population, while technically operating in a hypothetic superpopulation. We also construct estimators with a unique feature, in that they are robust against the misspecification of the regression error distribution in terms of variance structure, while all other nonparametric effects are estimated despite of the biased samples. We establish the asymptotic properties of the estimators and illustrate their finite sample performance through simulation studies, as well as through an empirical example on the relation between red meat consumption and heterocyclic amines. Our analysis verified the positive relationship between red meat consumption and two forms of HCA, indicating that increased red meat consumption leads to increased levels of MeIQA and PhiP, both being risk factors for colorectal cancer.

Speech 3: Quantile Decision Trees and Forest with its application for predicting the risk (Post-Traumatic Stress Disorder) PTSD after experienced an acute coronary syndrome

Speaker: Ying Wei (Columbia)
Abstract: Classification and regression trees (CART) are a classic statistical learning method that efficiently partitions the sample space into mutually exclusive subspaces with the distinctive means of an outcome of interest. It is a powerful tool for efficient subgroup analysis and allows for complex associations and interactions to achieve high prediction accuracy and stability. Hence, they are appealing tools for precision health applications that deal with large amounts of data from EMRs, genomics, and mobile data and aim to provide a transparent decision mechanism. Although there is a vast literature on decision trees and random forests, most algorithms identify subspaces with distinctive outcome means. The most vulnerable or high-risk groups for certain diseases are often patients with extremely high (or low) biomarker and phenotype values. However, means-based partitioning may not be effective for identifying patients with extreme phenotype values. We propose a new regression tree framework based on quantile regression \cite{KoenkerBassett1978} that partitions the sample space and predicts the outcome of interest based on conditional quantiles of the outcome variable. We implemented and evaluated the performance of the conditional quantile trees/forests to predict the risk of developing PTSD after experiencing an acute coronary syndrome (ACS), using an observational cohort data from the REactions to Acute Care and Hospitalization (REACH) study\cite{onge2017depressive} at New York Presbyterian Hospital. The results show that the conditional quantile based trees/forest have better discrimination power to identify patients with severe PTSD symptoms, in comparison to the classical mean based CART.

10. **Session title**: **Survival analysis with high-dimensional data**
Organizer: Ingrid Van Keilegom (KU Leuven)
Chair: Ricardo Cao (Universidade da Coruña)
Time: June 4th, 11:00am-12:30am
Location: VEC 903

Speech 1: Robust optimal treatment regime estimation with survival outcome
Speaker: Lan Wang (U Minnesota)
Abstract: We consider estimating the single-stage quantile-optimal treatment regime from data with right-censored outcome. The proposed method is directly applicable to individualized treatment decision making in medicine when survival time or time to event is used as the primary end point, and the produced estimated rule is easy to interpret since the quantile-type statistics is widely used in such context. We proposed a nonparametric estimator belonging to value search category in literature, which directly estimates the optimal rule from a class of practically useful treatment regimes without posing constraints on the way treatment interact with covariates. We studied the nonstandard asymptotics for the estimated parameter of the optimal rule using semiparametric M-estimation theories, which reveals how censoring is influencing the uncertainty in the learned rule. (Joint work with Yu Zhou and Rui Song).

Speech 2: Fine-Gray Competing Risks Model with High-Dimensional Covariates: Estimation and Inference
Speaker: Jelena Bradic (University of California, San Diego)

Abstract: The purpose of this paper is to construct confidence intervals for the regression coefficients in the Fine-Gray model for competing risks data with random censoring, where the number of covariates can be larger than the sample size. Despite strong motivation from biostatistics applications, highdimensional Fine-Gray model has attracted relatively little attention among the methodological or theoretical literatures. We fill in this blank by proposing first a consistent regularized estimator and then the confidence intervals based on the one-step bias-correcting estimator. We are able to generalize the partial likelihood approach for the Fine-Gray model under random censoring despite many technical difficulties. We lay down a methodological and theoretical framework for the one-step bias-correcting estimator with the partial likelihood, which does not have independent and identically distributed entries. We also handle for our theory the approximation error from the inverse probability weighting (IPW), proposing novel concentration results for time dependent processes. In addition to the theoretical results and algorithms, we present extensive numerical experiments and an application to a study of non-cancer mortality among prostate cancer patients using the linked Medicare-SEER data.

Speech 3: Envelopes for censored quantile regression
Speaker: Yue Zhao (KU Leuven)
Abstract: Quantile regression has emerged as a powerful tool for survival analysis with censored data. We propose an efficient estimator for the coefficients in quantile regression with censored data using the envelope model. The envelope model uses dimension reduction techniques to identify material and immaterial components in the data, and forms the estimator of the regression coefficient based only on the material component, thus reducing the variability of the estimation. We will derive asymptotic properties of the proposed estimator and demonstrate its efficiency gains compared to the traditional estimator for the quantile regression with censored data. The starting point of our technical analysis is the Z-estimation approach with local weighing (e.g., Wang & Wang 2009) that in particular involves the conditional Kaplan-Meier estimator. Traditionally, the Kaplan-Meier estimator is treated as an infinite dimensional nuisance parameter. We will instead invoke the i.i.d. representation of the Kaplan-Meier estimator, which leads to a re-writing of our objective function as a U-process indexed by only the Euclidean parameter in the envelope model. The modified Z-estimation problem then becomes much more amenable to analysis.

11. **Session title**: Recent advances of high-dimensional statistical learning.
Organizer: Xiaotong Shen (U of Minnesota)
Chair: Xiaotong Shen (U of Minnesota)
Time: June 4th, 1:45pm - 3:15pm
Location: Hammer LL109 A/B

Speech 1: Multiclass Probability Estimation with Support Vector Machines
Speaker: Helen Zhang (Arizona State University)
Abstract: Multiclass classification and probability estimation have important applications in data analytics. We propose a simple and scalable estimation framework for multiclass probabilities based on kernel SVMs. The new estimator does not rely on

any parametric assumption on the data distribution, and hence it is flexible and robust. Theoretically, we can show that the proposed estimator is asymptotically consistent. Computationally, the new procedure can be conveniently implemented using standard SVM softwares. Our numerical studies demonstrate competitive performance of the new estimator when compared with existing methods such as multiple logistic regression, linear discrimination analysis (LDA), tree-based methods, and random forest (RF), under various settings.

Speech 2: Minimizing Sum of Truncated Convex Functions and Its Applications
Speaker: Hui Jiang (UMich)
Abstract: We study a class of problems where the sum of truncated convex functions is minimized. In statistical applications, they are commonly encountered when L0-penalized models are fitted and usually lead to NP-Hard non-convex optimization problems. We propose a general algorithm for the global minimizer in low-dimensional settings. We also extend the algorithm to high-dimensional settings, where an approximate solution can be found efficiently. We introduce several applications where the sum of truncated convex functions is used, compare our proposed algorithm with other existing algorithms in simulation studies, and show its utility in edge-preserving image restoration on real data.

Speech 3: Uncertainty and Inference for High-Dimensional Models Using the Solution Paths
Speaker: Peng Wang (University of Cincinnati)
Abstract: Bootstrap based model inference has been well studied. However, such approaches will almost certainly fail for high-dimensional models due to the fact that the selection results are highly sensitive to the choice of tuning parameter and therefore are extremely unstable. We proposed to utilize the information of the entire solution paths to overcome this obstacle. In particular, we select the best model based on the entire solution paths for each bootstrap sample, and make inference about both model and parameters using the results from all the bootstrap samples. Moreover, we also develop tools that visualize and quantify model selection uncertainty. These tools would allow practitioners evaluate the validity of the estimated models in high-dimensional settings.

12. **Session title**: Modern Multivariate Statistics: Tensors and Networks
Organizer: Jacob Bien (Cornell)
Chair: Jacob Bien (Cornell)
Time: June 4th, 1:45pm - 3:15pm
Location: VEC 404

Speech 1: Computationally Efficient Tensor Completion with Statistical Optimality
Speaker: Dong Xia (columbia)
Abstract: We develop methods for estimating a low rank tensor from noisy observations on a subset of its entries to achieve both statistical and computational efficiencies. There have been a lot of recent interests in this problem of noisy tensor completion. Much of the attention has been focused on the fundamental computational challenges often associated with problems involving higher order tensors, yet very little is known about

their statistical performance. To fill in this void, in this article, we characterize the fundamental statistical limits of noisy tensor completion by establishing minimax optimal rates of convergence for estimating a k-th order low rank tensor which suggest significant room for improvement over the existing approaches. Furthermore, we propose a polynomial-time computable estimating procedure based upon power iteration and a second-order spectral initialization that achieves the optimal rates of convergence. Our method is fairly easy to implement and numerical experiments are presented to further demonstrate the practical merits of our estimator.

Speech 2: Structured shrinkage of tensor parameters
Speaker: Peter Hoff (Duke)
Abstract:
Tensor-valued parameters arise in many multivariate statistical models, such as network autoregression where the relationship between a pair of nodes is potentially dependent on that of any other pair. Parameters in such models are likely to be near, but not necessarily in, low-dimensional subspaces of the parameter space. In this talk we discuss some adaptive empirical Bayes methods for shrinking parameter estimates towards an appropriately chosen subspace.

Speech 3: Global Spectral Clustering for Dynamic Networks
Speaker: Patrick Perry (NYU)
Abstract: In this talk, we present a new method (PisCES) for finding time-varying community structure in dynamic networks. The method implements spectral clustering, with a smoothing penalty to promote similarity across time periods. We prove that this method converges to the global solution of a nonconvex optimization problem, which can be interpreted as the spectral relaxation of a smoothed K-means clustering objective. We also show that smoothing is applied in a time-varying and data-dependent manner; for example, when a drastic change point exists in the data, smoothing is automatically suppressed at the time of the change point. Finally, we show that the detected time-varying communities can be effectively visualized through the use of sankey plots.

13. Session title: **Flexible Statistical Learning and Inference**
Organizer: Yufeng Liu (UNC)
Chair: Siliang Gong (UNC)
Time: June 4th, 1:45pm - 3:15pm
Location: VEC 405

Speech 1: Multi-layered Graphical Models
Speaker: Min Jin Ha (MD Anderson)
Abstract:
Simultaneous modeling of data arising from multiple ordered layers provides insight into the holistic picture of the interactive system and the flow of information. Chain graphs have been used to model the layered architecture of networks where the vertices can be naturally partitioned into ordered layers that exhibit undirected and directed acyclic relations within and between the layers. We develop a multi-layered Gaussian graphical

model (mlGGM) to investigate conditional independence structures in probabilistic chain graphs. Our proposed model uses a Bayesian node-wise selection framework that coherently accounts for dependencies in the mlGGM. Using Bayesian variable selection strategies for each of the node-wise regressions allows for flexible modeling, sparsity and incorporation of edge-specific prior knowledge. Through simulated data generated from various scenarios, we demonstrate that our node-wise regression method outperforms other related multivariate regression-based methodologies. We apply mlGGM to identify integrative networks for key signaling pathways in kidney cancer and dynamic signaling networks using longitudinal proteomics data in breast cancer.

Speech 2: Variable Selection for Highly Correlated Predictors
Speaker: Fei Xue (UIUC)
Abstract: Penalty-based variable selection methods are powerful in selecting relevant covariates and estimating coefficients simultaneously. However, variable selection could fail to be consistent when covariates are highly correlated. The partial correlation approach has been adopted to solve the problem with correlated covariates. Nevertheless, the restrictive range of partial correlation is not effective for capturing signal strength for relevant covariates. In this paper, we propose a new Semi-standard PArtial Covariance (SPAC) which is able to reduce correlation effects from other predictors while incorporating the magnitude of coefficients. The proposed SPAC variable selection facilitates choosing covariates which have direct association with the response variable, via utilizing dependency among covariates. We show that the proposed method with the Lasso penalty (SPAC-Lasso) enjoys strong sign consistency in both finite-dimensional and high-dimensional settings under regularity conditions. Simulation studies and the 'HapMap' gene data application demonstrate that the proposed method outperforms the traditional Lasso, adaptive Lasso, SCAD, and Peter–Clark-simple (PC-simple) methods for highly correlated predictors.

Speech 3: Support Vector Machine with Confidence
Speaker: Xingye Qiao (SUNY Binghamton)
Abstract: Classification with confidence (Lei, 2014) is a new type of problem in statistical learning. Its goal, in the binary case, is to identify two regions with a specific coverage probability for each class. We propose a support vector classifier to achieve classification with confidence. The classifier has two boundaries. An observation outside of the two boundaries is deemed to be from one of the two classes (with certain confidence), while the region between the boundaries is an ambiguity region which could belong to either class. In the theoretical study, we show a Fisher consistency result and that, with high probability, the resulting classifier can control the non-coverage rates and minimize the ambiguity. Efficient algorithms are developed and numerical studies are conducted to illustrate the effectiveness of the proposed method. This is a joint work with Wenbo Wang.

14. **Session title**: Philosophy of Science and the New Paradigm of Data-Driven Science
Organizer: Todd Kuffner (Washington U)
Chair: Todd Kuffner (Washington U)
Time: June 4th, 1:45pm - 3:15pm
Location: VEC 902/903

Speech 1: Your Data-Driven Claims Must Still be Probed Severely
Speaker: Deborah Mayo (Virginia Tech)
Abstract: I analyse some of the philosophical claims from leaders of the "new paradigm" of data-driven science regarding the end of theory and the obsoleteness of scientific method. Science may be data intensive, but data are theory laden, so conceptual assumptions and biases must still be examined. Science may be question driven, but good science can't opt out of probing flaws in any potential answers to those questions.

Speech 2: On the replicability of scientific studies
Speaker: Ian McKeague (Columbia)
Abstract:

I will discuss a number of issues, both statistical and philosophical, related to the replicability and verification of scientific results. In particular, I will discuss a recent proposal of Munafo and Davey Smith (Nature, 2018) that such verification requires disparate lines of evidence, a technique that they call triangulation.

Speech 3: Conducting Highly Principled Data Science: A Statistician's Job and Joy
Speaker: Xiao-Li Meng (Harvard)
Abstract:

This talk is based on a contribution to the special issue on ``The Role of Statistics in the Era of Big Data" organized by *Statistics and Probability Letters,*

with the title above and the following abstract: "Highly Principled Data Science insists on methodologies that are: (1) scientifically justified; (2) statistically principled; and (3) computationally efficient. An astrostatistics collaboration, together with some reminiscences, illustrates the increased roles statisticians can and should play to ensure this trio, and to advance the science of data along the way. "

15. **Session title**: Advances in Bayesian methods for high-dimensional data
Organizer: Howard Bondell (U. of Melbourne)
Chair: Xuan Bi (Yale)
Time: June 4th, 1:45pm - 3:15pm
Location: VEC 1202/1203

Speech 1: The Graphical Horseshoe Estimator for Inverse Covariance Matrices
Speaker: Anindya Bhadra (Purdue)

Abstract: 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. This is joint work with Yunfan Li and Bruce Craig at Purdue.

