Conference on Statistical Learning and Data Science

• Abstracts •

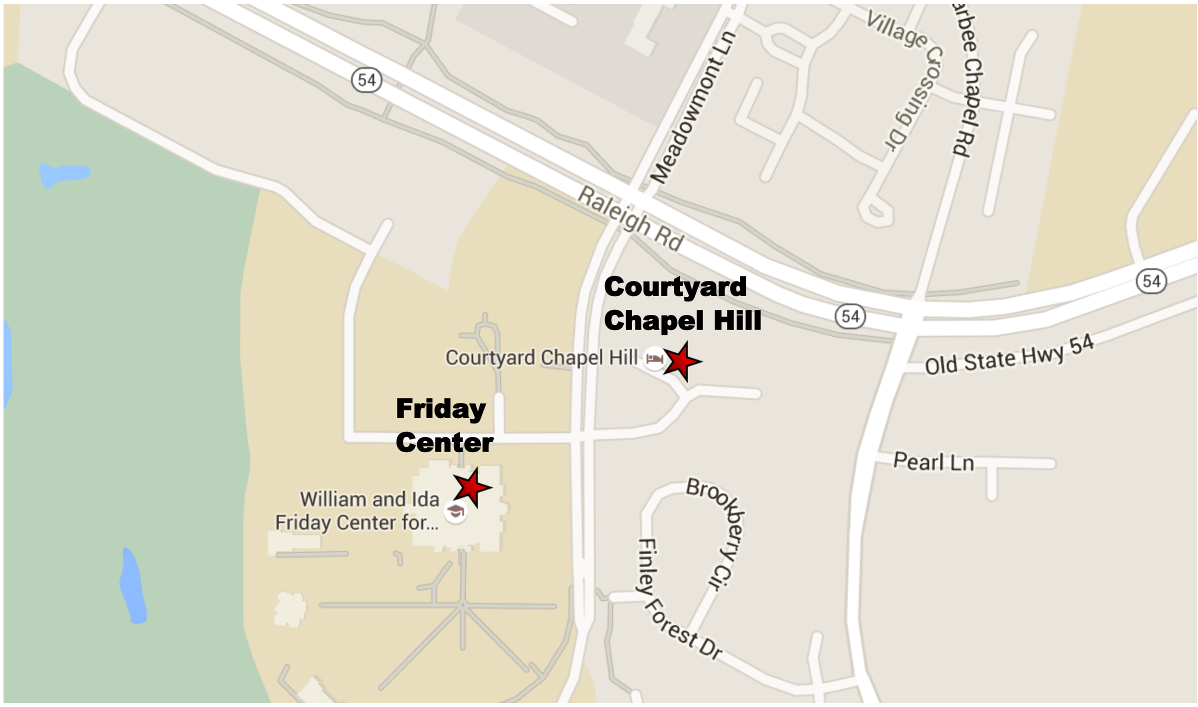
June 6-8, 2016

Chapel Hill, NC

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Contents

Map



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Sponsors

- The American Statistical Association Section on Statistical Learning and Data Science
- Institute for Operations Research and the Management Sciences (INFORMS) Data Mining Section
- Institute for Operations Research and the Management Sciences (INFORMS) Artificial Intelligence Section
- Statistical and Applied Mathematical Sciences Institute (SAMSI)
- National Science Foundation (Pending)
- Google Inc.
- RStudio
- SAS Institute Inc.
- UCB Biosciences Inc.
- Department of Statistics & Operations Research, Department of Biostatistics at the University of North Carolina at Chapel Hill

Program Chair

Yufeng Liu
University of North Carolina

Organizing Committee

- Yufeng Liu (University of North Carolina)
- Xingye Qiao (Binghamton University; Webmaster)
- Adam Rothman (University of Minnesota)
- Cynthia Rudin (MIT; INFORMS representative)
- Xiaotong Shen (University of Minnesota)

Program Committee

- Genevera Allen (Rice University)
- Yingying Fan (University of Southern California)
- Han Liu (Princeton University)
- Sahand Negahban (Yale University)
- Hernando Ombao (University of California, Irvine)
- Annie Qu (University of Illinois Urbana-Champaign)
- Karl Rohe (University of Wisconsin Madison)
- Ali Shojaie (University of Washington)
- Richard Smith (SAMSI & University of North Carolina)
- Matt Taddy (University of Chicago)
- Robert Warnock (UCB Biosciences Inc.)
- Yuying Xie (Michigan State University)
- Yan Xu (SAS Institute)
- Hao Helen Zhang (University of Arizona)
- Ji Zhu (University of Michigan)

Local Committee

- David Banks (Duke University)
- Shankar Bhamidi (University of North Carolina)
- Siliang Gong (University of North Carolina)
- Eric Laber (North Carolina State University)
- Jan Hannig (University of North Carolina)
- Quefeng Li (University of North Carolina)
- J. S. Marron (University of North Carolina)
- Andrew Nobel (University of North Carolina)
- Yichao Wu (North Carolina State University)
- Yin Xia (University of North Carolina)
- Donglin Zeng (University of North Carolina)
- Kai Zhang (University of North Carolina)

Invited talks

June 6 (Monday)

Parallel Sessions **9:00-10:30am**

Theoretical Statistical Learning **Dogwood AB**
(organized by Yin Xia, UNC; chaired by Jan Hannig, UNC)

Florentina Bunea (Cornell University) 9:00-9:30am
Minimax Optimal Variable Clustering in G-Models

The goal of variable clustering is to partition a random vector **X**∈R^p in sub-groups of similar probabilistic behavior. Popular methods such as hierarchical clustering or *K*-means are algorithmic procedures applied to observations on **X**, while no population level target is defined prior to estimation. We take a different view in this talk, where we discuss model based variable clustering. We consider three models, of increasing level of complexity, termed generically G-models, with G standing for the partition to be estimated. Motivated by the potential lack of identifiability of the G-latent models, which are currently used in problems involving variable clustering, we introduce two new classes of models, the G-exchangeable and the G-block covariance models. We show that both classes are identifiable, for any distribution of **X**, thereby providing well defined targets for estimation. Our focus is on clusters that are invariant with respect to unknown monotone transformations of the data, and that can be estimated in a computationally feasible manner. Both desiderata can be met if the clusters correspond to blocks in the copula correlation matrix of **X**, assumed to have a Gaussian copula distribution. This motivates the introduction of a new similarity metric for cluster membership, CORD, and of a homonymous method for cluster estimation. Central to our work is the derivation of the minimax rate of the CORD cluster separation for exact partition recovery. We obtain the surprising result that the CORD rate is of order (log(p)/n)^(1/2), irrespective of the number of clusters, or of the size of the smallest cluster. Our new procedure, CORD, available on CRAN, achieves this bound and has computational complexity that is polynomial in *p*. The CORD distance between two clusters is larger than the classical "within-between" correlation gap between clusters, and can be employed even when the latter is negative. However, in the particular case of a positive correlation GAP, the GAP minimax rate for exact recovery is (log(p)/mn)^(1/2), where m is the minimum cluster size. We show that while methods such as spectral clustering cannot, in general, recover the partition exactly at the minimax GAP separation level, convex algorithms can be near minimax optimal. Our results are further supported by extensive numerical studies and data examples.

Andrew Nobel (UNC Chapel Hill) 9:30-10:00am
Large Average Submatrices of a Gaussian Random Matrix: Behavior of Global and Local optima

The problem of finding large average submatrices of a real-valued matrix arises in the exploratory analysis of data from disciplines as diverse as genomics and social sciences. This talk will present several new theoretical results concerning large average submatrices of an n x n Gaussian random matrix that are motivated in part by previous work on this applied problem. We will begin by considering the average and distribution of the k x k submatrix having largest average value (the global maximum), and then turn our attention to submatrices with dominant row and column sums, which arise as the local maxima of a practical iterative search procedure for large average submatrices. These results characterize the value and joint distribution of a typical local maximum, and identify the limiting behavior of the number of local maxima. In the last part of the talk I will present some recent results on the overlap of k x k submatrices ordered by their average values. Joint work with Shankar Bhamidi (UNC), Partha S. Dey (UIUC), and James Wilson (USF).

Kai Zhang (UNC Chapel Hill) 10:00-10:30am
Packing Inference of Correlation for an Arbitrarily Large Number of Variables

We study the spherical cap packing problem with a probabilistic approach. Such probabilistic considerations result in an asymptotic universal uniform sharp bound on the maximal inner product between any set of unit vectors and a stochastically independent uniformly distributed unit vector. When the set of unit vectors are themselves independently uniformly distributed, we further develop the extreme value distribution limit of the maximal inner product, which characterizes its stochastic uncertainty around the bound. As applications of the above asymptotic results, we derive (1) an asymptotic universal uniform sharp bound on the maximal spurious correlation, as well as its uniform convergence in distribution when the explanatory variables are independently Gaussian; and (2) a sharp universal bound on the maximum norm of a low-rank elliptically distributed vector, as well as related limiting distributions. With these results, we develop a fast detection method for a low-rank in high dimensional Gaussian data without using the spectrum information.

Distributed Optimization Methods for Machine Learning **Azalea AB**

(organized and chaired by Yan Xu, SAS)

Alireza Yektamaram (Lehigh University and SAS) 9:00-9:30am
A Nonconvex Hessian-free Method for Deep Learning Problems

Krlyov-based non-convex optimization methods have gained interest as result of resurgence in Hessian-Free methods in deep learning. Current approaches are unable to solve the underlying Newton equations (even when positive-definite) to user-defined levels of accuracy; either as positive-definite approximations to the Hessian are used, or the system is restricted to work in a necessarily small dimensional Lanczos subspace which requires excessive amount of storage to update the weights. In this study we explore the use of a line-search approach, without the mentioned limitations and with trust-region strength convergence, and numerically demonstrate this method is effective for both Hessian and Generalized Gauss-Newton (GN) operators.