Speech 2: Scalable MCMC for Bayes shrinkage priors
Speaker: Anirban Bhattacharya (Texas A & M)
Abstract: Gaussian scale mixture priors are common in high-dimensional Bayesian analysis. While optimization algorithms for the extremely popular Lasso and elastic net scale to dimension in the hundreds of thousands, Bayesian computation by Markov chain Monte Carlo (MCMC) is limited to problems an order of magnitude smaller. This is due to high computational cost per step and growth of the variance of time-averages as a function of dimension. We propose an MCMC algorithm for computation in these models that combines block updating and approximations of the Markov kernel to combat both of these factors. Our algorithm gives orders of magnitude speedup over the best existing alternatives in high-dimensional applications. We give theoretical guarantees for the accuracy of the approximation. Scalability of the algorithm is illustrated in an application to a genome wide association study with $N=2,267$ observations and $p=98,385$ predictors. The empirical results show that the new algorithm yields estimates with lower mean squared error, intervals with better coverage, and elucidates features of the posterior often missed by previous algorithms, including bimodality of marginals indicating uncertainty about which covariates belong in the model. This latter feature is an important motivation for a Bayesian approach to testing and selection in high dimensions.(joint work with James Johndrow & Paulo Orenstein)

Speech 3: Clustering on the Sphere: State-of-the-art and a Poisson Kernel-Based Model
Speaker: Marianthi Markatou (U. at Buffalo)
Abstract: Many applications of interest involve data that can be analyzed as unit vectors on a d-dimensional sphere. Specific examples include text mining, biology, astronomy and medicine. We present a clustering method based on mixtures of Poisson-kernel based densities on the high-dimensional sphere. We study connections of the Poisson kernel-based densities with other distributions appropriate for the analysis of directional data, prove identifiability of mixtures of the Poisson kernel-based densities model, convergence of the associated EM-type algorithm, and study its operational characteristics. We further propose an *empirical densities distance plot* for estimating the number of clusters in a Poisson kernel-based densities model. Finally, we propose a method to simulated data from Poisson kernel-based densities and exemplify our methods

via application on real data sets and simulation experiments. Our experimental results show that the newly introduced model exhibits higher macro-precision and macro-recall than competing methods based on von Mises Fisher and Watson distributions. *This is joint work with Mojgan Golzy, Ph.D.*

16. **Session title**: High-dimensional machine learning methods
Organizer: Annie Qu (UIUC)
Chair: Annie Qu (UIUC)
Time: June 4th, 1:45pm - 3:15pm
Location: VEC 1302/1303

Speech 1: Latent Space Approaches to Community Detection in Dynamic Networks
Speaker: Yuguo Chen (UIUC)
Abstract: Embedding dyadic data into a latent space has long been a popular approach to modeling networks of all kinds. While clustering has been done using this approach for static networks, we give two methods of community detection within dynamic network data, building upon the distance and projection models previously proposed in the literature. Our proposed approaches capture the time-varying aspect of the data, can model directed or undirected edges, inherently incorporate transitivity and account for each actor's individual propensity to form edges. We provide Bayesian estimation algorithms, and apply these methods to a ranked dynamic friendship network and world export/import data.

Speech 2: Classification for High-Dimensional Functional Data
Speaker: Taps Maiti (MSU)
Abstract: The functional magnetic resonance imaging (fMRI) records signals coming from different areas in human brains, which show activities and states of brains. This measurements result in high-dimensional time series, and each dimension represents a region in brains. In this work, we propose a classification technique of this kind of high-dimensional time series data while we discuss several competitive approaches. In addition, we present two classification applications to demonstrate the performance of your method. One is an alcoholic condition detection with Electroencephalography (EEG) data collected from electrodes placed on subject's scalps, and the other is a resting state detection using resting state fMRI data from the OpenfMRI database. In both applications, we compare the performance of our method with some other classification methods.

Speech 3: A unified approach for censored quantile regression
Speaker: Naveen Narisetty (UIUC)
Abstract: In this talk, I will present a new and unified approach for estimation of quantile regression under censoring of arbitrary types. The proposed method is based on a variation of the data augmentation algorithm and easily adapts to different forms of censoring including doubly censored and interval censored data unlike existing methods which are mostly limited to single censoring. Moreover, the proposed estimators improve on the performance of the best known estimators with singly censored data. Empirical results demonstrating the fine performance of the proposed approach will be presented.

This is joint work with Xiaorong Yang and Xuming He.

17. **Session title**: Recent development of Statistical Neuroimaging Analysis
Organizer: Lexin Li (UC Berkeley)
Chair: Jason Lee (USC)
Time: June 4th, 1:45pm - 3:15pm
Location: VEC 1402/1403

Speech 1: Analyzing Non-Stationary High-Dimensional Time Series: Structural Break Detection and Parameter Estimation
Speaker: Ali Shojaie (U of Washington)
Abstract: Assuming stationarity is unrealistic in many time series applications, including neuroscience. A more realistic alternative is to assume piecewise stationarity, where the model is allowed to change at potentially many time points. We propose a three-stage procedure for consistent estimation of both structural change points and parameters of high-dimensional piecewise vector autoregressive (VAR) models. In the first step, we reformulate the change point detection problem as a high-dimensional variable selection one, and solve it using a penalized least square estimator with a total variation penalty. We show that the proposed penalized estimation method over-estimates the number of change points. We then propose a backward selection criterion to identify the change points. In the last step of our procedure, we estimate the VAR parameters in each of the segments. We show that the proposed procedure consistently detects the number of change points and their locations. We also show that the procedure consistently estimates the VAR parameters. The performance of the method is illustrated through several simulation studies, as well as an analysis of EEG data.

Speech 2: Tensor-on-tensor regression
Speaker: Eric Lock (U of Minnesota)
Abstract: In neuroimaging analysis and other fields, both predictors and outcomes can take the form of a multi-way array (i.e., a tensor). We propose a framework for the linear prediction of a multi-way array from another multi-way array of arbitrary dimension, using the contracted tensor product. This framework generalizes several existing approaches, including methods to predict a scalar outcome from a tensor, a matrix from a matrix, or a tensor from a scalar. We describe an approach that exploits the multiway structure of both the predictors and the outcomes by restricting the coefficients to have reduced CP-rank. We propose a general and efficient algorithm for penalized least-squares estimation, which allows for a ridge (L_2) penalty on the coefficients. The objective is shown to give the mode of a Bayesian posterior, which motivates a Gibbs sampling algorithm for inference. We illustrate the approach with an application to facial image data.

Speech 3: A Joint Modeling Approach for Baseline Matrix-valued Imaging Data and Treatment Outcome
Speaker: Bei Jiang (University of Alberta)
Abstract: In this talk we propose a unified Bayesian joint modeling framework for studying association between a binary treatment outcome and a baseline matrix-valued

predictor, such as imaging data. Under this framework, a theoretically implied relationship can be established between the treatment outcome and the matrix-valued imaging data, although the imaging data is not directly considered in the model. The proposed joint modeling approach provides a promising framework for both association estimation and prediction. Properties of this method are examined using simulated datasets. Finally, a detailed illustration of the proposed modeling approach is provided using a motivating depression study that aims to explore the association between the baseline EEG data and the probability of a favorable response to an antidepressant treatment.

18. **Session title**: Nonparameteric and Robust Statistical Methods for Imaging
Organizer: Hernando Ombao (KAUST)
Chair: Wei Pan(UMN)
Time: June 5th, 8:30am – 10:00am
Location: VEC 405

Speech 1: Nonparametric Collective Spectral Density Estimation and Clustering with Application to Brian Activities
Speaker: Mehdi Maadooliat (Marquette University)
Abstract: In this paper, we develop a method for the simultaneous estimation of spectral density functions (SDFs) for a collection of stationary time series that share some common features. Due to the similarities among the SDFs, the log-SDF can be represented using a common set of basis functions. The basis shared by the collection of the log-SDFs is estimated as a low-dimensional manifold of a large space spanned by a pre-specified rich basis. A collective estimation approach pools information and borrows strength across the SDFs to achieve better estimation efficiency. Also, each estimated spectral density has a concise representation using the coefficients of the basis expansion, and these coefficients can be used for visualization, clustering, and classification purposes. The Whittle pseudo-maximum likelihood approach is used to fit the model and an alternating blockwise Newton-type algorithm is developed for the computation. A web-based shiny Appfound at https://ncsde.shinyapps.io/NCSDE is developed for visualization, training and learning the SDFs collectively using the proposed technique. Finally, we apply our method to cluster similar brain signals recorded by the electroencephalogram for identifying synchronized brain regions according to their spectral densities.

Speech 2: A Flexible Non-parametric Framework for Imaging Genetics
Speaker: Zhaoxia Yu (UC Irvine)
Abstract: Data collected in many scientific areas are inherently high-dimensional and multi-way. While such data provides an excellent opportunity for us to conduct an integrative analysis of multiple data modalities, it is challenging to model associations between sets of massive, complexly structured, and high-dimensional data. Here we propose a flexible, easy-to-implement, and non-parametric framework to assessing the overall association between high-dimensional modalities. We illustrate how the methods are connected to classical regression-based methods. To take some important structure of brain imaging data into consideration, we also extend our method to non-Euclidean

space. The principles that we propose are applicable to various types of high dimensional data, such as genetic variants and brain connectivity.

Speech 3: Hybrid Principal Components Analysis For Region-Referenced Longitudinal Functional EEG Data
Speaker: Damla Senturk (UCLA)
Abstract: Electroencephalography (EEG) data possess a complex structure that includes regional, functional, and longitudinal dimensions. Our motivating example is a word segmentation paradigm in which typically developing (TD) children and children with Autism Spectrum Disorder (ASD) were exposed to a continuous speech stream. For each subject, continuous EEG signals recorded at each electrode were divided into one-second segments and projected into the frequency domain via Fast Fourier Transform. Following a spectral principal components analysis, the resulting data consist of region-referenced principal power indexed regionally by scalp location, functionally across frequencies and longitudinally by one-second segments. Standard EEG power analyses often collapse information across the longitudinal and functional dimensions by averaging power across segments and concentrating on specific frequency bands. We propose a hybrid principal components analysis (HPCA) for region-referenced longitudinal functional EEG data which utilizes both vector and functional principal components analyses and does not collapse information along any of the three dimensions of the data. The proposed decomposition only assumes weak separability of the higher-dimensional covariance process and utilizes a product of one dimensional eigenvectors and eigenfunctions, obtained from the regional, functional, and longitudinal marginal covariances, to represent the observed data, providing a computationally feasible non-parametric approach. A mixed effects framework is proposed to estimate the model components coupled with a bootstrap test for group level inference, both geared towards sparse data applications. Analysis of the data from the word segmentation paradigm leads to valuable insights about group-region differences among the TD and verbal and minimally verbal children with ASD. Finite sample properties of the proposed estimation framework and bootstrap inference procedure are further studied via extensive simulations.

19. **Session title**: Big Data of different forms and different challenges
Organizer: Regina Liu (Rutgers)
Chair: Heping Zhang(Yale)
Time: June 5th, 8:30am – 10:00am
Location: VEC 1202 /1203

Speech 1: Individualized Multilayer Learning with An Application in Breast Cancer Imaging
Speaker: Annie Qu (UIUC)
Abstract: This work is motivated by breast cancer imaging data produced by a multimodal multiphoton optical imaging technique. One unique aspect of breast cancer imaging is that different individuals might have breast imaging at different locations, which also creates a technical difficulty in that the imaging background could vary for different individuals. We develop a multilayer tensor learning method to predict disease status effectively through utilizing subject-wise imaging information. In particular, we

construct an individualized multilayer model which leverages an additional layer of individual structure of imaging in addition to employing a high-order tensor decomposition shared by populations. In addition, to incorporate multimodal imaging data for different profiling of tissue, cellular and molecular levels, we propose a higher order tensor representation to combine multiple sources of information at different modalities, so important features associated with disease status and clinical outcomes can be extracted effectively. One major advantage of our approach is that we are able to capture the spatial information of microvesicles observed in certain modalities of optical imaging through combining multimodal imaging data. This has medical and clinical significance since microvesicles are more frequently observed among cancer patients than healthy ones, and identification of microvesicles enables us to provide an effective diagnostic tool for early-stage cancer detection. This is joint work with Xiwei Tang and Xuan Bi.

Speech 2: Efficient estimation and fast algorithms for genetic microarray data with survival outcomes
Speaker: Catherine Chunling Liu (Polytech U of HK)
Abstract: In gene expression microarray studies, genetic and genomic data tend to be high- or ultrahigh- dimensional and are accompanied with random censored survival outcomes. To search out and evaluate influence features that will impact on the disease, it is imperative to develop new modeling, efficient estimation methodology, and feasible algorithms within such data setting. In this talk, we will discuss in three aspects. First of all, for ultrahigh dimensional data modeled by the proportional hazard model, we present a non-monotone proximal gradient algorithm with lasso-type initial value to do feature screening and variable selection; Next, we recommend a single index hazard model without specifying the functional form. Efficient estimation procedures for index coefficients will facilitate detection of significance of individual effects. Finally we consider jointly modeling the mean and intensity function involving multiple index structure and develop a unified methodology to conduct dimension reduction. A normal acute myeloid leukemia data is analyzed to demonstrate our approaches.

Speech 3: Nonparametric mean estimation for big-but-biased data
Speaker: Ricardo Cao (Universidade da Coruña)
Abstract: Some authors have recently warned about the risks of the sentence "with enough data, the numbers speak for themselves". Some of the problems coming from ignoring sampling bias in big data statistical analysis have been recently reported. The problem of nonparametric statistical inference in big data under the presence of sampling bias is considered in this work. The mean estimation problem is studied in this setup, in a nonparametric framework, when the biasing weight function is known (unrealistic) as well as for unknown weight functions (realistic). Two different scenarios are considered to remedy the problem of ignoring the weight function: (i) having a small sized simple random sample of the real population and (ii) having observed a sample from a doubly biased distribution. In both cases the problem is related to nonparametric density estimation. Asymptotic expressions for the mean squared error of the estimators proposed for scenario (i) are considered. This leads to asymptotic formulas for the optimal smoothing parameters. Some simulations illustrate the performance of the nonparametric

methods proposed in this work.

20. **Session title**: OODA: Manifold Data Integration
Organizer: Marron, James Stephen (UNC)
Chair: Anna Smith (Columbia)
Time: June 5th, 8:30am – 10:00am
Location: VEC 1302

Speech 1: Random Domain Decomposition for Kriging Riemannian Data
Speaker: Piercesare Secchi (Politecnico di Milano)
Abstract:

Investigation of object data distributed over complex domains gives rise to new challenges for spatial statistics. The linear methods of geostatistics cannot be directly applied when the embedding space for the observed data is non Hilbert. Moreover global assumptions about the stationarity of the random field generating the data are often unsuitable when the spatial domain is large, textured or convoluted, with holes or barriers. We here examine the Random Domain Decomposition computational approach proposed by Menafoglio et al. (2018) applied to the analysis and prediction of tensor data observed over a complex spatial domain. As an illustrative case study, we will analyse the covariances between dissolved oxygen and water temperature in the Chesapeake Bay.

This is joint work with Alessandra Menafoglio (Dept. of Mathematics, Politecnico di Milano) and Davide Pigoli (Dept. of Mathematics, King's College London)

Speech 2: Nonparametric K-Sample Test on Riemannian Manifolds with Applications to Analyzing Mitochondrial Shapes
Speaker: Ruiyi Zhang (Florida State)
Abstract: This paper develops a nonparametric approach for a k-sample test involving data lying on a Riemannian manifold, such as a shape manifold. The specific problem is to test the hypothesis that a factor (such as the subject, cell, or living conditions) significantly affects mitochondrial morphology as observed in images of skeletal muscles of mice. The fact that a shape space is non-Euclidean and infinite-dimensional rules out standard ANOVA decomposition and requires new ideas. In this work, we extend a metric-based approach, developed for Euclidean spaces previously and termed DISCO analysis, to the several shape representations of planar closed curves. This adaptation leads to a statistic for testing equality of distributions of across groups. We provide the underlying theory for one shape representation, but apply the test to several other shape metrics also. Since the data have a nested structure, we also develop a procedure to test a factor while it includes another significant factor. We analyze and present results for a mitochondria shape dataset, including an interesting result that a change in lifestyle alters shapes of some type of mitochondria.

Speech 3: High-Dimensional Manifold Data Clustering on Symmetric Spaces
Speaker: Chao Huang (UNC)
Abstract: Clustering is one of the fundamental tools in manifold learning, and it has been extensively studied in many applications. However, in many image analysis problems (e.g., directional data analysis, shape analysis), most existing clustering methods established in Euclidian space face several challenges including a symmetric space, a high dimensional feature space, and manifold data variation associated with some covariates. In order to address such challenges, a penalized model-based clustering framework is developed to cluster high dimensional manifold data in symmetric spaces. Specifically, a mixture of geodesic factor analyzers (MGFA) is proposed with mixing proportions de- fined through a logistic model and Riemannian normal distribution in each component for data in symmetric spaces. A geodesic factor analyzer is established to explicitly model the high dimensional features. Penalized likelihood approaches are used to real- ize variable selection procedures. Simulation studies are performed on data generated from unit sphere, and real data analysis are performed on the corpus callosum (CC) shape data from the ADNI study.

21. **Session title**: Advances in high-dimensional statistics
Organizer: Genevera Allen (Rice)
Chair: Genevera Allen (Rice)
Time: June 5th, 8:30am – 10:00am
Location: VEC 1303

Speech 1: Adaptive local estimation for high dimensional linear models
Speaker: Yufeng Liu (UNC)
Abstract: High dimensional linear models are commonly used in practice. In many applications, one is interested in linear transformations of regression coefficients such as prediction of the response. One common approach is the global technique which first estimates the coefficients, then plugs the estimator in the linear transformation for prediction. Despite its popularity, regression estimation can be difficult for high dimensional problems. Commonly used assumptions in the literature include that the signal of coefficients is sparse and predictors are weakly correlated. These assumptions, however, may not be easily verified, and can be violated in practice. When the coefficients are not sparse or predictors are strongly correlated, estimation of coefficients can be very difficult. In this talk, I will present a new adaptive local estimator for linear transformations of the coefficients. This new estimator greatly relaxes the common assumptions for high dimensional problems. Simulation and theoretical results demonstrate the competitive advantages of the proposed method for a wide range of problems.

Speech 2: Are Clusterings of Multiple Data Views Independent?
Speaker: Jacob Bien (Cornell)
Abstract: It has become increasingly common for scientists to collect more than one type of measurement on a single set of observations. For instance, a medical researcher might

gather both clinical measurements and DNA sequences on a single set of individuals. Many "multi-view clustering" methods have been developed, which use the information from the different data views to determine a clustering of the observations. In this talk, however, we explore a more basic question: is the clustering structure from each data view related or independent? We develop a hypothesis test for investigating this question on the basis of a set of data. This is joint work with Lucy Gao and Daniela Witten.

Speech 3: Regularized Robust Buckley-James method for AFT Model with General Loss Function
Speaker: Sijian Wang (Rutgers)
Abstract: In the last decade of genome research, one of the most popular topics is to relate large numbers of gene information to clinical survival phenotypes. As an alternative to the popular Cox's proportional hazard model, the Accelerated Failure Time (AFT) model specifies an association between survival time and covariates directly. Consequently, it has a simpler and possible more intuitive interpretation than Cox's model which is based on the hazard function. The Buckley-James method is a popular method to get the estimation for the AFT model. It iterates between the imputation of failure time for censored subjects and the estimation of regression coefficients. In the estimation step, the least square criterion with quadratic loss function is used to get the estimation. It is well known that, for regression with uncensored data, the traditional estimation resulting from the quadratic loss function may not be robust when the variability of response is high and/or there are outliers. In this talk, we proposed a regularized Robust Buckley-James method which can incorporate general loss functions including the absolute loss function, quantile loss function, Huber's loss function and Tukey's bisquare loss function. The proposed methods are demonstrated using simulation studies and analysis of a TCGA ovarian cancer dataset.

22. **Session title**: Causal Inference and Machine Learning
Organizer: Ryan Tibshirani (CMU)
Chair: Vincent Joseph Dorie (Columbia)
Time: June 5th, 8:30am – 10:00am
Location: VEC 1402

Speech 1: Nonparametric causal effects based on incremental propensity score interventions
Speaker: Edward Kennedy (CMU)
Abstract:

Most work in causal inference considers deterministic interventions that set each unit's treatment to some fixed value. However, under positivity violations these interventions can lead to non-identification, inefficiency, and effects with little practical relevance. Further, corresponding effects in longitudinal studies are highly sensitive to the curse of dimensionality, resulting in widespread use of unrealistic parametric models. We propose a novel solution to these problems: incremental interventions that shift propensity score values rather than set treatments to fixed values. Incremental interventions have several

crucial advantages. First, they avoid positivity assumptions entirely. Second, they require no parametric assumptions and yet still admit a simple characterization of longitudinal effects, independent of the number of timepoints. For example, they allow longitudinal effects to be visualized with a single curve instead of lists of coefficients. After characterizing these incremental interventions and giving identifying conditions for corresponding effects, we also develop general efficiency theory, propose efficient nonparametric estimators that can attain fast convergence rates even when incorporating flexible machine learning, and propose a bootstrap-based confidence band and simultaneous test of no treatment effect. Finally we explore finite-sample performance via simulation, and apply the methods to study time-varying sociological effects of incarceration on entry into marriage.

Speech 2: Quasi-Oracle Estimation of Heterogeneous Causal Effects
Speaker: Stefan Wager (Stanford)
Abstract: Many scientific and engineering challenges, ranging from personalized medicine to customized marketing recommendations, require an understanding of treatment effect heterogeneity. In this paper, we develop a class of two-step algorithms for heterogeneous treatment effect estimation in observational studies. We first estimate marginal effects and treatment propensities to form an objective function that isolates the heterogeneous treatment effects, and then optimize the learned objective. This approach has several advantages over existing methods. From a practical perspective, our method is very flexible and easy to use: In both steps, we can use any method of our choice, e.g., penalized regression, a deep net, or boosting; moreover, these methods can be fine-tuned by cross-validating on the learned objective. Meanwhile, in the case of penalized kernel regression, we show that our method has a quasi-oracle property, whereby even if our pilot estimates for marginal effects and treatment propensities are not particularly accurate, we achieve the same regret bounds as an oracle who has a-priori knowledge of these nuisance components. We implement variants of our method based on both penalized regression and convolutional neural networks, and find promising performance relative to existing baselines.

Speech 3: Off-policy Learning in Theory and in the Wild
Speaker: Yu-Xiang Wang (Amazon/UCSB)
Abstract: The talk considers the problem of offline policy learning for automated decision systems under the contextual bandits model, where we aim at evaluating the performance of a given policy (a decision algorithm) and also learning a better policy using logged historical data consisting of context, actions, rewards and probabilities of the actions taken. This is a generalization of the Average Treatment Effect (ATE) estimation problem and has some interesting new set of desiderata to consider.

In the first part of the talk, I will compare and contrast off-policy evaluation and ATE estimation and clarify how different assumptions change the corresponding minimax risk in estimating the ``causal effect''. In addition, I will talk about how one can achieve

significantly better finite sample performance than asymptotically optimal estimators through the SWITCH estimator.

In the second part of the talk, I will talk about off-policy evaluation and learning in a real industry environment. I will highlight several intersting challenges there including partially logged probilities, unobserved decision variables (Simpson's paradox), effect of model bias and so on. We then propose and recommend practical ways to deal with these challenges under different circumstances.

23. **Session title**: Decision making, operations research and statistical learning
Organizer: Cynthia Rudin (Duke)
Chair: Cynthia Rudin (Duke)
Time: June 5th, 8:30am – 10:00am
Location: VEC 1403

Speech 1: Online Learning of Buyer Behavior under Realistic Pricing Restrictions
Speaker: Theja Tulabandhula (UIC)
Abstract: We propose a new efficient online algorithm to learn the parameters governing the purchasing behavior of a utility maximizing buyer, who responds to prices, in a repeated interaction setting. The key feature of our algorithm is that it can learn even non-linear buyer utility while working with arbitrary price constraints that the seller may impose. This overcomes a major shortcoming of previous approaches, which use unrealistic prices to learn these parameters making them unsuitable in practice.

Speech 2: Smart "Predict, then Optimize"
Speaker: Adam Elmachtoub (Columbia)
Abstract: We consider a class of optimization problems where the objective function is not explicitly provided, but contextual information can be used to predict the objective based on historical data. A traditional approach would be to simply predict the objective based on minimizing prediction error, and then solve the corresponding optimization problem. Instead, we propose a prediction framework that leverages the structure of the optimization problem that will be solved given the prediction. We provide theoretical, algorithmic, and computational results to show the validity and practicality of our framework. This is joint work with Paul Grigas (UC Berkeley).

Speech 3: P-splines with an l1 penalty for repeated measures
Speaker: Brian Segal (Flatiron Health)
Abstract: P-splines are penalized B-splines, in which finite order differences in coefficients are typically penalized with an l2 norm. P-splines can be used for semiparametric regression and can include random effects to account for within-subject variability. In addition to l2 penalties, l1-type penalties have been used in nonparametric and semiparametric regression to achieve greater flexibility, such as in locally adaptive regression splines, l1 trend filtering, and the fused lasso additive model. However, there has been less focus on using l1 penalties in P-splines, particularly for estimating conditional means. We demonstrate the potential benefits of using an l1 penalty in P-splines with an emphasis on fitting non-smooth functions. We propose an estimation

procedure using the alternating direction method of multipliers and cross validation, and provide degrees of freedom and approximate confidence bands based on a ridge approximation to the l1 penalized fit. We also demonstrate potential uses through simulations and an application to electrodermal activity data collected as part of a stress study.

24. **Session title**: Novel inference approaches for complex data setting
Organizer: Regina Liu (Rutgers)
Chair: Junhui Wang (City U. of Hong Kong)
Time: June 5th, 1:15pm - 2:45pm
Location: VEC 1202/1203

Speech 1: Connecting pairwise spheres by depth : DCOPS
Speaker: Ricardo Fraiman (Universidad de la República de Uruguay)
Abstract: We extend the notion of spherical depth to the important setup of complex data in Riemannian manifolds. We show that this depth notion shares the set of desirable properties. For the empirical version of this depth function both uniform consistency and the asymptotic distribution are studied. The behaviour of the depth is illustrated through several examples in Riemannian manifolds.

Speech 2: Stein Discrepancy Methods for Robust Estimation and Regression
Speaker: Emre Barut (George Washington University)
Abstract: All statistical procedures highly depend on the modeling assumptions and how close these assumptions are to reality. This dependence is critical: Even the slightest deviation from assumptions can cause major instabilities during statistical estimation.

In order to mitigate issues arising from model mismatch, numerous methods have been developed in the area of robust statistics. However, these approaches are aimed at specific problems, such as heavy tailed or correlated errors. The lack of a holistic framework in robust regression results in a major problem for the data practitioner. That is, in order to build a robust statistical model, possible issues in the data have to be found and understood before conducting the analysis. In addition, the practitioner needs to have an understanding of which robust models can be applied in which situations.

In this talk, we propose a new framework for robust parameter estimation to address these issues. The new method relies on the Stein Discrepancy Measure, and the estimate is given as the empirical minimizer of a second order U-statistic. The approach provides a "silver bullet" that can be used in a range of problems. When estimating parameters in the exponential family, the estimate can be obtained by solving a convex problem. For parameter estimation, our approach significantly improves upon MLE when outliers are present, or when the model is misspecified. Furthermore, we show how the new estimator can be used for robust high dimensional covariance estimation. Extensions of the method for regression problems and its efficient computation by subsampling are also discussed.

Speech 3: Estimating a covariance function from fragments of functional data
Speaker: Aurore Delaigle (U of Melbourne)
Abstract: Functional data are often observed only partially, in the form of fragments. In that case, the standard approaches for estimating the covariance function do not work because entire parts of the domain are completely unobserved. In previous work, Delaigle and Hall (2013, 2016) have suggested ways of estimating the covariance function, based for example on Markov assumptions. In this work we take a completely different approach which does not rely on such assumptions. We show that, using a tensor product approach, it is possible to reconstruct the covariance function using observations located only on the diagonal of its domain.