Jorge Silva (SAS) 9:30-10:00am
Learning Good Ensembles: New Approaches

Stacked generalization models are flexible ensemble models that combine multiple machine learning classifiers by using their linear combination. The weights for each classifier are found by least squares regression. To ovoid overfitting, stacking methods use cross validated predictions for each classifier. It has been shown by previous studies that even when cross validated predictions are used, stacking models tend to result in overfitting models. In this paper we demonstrate that the overfitting can be further reduced, and hence the predictive accuracy can be increased, by using various penalties on the size of the regression coefficients including L1, L2 penalty, and penalty on the summation of the regression weights.

Patrick Koch (SAS) 10:00-10:30am
Local Search Optimization for Hyper-Parameter Tuning

Facilitating effective decision making requires the transformation of relevant data to high quality descriptive and predictive models. Machine learning modeling algorithms are commonly used to find hidden value in big data. These algorithms are governed by ‘hyper-parameters’ with no clear defaults agreeable to a wide range of applications. Ideal settings for these hyper-parameters can significantly influence the resulting accuracy of the predictive models. To move from a manual, random or expert user based adjustment of tuning parameters, most often a rough grid search approach has been taken. More recently intelligent optimization strategies have been applied. In this talk we discuss the use of local search optimization (LSO) for machine learning algorithm hyper-parameter tuning. As a complex black-box to the tuning process, machine learning algorithms create a challenging class of optimization problems. The corresponding objective functions involved tend to be nonsmooth, discontinuous, unpredictably computationally expensive, and require support for continuous, categorical, and integer variables. Further evaluations can fail for a variety of reasons such as early termination due to node failure or time out on a busy network/grid installation. Additionally, not all hyperparameter combinations are compatible (creating so called "hidden constraints"). In this context, we apply a parallel hybrid derivative-free optimization strategy that can tune predictive models despite these difficulties, providing significantly improved results over default settings with minimal user interaction. We also address efficient parallel paradigms for different types of machine learning problems, while exploring the importance of validation to avoid overfitting and emphasizing that even for small data problems, the computation expense of performing cross validation can benefit from a distributed/threaded environment.

Applied Learning and Analysis — Mountain Laurel AB

(organized and chaired by Cynthia Rudin, MIT)

John Guerard (McKinley Capital Mgt., LLC) 9:00-9:30am
Robust Regression and Data Mining of Financial Data

In our analysis of Global data, 1997-9/2015, we apply several Data Mining Corrections tests to establish that our statistically-based regression analysis of financial data is not due to chance. We create a universe of 7500 stocks and analyze 28 financial variables to establish highly predictive statistically-based stock selection models using several methods of Tukey and Yohai-estimated optimal influence efficiency measures to robust regression techniques. We address issues of outliers and multicollinearity to estimate models by applying the least angular regression (LAR), LASSO, and Weighted Latent Root Regression (WLRR.). We study how to apply the cross-validation to choose among LAR, LASSO, and the WLRR regression techniques. We create Markowitz Efficient Frontiers using individual and composite model strategies to apply the Markowitz-Xu Data Mining Corrections test. The LAR, LASSO, and WLRR stock selection models produce highly statistically significant stock selection and are statistically significantly different from average models such that we reject the null hypothesis of data mining.

Edward McFowland III (University of Minnesota) 9:30-10:00am
Efficient Identification of Heterogeneous Treatment Effects in Randomized Experiments via Anomalous Pattern Detection

The randomized experiment has been an important tool for inferring the causal impact of an intervention; the most common analysis conducted in this context is the average treatment effect (ATE). However, the recent heterogeneous treatment effects literature is showing the utility in estimating the marginal conditional average treatment effect (MCATE), which estimates a treatment effect for a subpopulation– i.e., respondents who share a particular subset of covariates. Additionally, the literature proposes the use of data mining methods to estimate the exponential number (in covariate size) of MCATEs that exist in the data. However, each proposed method makes its own set of (restrictive) assumptions about the intervention's affect, the underlying data generating processes, and which subpopulations (MCATEs) to explicitly estimate. Moreover, the majority of the literature provides no mechanism to identify which subpopulations are the most affected–beyond manual inspection–and provide little guarantee for the estimation error of the specific MCATE estimates. Therefore, we propose Treatment Effect Subset Scan (TESS), a new method for identifying which subpopulation in a randomized experiment is most affected by a treatment. We frame the affected subpopulation identification challenge as a pattern detection problem where we maximize a nonparametric scan statistic (measurement of distributional divergence) over all subpopulations, while being extremely parsimonious in which specific subpopulations' effects to estimate. Furthermore, we identify the subpopulation which experiences the largest distributional change as a result of the intervention, while making minimal assumptions about the intervention's affect or the underlying data generating process. In addition to the algorithm, we provide non-asymptotic statistical bounds on its error, detection power, and provide sufficient conditions for detection consistency–i.e., exact identification of affected subpopulation. Finally, we validate the efficacy of the method by identifying heterogenous treatment effects in simulations and in well-known program evaluations studies.

Stefano Traca (MIT) 10:00-10:30am
Regulating Greed Over Time

In retail, there are predictable yet dramatic time-dependent patterns in customer behavior, such as periodic changes in the number of visitors, or increases in visitors just before major holidays (e.g., Christmas). The current paradigm of multi-armed bandit analysis does not take these known patterns into account, which means that despite the firm theoretical foundation of these methods, they are fundamentally flawed when it comes to real applications. This work provides a remedy that takes the time-dependent patterns into account, and we show how this remedy is implemented in the UCB and ε-greedy methods. In the corrected methods, exploitation (greed) is regulated over time, so that more exploitation occurs during higher reward periods, and more exploration occurs in periods of low reward. In order to understand why regret is reduced with the corrected methods, we present a set of bounds that provide insight into why we would want to exploit during periods of high reward, and discuss the impact on regret. Our proposed methods have excellent performance in experiments, and were inspired by a high-scoring entry in the Exploration and Exploitation 3 contest using data from Yahoo! Front Page. That entry heavily used time-series methods to regulate greed over time, which was substantially more effective than other contextual bandit methods.

Network Inference — Bellflower AB

(organized by Karl Rohe, University of Wisconsin Madison; chaired by Bailey Fosdick, Colorado State University)

Can Le (University of Michigan) 9:00-9:30am
Structure of sparse random networks

Network analysis has become an important area in many research domains. A common way to study real-world networks is to model them as random graphs whose structure is encoded in the expectation matrix. We consider a general model of networks on n nodes, known as the inhomogeneous Erdos-Renyi model, where edges between nodes are formed independently and possibly with different probabilities. We study the behavior of such random networks through the concentration of their adjacency and Laplacian matrices in the spectral norm. Sparse random networks whose expected average degrees grow slower than log(n) fail to concentrate. The obstruction is caused by vertices with abnormally high and low degrees. We show that concentration can be restored if we regularize the degrees of such vertices, and one can do this in various ways. As an immediate consequence, we establish the validity of one of the simplest and fastest approaches to community detection – regularized spectral clustering, under the stochastic block model. We also discuss how to choose the regularization parameter and estimate the number of communities.

Daniel Sussman (Harvard University) 9:30-10:00am
Unbiased Estimation of Causal Effects under Network Interference

In online experiments, frequently the treatment of one unit may cause effects on neighboring units according to a network structure. In this talk, we investigate a series of assumptions about network interference for causal inference. We will answer the question of when unbiased estimators exist and provide a method to choose among them.

Alexander Volfovsky (Harvard/ Duke) 10:00-10:30am
Testing and estimation for relational data

Recent years have seen a dramatic rise in social media, networks, and other settings in which the relationships and interactions between individuals, countries or objects are observed. These types of relational data are often represented as a square matrix, the entries of which record the relationships between pairs of objects. Statistical methods for such data range from network regression where entries are frequently assumed to be independent to latent space methods that assume some degree of similarity or dependence between objects in terms of the way they relate to each other. However, formal tests for such dependence have not been developed. First, we provide a test (based on the observation of a single relational data matrix) for such dependence using the framework of the matrix normal model, a type of multivariate normal distribution parameterized in terms of row- and column-specific covariance matrices. Second, we develop an estimation procedure (still based on the observation of a single relational data matrix) that captures the variability in such data by leveraging the identical index sets of the rows and columns.

Plenary Talk **11:00am-12:00pm**
(chaired by Richard Smith, SAMSI & UNC) **Dogwood AB**

Bin Yu (UC Berkeley)
Unveiling the mysteries in spatial gene expression

Genome-wide data reveal an intricate landscape where gene activities are highly differentiated across diverse spatial areas. These gene actions and interactions play a critical role in the development and function of both normal and abnormal tissues. As a result, understanding spatial heterogeneity of gene networks is key to developing treatments for human diseases. Despite the abundance of recent spatial gene expression data, extracting meaningful information remains a challenge for local gene interaction discoveries. In response, we have developed staNMF, a method that combines a powerful unsupervised learning algorithm, nonnegative matrix factorization (NMF), with a new stability criterion that selects the size of the dictionary. Using staNMF, we generate biologically meaningful Principle Patterns (PP), which provide a novel and concise representation of Drosophila embryonic spatial expression patterns that correspond to pre-organ areas of the developing embryo. Furthermore, we show how this new representation can be used to automatically predict manual annotations, categorize gene expression patterns, and reconstruct the local gap gene network with high accuracy. Finally, we discuss on-going crispr/cas9 knock-out experiments on Drosophila to verify predicted local gene-gene interactions involving gap-genes. An open-source software is also being built based on SPARK and Fiji.