25. **Session title**: New development for analyzing biomedical complex data
Organizer: Zhezhen Jin (Columbia)
Chair: Peng Wang (University of Cincinnati)
Time: June 5th, 1:15pm - 2:45pm
Location: VEC 1302

Speech 1: New methods for estimating follow-up rates in cohort studies
Speaker: Xiaonan Xue (Albert Einstein College of Medicine)
Abstract:
Background: The follow-up rate, a standard index of the completeness of follow-up, is important for assessing the validity of a cohort study. A common method for estimating the follow-up rate, the "Percentage Method", defined as the fraction of all enrollees who developed the event of interest or had complete follow-up, can severely underestimate the degree of follow-up. Alternatively, the median follow-up time does not indicate the completeness of follow-up, and the reverse Kaplan-Meier based method and Clark's Completeness Index (CCI) also have limitations.
Methods: We propose a new definition for the follow-up rate, the Person-Time Follow-up Rate (PTFR), which is the observed person-time divided by total person-time assuming no dropouts. The PTFR cannot be calculated directly since the event times for dropouts are not observed. Therefore, two estimation methods are proposed: a formal person-time method (FPT) in which the expected total follow-up time is calculated using the event rate estimated from the observed data, and a simplified person-time method (SPT) that avoids estimation of the event rate by assigning full follow-up time to all events. Simulations were conducted to measure the accuracy of each method, and each method was applied to a prostate cancer recurrence study dataset.
Results: Simulation results showed that the FPT has the highest accuracy overall. In most situations, the computationally simpler SPT and CCI methods are only slightly biased. When applied to a retrospective cohort study of cancer recurrence, the FPT, CCI and SPT showed substantially greater 5-year follow-up than the Percentage Method (92%, 92% and 93% vs 68%).
Conclusions: The Person-time methods correct a systematic error in the standard Percentage Method for calculating follow-up rates. The easy to use SPT and CCI methods can be used in tandem to obtain an accurate and tight interval for PTFR. However, the FPT is recommended when event rates and dropout rates are high.

Speech 2: Mediation analysis with time-to-event mediator
Speaker: Mengling Liu (New York University)
Abstract: Mediation analysis is often employed in social and biomedical sciences to facilitate understanding of the effects that an intervention exerts over an outcome either directly or through a mediator. A motivating question is to study the pathway from circulating level of anti-Mullerian hormone (AMH, the exposure) to age at menopause (the mediator) and then to post-menopausal breast cancer (the outcome). However, challenge in the AMH mediation analysis is that the mediator variable, age at menopause, is a time-to-event variable and subject to censoring. Statistical methods to handle censored time-to-event mediator, or incompletely observed mediator in general, are lacking. We propose a series of statistical inference approaches for mediation analysis with a censored time-to-event mediator. Specifically, an estimating equation-based method is proposed for continuous outcomes and incorporates censored mediator through mean residual life (MRL) modeling; and a likelihood-based approach is proposed for categorical outcomes and estimates mediation effects through Monte Carlo methods to handle censored mediator.

Speech 3: Adjustment for covariates in genome-wide association study
Speaker: Tao Wang (Albert Einstein College of Medicine)
Abstract: Genome-wide association study (GWAS) has become a popular approach for identifying common genetic variants associated with complex diseases and quantitative traits. For continuous traits, linear regression model is a standard approach widely used in GWAS. Adjustment of covariates is often used to identify the direct effects of genetic variants or to improve statistical power by reducing variability of the trait. However, it is problematic to adjust for heritable covariates. Here, we propose a new method for adjusting covariates by incorporating prior GWAS summary statistics for inferring the direct biological influence on a given trait and improve statistical power. Using simulation studies, the proposed methodology remains a good control of type I error rate under various situations and can achieve high power than a simple linear regression. The method is illustrated by an application to a GWAS analysis.

26. **Session title**: New Statistical Machine Learning Tools
Organizer: Liu, Yufeng (UNC)
Chair: Liu, Yufeng (UNC)
Time: June 5th, 1:15pm - 2:45pm
Location: VEC 201

Speech 1: Inference, Computation, and Visualization for Convex Clustering and Biclustering
Speaker: Genevera Allen (Rice)
Abstract: Hierarchical clustering enjoys wide popularity because of its fast computation, ease of interpretation, and appealing visualizations via the dendogram and cluster heatmap. Recently, several have proposed and studied convex clustering and biclustering which, similar in spirit to hierarchical clustering, achieve cluster merges via convex fusion penalties. While these techniques enjoy superior statistical performance, they suffer from slower computation and are not generally conducive to representation as

a dendogram. In the first part of the talk, we present new convex (bi)clustering methods and fast algorithms that inherit all of the advantages of hierarchical clustering. Specifically, we develop a new fast approximation and variation of the convex (bi)clustering solution path that can be represented as a dendogram or cluster heatmap. Also, as one tuning parameter indexes the sequence of convex (bi)clustering solutions, we can use these to develop interactive and dynamic visualization strategies that allow one to watch data form groups as the tuning parameter varies. In the second part of this talk, we consider how to conduct inference for convex clustering solutions that addresses questions like: Are there clusters in my data set? Or, should two clustersbe merged into one? To achieve this, we develop a new geometric representation of Hotelling's T^2-test that allows us to use the selective inference paradigm to test multivariate hypotheses for the first time. We can use this approach to test hypotheses and calculate confidence ellipsoids on the cluster means resulting from convex clustering. We apply these techniques to examples from text mining and cancer genomics. This is joint work with John Nagorski, Michael Weylandt, and Frederick Campbell.

Speech 2: High-dimensional Cost-constrained Regression via Non-convex Optimization
Speaker: Guan Yu (SUNY Buffalo)
Abstract: In modern predictive modeling process, budget constraints become a very important consideration due to the high cost of collecting data using new techniques such as brain imaging and DNA sequencing. This motivates us to develop new and efficient high-dimensional costconstrained predictive modeling methods. In this paper, to address this challenge, we first study a new non-convex high-dimensional cost-constrained linear regression problem, that is, we aim to find the cost-constrained regression model with the smallest expected prediction error among all models satisfying a budget constraint. The non-convex budget constraint makes this problem NP-hard. In order to estimate the regression coefficient vector of the cost-constrained regression model, we propose a new discrete extension of recent first-order continuous optimization methods. In particular, our method delivers a series of estimates of the regression coefficient vector by solving a sequence of 0-1 knapsack problems that can be addressed by many existing algorithms such as dynamic programming efficiently. Next, we show some extensions of our proposed method for statistical learning problems using loss functions with Lipschitz continuous gradient. It can be also extended to problems with groups of variables or multiple constraints. Theoretically, we prove that the series of the estimates generated by our iterative algorithm converge to a first-order stationary point, which can be a globally optimal solution to the nonconvex high-dimensional cost-constrained regression problem. Computationally, our numerical studies show that the proposed method can solve problems of fairly high dimensions and has promising estimation, prediction, and model selection performance.

Speech 3: Modeling Hybrid Traits for Comorbidity and Genetic Studies of Alcohol and Nicotine Co-Dependence
Speaker: Heping Zhang (Yale)
Abstract: I will present a novel multivariate model for analyzing hybrid traits and

identifying genetic factors for comorbid conditions. Comorbidity is a common phenomenon in mental health in which an individual suffers from multiple disorders simultaneously. For example, in the Study of Addiction: Genetics and Environment (SAGE), alcohol and nicotine addiction were recorded through multiple assessments that we refer to as hybrid traits. Statistical inference for studying the genetic basis of hybrid traits has not been well-developed. Recent rank-based methods have been utilized for conducting association analyses of hybrid traits but do not inform the strength or direction of effects. To overcome this limitation, a parametric modeling framework is imperative. Although such parametric frameworks have been proposed in theory, they are neither well-developed nor extensively used in practice due to their reliance on complicated likelihood functions that have high computational complexity. Many existing parametric frameworks tend to instead use pseudo-likelihoods to reduce computational burdens. Here, we develop a model fitting algorithm for the full likelihood. Our extensive simulation studies demonstrate that inference based on the full likelihood can control the type-I error rate, and gains power and improves the effect size estimation when compared with several existing methods for hybrid models. These advantages remain even if the distribution of the latent variables is misspecified. After analyzing the SAGE data, we identify three genetic variants (rs7672861, rs958331, rs879330) that are significantly associated with the comorbidity of alcohol and nicotine addiction at the chromosome-wide level. Moreover, our approach has greater power in this analysis than several existing methods for hybrid traits. Although the analysis of the SAGE data motivated us to develop the model, it can be broadly applied to analyze any hybrid responses.

27. **Session title**: **Functional Data Analysis in Action**
Organizer: **Kehui Chen (U of Pitt)**
Chair: Kehui Chen (U of Pitt)
Time: June 5th, 1:15pm - 2:45pm
Location: VEC 1402

Speech 1: Brain Functional Connectivity -- The FDA Approach
Speaker: Jane-Ling Wang (UC Davis)
Abstract: Functional connectivity refers to the connectivity between brain regions that share functional properties. It can be defined through statistical association or dependency among two or more anatomically distinct brain regions. In functional magnetic resonance imaging (fMRI), a standard way to measure brain functional connectivity is to assess the similarity of fMRI time courses for anatomically separated brain regions. Due to the temporal nature of fMRI data, tools of functional data analysis (FDA) are intrinsically applicable to such data. However, standard functional data techniques need to be modified when the goal is to study functional connectivity. We discuss two examples, where a new functional data approach is employed to study brain functional connectivity.

Speech 2: Functional Data Methods for Replicated Point Processes
Speaker: Daniel Gervini (U of Wisconsin at Milwaukee)
Abstract: Functional Data Analysis has traditionally focused on samples of smooth functions. However, many functional data methods can be extended to discrete point processes which are driven by smooth intensity functions. We will review some models that can be used for principal component analysis, joint modelling of discrete and continuous processes, and clustering of spatio-temporal point processes. We will apply these approaches to the analysis of spatio-temporal patterns in the distribution of crime and in the use of the shared-bicycle system in the city of Chicago.

Speech 3: Frechet Regression for Time-Varying Covariance Matrices: Assessing Regional Co-Evolution in the Developing Brain
Speaker: Hans Mueller (UC Davis)
Abstract: Frechet Regression provides an extension of Frechet means to the case of conditional Frechet means and is of interest for samples of random objects in a metric space (Petersen & Müller 2018). A specific application is encountered in cross-sectional studies where one observes p-dimensional vectors at one or a few random time points per subject and is interested in the p x p covariance or correlation matrix as a function of time. A challenge is that at each observation time one observes only a p-vector of measurements but not a covariance or correlation matrix. For a given metric on the space of covariance matrices, Frechet regression then generates a matrix function where at each fixed time the matrix is a non-negative definite covariance matrix. We demonstrate how this approach can be applied to MRI-extracted measurements of the myelin contents of various brain regions in small infants, aiming to quantify the regional co-evolution of myelination in the developing brain. Based on joint work with Alex Petersen and Sean Deoni.

28. **Session title**: Statistical Learning and Genomics
Organizer: Ji Zhu (Umich)
Chair: Bing Li (Penn State)
Time: June 5th, 1:15pm - 2:45pm
Location: VEC 1403

Speech 1: Proteomics and Genomics Integration for Translational Cancer Research
Speaker: Umut Ozbek (Mount Sinai)
Abstract: Advances in biomedical research bring the opportunity to gather data in various platforms. Integrating those diverse data to understand complex biological systems has been a big challenge for statisticians. We propose a novel statistical tool, spaceMap; a conditional graphical model, which learns the conditional dependency relationships between two types of high dimensional omic profiles through a penalized multivariate regression framework. spaceMap infers an undirected graph among response variables in tandem with a directed graph encoding perturbations from predictor variables on the response network. In addition, it utilizes cross-validation and model aggregation to reduce the false
discovery rate and consequently to improve reproducibility. We applied spaceMap to the copy number alterations, gene expression and proteomics datasets from CPTAC-TCGA

ovarian cancer study. The results help to pinpoint crucial cancer genes and provide insights on the functional consequences of important CNA in the disease.

Speech 2: What can we gain from proteogenomics prediction? The downstream analysis of NCI-CPTAC Proteogenomics DREAM Challenge
Speaker: Xiaoyu Song (MSSM)
Abstract: Background: Proteins are complex macromolecules responsible for nearly every task of cellular life, and thus play an essential role in the formation, progression and metastasis of cancer. A community-based collaborative competition, NCI-CPTAC DREAM Proteogenomics Challenge, is developing computational tools to answer "Can one predict abundance of any given protein from mRNA and genetic data?" in sub-challenge 2 and "Can one predict phosphoprotein abundances from protein abundance?" and in sub-challenge 3. Methods and Results: In pairwise Pearson correlation analyses, we found the correlation between true protein/phosphoprotein abundances and their predicted scores from the top performing models varies dramatically from protein to protein. Therefore, we investigated the biological factors that influence the performance of predictions. We also applied the well-predicted proteins/phosphosites to 317 independent samples of ovarian cancer in TCGA for protein prediction and to 105 independent samples of ovarian cancer in CPTAC for phosphoprotein prediction. We found that the most significant overall survival associated pathways were repeatedly identified in the top performing models and both in training samples of the models and the independent samples. The utility of these prediction models in drug sensitivity analyses, cell lines and trans-cancer models have also been investigated. Conclusion: Proteogenomics prediction is promising to improve our understanding of molecular mechanisms of human cancer.

Speech 3: An empirical comparison of deep neural networks and other supervised learning methods
Speaker: Wei Pan (U of Minnesota)
Abstract: Deep convolutional neural networks (CNN) have been proposed for supervised classification of high-throughput microscopy images to predict protein subcellular localization. We consider several CNN architectures in addition to a few traditional supervised learning methods such as random forest and gradient boosting. We compare their empirical performance when applied to a large dataset.

29. **Session title**: Recent advances in high-dimensional data
Organizer: Cunhui Zhang (Rutgers)
Chair: Sijian Wang (Rutgers)
Time: June 5th, 3:15pm - 4:45pm
Location: VEC 1202

Speech 1: The noise barrier and the large signal bias of the Lasso and other convex estimators
Speaker: Pierre Bellec (Rutgers)
Abstract: Convex estimators such as the Lasso, the matrix Lasso and the group Lasso have been studied extensively in the last two decades, demonstrating great success in

both theory and practice. This paper introduces two quantities, the noise barrier and the large scale bias, that provides novel insights on the performance of these convex regularized estimators. In sparse linear regression, it is now well understood that the Lasso achieves fast prediction rates, provided that the correlations of the design satisfy some Restricted Eigenvalue or Compatibility condition, and provided that the tuning parameter is at least larger than some universal threshold. Using the two quantities introduced in the paper, we show that the compatibility condition on the design matrix is actually unavoidable to achieve fast prediction rates with the Lasso. In other words, the ℓ_1-regularized Lasso must incur a loss due to the correlations of the design matrix, measured in terms of the compatibility constant. This results holds for any design matrix, any active subset of covariates, and any positive tuning parameter.It is now well known that the Lasso enjoys a dimension reduction property: if the target vector is s-sparse, the prediction rate of the Lasso with tuning parameter λ is of order \lambda\sqrt s$, even if the ambient dimension p is much larger than p. Such results require that the tuning parameters is greater than some universal threshold. We characterize sharp phase transitions for the tuning parameter of the Lasso around a critical threshold dependent on s. If λ is equal or larger than this critical threshold, the Lasso is minimax over s-sparse target vectors. If λ is equal or smaller than critical threshold, the Lasso incurs a loss of order $\sigma\sqrt s$ --which corresponds to a model of size s-- even if the target vector is more sparse than s.Remarkably, the lower bounds obtained in the paper also apply to random, data-driven tuning parameters. Additionally, the results extend to convex penalties beyond the Lasso.