This talk is based on collaborative work of a multi-disciplinary team (co-lead Erwin Frise) from the Yu group (statistics) at UC Berkeley, the Celniker group (biology) at the Lawrence Berkeley National Lab (LBNL), and the Xu group (computer science) at Tsinghua Univ.

Parallel Sessions **1:30-3:00pm**

New Regularization Techniques **Dogwood AB**
(organized by Ji Zhu, Univ. Michigan; chaired by Annie Qu, UIUC)

Cun-Hui Zhang (Rutgers University) 1:30-2:00pm
Nonparametric Shrinkage Estimation

We revisit the classical problem of estimating the mean vector of three or more uncorrelated observations with a known common variance. James and Stein (1962) claimed that when the fourth moment of the noise has a known upper bound, a shrinkage estimator of Stein (1956) always has a smaller average mean squared error than the common variance, provided that certain parameters of the shrinkage factor are properly specified. James and Stein (1962) further commented “It would be desirable to obtain explicit formulas for estimators one can seriously recommend” in this nonparametric setting.

Fang Han (University of Washington) 2:00-2:30pm
Optimal Structure-Induced Network Estimation

The problem of studying connections between n nodes finds many applications, in neuroscience and social network analysis among others. This paper focuses on optimal estimation of such connections. In contrast to Gao, Lu, and Zhou (2015), we assume a "Universal Law of Geometry", indicating nodes are more possible to be connected when they are geometrically closer. In the simplest case, this resembles the problems of change point detection (CPD) and piecewise linear function estimation (PLFE). For this, we show, assuming k pieces in the model, a phase transition exists: When k increases from 2 to higher, the minimax rate of convergence increases from loglog(n)/n to klog(en/k)/n. Such a result is absent in CPD and PLFE, and could motivate retrospect on the "sparsity assumption" popular in the high dimensional statistics literature. An extension to shape constraint models and an exploration of Kiefer's Theorem will also be discussed if time permits.

Qing Mai (Florida State University) 2:30-3:00pm
Multiclass Sparse Discriminant Analysis

In recent years many sparse linear discriminant analysis methods have been proposed for high-dimensional classification and variable selection. However, most of these proposals focus on binary classification and they are not directly applicable to multiclass classification problems. There are two sparse discriminant analysis methods that can handle multiclass classification problems, but their theoretical justifications remain unknown. In this talk, we propose a new multiclass sparse discriminant analysis method that estimates all discriminant directions simultaneously. We show that when applied to the binary case our proposal yields a classification direction that is equivalent to those by two successful binary sparse LDA methods in the literature. An efficient algorithm is developed for computing our method with high-dimensional data. Variable selection consistency and rates of convergence are established under the ultrahigh dimensionality setting. We further demonstrate the superior performance of our proposal over the existing methods on simulated and real data.

Machine Learning on Big Data Azalea AB

(organized and chaired by Matt Taddy, University of Chicago)

Daniel Roy (University of Toronto) 1:30-2:00pm
Sparse Random Graphs arising from Exchangeable Random Measures

The statistical analysis of network data rests on a foundation of random graph models, yet existing models of sparse graphs are inadequate for many purposes. We introduce a new class of random graphs on the reals R defined by the exchangeability of their vertices. A straightforward adaptation of a result by Kallenberg yields a representation theorem: every such random graph is characterized by three (potentially random) components: a nonnegative real I in R+, an integrable function S: R+ to R+, and a symmetric measurable function W: R+^2 to [0,1] that satisfies several weak integrability conditions. We call the triple (I,S,W) a graphex, in analogy to graphons, which characterize the (dense) exchangeable graphs on N. I will present some results about the structure and consistent estimation of these random graphs, and what role they can play in the analysis of real-world networks. Joint work with Victor Veitch.

Rebecca Steorts (Duke University) 2:00-2:30pm
Why infinite exchangeable mixture models fail for sparse data sets yet microclustering succeeds

Record linkage merges together large, potentially noisy databases to remove duplicate entities. Community detection is the process of placing entities into similar partitions or “communities”. Both applications are important to applications in author disambiguation, genetics, official statistics, human rights conflict, and others. It is common to treat record linkage and community detection as clustering tasks. In fact, generative models for clustering implicitly assume that the number of data points in each cluster grows linearly with the total number of data points. Finite mixture models, Dirichlet process mixture models, and Pitman--Yor process mixture models make this assumption. For example, when performing record linkage, the size of each cluster is often unrelated to the size of the data set. Consequently, each cluster contains a negligible fraction of the total number of data points. Such tasks require models that yield clusters whose sizes grow sublinearly with the size of the data set. We address this requirement by defining the *micro clustering property* and discussing a new model that exhibits this property. We illustrate this on real and simulated data. This is joint work with Jeff Miller, Brenda Betancourt, Abbas Zaidi (Duke University) and Hanna Wallach (UMass Amherst and Microsoft Research).

Yichao Wu (NCSU) 2:30-3:00pm
Principal Weighted Support Vector Machines for Sufficient Dimension Reduction in Binary Classification

Sufficient dimension reduction is an efficient tool and has been popular for reducing data dimensionality without relying on stringent model assumptions. However, most existing methods primarily target at regression with continuous responses and may not work well for binary classification. For example, sliced inverse regression (Li, 1991) can estimate at most one direction if the response is binary. In this paper, we propose principal weighted support vector machine, a unified framework for both linear and nonlinear sufficient dimension reduction in binary classification. Asymptotic properties are studied and an efficient computing algorithm is proposed for the proposed method. Numerical examples demonstrate its competitive performance in binary classification.

High Dimensional Learning Methods and Theory Mountain Laurel AB

(organized by Xiaotong Shen, Univ. of Minnesota; chaired by Wei Sun, Fred Hutchinson Cancer Research Center)

Jan Hannig (UNC Chapel Hill) 1:30-2:00pm
Generalized fiducial inference for high-dimensional data

In recent years, the ultrahigh-dimensional linear regression problem has attracted enormous attention from the research community. Under the sparsity assumption, most of the published work is devoted to the selection and estimation of the predictor variables with nonzero coefficients. This article studies a different but fundamentally important aspect of this problem: uncertainty quantification for parameter estimates and model choices. To be more specific, this article proposes methods for deriving a probability density function on the set of all possible models, and also for constructing confidence intervals for the corresponding parameters. These proposed methods are developed using the generalized fiducial methodology, which is a variant of Fisher's controversial fiducial idea. Theoretical properties of the proposed methods are studied, and in particular it is shown that statistical inference based on the proposed methods will have correct asymptotic frequentist property. In terms of empirical performance, the proposed methods are tested by simulation experiments and an application to a real dataset. Finally, this work can also be seen as an interesting and successful application of Fisher's fiducial idea to an important and contemporary problem. To the best of the authors' knowledge, this is the first time that the fiducial idea is being applied to a so-called "large p small n" problem. A connection to objective Bayesian model selection and non-local priors is also discussed.

Yiyuan She (Florida State University) 2:00-2:30pm
Indirect Gaussian Graph Learning beyond Gaussianity

This talk discusses how to capture parsimonious dependency structure between random variables in multivariate data analysis. Given an arbitrary continuously differentiable loss function, not necessarily derived from a particular distribution, we use a new technique called additive over-parametrization to come up with a multivariate criterion that builds the dependency into statistical modeling. The method provides good asymptotics and offers great ease in computation. After linearization, the optimization problem degenerates to Gaussian graph learning. An iterative algorithm is proposed and statistical analysis shows its fast convergence rate. Examples are shown in Bernoullli, Poisson and robust settings.

Zhigen Zhao (Temple University) 2:30-3:00pm
A New Approach to Multiple Testing of Grouped Hypotheses

A new approach to multiple testing of grouped hypotheses controlling false discoveries is proposed. It decomposes a posterior measure of false discoveries across all hypotheses into within- and between-group components allowing a portion of the overall FDR level to be used to maintain control over within-group false discoveries. Numerical calculations performed under certain model assumption for the hidden states of the within-group hypotheses show its superior performance over its competitors that ignore the group structure, especially when only a few of the groups contain the signals, as expected in many modern applications. We offer data-driven version of our proposed method under our model by estimating the parameters using EM algorithms and provide simulation evidence of its favorable performance relative to these competitors. Real data applications have also produced encouraging results for the proposed approach.

New Mining Tools for Complex Data Bellflower AB

(organized by Hao Helen Zhang, U. Arizona; chaired by Xiaoli Gao, UNC-Greensboro)

Ping Ma (University of Georgia) 1:30-2:00pm
Smoothing spline ANOVA for super-large samples

In the current era of big data, researchers routinely collect and analyze data of super-large sample sizes. Data-oriented statistical methods have been developed to extract information from super-large data. Smoothing spline ANOVA (SSANOVA) is a promising approach for extracting information from noisy data; however, the heavy computational cost of SSANOVA hinders its wide application. In this talk, I will present some algorithms for fitting SSANOVA models to super-large sample data. Our results reveal that our methods enable researchers to fit nonparametric regression models to very large samples within a few seconds using a standard laptop or tablet computer.

Junming Yin (University of Arizona) 2:00-2:30pm
Latent Space Inference of Internet-Scale Networks

The rise of social and Internet networks with hundreds of millions to billions of nodes, presents new challenges for scaling up statistical network models to execute in a reasonable amount of time on Internet-scale networks. By applying a succinct representation of networks as a bag of triangular motifs, developing a parsimonious statistical model, deriving an efficient stochastic variational inference algorithm, and implementing it as a distributed cluster program, we demonstrate latent space inference and overlapping community detection on very large networks with over 100 million nodes on just a few cluster machines. Compared to other state-of-the-art probabilistic network approaches, our method is several orders of magnitude faster, with competitive or improved accuracy at overlapping community detection. This is joint work with Qirong Ho and Eric P. Xing.