Speech 2: Factor-Driven Two-Regime Regression
Speaker: Yuan Liao (Rutgers)
Abstract: We propose a novel two-regime regression model, where the switching between the regimes is driven by a vector of possibly unobservable factors that are estimated from a much larger panel data set. Estimating this model brings new challenges in terms of both computationand asymptotic theory. We show that our optimization problem can be reformulated as Mixed Integer Optimization and present two alternative computational algorithms. We also derive the asymptotic distribution of the resulting estimator and find that the effectof estimating the factors results in a phase transition on the rates of convergence and asymptotic distributions. (Joint with Lee S, Seo M, and Shin Y)

Speech 3: Network Analysis by SCORE
Speaker: Jiashun Jin (CMU)
Abstract: We have collected a data set for the networks of statisticians, consisting of titles, authors, abstracts, MSC numbers, keywords, and citation counts of papers published in representative journals in statistics and related fields. In Phase I of our study, the data set covers all published papers from 2003 to 2012 in Annals of Statistics, Biometrika, JASA, and JRSS-B. In Phase II of our study, the data set covers all published papers in 36 journals in statistics and related fields, spanning 40 years. The data sets motivate an array of interesting problems, and for the talk, I will focus on two closely related problems: network community detection, and network membership estimation. We tackle these problems with the recent approach of Spectral Clustering On Ratioed

Eigenvectors (SCORE), reveal a surprising simplex structure underlying the networks, and explain why SCORE is the right approach. We use the methods to investigate the Phase I data and report some of the results.We also report some Exploratory Data Analysis (EDA) results including productivity, journal-journal citations, and citation patterns. This part of result is based on Phase II of our data set (ready for use not very long ago).

30. **Session title**: Interpretable modeling and understanding variables
Organizer: Cynthia Rudin (Duke)
Chair: Cynthia Rudin (Duke)
Time: June 5th, 3:15pm - 4:45pm
Location: VEC 1203

Speech 1: Model Class Reliance: Variable Importance when all Models are Wrong, but *Many* are Useful.
Speaker: Aaron Fisher (Harvard)
Abstract: Variable importance (VI) tools are typically used to examine the inner workings of prediction models. However, many existing VI measures are not comparable across model types, can obscure implicit assumptions about the data generating distribution, or can give seemingly incoherent results when multiple prediction models fit the data well. In this paper we propose a framework of VI measures for describing how much any model class (e.g. all linear models of dimension p), any model-fitting algorithm (e.g. Ridge regression with fixed regularization parameter), or any individual prediction model (e.g. a single linear model with fixed coefficient vector), relies on covariate(s) of interest. The building block of our approach, Model Reliance (MR), compares a prediction model's expected loss with that model's expected loss on a pair of observations in which the value of the covariate of interest has been switched. Expanding on MR, we propose Model Class Reliance (MCR) as the upper and lower bounds on the degree to which any well-performing prediction model within a class may rely on a variable of interest, or set of variables of interest. Thus, MCR describes reliance on a variable while accounting for the fact that many prediction models, possibly of different parametric forms, may fit the data well. We give probabilistic bounds for MR and MCR, using existing results for U-statistics. These bounds can be generalized to create finite-sample confidence regions for the best-performing models from any class. We also illustrate connections between MR, conditional causal effects, and linear regression coefficients. We then apply MR & MCR in a public dataset of Broward County criminal records to study the reliance of recidivism prediction models on sex and race.

Speech 2: Feature-Efficient Multi-value Rule Sets for Interpretable Classification
Speaker: Tong Wang (U Iowa)
Abstract:

We present Multi-vAlue Rule Set (MARS) models for interpretable classification with feature efficient presentations. Compared to rule sets built from single-valued rules, MARS introduces a more generalized form of association rules that allows multiple values in a condition. Rules of this form are more concise than classical single-valued

rules in capturing and describing patterns in data. Our formulation also pursues a higher efficiency of feature utilization, which reduces possible cost in data collection and storage. We propose a Bayesian framework for formulating a MARS model and propose an efficient inference method for learning a maximum a posteriori, incorporating theoretically grounded bounds to iteratively reduce the search space and improve the search efficiency. Experiments on synthetic and real-world data demonstrate that MARS models have significantly smaller complexity and fewer features than baseline models while being competitive in predictive accuracy. We conducted a usability study with human subjects and the results show that MARS is the easiest to understand compared with other competing rule-based models. We apply MARS to a real-world application to predict in-hospital mortality rate.

Speech 3: Recent Work on Interpretable Machine Learning Models
Speaker: Cynthia Rudin (Duke)
Abstract: I will overview some work on interpretable machine learning models including: (i) one-sided decision trees (rule lists) that provably minimize accuracy and sparsity, (ii) falling rule lists, which are constrained one-sided decision trees, (iii) deep neural networks with an interpretable prototype layer, (iv) matching methods for causal inference that use machine learning to gain both interpretability and accuracy.

31. **Session title**: Statistical Inference for High-Dimensional Data
Organizer: Jeff Simonoff (NYU)
Chair: Jeff Simonoff (NYU)
Time: June 5th, 3:15pm - 4:45pm
Location: VEC 1302

Speech 1: Quantile Regression for big data with small memory
Speaker: Xi Chen (NYU)
Abstract: In this talk, we discuss the inference problem of quantile regression for a large sample size n (and a diverging dimensionality p) but under a limited memory constraint, where the memory can only store a small batch of data of size m. A popular approach, the naive divide-and-conquer method, only works when $n=o(m^2)$ and is computationally expensive. This talk proposes a novel inference approach and establishes the asymptotic normality result that achieves the same efficiency as the quantile regression estimator computed on all the data. Essentially, our method can allow arbitrarily large sample size n as compared to the memory size m. Our method can also be applied to address the quantile regression under distributed computing environment (e.g., in a large-scale sensor network) or for real-time streaming data.

Speech 2: Inference after cross-validation
Speaker: Joshua Loftus (NYU)
Abstract: We described a method for performing inference on models chosen by cross-validation. When the test error being minimized in cross-validation is a residual sum of squares it can be written as a quadratic form. This allows us to apply the inference framework in Loftus et al. (2016) for models determined by quadratic constraints to the model that minimizes CV test error. Our only requirement on the model training

procedure is that its selection events are regions satisfying linear or quadratic constraints. This includes both Lasso and forward stepwise, which serve as our main examples throughout. We do not require knowledge of the error variance. The overall procedure is a computationally intensive method of selecting models adaptively and performing inference for the selected model.

Speech 3: Optimal estimation of Gaussian mixtures via denoised method of moments
Speaker: Yihong Wu (Yale)
Abstract: The Method of Moments is one of the most widely used methods in statistics for parameter estimation, obtained by solving the system of equations that match the population and estimated moments. However, in practice and especially for the important case of mixture models, one frequently needs to contend with the difficulties of non-existence or non-uniqueness of statistically meaningful solutions, as well as the high computational cost of solving large polynomial systems. Moreover, theoretical analysis of method of moments are mainly confined to asymptotic normality style of results established under strong assumptions.

32. **Session title**: New Development on Neuroimage Data Analysis
Organizer: Zhu, Hongtu (MD Anderson)
Chair: Xuan Bi(Yale)
Time: June 5th, 3:15pm - 4:45pm
Location: VEC 1303

Speech 1: A Low-Rank Multivariate General Linear Model forMulti-Subject fMRI Data and a Non-Convex Optimization Algorithm for Brain Response Comparison
Speaker: Tingting Zhang (UVA)
Abstract:
The focus of this paper is on evaluating brain responses to different stimuli and identifying brain regions with different responses using multi-subject, stimulus-evoked functional magnetic resonance imaging (fMRI) data. To jointly model many brain voxels' responses to designed stimuli, we present a new low-rank multivariate general linear model (LRMGLM) for stimulus-evoked fMRI data. The new model not only is flexible to characterize variation in hemodynamic response functions (HRFs) across different regions and stimulus types, but also enables information "borrowing" across voxels and uses much fewer parameters than typical nonparametric models for HRFs. To estimate the proposed LRMGLM, we introduce a new penalized optimization function, which leads to temporally and spatially smooth HRF estimates. We develop an efficient optimization algorithm to minimize the optimization function and identify the voxels with different responses to stimuli. We show that the proposed method can outperform several existing voxel-wise methods by achieving both high sensitivity and specificity. We apply the proposed method to the fMRI data collected in an emotion study, and identify anterior dACC to have different responses to a designed threat and control stimuli.

Speech 2: Nonparametric Bayes Models of Fiber Curves Connecting Brain Regions

Speaker: Zhengwu Zhang (Rochester)
Abstract:

In studying structural inter-connections in the human brain, it is common to first estimate fiber bundles connecting different regions relying on diffusion MRI. These fiber bundles act as highways for neural activity. Current statistical methods reduce the rich information into an adjacency matrix, with the elements containing a count of fibers or a mean diffusion feature along the fibers. The goal of this article is to avoid discarding the rich geometric information of fibers, developing flexible models for characterizing the population distribution of fibers between brain regions of interest within and across different individuals. We start by decomposing each fiber into a rotation matrix, shape and translation from a global reference curve. These components are viewed as data lying on a product space composed of different Euclidean spaces and manifolds. To non-parametrically model the distribution within and across individuals, we rely on a hierarchical mixture of product kernels specific to the component spaces. Taking a Bayesian approach to inference, we develop efficient methods for posterior sampling. The approach automatically produces clusters of fibers within and across individuals. Applying the method to Human Connectome Project data, we find an interesting relationship between brain fiber geometry and reading ability.

Speech 3: Supervised Principal Component Regression for Functional Data with High Dimensional Predictors
Speaker: Dehan Kong (U Toronto)
Abstract: Motived by functional magnetic resonance imaging studies, we study the effect of clinical/demographic variables on the dynamic functional structures, which plays a key role in understanding brain functionality. To this end, we propose the supervised principal component regression for functional data with possibly high dimensional clinical variables. Compared with its classical counterpart, the principal component regression, the proposed methodology relies on a newly proposed integrated residual sum of squares for functional data and makes use of the association information directly. It can be formulated as a sequence of nonconvex generalized Rayleigh quotient optimization problems, which turn out to be NP-hard and thus computational intractable. Utilizing the invariance property of linear subspaces under rotations, we then reformulate the problem into a simultaneous sparse linear regression problem. Formally, we show that the reformulated problem can recover the same subspace as if the original sequence of nonconvex problems were solved. Statistically, the rate of convergence for the obtained estimators is established. Numerical studies and an application to the Human Connectome Project lend further support to the proposed methodology. (Joint work with Xinyi Zhang and Qiang Sun)

33. **Session title**: **Spectral Clustering, Functional Graphical Models, and Hierarchical Interactions**
Organizer: Lingzhou Xue (Penn State)
Chair: Kuang-Yao Lee (Temple)
Time: June 5th, 3:15pm - 4:45pm

Location: VEC 1402

Speech 1: Spectral clustering based on learning similarity matrix
Speaker: Hongyu Zhao (Yale)
Abstract: Single-cell RNA-sequencing (scRNA-seq) technology can generate genome-wide expression data at the single-cell levels. One important objective in scRNA-seq analysis is to cluster cells where each cluster consists of cells belonging to the same cell type based on gene expression patterns. In this presentation, we will introduce a novel spectral clustering framework that imposes sparse structures on a target matrix. Specifically, we utilize multiple doubly stochastic similarity matrices to learn a similarity matrix, motivated by the observation that each similarity matrix can be a different informative representation of the data. We impose a sparse structure on the target matrix followed by shrinking pairwise differences of the rows in the target matrix, motivated by the fact that the target matrix should have these structures in the ideal case. We solve the proposed non-convex problem iteratively using the ADMM algorithm and show the convergence of the algorithm. We evaluate the performance of the proposed clustering method on various simulated as well as real scRNA-seq data, and show that it can identify clusters accurately and robustly. This is joint work with Seyoung Park.

Speech 2: Copula Gaussian Graphical Models for Functional Data
Speaker: Bing Li (Penn State)
Abstract: We consider the problem of constructing statistical graphical models for functional data; that is, the observations on the vertices are random functions. This types of data are common in medical applications such as EEG and fMRI. Recently published functional graphical models rely on the assumption that the random functions are Hilbert-space-valued Gaussian random elements. We relax this assumption by introducing a copula Gaussian random elements Hilbert spaces, leading to what we call the Functional Copula Gaussian Graphical Model (FCGGM). This model removes the marginal Gaussian assumption but retains the simplicity of the Gaussian dependence structure, which is particularly attractive for large data. We develop four estimators, together with their implementation algorithms, for the FCGGM. We establish the consistency and the convergence rates of one of the estimators under different sets of sufficient conditions with varying strengths. We compare our FCGGM with the existing functional Gaussian graphical model by simulation, under both non-Gaussian and Gaussian graphical models, and apply our method to an EEG data set to construct brain networks. This is a joint work with Eftychia Solea.

Speech 3: Learning Nonconvex Hierarchical Interactions
Speaker: Lingzhou Xue (Penn State)
Abstract: In this talk, we will focus on learning nonconvex hierarchical interactions in high-dimensional statistical models. We first use the affine sparsity constraints to provide a precise characterization of both strong and weak hierarchical interactions. However, these affine sparsity constraints do not lead to a closed feasible region. To address this issue, we derive the explicit closure of the affine sparsity constraint for learning nonconvex hierarchical interactions. We prove that the global solution can be found by solving a sequence of folded concave penalized estimation and the desired strong or weak

hierarchy holds with probability one. Furthermore, we study the local convergence theory for learning hierarchical interactions using the folded concave penalized estimation. Numerical studies are used to demonstrate the power of our proposed methods.

34. **Session title**: Data Science in IT Industries
Organizer: David Banks (Duke)
Chair: Genevera Allen(Rice)
Time: June 5th, 3:15pm - 4:45pm
Location: VEC 1403

Speech 1 : Using Data Science to Improve Streaming Quality at Netflix
Speaker: Julie Novak (Netflix)
Abstract: In this talk, I will begin by giving an overview of the data science challenges involved in providing an optimal streaming service at Netflix. There are many dimensions to this problem, including selecting best picture quality based on network speed, determining proper content to cache on Netflix's Content Delivery Networks (CDN), and improving each customer's Quality of Experience (QoE). The talk will then dive deeper into the notion of QoE by explaining how to use statistical tools to measure and gain deeper understanding of it in the context of A/B testing.

Speech 2: Random Forests, Decision Trees, and Categorical Predictors: The "Absent Levels" Problem
Speaker: Tim Au (Google)
Abstract: One advantage of decision tree based methods like random forests is their ability to natively handle categorical predictors without having to first transform them (e.g., by using one-hot encoding). However, in this talk, we show how this capability can lead to an inherent "absent levels" problem for decision tree based methods that has never been thoroughly discussed, and whose consequences have never been carefully explored. This problem occurs whenever there is an indeterminacy over how to handle an observation that has reached a categorical split which was determined when the observation in question's level was absent during training. Although these incidents may appear to be innocuous, by using Leo Breiman and Adele Cutler's random forests FORTRAN code and the randomForest R package (Liaw and Wiener, 2002) as motivating case studies, we study how overlooking the absent levels problem can systematically bias a model. Furthermore, by using three real data examples, we illustrate how absent levels can dramatically alter a model's performance in practice, and we empirically demonstrate how some simple heuristics can be used to help mitigate the effects of the absent levels problem until a more robust theoretical solution is found.

Speech 3: The Challenge of Educating Data Scientists for Industry
Speaker: David Banks (Duke University and SAMSI)
Abstract: The statistical world is changing quickly, and our graduate programs are (generally) not keeping pace. This talk reviews some of the structural and cultural barriers that we need to overcome. Besides proposing a model curriculum, it also addresses ways in which our publication processes now longer serve the interests of our

profession, and it discusses the commodification of analysis.

35. **Session title**: Machine Learning and Precision Medicine
Organizer: Haoda Fu (Eli Lilly)
Chair: Genevera Allen(Rice)
Time: June 6th, 8:30am – 10:00am
Location: VEC 404 /405

Speech 1: Support vector machines for learning optimal individualized treatment rules with multiple treatments
Speaker: Donglin Zeng (UNC)
Abstract: Support vector machine (SVM) methods have been proposed to estimate optimal individualized treatment rules when treatment is binary. Extending SVM to more than two treatment options remains an open and challenging problem. In this work, we propose a novel and efficient algorithm to generalize SVM-based outcome weighted learning to a multi-treatment setting. The proposed method sequentially solves binary SVM problems. Theoretically, we show that the resulting treatment rule is Fisher consistent and derive the convergence rate for the estimated value function. We conduct extensive simulation studies to demonstrate that the proposed method has superior performance to competing methods.

Speech 2: Individualized Treatment Recommendation (ITR) for SurvivalOutcomes
Speaker: Haoda Fu (Eli Lilly)
Abstract: 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

1. Instead of fitting the data, our method directly search the optimal treatment policy which improvesthe efficiency
2. 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
3. Our method handles multiple treatments with intuitive geometry explanations
4. 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.

Speech 3: Estimation and Evaluation of Linear Individualized Treatment Rules to Guarantee Performance
Speaker: Yuanjia Wang (Columbia)
Abstract: In clinical practice, an informative and practically useful treatment rule should be simple and transparent. However, because simple rules are likely to be far from optimal, effective methods to construct such rules must guarantee performance, in terms of yielding the best clinical outcome (highest reward) among the class of simple rules under consideration. Furthermore, it is important to evaluate the benefit of the derived rules on the whole sample and in pre-specified subgroups (e.g., vulnerable patients). To achieve both goals, we propose a robust machine learning method to estimate a linear treatment rule that is guaranteed to achieve optimal reward among the class of all linear rules. We then develop a diagnostic measure and inference procedure to evaluate the benefit of the obtained rule and compare it with the rules estimated by other methods. We provide theoretical justification for the proposed method and its inference procedure, and we demonstrate via simulations its superior performance when compared to existing methods. Lastly, we apply the method to the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial on major depressive disorder and show that the estimated optimal linear rule provides a large benefit for mildly depressed and severely depressed patients but manifests a lack-of-fit for moderately depressed patients.

36. **Session title**: Advances in Nonparametric Statistics and their Applications
Organizer: Narisetty, Naveen (UIUC)
Chair: Fei Xue (UIUC)
Time: June 6th, 8:30am – 10:00am
Location: VEC 902/903

Speech 1: Fly-By-Night Life Insurance and the NPMLE for Weibull Frailty Models
Speaker: Roger Koenker (UIUC and UCL)
Abstract: Classical parametric statistical models are often quite fragile when confronted with data that fails to conform precisely to their idealized conditions. In survival analysis more flexible models have been constructed as mixtures, yielding estimators with improved robustness properties. These frailty models can themselves be parametric, but recent work has stressed non-parametric mixtures. This talk will describe an application of the latter approach to the study of mortality rates for a large sample of mediterranean fruit flies. In addition to revealing some rather surprising biological findings, the methods illustrate a general approach to estimation based on the Kiefer Wolfowitz non-parametric MLE for mixture models and its use in a somewhat fanciful compound decision problem.

Speech 2: Learning from Dr. Martin Luther King Jr: Text analysis and statistical approaches for civil rights
Speaker: Christopher Kinson (UIUC)
Abstract: The recent shifts in the political landscape have heightened the importance of education and advancement for minorities, especially those underrepresented in STEM disciplines. Minority students have been battling for equal access and rights to education at institutions of higher learning and the resources therein throughout the history of the

US. We as a scientific community must boldly step up to improve the outcomes of underrepresented students. The objective of this project is to educate and recruit Black and African students into the fields of statistics and data science. We create this research cohort and lab environment for the students to contribute to their own knowledge and to the body of knowledge within data science and the digital humanities. This lab is a creative and interdisciplinary space where the subject is one of the most important figures in Black history, Rev. Dr. Martin Luther King Jr. Specifically, we employ text analysis of King's writings and speeches as well as articles written about him. Additionally, we educate students about statistics and introduce them to statistical programming in R. The project tackles several challenges in broadening the participation of underrepresented students in STEM and celebrates the achievements made in that process.

Speech 3: Inference on the dependence structure of time series extremes
Speaker: Stanislav Volgushev (U Toronto)
Abstract: Many natural phenomena such as extreme precipitation or heat waves can be described as maxima over blocks of time series. If the dependence structure of such extremes is of interest, component-wise maxima of vector-valued time series need to be considered. Under suitable assumptions on a vector-valued time series, properly standardized component-wise maxima are known to converge to a multivariate extreme-value distribution. In the present talk, we discuss functional central limit theorems and resulting inference procedures for the dependence structure of this limiting distribution. We propose several improvements over existing approaches which allow to reduce both bias and variance. We also contrast our approach (which is based on taking block maxima) with the peak-over-threshold approach which is a popular tool in the analysis of extremes.

37. **Session title**: Recent advances in spectral methods for complex data
Organizer: Yuekai Sun (UMich)
Chair: Edgar Dobriban (Upenn)
Time: June 6th, 8:30am – 10:00am
Location: VEC 1202/1203

Speech 1: Analyzing Developmental Processes with Optimal Transport
Speaker: Geoff Schiebinger (MIT)
Abstract: Understanding the molecular programs that guide cellular differentiation during development is a major goal of modern biology. Here, we introduce an approach, based on the mathematics of optimal transport, for inferring developmental landscapes, probabilistic cellular fates and dynamic trajectories from large-scale single-cell RNA-seq (scRNA-seq) data collected along a time course. Our approach, Waddington-OT is based on a novel framework for analyzing stochastic processes whose instantaneous temporal couplings agree with optimal transport. We demonstrate the power of WADDINGTON-OT by applying the approach to study 65,781 scRNA-seq profiles collected at 10 time points over 16 days during reprogramming of fibroblasts to iPSCs.

We construct a high-resolution map of reprogramming that rediscovers known features; uncovers new alternative cell fates including neural- and placental-like cells; predicts the

origin and fate of any cell class; highlights senescent-like cells that may support reprogramming through paracrine signaling; and implicates regulatory models in particular trajectories. Of these findings, we highlight Obox6, which we experimentally show enhances reprogramming efficiency. Our approach provides a general framework for investigating cellular differentiation.

Speech 2: How to select the number of components in PCA and factor analysis? Understanding and improving permutation methods
Speaker: Edgar Dobriban (Wharton)
Abstract: Selecting the number of components in PCA and factor analysis is a key problem facing practitioners of data science. One of the most popular methods is a permutation approach that randomly scrambles the elements of each feature. It selects the components whose singular values are large compared to the permuted data. This method (also known as parallel analysis) is recommended in many textbooks and review papers, and used in genomics by leading applied statisticians including T Hastie, M Stephens, J Storey, R Tibshirani and WH Wong. However, it is poorly understood. In this talk, we develop a theoretical understanding and propose improvements.

Speech 3: Higher-order spectral graph clustering with motifs
Speaker: Austin Benson (Cornell)
Abstract: Networks are typically described by lower-order connectivity patterns that are captured at the level of individual nodes and edges. However, higher-order connectivity patterns captured by small subgraphs, or network motifs, describe the fundamental structures that control and mediate the behavior of many complex systems. In this talk, I will discuss a higher-order spectral graph clustering framework that finds groups of nodes that participate in many instances of a given motif. I will also show applications of this framework in ecology, biology, transportation, and social networks.

38. **Session title**: New machine learning methods
Organizer: Annie Qu (UIUC)
Chair: Annie Qu (UIUC)
Time: June 6th, 8:30am – 10:00am
Location: VEC 1302

Speech 1: Generalized self-concordant optimization and its applications in statistical learning
Speaker: Quoc Tran-Dinh (UNC)
Abstract: Many statistics and machine learning applications can be cast into a composite convex minimization problem. Well-known examples include sparse logistic regression, SVM, and inverse covariance estimation. These problems are well studied and can efficiently be solved by several state-of-the-arts. Recent development in first-order, second-order, and stochastic gradient-type methods has brought a new opportunity to solve many other classes of convex optimization problems in large-scale settings. Unfortunately, so far, such methods require the underlying models to satisfy some structural assumptions such as Lipschitz gradient and restricted strong convexity, which may be failed to hold or may be hard to

check.
In this talk, we demonstrate how to exploit an analytical structure hidden in convex optimization for developing solution methods. Our key idea is to generalize a powerful concept so-called "self-concordance" introduced by Y. Nesterov and A. Nemirovskii to a broader class of convex functions. We show that this structure covers many applications in statistics and machine learning. Then, we develop a unified theory for designing numerical methods. We illustrate our theory through Newton-type and proximal Newton-type methods. We note that the proposed theory can further be applied to develop other methods as long as the underlying model is involved with a "generalized self-concordant structure". We provide some numerical examples in different fields to illustrate our theoretical development.

Speech 2: On Scalable Inference with Stochastic Gradient Descent
Speaker: Yixin Fang (New Jersey Institute of Technology)
Abstract: In many applications involving large dataset or online updating, stochastic gradient descent (SGD) provides a scalable way to compute parameter estimates and has gained increasing popularity due to its numerical convenience and memory efficiency. While the asymptotic properties of SGD-based estimators have been established decades ago, statistical inference such as interval estimation remains much unexplored. The traditional resampling method such as the bootstrap is not computationally feasible since it requires to repeatedly draw independent samples from the entire dataset. The plug-in method is not applicable when there are no explicit formulas for the covariance matrix of the estimator. In this paper, we propose a scalable inferential procedure for stochastic gradient descent, which, upon the arrival of each observation, updates the SGD estimate as well as a large number of randomly perturbed SGD estimates. The proposed method is easy to implement in practice. We establish its theoretical properties for a general class of models that includes generalized linear models and quantile regression models as special cases. The finite-sample performance and numerical utility is evaluated by simulation studies and two real data applications.

Speech 3: Scalable Kernel-based Variable Selection with Sparsistency
Speaker: Junhui Wang (City U. of Hong Kong)
Abstract: Variable selection is central to sparse modeling, and many methods have been proposed under various model assumptions. In this talk, we will present a scalable framework for model-free variable selection in reproducing kernel Hilbert space (RKHS) without specifying any restrictive model. As opposed to most existing model-free variable selection methods requiring fixed dimension, the proposed method allows dimension p to diverge with sample size n. The proposed method is motivated from the classical hard-threshold variable selection for linear models, but allows for general variable effects. It does not require specification of the underlying model for the response, which is appealing in sparse modeling with a large number of variables. The proposed method can also be adapted to various scenarios with specific model assumptions, including linear models, quadratic models, as well as additive models. The asymptotic estimation and variable selection consistencies of the proposed method are established in all the scenarios. If time permits, the extension of the proposed method beyond mean regression will also be

discussed.

39. **Session title**: New directions in functional data analysis.
Organizer: Tailen Hsing (UMich)
Chair: : Vincent Joseph Dorie (Columbia)
Time: June 6th, 8:30am – 10:00am
Location: VEC 1402

Speech 1: Nonparametric covariance estimation for mixed longitudinal studies
Speaker: Kehui Chen (U of Pitt)
Abstract: Motivated by applications of mixed longitudinal studies, where a group of subjects entering the study at different ages (cross-sectional) are followed for successive years (longitudinal), we consider nonparametric covariance estimation with samples of noisy and partially observed functional trajectories. In this talk, we will introduce a novel sequential aggregation scheme, which works for both dense regular and sparse irregular observations. We will present numerical experiment results and applications a midlife women's working memory study. We will also discuss the details of identifiability and estimation consistency.

Speech 2: Functional Data Analysis with Highly Irregular Designs with Applications to Head Circumference Growth
Speaker: Matthew Reimherr (Penn State)
Abstract: Functional Data Analysis often falls into one of two branches, either sparse or dense, depending on the sampling frequency of the underlying curves. However, methods for sparse FDA often still rely on having a growing number of observations per subject as the sample size grows. Practically, this means that for very large sample sizes with infrequently or irregularly sampled curves, common methods may still suffer a non-negligible bias. This becomes especially true for nonlinear models, which are often defined based on complete curves. In this talk I will discuss how this issue can be fixed to obtain valid statistical inference regardless of the sampling frequency of the curves. This work is motivated by a study by Dr. Carrie Daymont from Hershey medical school that examines pathologies related to head circumference growth in children. In her study, tens of thousands of children are sampled, but with widely varying frequency.