Boxiang Wang (University of Minnesota) 2:30-3:00pm
Another Look at DWD

Distance weighted discrimination (DWD) is a modern margin-based classifier with an interesting geometric motivation. Despite many recent papers on DWD, DWD is far less popular compared with the support vector machine, mainly due to computational and theoretical reasons. In this work, we greatly advance the current DWD methodology and its learning theory. We propose a novel efficient algorithm for solving DWD, and our algorithm can be several hundred times faster than the existing state-of-the-art algorithm based on the second order cone programming (SOCP). In addition, our algorithm can handle the generalized DWD, while the SOCP algorithm only works well for a special DWD but not the generalized DWD. Furthermore, we formulate a natural kernel DWD in a reproducing kernel Hilbert space and then establish the Bayes risk consistency of the kernel DWD using a universal kernel such as the Gaussian kernel. This result solves an open theoretical problem in the DWD literature. We compare DWD and the support vector machine on several benchmark data sets and show that the two have comparable classification accuracy, but DWD equipped with our new algorithm can be much faster to compute than the support vector machine.

Parallel Sessions 3:30-5:00pm

Machine Learning for Imaging and Medical Applications Dogwood AB

(organized and chaired by Donglin Zeng, UNC)

Dinggang Shen (UNC Chapel Hill) 3:30-4:00pm
Machine Learning in Medical Imaging Analysis

This talk will summarize our recently developed machine learning techniques, including sparse learning and deep learning, for various applications in medical imaging. Specifically, in neuroimaging field, we have developed an automatic tissue segmentation method for the first-year brain images with the goal of early detection of autism such as before 1-year-old, and also a novel multivariate classification method for early diagnosis of Alzheimer’s Disease (AD) with the goal of potential early treatment. In image reconstruction field, we have developed a sparse learning method for reconstructing 7T-like MRI from 3T MRI for enhancing image quality, and also another novel sparse learning technique for estimation of standard-dose PET image from low-dose PET and MRI data. Finally, in cancer radiotherapy field, we have developed an innovative regression-guided deformable model to automatically segment pelvic organs from single planning CT which is currently done manually, as well as a novel image synthesis technique for estimating CT from MRI for current new direction of MRI-based dose planning (and also for PET attenuation correction in the case of using PET/ MRI scanner). All these techniques and applications will be discussed in this talk.

Yuanjia Wang (Columbia University) 4:00-4:30pm
Support Vector Hazards Machine: A Counting Process Framework for Learning Risk Scores for Censored Outcomes

Learning risk scores to predict dichotomous or continuous outcomes using machine learning approaches has been studied extensively. However, how to learn risk scores for time-to-event outcomes subject to right censoring has received little attention until recently. Existing approaches rely on inverse probability weighting or rank-based regression, which may be inefficient. In this paper, we develop a new support vector hazards machine (SVHM) approach to predict censored outcomes. Our method is based on predicting the counting process associated with the time-to-event outcomes among subjects at risk via a series of support vector machines. Introducing counting processes to represent time-to-event data leads to a connection between support vector machines in supervised learning and hazards regression in standard survival analysis. To account for different at risk populations at observed event times, a time-varying offset is used in estimating risk scores. The resulting optimization is a convex quadratic programming problem that can easily incorporate non- linearity using kernel trick. We demonstrate an interesting link from the profiled empirical risk function of SVHM to the Cox partial likelihood. We then formally show that SVHM is optimal in discriminating covariate-specific hazard function from population average hazard function, and establish the consistency and learning rate of the predicted risk using the estimated risk scores. Simulation studies show improved prediction accuracy of the event times using SVHM compared to existing machine learning methods and standard conventional approaches. Finally, we analyze two real world biomedical study data where we use clinical markers and neuroimaging biomarkers to predict age-at-onset of a disease, and demonstrate superiority of SVHM in distinguishing high risk versus low risk subjects.

Ming Yuan (University of Wisconsin Madison) 4:30-5:00pm
Structured Correlation Detection with Application to Colocalization Analysis in Dual-Channel Fluorescence Microscopic Imaging

Motivated by the problem of colocalization analysis in fluorescence microscopic imaging, we study in this paper structured detection of correlated regions between two random processes observed on a common domain. We argue that although intuitive, direct use of the maximum log-likelihood statistic suffers from potential bias and substantially reduced power, and introduce a simple size-based normalization to overcome this problem. We show that scanning with the proposed size-corrected likelihood ratio statistics leads to optimal correlation detection over a large collection of structured correlation detection problems.

High-dimensional Inference Azalea AB

(organized and chaired by Kai Zhang, UNC)

Anru Zhang (University of Wisconsin at Madison) 3:30-4:00pm
Rate-Optimal Perturbation Bounds for Singular Subspaces with Applications to High-Dimensional Statistics

Perturbation bounds for singular spaces, in particular Wedin's sinΘ theorem, are a fundamental tool in many fields including high-dimensional statistics, machine learning, and applied mathematics. In this paper, we establish new perturbation bounds, measured in both spectral and Frobenius sinΘ distances, separately for the left and right singular subspaces. Lower bounds, which show that the individual perturbation bounds are rate-optimal, are also given. The new perturbation bounds are applicable to a wide range of problems. In this paper, we consider in detail applications to low-rank matrix denoising and singular space estimation, high-dimensional clustering, and canonical correlation analysis (CCA). In particular, separate matching upper and lower bounds are obtained for estimating the left and right singular spaces. To the best of our knowledge, this is the first result that gives different optimal rates for the left and right singular spaces under the same perturbation. In addition to these problems, applications to high-dimensional problems such as community detection in bipartite networks and multidimensional scaling are also discussed.

Han Liu (Princeton University) 4:00-4:30pm
Combinatorial Inference

We introduce a new research area named combinatorial inference which explores the topological structures of high dimensional graphical models. In particular, the combinatorial inference aims to test complex structural hypotheses including connectivity, hub detection, perfect matching, etc. We propose a shortest self-returning path approach to prove the general optimality of our proposed methods. Our methods are applied to the neuroscience by discovering hub voxels contributing to visual memories. This is based on joint work with Junwei Lu, Matey Neykov, and Kean Ming Tan.

Shu Lu (UNC Chapel Hill) 4:30-5:00pm
Confidence Regions and Intervals for Sparse Penalized Regression Using Variational Inequality Techniques

With the abundance of large data, sparse penalized regression techniques are commonly used in data analysis due to the advantage of simultaneous variable selection and prediction. In this talk, we discuss a framework to construct confidence intervals for sparse penalized regression with a wide range of penalties including the LASSO penalty. We study the inference for two types of parameters: the parameters under the population version of the penalized regression and the parameters in the underlying linear model. We present convergence properties of the proposed methods as well as results for simulated and real data examples. This is based on joint work with Yufeng Liu, Liang Yin and Kai Zhang.

Topics on High Dimensional Learning and Inference — Mountain Laurel AB

(organized and chaired by Genevera Allen, Rice University)

Johannes Lederer (University of Washington) 3:30-4:00pm
Efficient Feature Selection With Big Data

Big Data relates aspects of very different fields, such as statistics, computer science, informatics, and mathematics. In this talk, we will cover these different aspects and present novel, easy-to-use methods for feature selection with Big Data.

Weijie Su (Stanford University) 4:00-4:30pm
Multiple Testing and Adaptive Estimation via the Sorted L-One Norm

In many real-world statistical problems, we observe a large number of potentially explanatory variables of which a majority may be irrelevant. For this type of problem, controlling the false discovery rate (FDR) guarantees that most of the discoveries are truly explanatory and thus replicable. In this talk, we propose a new method named SLOPE to control the FDR in sparse high-dimensional linear regression. This computationally efficient procedure works by regularizing the fitted coefficients according to their ranks: the higher the rank, the larger the penalty. This is analogous to the Benjamini -Hochberg procedure, which compares more significant p-values with more stringent thresholds. Whenever the columns of the design matrix are not strongly correlated, we show empirically that SLOPE obtains FDR control at a reasonable level while offering substantial power. Although SLOPE is developed from a multiple testing viewpoint, we show the surprising result that it achieves optimal squared errors under Gaussian random designs over a wide range of sparsity classes. An appealing feature is that SLOPE does not require any knowledge of the degree of sparsity. This adaptivity to unknown sparsity has to do with the FDR control, which strikes the right balance between bias and variance. The proof of this result presents several elements not found in the high-dimensional statistics literature.

Stefan Wager (Stanford University) 4:30-5:00pm
Causal Inference with Random Forests

Many scientific and engineering challenges---ranging from personalized medicine to customized marketing recommendations---require an understanding of treatment heterogeneity. We develop a non-parametric causal forest for estimating heterogeneous treatment effects that extends Breiman's widely used random forest algorithm. Given a potential outcomes framework with unconfoundedness, we show that causal forests are pointwise consistent for the true treatment effect, and have an asymptotically Gaussian and centered sampling distribution. We also propose a practical estimator for the asymptotic variance of causal forests. In both simulations and an empirical application, we find causal forests to be substantially more powerful than classical methods based on nearest-neighbor matching, especially as the number of covariates increases. Our theoretical results rely on a generic asymptotic normality theory for a large family of random forest algorithms. To our knowledge, this is the first set of results that allows any type of random forest, including classification and regression forests, to be used for valid statistical inference.

Biopharmaceutical Applications — Bellflower AB

(organized and chaired by Robert Warnock, UCB Biosciences Inc.)

Bhargav Reddy (UCB Biosciences Inc.) 3:30-4:00pm
Predicting Disease State in Crohn’s Patients using Clinical Trial Data

Crohn’s disease impacts approximately 1.4 million people in the US and around 2.2 million people in Europe. Research shows that there are several factors that are causing the disease. With several clinical trials conducted already across the globe, there is still a large amount of research that could be done in understanding the disease and helping patients. In the current research, 3 predictive modeling methodologies have been applied to identify the predictors of the disease. In addition to understanding the predictors, we have predicted the disease state for a patient using several criteria including demographic information, baseline lab parameters, prior disease state, and medical history and life style habits.

Holger Frohlich (UCB Biosciences Inc.) 4:00-4:30pm
Re-Use of Randomized Clinical Trials Data for Predictive Modeling in Epilepsy and Systemic Lupus Erythematosus

Personalized patient treatment is a topic of increasing interest in academic research as well as the pharmaceutical industry. In that context complex machine learning models are built from larger datasets. Here we demonstrate these efforts via two examples: The first example focuses on predicting Briviact® drug effectiveness in epileptic patients. The second example is on predicting placebo response in systemic lupus erythematosus patients at baseline. In both cases models were based on data collected during randomized clinical trials. Potential predictor variables comprised general patient characteristics, current disease state, medical history and co-medications. A rigorous model comparison via repeated, nested cross-validation was followed by application of the best model on independent test data and in-depth investigation of model characteristics. In summary our work highlights the potential, but also the limitations of predictive modeling approaches for personalized patient treatment based on randomized clinical trials data.

Scott Clark (Eli Lilly) 4:30-5:00pm
Discussion and vision of future developments in pharmaceutical research

June 7 (Tuesday)

Parallel Sessions 8:30-10:00am

The Challenges of machine learning methods and computing tools for large-scale data Dogwood AB

(organized by Annie Qu, UIUC; chaired by Yufeng Liu, UNC)

Heping Zhang (Yale University) 8:30-9:00am
Inference with unequal knowledge: nuisance penalized regression, conditional distance correlation, and prior LASSO

Analysis of high dimensional data has received considerable and increasing attention in statistics. Most of the work focuses on the challenges resulting from the high dimensionality; however, in applications, we may know certain aspects of the data and not necessarily be interested in every variable that is observed. Without going into the details, I will present several advances in statistical inference when there exists prior acknowledge about the data. For examples, in genetic studies we may know that certain genes are associated with a disease, and we want to confirm those genes and/or identify additional novel genes. In precision medicine, we may want to evaluate treatment effects while we need to consider other factors. I will first present a nuisance penalized regression framework for efficient inference on the parameters of interest, in the presence of high dimensional nuisance parameters. Then, I will introduce the concept of conditional correlation which enables us to evaluate the correlation between two sets of random variables conditional on the third set. Finally, I will discuss prior LASSO that performs variable selection when a prespecified set of predictors are documented to be associated with the response variable. Whenever possible, I will use numerical examples and real applications to illustrate and further demonstrate the utility of our concept and methodology.

Annie Qu (UIUC) 9:00-9:30am
A Group-Specific Recommender System

In recent years, there has been a growing demand to develop efficient recommender systems which track users' preferences and recommend potential items of interest to users. In this paper, we propose a group-specific method to utilize dependency information from users and items which share similar characteristics under the singular value decomposition framework. The new approach is effective for the "cold-start" problem, where, in the testing set, majority responses are obtained from new users or for new items, and their preference information is not available from the training set. One advantage of the proposed model is that we are able to incorporate information from the missing mechanism and group-specific features through clustering based on the numbers of ratings from each user and other variables associated with missing patterns. In addition, since this type of data involves large-scale customer records, traditional algorithms are not computationally scalable. To implement the proposed method, we propose a new algorithm that embeds a back-fitting algorithm into alternating least squares, which avoids large matrices operation and big memory storage, and therefore makes it feasible to achieve scalable computing. Our simulation studies and MovieLens data analysis both indicate that the proposed group-specific method improves prediction accuracy significantly compared to existing competitive recommender system approaches. This is joint work with Xuan Bi, Junhui Wang and Xiaotong Shen.

Yuan Zhang (University of Michigan) 9:30-10:00am
Estimating network edge probabilities by neighborhood smoothing

The problem of estimating probabilities of network edges from the observed adjacency matrix has important applications to predicting missing links and network denoising. It has usually been addressed by estimating the graphon, a function that determines the matrix of edge probabilities, but is ill-defined without strong assumptions on the network structure. Here we propose a novel computationally efficient method based on neighborhood smoothing to estimate the expectation of the adjacency matrix directly, without making the strong structural assumptions graphon estimation requires. The neighborhood smoothing method requires little tuning, has a competitive mean-squared error rate, and outperforms many benchmark methods on the task of link prediction in both simulated and real networks. This is a joint with Elizaveta Levina and Ji Zhu.

Inference and Estimation in Statistical Machine Learning Azalea AB

(organized and chaired by Han Liu, Princeton)

Adel Javanmard (USC) 8:30-9:00am
Online Rules for Control of False Discovery Rate

Multiple hypothesis testing is a core problem in statistical inference and arises in almost every scientific field. Given a set of null hypotheses, Benjamini and Hochberg introduced the false discovery rate (FDR), which is the expected proportion of false positives among rejected null hypotheses, and proposed a testing procedure that controls FDR below a pre-assigned significance level. In this talk, we consider the problem of controlling FDR in an "online manner". Concretely, we consider an ordered --possibly infinite-- sequence of null hypotheses where at each step the statistician must decide whether to reject current null hypothesis having access only to the previous decisions. We introduce a class of "generalized alpha-investing" procedures and prove that any rule in this class controls FDR in online manner, provided distinct p-values are independent. Next, we propose some modifications to the proposed rules to control FDR in the presence of p-values dependencies.

Zhao Ren (University of Pittsburgh) 9:00-9:30am
Robust Covariance/Scatter Matrix Estimation via Matrix Depth

Covariance matrix estimation is one of the most important problems in statistics. To deal with modern complex data sets, not only do we need estimation procedures to take advantage of the structural assumptions of the covariance matrix, it is also important to design methods that are resistant to arbitrary source of outliers. In this paper, we define a new concept called matrix depth and propose to maximize the empirical matrix depth function to obtain a robust covariance matrix estimator. The proposed estimator is shown to achieve minimax optimal rate under Huber's ε-contamination model for estimating covariance/scatter matrices with various structures such as bandedness and sparsity. Competitive numerical results are presented for both simulated and real data examples. This is a joint work with Mengjie Chen and Chao Gao.

Zhaoran Wang (Princeton University) 9:30-10:00am
Probing the Pareto Frontier of Computational-Statistical Tradeoffs

In this talk, we discuss the fundamental tradeoffs between computational efficiency and statistical accuracy that arise in big data. In particular, we are interested in probing the Pareto frontier of such tradeoffs from the following two aspects. (i) Fundamental limits: Based on an oracle computational model, we introduce a systematic approach for developing minimax lower bounds under computational budget constraints. This approach mirrors the classical Le Cam’s method, and draws explicit connections between algorithmic complexity and geometric structures of parameter space. The resulting computational lower bounds cover most popular algorithms, and do not rely on computational hardness hypotheses. (ii) Efficient algorithms: We focus on developing and analyzing nonconvex statistical optimization algorithms, which exhibit superior computational efficiency and statistical accuracy to their convex counterparts. In particular, to understand the success of the nonconvex approaches and respectively establish provable guarantees, we discuss how to explore the hidden convex structures by incorporating the underlying statistical models.

Network Analysis and Inference tools Mountain Laurel AB

(organized by Kai Zhang, UNC; chaired by Mu Zhu, Univ. Waterloo)

Shankar Bhamidi (UNC Chapel Hill) 8:30-9:00am
Change Point Detection in Evolving Network Models

The last few years have seen an explosion in the amount of data on real world networks, including networks that evolve over time. A number of mathematical models have been proposed to understand the evolution of such networks and explain the emergence of a wide array of structural features such as heavy tailed degree distribution and small world connectivity of real networks. In this paper we consider one famous class of such models, the preferential attachment model. We formulate and study the regime where the network transitions from one evolutionary scheme to another. In the large network limit we derive asymptotics for various functionals of the network including degree distribution and maximal degree. We study functional central limit theorems for the evolution of the degree distribution which feed into proving consistency of a proposed estimator of the change point.

Xi Luo (Brown University) 9:00-9:30am
Network Communities and Variable Clustering: A Covariance Matrix Approach

Clustering is one of the first tools to explore big data. Classical techniques, such as hierarchical clustering and k-means, are usually based on the closeness between two data points or between a data point to a cluster centroid. In this talk, we examine the theoretical and methodological aspects of clustering using a covariance matrix approach. We introduce a new class of models, the G-Models, for partitioning variables into communities with exchangeable behavior, defined on the whole covariance matrix. These models are motivated by three different but inter-related concepts: the exchangeability of variables, block covariance structures, and latent variable covariance matrices. These concepts will lead to the same clustering partitions under certain regularity conditions. We introduce a computationally fast method to recover the partitions. Theoretical analysis shows that our method recovers the partition with high probability and is also minimax optimal. We will illustrate the numerical merits using simulated data and two real datasets: stock returns and functional MRI. This is joint work with Florentina Bunea and Christophe Giraud.

Pingshou Zhong (Michigan State University) 9:30-10:00am
Tests for Covariance Structures with High-dimensional Repeated Measurements

In regression analysis with repeated measurements, such as longitudinal data and panel data, structured covariance matrices characterized by a small number of parameters have been widely used and play an important role in parameter estimation and statistical inference. To assess the adequacy of a specified covariance structure, one often adopts the classical likelihood-ratio test when the dimension of the repeated measurements (p) is smaller than the sample size (n). However, this assessment becomes quite challenging when p is bigger than n, since the classical likelihood-ratio test is no longer applicable. This paper proposes an adjusted goodness-of-fit test to examine a broad range of covariance structures under the scenario of “large p, small n.” Analytical examples are presented to illustrate the effectiveness of the adjustment. In addition, large sample properties of the proposed test are established. Moreover, simulation studies and a real data example are provided to demonstrate the finite sample performance and the practical utility of the test.