Speech 3: Supervised Learning on the Path Space and its Applications
Speaker: Hao Ni (UCL)
Abstract: Regression analysis aims to use observational data from multiple observations to develop a functional relationship relating explanatory variables to response variables, which is important for much of modern statistics, and econometrics, and also the field of machine learning. In this talk, we consider the special case where the explanatory variable is a data stream. We provide an approach based on identifying carefully chosen features of the stream which allows linear regression to be used to characterise the functional relationship between explanatory variables and the conditional distribution of the response; the methods used to develop and justify this approach, such as the signature of a stream and the shuffle product of tensors, are standard tools in the theory of rough paths and provide a unified and non-parametric approach with potential significant

dimension reduction. To further improve the efficiency of the signature method, we can combine the non-linear regression method (e.g. neural network) with the signature feature set. Numerical examples are provided to show the superior performance of the proposed method. Lastly I will show that the signature based method have achieved the state-of-the-art results in online handwritten text recognition and action recognition.

40. **Session title**: Modern Approaches for Inference and Estimation
Organizer: Genevera Allen (Rice)
Chair: Genevera Allen (Rice)
Time: June 6th, 1:15pm - 2:45pm
Location: VEC 404/405

Speech 1: High-Dimensional Propensity Score Estimation via Covariate Balancing
Speaker: Yang Ning (Cornell)
Abstract: In this paper, we address the problem of estimating the average treatment effect (ATE) and the average treatment effect for the treated (ATT) in observational studies when the number of potential confounders is possibly much greater than the sample size. In particular, we develop a robust method to estimate the propensity score via covariate balancing in high-dimensional settings. Since it is usually impossible to obtain the exact covariate balance in high dimension, we propose to estimate the propensity score by balancing a carefully selected subset of covariates that are predictive of the outcome under the assumption that the outcome model is linear and sparse. The estimated propensity score is, then, used for the Horvitz-Thompson estimator to infer the ATE and ATT. We prove that the proposed methodology has the desired properties such as sample boundedness, root-n consistency, asymptotic normality, and semiparametric efficiency. We then extend these results to the case where the outcome model is a sparse generalized linear model. More importantly, we show that the proposed estimator is robust to model misspecification. Finally, we conduct simulation studies to evaluate the finite-sample performance of the proposed methodology, and apply it to estimate the causal effects of college attendance on adulthood political participation. Open-source software is available for implementing the proposed methodology. This is the joint work with Peng and Imai.

Speech 2: Interactive algorithms for graphical model selection
Speaker: Gautam Dasarthy (Rice University)
Abstract: With rapid progress in our ability to acquire, process, and learn from data, the true democratization of data-driven intelligence has never seemed closer. Unfortunately, there is a catch. Machine learning algorithms have traditionally been designed independently of the systems that acquire data. As a result, there is a fundamental disconnect between their promise and their real-world applicability. An urgent need has therefore emerged for integrating the design of learning and acquisition systems.
In this talk, I will present an approach for addressing this learning-acquisition disconnect using interactive machine learning methods. In particular, I will consider the problem of learning graphical model structure in high dimensions. This will highlight how traditional (open loop) methods do not take into account data acquisition constraints that arise in applications ranging from sensor networks to calcium imaging of the brain. I will then

demonstrate how one can close this loop using techniques from interactive machine learning. I will conclude by discussing several connections to post-selection inference in this context.

Speech 3: AdaPT: An interactive procedure for multiple testing with side information
Speaker: Will Fithian (UCB)
Abstract: We consider the problem of multiple hypothesis testing with generic side information: for each hypothesis we observe both a p-value and some predictor encoding contextual information about the hypothesis. For large-scale problems, adaptively focusing power on the more promising hypotheses (those more likely to yield discoveries) can lead to much more powerful multiple testing procedures. We propose a general iterative framework for this problem, called the Adaptive p-value Thresholding (AdaPT) procedure, which adaptively estimates a Bayes-optimal p-value rejection threshold and controls the false discovery rate (FDR) in finite samples. At each iteration of the procedure, the analyst proposes a rejection threshold and observes partially censored p-values, estimates the false discovery proportion (FDP) below the threshold, and either stops to reject or proposes another threshold, until the estimated FDP is below α. Our procedure is adaptive in an unusually strong sense, permitting the analyst to use any statistical or machine learning method she chooses to estimate the optimal threshold, and to switch between different models at each iteration as information accrues.
This is joint work with Lihua Lei.

41. **Session title**: Functional and high dimensional data
Organizer: Aurore Delaigle (U of Melbourne)
Chair: Aurore Delaigle (U of Melbourne)
Time: June 6th, 1:15pm - 2:45pm
Location: VEC 1403

Speech 1: Small Sample Confidence Intervals for the ACL (Abduskhurov, Cheng, and Lin) Estimators Under the Proportional Hazards Model
Speaker: Emad Abdurasul (James Madison University)
Abstract: We develop a saddlepoint- based method for generating small sample confidence bands for the population survival function from the ACL estimators, under the proportional hazards model. In the process, we derive the exact distribution of this estimator and develop mid-population tolerance bands for saddlepoint estimators. Our method depends upon the Mellin transform of the zero-truncated survival estimator. This transform is inverted via saddlepoint approximation to yield a highly accurate approximation to the cumulative distribution function of the respective cumulative hazard function. This distribution function is then inverted to produce our saddlepoint onfidence bands. Then we compare our saddlepoint confidence bands with those obtained from competing large sample methods as well as with those obtained from the exact distribution. In our simulation study, we found that the saddlepoint confidence bands are very close to the confidence bands derived from the exact distribution. In addition being close, it is easier to compute, and it outperforms the large sample methods in terms probability convergence.

Speech 2: Binary functional linear models in a stratified sampling setting
Speaker: Sophie Dabo-Niang (Université Lille 3)
Abstract: A functional binary choice model is explored in a stratified sample design context. In other words, a model is considered in which the response is binary, the explanatory variable is functional, and the sample is stratified with respect to the values of the response variable. A dimension reduction of the space of the explanatory random function based on a Karhunen– Loève expansion is used to define a conditional maximum likelihood estimate of the model. Based on this formulation, several asymptotic properties are given. Numerical studies are used to compare the proposed method with the ordinary maximum likelihood method, which ignores the nature of the sampling. The proposed model yields encouraging results. The potential of the functional sampling model for integrating special non-random features of the sample, which would have been difficult to see otherwise, is also outlined.

Speech 3: Functional CLT and sharp bounds for some (conditional Poisson) survey sampling plans with applications to big (tall) data
Speaker: Patrice Bertail (Université Paris Nanterre)
Abstract: Subsampling methods as well as general sampling methods appear as natural tools to handle very large database (big data in the indivual dimension) when traditionnal statistical methods or statistical learning algorithms fail to be implemented on too large datasets. The choice of the weights of the survey sampling scheme may reduce the loss implied by the choice of a much more smaller sampling size (according to the problem of interest). We will first recall some asymptotic results for general survey sampling based empirical processes indexed by class of functions (see Bertail and Clémençon, 2017, Scandinavian Journal of Statistics), for Poisson type and conditional Poisson (rejective) survey samplings. These results may be extended to a large class of survey sampling plans via the notion of negative association of most survey sampling plans (Bertail, Rebecq, 2018) Then in the perspective to generalize some statistical learning tasks to sampled data, we will obtain exponential bounds for the probabilities of deviation of a Horvitz Thompson sum from its expectation when the variables involved in the summation are obtained by sampling in a finite population according to associated or rejective scheme.

42. **Session title**: Machine learning, classification and designs
Organizer: Annie Qu (UIUC)
Chair: Annie Qu (UIUC)
Time: June 6th, 1:15pm - 2:45pm
Location: VEC 1202 /1203

Speech 1: Efficient Gaussian Process Modeling using Experimental Design-based Subagging
Speaker: Ying Hung (Rutgers)
Abstract: We address two important issues in Gaussian process (GP) modeling. One is how to reduce the computational complexity in GP modeling and the other is how to simultaneous perform variable selection and estimation for the mean function of GP

models. Estimation is computationally intensive for GP models because it heavily involves manipulations of an n-by-n correlation matrix, where n is the sample size. Conventional penalized likelihood approaches are widely used for variable selection. However the computational cost of the penalized likelihood estimation (PMLE) or the corresponding one-step sparse estimation (OSE) can be prohibitively high as the sample size becomes large, especially for GP models. To address both issues, we propose an efficient subsample aggregating (subagging) approach with an experimental design-based subsampling scheme. The proposed method is computationally cheaper, yet it can be shown that the resulting subagging estimators achieve the same efficiency as the original PMLE and OSE asymptotically. The finite-sample performance is examined through simulation studies. Application of the proposed methodology to a data center thermal study reveals some interesting information, including identifying an efficient cooling mechanism.

Speech 2: Structural Learning and Integrative Decomposition of Multi-View Data
Speaker: Irina Gaynanova (Texas A&M)
Abstract: The increased availability of the multi-view data (data on the same samples from multiple sources) has led to strong interest in models based on low-rank matrix factorizations. These models represent each data view via shared and individual components, and have been successfully applied for exploratory dimension reduction, association analysis between the views, and further learning tasks such as consensus clustering. Despite these advances, there remain significant challenges in modeling partially-shared components, and identifying the number of components of each type (shared/partially-shared/individual). In this work, we formulate a novel linked component model that directly incorporates partially-shared structures. We call this model SLIDE for Structural Learning and Integrative DEcomposition of multi-view data. We prove the existence of SLIDE decomposition and explicitly characterize the identifiability conditions. The proposed model fitting and selection techniques allow for joint identification of the number of components of each type, in contrast to existing sequential approaches. In our empirical studies, SLIDE demonstrates excellent performance in both signal estimation and component selection. We further illustrate the methodology on the breast cancer data from The Cancer Genome Atlas repository. This is joint work with Gen Li.

Speech 3: Shrinking characteristics of precision matrix estimators
Speaker: Adam Rothman (U. of Minnesota)
Abstract: We propose a framework to shrink a user-specified characteristic of a precision matrix estimator that is needed to fit a predictive model. Estimators in our framework minimize the Gaussian negative log-likelihood plus an L1 penalty on a linear function evaluated at the optimization variable corresponding to the precision matrix. We establish convergence rate bounds for these estimators and we propose an alternating direction method of multipliers algorithm for their computation. Our simulation studies show that our estimators can perform better than competitors when they are used to fit predictive models. In particular, we illustrate cases where our precision matrix estimators

perform worse at estimating the population precision matrix while performing better at prediction. This is joint work with Aaron Molstad.

43. **Session title**: Statistics in neuroscience and microbiome research at the Flatiron Institute
Organizer: Christian L. Müller (Flatiron Institute, Simons Foundation)
Chair: Christian L. Müller (Flatiron Institute, Simons Foundation)
Time: June 6th, 1:15pm - 2:45pm
Location: VEC 1303

Speech 1: Neural representation learning as kernel alignment
Speaker: Cengiz Pehlevan (Simons Foundation)
Abstract: What are the brain's learning cost functions? I show that some kernel alignment cost functions can be minimized by biologically plausible neural learning algorithms. Starting from such cost functions, I derive neural networks for various biologically motivated unsupervised learning tasks, such as soft-clustering and manifold disentangling. I discuss applications of these ideas to circuits of the brain.

Speech 2: Robust regression with compositional covariates
Speaker: Aditya Mishra (Flatiron Institute)
Abstract: With the large-scale efforts in 16S ribosomal RNA sequencing in microbiome study related to human gut or marine ecosystem, we have relative abundance/compositional data of the group of microbial taxa at different taxonomic levels. A problem of interest is to model phenotype/response using these compositional covariates. Often we have observed that there is presence of either outlier or leveraged observation in data. Hence, we propose a robust regression model with compositional covariates. Subcompositional coherence of the model estimates are satisfied via linear constraint to the linear logcontrast model. In order for model to be robust, we add a mean shift parameter to each n instance of the data. The estimation is performed via penalized regression approach with regularization enforcing sparsity in mean shift parameter. We have investigated the model with both convex (l1 and adaptive l1) and non-convex (l0) penalty on mean shift. For the purpose of initialization in latter case and weight constriction in adaptive l1 penalty case, we propose an algorithm extending the idea of S-estimation in case of linear model with compositional covariates. Our approach has only one tuning parameter which is selected via a modified BIC selection criterion. Our theoretical analysis focus on non-asymptotic prediction error bound revealing interesting finitesample behaviors of the estimators. We have demonstrated the efficacy of the approach using various simulation studies and an application relating body mass index to human gut microbiome data.

Speech 3: Online deconvolution and demixing of calcium imaging data in real time
Speaker: Eftychios Pnevmatikakis (Simons Foundation)
Abstract: Optical imaging methods using calcium indicators enable monitoring the activity of large neuronal populations in vivo. Imaging experiments typically generate a large amount of data that needs to be processed to extract the activity of the imaged neuronal sources. While deriving such processing algorithms is an active area of research, most existing methods require the processing of large amounts of data rendering them

vulnerable to the volume of the recorded data, and preventing real-time experimental interrogation. Here we introduce CaImAn, an open source suite of tools for the online analysis of calcium imaging data, including i) motion artifact correction, ii) neuronal source extraction, and iii) activity denoising and deconvolution. Our approach combines and extends previous work on online dictionary learning and calcium imaging data analysis, to deliver an automated pipeline that can discover and track the activity of hundreds of cells in real time. We benchmark the performance of our algorithm on manually annotated data, and show that it outperforms popular offline approaches.

44. **Session title**: Recent Avances in Statistical Network, Functional and High-dimensional Data Analysis
Organizer: Ji Zhu (Umich)
Chair: Yujia Deng (UIUC)
Time: June 6th, 1:15pm - 2:45pm
Location: VEC 1402

Speech 1: Factor Augmented Vector Autoregressive Models under High Dimensional Scaling
Speaker: George Michailidis (U of Florida)
Abstract: Vector Autoregressive Models (VAR) are widely used in applied economics and finance. In this talk, we consider a VAR model augmented with dynamically evolving factors. The time series modeled as a VAR, together with the dynamic factors relate to a large number of other time series hat aid in the identifiability of the model parameters. We investigate the identifiability of such models, as well as estimation and inference issues under high-dimensional scaling. The performance of the proposed methods is assessed through synthetic data and the methodology is illustrated on a economic data set.

Speech 2: Model-assisted design of experiments on networks and social media platforms
Speaker: Edoardo Airoldi (Harvard)
Abstract: Classical approaches to causal inference largely rely on the assumption of "lack of interference", according to which the outcome of an individual does not depend on the treatment assigned to others, as well as on many other simplifying assumptions, including the absence of strategic behavior. In many applications, however, such as evaluating the effectiveness of healthcare interventions that leverage social structure, assessing the impact of product innovations and ad campaigns on social media platforms, or experimentation at scale in large IT companies, assuming lack of interference and other simplifying assumptions is untenable. Moreover, the effect of interference itself is often an inferential target of interest, rather than a nuisance. In this talk, we will formalize technical issues that arise in estimating causal effects when interference can be attributed to a network among the units of analysis, within the potential outcomes framework. We will introduce and discuss several strategies for experimental design in this context centered around a judicious use statistical models, which we refer to as "model-assisted" design of experiments. In particular, we wish for certain finite-sample properties of the estimator to hold even if the model catastrophically fails, while we would like to gain

efficiency if certain aspects of the model are correct. We will then contrast design-based, model-based and model-assisted approaches to experimental design from a decision theoretic perspective.