Discovery of Features and Patterns Bellflower AB

(organized by Cynthia Rudin MIT; chaired by Yiyuan She, Florida State Univ.)

Genevera Allen (Rice University) 8:30-9:00am
Algorithmic Regularization Paths: A New Approach to Variable Selection for High-Dimensional, Highly Correlated Data

Variable selection has become a cornerstone of statistical machine learning and is ubiquitously used for the analysis of high-dimensional data. While there is a robust literature on providing guarantees for the performance of variable selection techniques, there is one setting for which there are no such guarantees and existing methods perform poorly: that of high-dimensional high-correlation (HDHC) data. Such HDHC data commonly arises from high-throughput biomedical technologies and in graph selection problems for highly connected graphs. In this paper, we develop a radically different type of variable selection method that will prove to be superior in HDHC settings. Our so-called Algorithmic Regularization Paths generate a sequence of sparse models as the iterates of an algorithm inspired by the Alternating Direction Methods of Multiplers (ADMM) algorithm for solving the Lasso. In this paper, we introduce our method, discuss its origin, study its theoretical properties to better understand how it works, and draw connections to existing regularization techniques. Empirical studies show that our Algorithmic Regularization Paths yield better performance in terms of prediction accuracy, variable selection, model selection, and computing time than all existing approaches in HDHC settings. Joint work with Yue Hu and Michael Weylandt.

Lauren Hannah (Columbia University) 9:00:9:30am
Statistically Summarizing Labeled Text Data

Label generating models, such as k-means, mixture models, and topic models, are standard for exploratory analysis of text data. The labels indicate similarities between texts and the contents associated with the labels can often be summarized by other objects generated by the model, such as mean or topic vectors. Most summaries, however, are rooted in the bag-of-words framework used to generate the model. We present a new way to summarize data associated with a label by including both words and statistically significant phrases in a two-step process. Our summaries are purely post-processing and can be used with any label-generating model. First, we use Bayes Factors to generate candidate phrases, then we use a mutual-information based metric to assign label specific but common phrases to each label summary. We apply this method to several corpora, and discuss sensitivities to regularization and tunable parameters.

Shawn Mankad (Cornell University) 9:30-10:00am
Single Stage Prediction with Text Data using Dimension Reduction Techniques

Text data is playing an increasingly important role within the business world for economic analyses, improved marketing, operations management, and financial trading. The most common approach is to follow a two-stage procedure, where one first derives text features through dimension reduction techniques (e.g., topic modeling) and subsequently applies OLS or another statistical model for prediction and inference. In principle there are many ways to perform this two-stage procedure, both in terms of generating text features and properly combining them within a statistical model. This paper addresses this issue by integrating both steps together using a constrained matrix factorization framework that leverages the relationship between term frequency and a given response variable in addition to co-occurrences between terms to recover topics that are both predictive of the outcome of interest and useful for understanding the underlying textual themes. To validate our approach we show that the proposed methodology recovers topics that have improved out of sample prediction accuracy using online hotel reviews from TripAdvisor.

Plenary Talk **10:30-11:30am**
(chaired by Michael Kosorok, UNC) **Dogwood AB**

Susan A. Murphy (University of Michigan)
Assessing Time-Varying Causal Effect Moderation in Intensive Time-Varying Treatment

In this talk we provide a definition for moderated treatment effects in terms of potential outcomes; this definition is particularly suited to settings in which treatment occasions are numerous, individuals are not always available for treatment, and potential moderators might be influenced by past treatment. Methods for estimating moderated effects are developed and compared. The proposed approach is illustrated with data from mobile health interventions.

Parallel Sessions **1:00-2:30pm**

Network and Graphical Models **Dogwood AB**
(organized by Hernando Ombao, UC Irvine; chaired by Yunzhang Zhu, Ohio State)

Ali Shojaie (University of Washington) 1:00-1:30pm
Network Reconstruction from High Dimensional Ordinary Differential Equations

We consider the task of learning a dynamical system from high-dimensional time-course data. For instance, we might wish to estimate a gene regulatory network from gene expression data measured at discrete time points. We model the dynamical system non-parametrically as a system of additive ordinary differential equations. Most existing methods for parameter estimation in ordinary differential equations estimate the derivatives from noisy observations. This has been shown to be challenging and inefficient. We propose a novel approach that does not involve derivative estimation. We show that the proposed method can consistently recover the true network structure even in high dimensions, and we demonstrate empirical improvement over competing approaches.

Lina Lin (University of Washington) 1:30-2:00pm
Estimation of High-dimensional Graphical Models using Regularized Score Matching

Graphical models are widely used to model stochastic dependences among large collections of variables. We introduce a new method of estimating undirected conditional independence graphs based on the score matching loss, introduced by Hyvärinen (2005), and subsequently extended in Hyvärinen (2007). The regularized score matching method we propose applies to settings with continuous observations and allows for computationally efficient treatment of possibly non-Gaussian exponential family models. In the well-explored Gaussian setting, regularized score matching avoids issues of asymmetry that arise when applying the technique of neighborhood selection, and compared to existing methods that directly yield symmetric estimates, the score matching approach has the advantage that the considered loss is quadratic and gives piecewise linear solution paths under L1 regularization. Under suitable irrepresentability conditions, we show that L1-regularized score matching is consistent for graph estimation in sparse high-dimensional settings. Through numerical experiments and an application to RNAseq data, we confirm that regularized score matching achieves state-of-the-art performance in the Gaussian case and provides a valuable tool for computationally efficient estimation in non-Gaussian graphical models.

Shuo Chen (University of Maryland) 2:00-2:30pm
Network induced large covariance matrix estimation

In this paper, we consider to estimate network/community induced large correlation/covariance matrices. The massive biomedical data (e.g. gene expression or neuroimaging data) often include latent networks where features are highly correlated with each other, and thus the covariance matrix in a complex yet organized topology. Although, current large covariance matrix and precision matrix estimation methods using thresholding or shrinkage strategies could provide satisfactory overall covariance/precision matrix estimation, they are limited for automatic network/topology detection and network induced correlation matrix estimation. To fill the gap, we propose a new network induced correlation estimation method (NICE) to simultaneously detect highly correlated networks and estimate the covariance matrix by leveraging an adaptive and graph topology oriented thresholding strategy. The novel thresholding strategy can reduce both false positive and false negative discovery rates by using the whole graph topological information which allows edges to borrow with each other. Moreover, we propose a novel `quality and quantity' objective function to shrink the covariance matrix towards a parsimonious model while retaining most of the information. Simulation study results show that our method outperform the competing thresholding and shrinkage methods. We further illustrate the application of our new method by analysis of a serum mass spectrometry proteomics data set.

Flexible Methods for Genomic data **Azalea AB**

(organized by Yufeng Liu, UNC; chaired by Guan Yu, UNC)

Wei Sun (Fred Hutchinson) 1:00-1:30pm

A Two-Step Approach to Estimate the Skeletons of High-Dimensional Directed Acyclic Graphs

Estimation of the skeleton of a directed acyclic graph (DAG) is of great importance for understanding the underlying DAG and causal effects can be assessed from the skeleton when the DAG is not identifiable. We propose a novel method named PenPC to estimate the skeleton of a high-dimensional DAG by a two-step approach. We first estimate the nonzero entries of a concentration matrix using penalized regression, and then fix the difference between the concentration matrix and the skeleton by evaluating a set of conditional independence hypotheses. For high-dimensional problems where the number of vertices p is in polynomial or exponential scale of sample size n, we study the asymptotic property of PenPC on two types of graphs: traditional random graphs where all the vertices have the same expected number of neighbors, and scale-free graphs where a few vertices may have a large number of neighbors. As illustrated by extensive simulations and applications on gene expression data of cancer patients, PenPC has higher sensitivity and specificity than the state-of-the-art method, the PC-stable algorithm.

Yuying Xie (Michigan State University) 1:30-2:00pm

Joint Estimation of Multiple Dependent Gaussian Graphical Models with Applications to Mouse Genomics

Gaussian graphical models are widely used to represent conditional dependence among random variables. In this paper we propose a novel estimator for data arising from a group of Gaussian graphical models that are themselves dependent. A motivating example is that of modeling gene expression collected on multiple tissues from the same individual: a multivariate outcome that is affected by dependencies at the level of both the tissue and the whole body, and existing 20 methods that assume independence among graphs are not applicable. To estimate multiple de- pendent graphs, we decompose the problem into two graphical layers: the systemic layer, which is the network affecting all outcomes and thereby inducing cross-graph dependency, and the category-specific layer, which represents the graph-specific variation. We propose a graphical EM technique that estimates the two layers jointly, establish the estimation consistency and se- 25 lection sparsistency of the proposed estimator, and confirm by simulation that the EM method is superior to a simple one-step method. Lastly, we apply our graphical EM technique to mouse genomics data and obtain biologically plausible results.

Dongmei Li (University of Rochester) 2:00-2:30pm
An evaluation of statistical methods for RNA-Seq data analysis

RNA-Seq is a high-throughput sequencing technology widely used for gene transcript discovery and quantification under different biological or biomedical conditions. A fundamental research question in many RNA-Seq experiments is the identification of differentially expressed genes among experimental conditions or sample groups. Numerous statistical methods for RNA-Seq differential analysis have been proposed since the emergence of the RNA-Seq assay. To evaluate popular statistical methods used in the open source R and Bioconductor packages, we conducted multiple simulation studies to compare the performance of eight methods used in RNA-Seq data analysis (edgeR, DESeq, DESeq2, baySeq, EBSeq, NOISeq, SAMSeq, Voom) across different scenarios. We measured performance using false discovery rate control, power, and stability. Real RNA-Seq experimental data were also used to compare the apparent test power and stability of each method.