Speech 3: Correcting Selection Bias via Functional Empirical Bayes
Speaker: Gareth James (USC)
Abstract: Selection bias results from the selection of extreme observations and is a well recognized issue for standard scalar or multivariate data. Numerous approaches have been proposed to address the issue, dating back at least as far as the James-Stein shrinkage estimator. However, the same potential issue arises, albeit with additional complications, for functional data. Given a set of observed functions, one may wish to select for further analysis those which are most extreme according to some metric such as the average, maximum, or minimum value of the function. However, given the functions are often noisy realizations of some underlying mean process, these outliers are likely to generate biased estimates of the quantity of interest. In this talk I propose an Empirical Bayes approach, using Tweedie's formula, to adjust such functional data to generate approximately unbiased estimates of the true mean functions. The approach has several advantages. It is non-parametric in nature, but is capable of automatically shrinking back towards a James-Stein type estimator in low signal situations. It is also computationally efficient and possesses desirable theoretical properties. Furthermore, I demonstrate through extensive simulations that the approach can produce significant improvements in prediction accuracy relative to possible competitors. It is joint work with Joshua Derenski and Yingying Fan.

45. **Session title**: New insights into classical statistical methods
Organizer: Qing Mai (Florida State U)
Chair: Qing Mai (Florida State U)
Time: June 6th, 3:15pm - 4:45pm
Location: VEC 404/405

Speech 1: Rank-constrained inherent clustering paradigm for supervised and unsupervised learning
Speaker: Yiyuan She (Florida State U)
Abstract: Modern clustering applications are often faced with challenges from high dimensionality and/or nonconvex clusters. This paper gives a mathematical formulation of clustering with concurrent dimension reduction and proposes an optimization-based inherent clustering framework. Inherent clustering enjoys a kernel property to work on similarity matrices and can be extended to supervised learning. A simple-to-implement iterative algorithm is developed by use of linearization and block coordinate descent. Non-asymptotic analysis shows the tight error rate of inherent clustering in the supervised setting. Extensive simulations, as well as real-data experiments in network community detection and learning, demonstrate the excellent performance of the proposed approach.

Speech2: Fast and Optimal Bayesian Inference via Variational Approximations
Speaker: Yun Yang (Florida State U)
Abstract: We propose a variational approximation to Bayesian posterior distributions,

called α-VB, with provable statistical guarantees for models with and without latent variables. The standard variational approximation is a special case of α-VB with $\alpha=1$. When $\alpha \in(0,1)$, a novel class of variational inequalities are developed for linking the Bayes risk under the variationalapproximation to the objective function in the variational optimization problem, implying that maximizing the evidence lower bound in variational inference has the effect of minimizing the Bayes risk within the variational density family. Operating in a frequentist setup, the variational inequalities imply that point estimates constructed from the α-VB procedure converge at an optimal rate to the true parameter in a wide range of problems. We illustrate our general theory with a number of examples, including the mean-field variational approximation to (low)-high-dimensional Bayesian linear regression with spike and slab priors, mixture of Gaussian models, latent Dirichlet allocation, and (mixture of) Gaussian variational approximation in regular parametric models.

Speech 3: An Iterative Penalized Least Squares Approach to Sparse Canonical Correlation Analysis
Speaker: Xin Zhang (Florida State U)
Abstract: It is increasingly interesting to model the relationship between two sets of measurements when both of them are high-dimensional. Canonical correlation analysis (CCA) is a classical tool that explores the dependency of two multivariate random variables and extracts canonical pairs of highly correlated linear combinations. Driven by applications in genomics, text mining and imaging research among others, many recent studies generalize CCA to high-dimensional settings. However, most of them either rely on strong assumptions on the covariance matrices, or does not produce nested solutions. We propose a new sparse CCA (SCCA) method that recasts high-dimensional CCA as an iterative penalized least squares problem. Thanks to the new penalized least squares formulation, our SCCA method directly penalizes and estimates the sparse CCA directions with efficient algorithms. Therefore, in contrast to some existing methods, the new SCCA does not impose any sparsity assumptions on the covariance matrices. The proposed SCCA is also very flexible in the sense that it can be easily combined with properly chosen penalty functions to perform structured variable selection or to incorporate prior information. Moreover, our proposal of SCCA produces nested solutions, which provides great convenient in practice. Theoretical results show that SCCA can consistently estimate the true canonical pairs with an overwhelming probability in ultra-high dimensions. Numerical results also demonstrate the competitive performance of SCCA.

46. **Session title**: New developments for large complex data
Organizer: Annie Qu (UIUC)
Chair: Annie Qu (UIUC)
Time: June 6th, 3:15pm - 4:45pm
Location: VEC 902

Speech 1: Point and Interval Estimations for Individualized MCID
Speaker: Jiwei Zhao (SUNY, Buffalo)

Abstract: The minimal clinically important difference (MCID) is the smallest change in a treatment outcome that an individual patient would identify as important. In the era of precision medicine, it is of particular interest to study both point and interval estimations for the individualized MCID. The motivating example of this work is the ChAMP trial, which is a randomized controlled trial to compare debridement to observation of chondral lesions encountered during partial meniscectomy. In this trial, the primary outcome is the patient reported pain score one year after the surgery and we are interested in estimating the individualized MCID so that the treatment effect can be further studied. In this paper, we formulize this problem in a classification setting where nonconvex minimization technique is needed for the optimization. Furthermore, we develop the Bahadur representation of the individualized MCID so that its confidence interval can be derived. The proposed method is illustrated via comprehensive simulation studies. We also apply our proposed methodology to the ChAMP trial analysis.

Speech 2: Robust Probabilistic Classification for Irregularly Sampled Functional Data

Speaker: Doug Simpson (UIUC)

Abstract: Motivated by research on diagnostic ultrasound to evaluate tissue regions of interest, we present a robust probabilistic classifier for functional data that predicts the membership for given input and provides reliable probability estimates for class memberships. This method combines Bayes classifier and semi-parametric mixed effects model with robust tuning parameter. We aim to make the method robust to outlying curves especially in providing a robust estimate of certainty in prediction, which is crucial in medical diagnosis. This approach is applicable to various structures, such as samples observed over varying intervals or repeatedly measured curves retaining between-curve correlation, with no parametric assumption on within curve covariance. We conduct simulation studies to investigate the operating characteristics of the probability estimates in the presence of ideal data and data with outlying curves and compare with other functional classification procedures. We illustrate the methodology in classification of quantitative ultrasound data and other applications.

Speech 3: A new method for constructing gene co-expression networks based on samples with tumor purity heterogeneity

Speaker: Francesca Petralia (Mount Sinai)

Abstract: Tumor tissue samples often contain an unknown fraction of stromal cells. This problem well known as tumor purity heterogeneity (TPH) was recently recognized as a severe issue in omics studies. Specifically, if TPH is ignored when inferring co-expression networks, edges are likely to be estimated among genes with mean shift between non-tumor and tumor cells rather than among gene pairs interacting with each other in tumor cells. To address this issue, we propose TSNet a new method which constructs tumor-cell specific gene/protein co-expression networks based on gene/protein expression profiles of tumor tissues. TSNet treats the observed expression profile as a mixture of expressions from different cell types and explicitly models tumor purity percentage in each tumor sample. Using extensive synthetic data experiments, we demonstrate that TSNet outperforms a standard graphical model not accounting for tumor-purity heterogeneity. We then apply TSNet to estimate tumor specific gene co-

expression networks based on TCGA ovarian cancer RNAseq data. We identify novel co-expression modules and hub structure specific to tumor cells.

47. **Session title**: Statistical inference and complex data structures
Organizer: Eric Laber (NCST)
Chair: Yubai Yuan (UIUC)
Time: June 6th, 3:15pm - 4:45pm
Location: VEC 1303

Speech 1: Inter-modal Coupling: A Class of Measurements for Studying Local Covariance Patterns Among Multiple Imaging Modalities
Speaker: Kristin Linn (UPenn)
Abstract: Local cortical coupling was recently introduced as a subject-specific measure for studying localized relationships between cortical thickness and sulcal depth. Although a promising first step towards understanding local covariance patterns that are present between these two specific neuro-anatomical measurements, local cortical coupling suffers from a limited scope of imaging modalities that can be analyzed within the framework. We generalize and improve this local coupling measure by proposing an analogue in volumetric space that can be used to produce subject-level feature images among an arbitrary number of volumetric imaging modalities. Our proposed class of measures, collectively referred to as inter-modal coupling (IMCo), is based on a locally weighted regression framework. In this work, we study IMCo between cerebral blood flow and gray matter density using a sample of youths ages 8-21 from the Philadelphia Neurolodevelopmental Cohort. We describe how these two modalities covary spatially throughout the brain find evidence of significant developmental effects in several notable regions. We also give an overview of other applications where we are applying IMCo to study relationships between multiple types of images.

Speech 2: Modeling Heterogeneity in Motor Learning using Heteroskedastic Functional Principal Components
Speaker: Jeff Goldsmith (Columbia University)
Abstract: We propose a novel method for estimating population-level and subject-specific effects of covariates on the variability of functional data. We extend the functional principal components analysis framework by modeling the variance of principal component scores as a function of covariates and subject-specific random effects. In a setting where principal components are largely invariant across subjects and covariate values, modeling the variance of these scores provides a flexible and interpretable way to explore factors that affect the variability of functional data. Our work is motivated by a novel dataset from an experiment assessing upper extremity motor control, and quantifies the reduction in motion variance associated with skill learning.

Speech 3: Prior Adaptive Semi-supervised Learning with Application to Electronic Health Records Phenotyping
Speaker: Yichi Zhang (Harvard)
Abstract: Electronic Health Records (EHR) provides large and rich data sources for biomedical researches, and EHR data have been successfully used to gain novel insights

into several diseases. However, the usage of EHR data remains quite limited, because extracting precise phenotype for individual patient requires labor intensive medical chart review and such a manual process is not scalable. To facilitate an automatic procedure for accurate phenotyping, we formulate the problem in a high dimensional setting and propose a semi-supervised method that combine information from chart reviewed records with some data-driven prior knowledge derived from the entire dataset. The proposed estimator, Prior Adaptive Semi-supervised (PASS) estimator, enjoys nice theoretical properties including efficiency and robustness, and applies to a broad class of problems beyond EHR applications. The finite sample performance is evaluated via simulation studies and a real dataset on rheumatoid arthritis phenotyping. Further improvements involving word embedding and selective sampling are discussed.

48. **Session title**: Causal inference and statistical learning
Organizer: Cynthia Rudin (Duke)
Chair: Cynthia Rudin (Duke)
Time: June 6th, 3:15pm - 4:45pm
Location: VEC 1402

Speech 1: Teaching History and Ethics of Data, with Python
Speaker: Chris Wiggins (Columbia & NY Times)
Abstract: Data-empowered algorithms are reshaping our professional, personal, and political realities. However, existing curricula are predominantly designed either for future technologists, focusing on functional capabilities; or for future humanists, focusing on critical and rhetorical context surrounding data. "Data: Past, Present, and Future" is a new course at Columbia which seeks to define a curriculum at present taught to neither group, yet of interest and utility to future statisticians, CEOs, and senators alike. The course has been co-developed by Matt Jones, Professor at Columbia's History department, and myself. The intellectual arc traces from the 18th century to present day, beginning with examples of contemporary technological advances, disquieting ethical debates, and financial success powered by panoptic persuasion architectures. The weekly cadence of the course pairs primary and secondary readings with Jupyter notebooks in Python, engaging directly with the data and intellectual advances under study. Throughout, these intellectual technical advances are paired with critical inquiry into the forces which encouraged and benefited from these new capabilities, i.e., the political dimension of data and technology. In this talk I will give an overview of lessons learned from teaching the class, and argue that 1) the material can be engaged by students from a wide variety of curricular backgrounds and 2) the structure of the class -- using history to make the present strange, then critiquing the ethics of the technology-enabled future we are building -- can be useful for a variety of subjects. Syllabus, Jupyter notebooks, and additional info can be found via https://urldefense.proofpoint.com/v2/url?u=https-3A__data-2Dppf.github.io_&d=DwIGaQ&c=imBPVzF25OnBgGmVOlcsiEgHoG1i6YHLR0Sj_gZ4adc&r=tOZZtjNyrCSrR8o-Z8CHQgSAixSz_BAEnVZS6kAcAqM&m=F8OpY9yVfjfDb4WjIn47RbBGz-fwLvStzUaysCj_eBs&s=nWjj0uj0jEKmFsD0m3J_SPilwvfY3Fs5ZxA0HMUKPao&e= " Data: Past, Present, and Future" is supported by the Columbia University Collaboratory

Fellows Fund. Jointly founded by Columbia University's Data Science Institute and Columbia Entrepreneurship, The Collaboratory@Columbia is a university-wide program dedicated to supporting collaborative curricula innovations designed to ensure that all Columbia University students receive the education and training that they need to succeed in today's data rich world.

Speech 2: Bayesian optimization and A/B tests
Speaker: Ben Letham (Facebook data science)
Abstract: Randomized experiments provide a direct, albeit time-consuming and noisy, measurement of the effect of changes to a system. We often want to optimize the parameters of systems that can only be evaluated via noisy experiments. I will describe how Bayesian optimization is used in this setting at Facebook, such as for optimizing web server compiler flags. I will then discuss our current efforts to expand the scope of optimization via field experiments.

Speech 3: Causal inference from complex observational data
Speaker: Alex Volfovsky (Duke)
Abstract: A classical problem in causal inference is that of matching treatment units to control units. Some of the main challenges in developing matching methods arise from the tension among (i) inclusion of as many covariates as possible in defining the matched groups, (ii) having matched groups with enough treated and control units for a valid estimate of Average Treatment Effect (ATE) in each group, (iii) computing the matched pairs efficiently for large datasets, and (iv) dealing with complicating factors such as non-independence among units. We propose the Fast Large-scale Almost Matching Exactly (FLAME) framework to tackle these problems. At its core this framework proposes an optimization objective for match quality that captures covariates that are integral for making causal statements while encouraging as many matches as possible. This objective can then be augmented to capture common complicating factors.