Computational Methods in Statistics Mountain Laurel AB

(organized and chaired by Sahand Neghaban, Yale University)

Constantine Caramanis (UT Austin) 1:00-1:30pm
High-dimensional EM algorithm

The popular EM algorithm and its variants, is a much used algorithmic tool; yet our rigorous understanding of its performance is highly incomplete. Recent work has demonstrated that for an important class of problems, EM exhibits linear local convergence. In the high-dimensional setting, however, the M-step may not be well defined. We address precisely this setting through a unified treatment using regularization. The iterative EM algorithm requires a careful balancing of making progress towards the solution while identifying the right structure (e.g., sparsity or low-rank). Our algorithm and analysis are linked in a way that reveals the balance between optimization and statistical errors. We specialize our general framework to sparse gaussian mixture models, high-dimensional mixed regression, and regression with missing variables, obtaining statistical guarantees for each of these examples. Joint work with Xinyang Yi.

Amin Karbas (Yale University) 1:30-2:00pm
Data as a Computational Resource: The Power of Data Summarization

Faced with massive data, is it possible to trade off statistical risk and computational time? This challenge lies at the heart of large-scale machine learning. I will show in this talk that we can indeed achieve such risk-time tradeoffs by strategically summarizing the data, in the unsupervised learning problem of probabilistic k-means, i.e. vector quantization. In particular, there exist levels of summarization for which as the data size increases, the running time decreases, while a given risk is maintained. Furthermore, there exists a constructive algorithm that provably finds such tradeoff levels. The summarization in question is based on coreset constructions from computational geometry. I will also show that these tradeoffs exist and may be harnessed for a wide range of real data. This adds data summarization to the list of methods, including stochastic optimization, that allow us to perceive data as a resource rather than an impediment.

Garvesh Raskutti (University of Wisconsin Madison) 2:00-2:30pm
High-dimensional Poisson auto-regressive models for dynamic network modeling

Vector autoregressive models are of particular interest when a researcher is observing counts of actions taken by nodes in a network and counts at one time point can help predict counts at future times. Such data are common in spike train observations of biological neural networks, interactions within a social network, and pricing changes within financial networks. This paper addresses the inference of the network structure and autoregressive parameters from such data. A sparsity-regularized maximum likelihood estimator is proposed for a Poisson autoregressive process. While sparsity-regularization is well-studied in the statistics and machine learning communities, common assumptions from that literature are difficult to verify because of the correlations and heteroscedasticity within the observations. Novel performance guarantees characterize how much data must be collected to ensure reliable inference depending on the size and sparsity of the autoregressive parameters, and these bounds are supported by several simulation studies.

New developments for analyzing complex data — Bellflower AB

(organized and chaired by Xingye Qiao, SUNY Binghamton)

Xi Chen (NYU) 1:00-1:30pm
Optimal Stopping and Worker Selection in Crowdsourcing: an Adaptive Sequential Probability Ratio Test Framework

In this talk, we propose an adaptive sequential probability ratio test (Ada-SPRT) that obtains the optimal experiment selection rule, stopping time, and final decision rule under a single Bayesian decision framework. Our motivating application comes from binary labeling tasks in crowdsourcing, where the requestor needs to simultaneously decide which worker to choose to provide the label and when to stop collecting labels to save for budget. We characterize the structure of the optimal adaptive sequential design that minimizes the Bayes risk through log-likelihood ratio statistic and develop dynamic programming based algorithms for both non-truncated and truncated tests. We further propose to adopt empirical Bayes approach for estimating class priors and an EM algorithm for estimating workers' quality. This is a joint work with Xiaoou Li, Yunxiao Chen, Jingchen Liu and Zhiliang Ying.

Jacob Bien (Cornell University) 1:30-2:00pm
Lag Structured Modeling for High Dimensional Vector Autoregression

Vector autoregression (VAR) is a fundamental tool for modeling the joint dynamics of multivariate time series. However, as the number of component series is increased, the VAR model quickly becomes over-parameterized, making reliable estimation difficult in high dimensional settings. A common assumption in time series is that the dynamic dependence among components is short-range, leading to the common practice of lag order selection. We propose a new class of regularized VAR models that embeds the notion of lag selection into a convex regularizer. The key convex modeling tool is a group lasso with hierarchically nested groups that guarantees that the sparsity pattern of autoregressive lag coefficients honors the ordered structure inherent to VAR. We provide computationally efficient algorithms for solving this problem and demonstrate improved performance in forecasting and lag order selection over previous approaches. A macroeconomic application highlights the convenient, interpretable output of our method. This is joint work with David Matteson and William Nicholson.

Ganggang Xu (Binghamton University) 2:00-2:30pm
A simple averaged post-model-selection confidence interval

The naive post-model-selection confidence interval suffers from under-coverage since it does not take into account the model selection uncertainty. To improve the coverage probability, we propose to locate the center of the confidence interval at a weighted average of two stimators and simultaneously adjusting the length of the interval stochastically. The proposed method is conceptually simple yet effective and is applicable for any reasonably good model selection tools.

Parallel Sessions 3:00-4:30pm

Causal Inference Dogwood AB

(organized and chaired by Eric Laber, NCSU)

Tyler McCormick (University of Washington) 3:00-3:30pm
Standard errors for exchangeable relational arrays

Social networks represent relationships between pairs of interconnected individuals and are widely used to understand complex social phenomena. In this work, we consider inference on regression coefficients in network regression models, where the presence/absence (or strength) of a connection between two individuals is modeled as a linear function of observable covariates and structured, network dependent, error. We leverage a joint exchangeability assumption, nearly ubiquitous in the statistics literature on networks but not previously considered in the estimating equations formulation for network regressions, to derive parsimonious estimators of the covariance within the network. We demonstrate our proposed estimators using simulation and multiple observed network datasets. This is joint work with Bailey Fosdick and Frank Marrs at Colorado State University.

Long Nguyen (University of Michigan) 3:30-4:00pm
Bayesian Nonparametric Multilevel Clustering with Group-Level Contexts

We present a Bayesian nonparametric framework for multilevel clustering which utilizes group-level context information to simultaneously discover low-dimensional structures of the group contents and partitions groups into clusters. Using the Dirichlet process as the building block, our model constructs a product base-measure with a nested structure to accommodate content and context observations at multiple levels. The proposed model possesses properties that link the nested Dirichlet processes (nDP) and the Dirichlet process mixture models (DPM) in an interesting way: integrating out all contents results in the DPM over contexts, whereas integrating out group-specific contexts results in the nDP mixture over content variables. We provide a Polya-urn view of the model and an efficient collapsed Gibbs inference procedure. Extensive experiments on real-world datasets demonstrate the advantage of utilizing context information via our model in both text and image domains. This is joint work with Vu Nguyen, Dinh Phung, Svetha Venkatesh and Hung Bui.

Cynthia Rudin (MIT) 4:00-4:30pm
Causal Falling Rule Lists

A Causal Falling Rule List is a sequence of IF-THEN rules that specifies heterogeneous treatment effects. In this model, (a) the order of rules determines the treatment effect subgroup that a subject belongs to, (b) the treatment effect decreases monotonically down the list. For example, a Causal Falling Rule List might say that if a person is below 60 years, then they are in the highest treatment effect subgroup, such that administering a drug will result in a 20 unit increase in good cholesterol levels. Otherwise, if they are regular exercisers, then taking the drug will result in a 15 unit increase in cholesterol level. Finally, if they satisfy neither of these rules, they are in the default treatment subgroup, such that the drug will result in only a 2 unit increase. The collection of rules, their sequence, and the treatment effects are learned from data. This is joint work with Fulton Wang at MIT.

Machine Learning for Structured Data Azalea AB

(organized by Xiaotong Shen, Univ. Minnesota; chaired by Shu Lu, UNC)

Shuheng Zhou (University of Michigan) 3:00-3:30pm
High dimensional statistical modeling and estimation with matrix variate data

In this talk, I will discuss several new methods for the modeling and estimation of graphical structures and underlying population parameters from data with two-way dependence. Under sparsity conditions, I will show that one is able to recover the low-rank mean matrix, as well as graphs and covariance matrices with a single random matrix from the matrix variate distributions. Time allowing, I will discuss an errors-in-variables model where the covariates in the data matrix are contaminated with random noise. This model is significantly different from those analyzed in the literature in the sense that we allow the measurement error for each covariate to be dependent across its observations. Such error structures appear in the science literature, for example, when modeling the trial-to-trial fluctuations in response strength shared across a set of neurons. We provide real-data examples and simulation evidence showing that we can recover the mean matrix, graphical structures as well as estimating the precision matrices and the regression coefficients for these two classes of problems. This talk is based on joint work with Michael Hornstein, Mark Rudelson, and Kerby Shedden.

Xingye Qiao (Binghamton University) 3:30-4:00pm
Noncrossing Ordinal Classification

Ordinal data are often seen in real applications. Regular multicategory classification methods are not designed for this data type and a more proper treatment is needed. We consider a framework of ordinal classification which pools the results from binary classifiers together. An inherent difficulty of this framework is that the class prediction can be ambiguous due to boundary crossing. To fix this issue, we propose a noncrossing ordinal classification method which materializes the framework by imposing noncrossing constraints. An asymptotic study of the proposed method is conducted. We show by simulated and data examples that the proposed method can improve the classification performance for ordinal data without the ambiguity caused by boundary crossings.

Yunzhang Zhu (Ohio State University) 4:00-4:30pm
High-dimensional Multivariate Regression

In this talk, I present two convex optimization formulations for high-dimensional multivariate linear regression under general error covariance structure. The main difficulty for simultaneous estimation of the regression coefficients and the error covariance lies in that the negative log-likelihood function is not jointly convex. To overcome this difficulty, two new parameterizations are proposed, under which the negative log-likelihood function is convex. It will be demonstrated that one parameterization is particularly useful for the task of estimating the regression coefficient matrix; and the other parameterization is more useful for covariate-adjusted graphical modeling. The proposed methods compare favorably to existing high-dimensional multivariate linear regression methodologies that are based either on minimizing non-convex criteria or certain two-step procedures. Finally I present some theoretical properties and applications to multi-task learning and gene network analysis.

Inference for regularized estimation in high dimensions — Mountain Laurel AB

(organized and chaired by Ali Shojaie, Univ. Washington)

Max G'Sell (CMU) 3:00-3:30pm
Model selection via sequential goodness-of-fit testing

We consider goodness-of-fit testing for sequential model selection procedures. This leads to a multiple hypothesis testing setting where the hypotheses are ordered and one must reject an initial contiguous block, H_1,..., H_{\hat{k}}, of hypotheses. A rejection rule in this setting amounts to a procedure for choosing the stopping time \hat{k}. We will discuss a multiple hypothesis testing procedure for FDR control in this setting. We will also introduce recent results for goodness-of-fit testing for the graphical lasso as an illustration of this approach.

Mladen Kolar (University of Chicago) 3:30-4:00pm
Post-Regularization Confidence Bands for High-Dimensional Nonparametric Models with Local Sparsity

We propose a novel high dimensional nonparametric model named ATLAS which is a generalization of the sparse additive model. The ATLAS model assumes the high dimensional regression function can be locally approximated by a sparse additive function, while such an approximation may change from the global perspective. We aim to estimate high dimensional function using a novel kernel-sieve hybrid regression estimator that combines the local kernel regression with the B-spline basis approximation. We show the estimation rate of true function in the supremum norm. We also propose two types of confidence bands for true function. Both procedures proceed in two steps: (1) a novel bias correction method is applied to remove the shrinkage introduced by the model selection penalty and (2) quantiles of the normalized de-biased estimator are approximated by quantiles of the limiting distribution or a Gaussian multiplier bootstrap. We further show that the covering probability of the bootstrap confidence bands converges to the nominal one at a polynomial rate. Joint work with Junwei Lu and Han Liu.

Sen Zhao (University of Washington) 4:00-4:30pm
High-Dimensional Hypothesis Testing with the Lasso

We consider the problem of hypothesis testing in a high-dimensional linear model using the lasso. We show that, under some standard assumptions, the set of variables selected by the lasso is almost surely fixed, and contains all of the variables that have non-zero regression coefficients. These theoretical results are applied in order to justify restricting our attention to the set of features selected by the lasso. We then apply classical Wald and score tests on the reduced data set. Because the lasso-selected set is almost surely fixed, distribution truncation is not required in order to obtain asymptotically valid inference on the population regression coefficients; this is in contrast to the recently-proposed exact post-selection testing framework. We also establish connections between our proposals and the debiased lasso tests and investigate their differences. Finally, we perform extensive numerical studies in support of our methods.

New learning tools for complex data and beyond Bellflower AB

(organized by Yufeng Liu, UNC; chaired by David Pritchard UNC)

J. Paul Brooks (Virginia Commonwealth University) 3:00-3:30pm
Estimating L1-Norm Best-Fit Lines

Recently, the L1-norm best-fit line problem has been shown to be NP-hard, but it has a long tradition in the context of statistical learning and location theory. We present properties of L1-norm projection of points onto lines. A highly-effective heuristic in general dimensions results from adapting an algorithm for the best-fit line in two dimensions. The adaptation is a solution to the best-fit line problem under certain conditions. We demonstrate that our method is a powerful tool for generating estimates of the L1-norm best-fit line, and that the results are robust to outliers. The results can be used for robust principal component analysis.

Chengyong Tang (Temple University) 3:30-4:00pm
Precision Matrix Estimation by Inverse Principal Orthogonal Decomposition

We consider a parsimonious approach for modeling a large precision matrix in a factor model setting. The approach is developed by inverting a principal orthogonal decomposition (IPOD) that disentangles the systematic component from the idiosyncratic component in the target dynamic system of interest. In the IPOD approach, the impact due to the systematic component is captured by a low-dimensional factor model. Motivated by practical considerations for parsimonious and interpretable methods, we propose to use a sparse precision matrix to capture the contribution from the idiosyncratic component to the variation in the target dynamic system. Conditioning on the factors, the IPOD approach has an appealing practical interpretation in the conventional graphical models for informatively investigating the associations between the idiosyncratic components. We discover that the large precision matrix depends on the idiosyncratic component only through its sparse precision matrix, and show that IPOD is convenient and feasible for estimating the large precision matrix in which only inverting a low-dimensional matrix is involved. We formally establish the estimation error bounds of the IPOD approach under various losses and find that the impact due to the common factors vanishes as the dimensionality of the precision matrix diverges. Extensive numerical examples including real data examples in practical problems demonstrate the merits of the IPOD approach in its performance and interpretability. This is a joint work with Yingying Fan.

Mu Zhu (University of Waterloo) 4:00-4:30pm
Networks, Small G Proteins, and Basketball Games

A protein molecule is made up of a sequence of amino acid residues. An important task toward understanding the 3D structures of a protein molecule is the quantification of direct-coupling strengths between any two residues. This type of analysis often results in a residue-residue network, in which each node corresponds to a residue, and an edge is drawn between two nodes if their direct-coupling strength is high. Such networks are also important for the study of allostery, the process by which conformational changes at one residue site affect the conformation of another. In the first part of my talk, I will describe my joint work with Laleh Soltan-Ghoraie and Forbes Burkowski, in which we used kernelized partial canonical correlation analysis to perform this task on small G proteins. Network data analysis is an emerging area of statistics. Often, each node in a network can belong to one of many different communities. Many researchers have used a so-called stochastic block model (SBM) to detect these communities. Early works often do not account for time and treat the network as if it were static. Recently, researchers have started to consider network dynamics over time, but they often treat the time variable in a discrete manner. We have been studying a continuous -time extension of the SBM. In the second part of the talk, I will briefly describe my joint work with Lu Xin and Hugh Chipman, in which we used our continuous-time SBM to analyze basketball games played in the NBA.

June 8 (Wednesday)

Parallel Sessions 8:30-10:00am

Machine learning for precision medicine Dogwood AB

(organized by Yufeng Liu, UNC; chaired by Dacheng Liu, Boehringer Ingelheim)

Donglin Zeng (UNC Chapel Hill) 8:30-9:00am
Estimating Personalized Diagnostic Rules via Weighted Support Vector Machines

There is an increasing demand for personalization of disease screening based on assessment of patient risk and other characteristics. For example, in breast cancer screening, advanced imaging technologies have made it possible to move away from "one-size-fits-all" screening guidelines to targeted risk-based screening for those who are in need. Since diagnostic performance of various imaging modalities may vary across subjects, applying the most accurate modality to the patients who would benefit the most requires personalized strategy. To address these needs, we propose novel machine learning methods to estimate personalized diagnostic rules for medical screening or diagnosis by maximizing a weighted combination of sensitivity and specificity across subgroups of subjects. We first develop methods that can be applied to estimate personalized diagnostic rules where competing modalities or screening strategies are observed on the same subject (paired design). Next, we present methods for studies where not all subjects receive both modalities (unpaired design). We study theoretical properties including consistency and risk bound of the personalized diagnostic rules, and conduct simulation studies to demonstrate effectiveness of the proposed methods. Lastly, we analyze data collected from a brain imaging study of Parkinson's disease using positron emission tomography (PET) or diffusion tensor imaging (DTI) modalities with paired and unpaired designs.

Yingqi Zhao (Fred Hutchinson) 9:00-9:30am
Develop Parsimonious and Robust Treatment Strategies for Target Populations

Individualized treatment rules based on individual patient characteristics are becoming important in clinical practice. Properly planned and conducted randomized clinical trials are ideal for constructing individualized treatment rules. However, it is often a concern that trial subjects lack representativeness, which limits the applicability of the derived rules to a future large population. Furthermore, to inform clinical practice, it is crucial to provide rules that are easy to interpret and disseminate. To tackle these issues, we use data from a single trial study to propose a two-stage procedure to derive a robust and parsimonious rule to maximize the proportion of future patients receiving their optimal treatments. The procedure allows a wide range of possible covariate distributions in the target population, with minimal assumptions on the mean and covariance of the patients who benefit from each treatment. The practical utility and favorable performance of the methodology are demonstrated using extensive simulations and a real data application.

Haoda Fu (Eli Lily) 9:30-10:00am
Personalized Medicine, Machine Learning and Artificial Intelligence: Challenges and Opportunities

With new treatments and novel technology available, personalized medicine has become an important piece in the new era of medical product development. Most recently, deep reinforcement learning has achieved great success in the field of artificial intelligence. In this talk, we will share some perspectives on how these methods can potentially help to choose right treatments, selecting the right dose, and making optimal treatment transitions. Besides sharing some of solutions, we will also share of some challenges to invite collaborations.

New developments on sufficient dimension reduction and envelope estimation

Azalea AB

(organized and chaired by Yichao Wu, NCSU)

Andreas Artemiou (Cardiff University) 8:30-9:00am
Robustifying sufficient dimension reduction against inliers and outliers

Sufficient dimension reduction has been extensively used to reduce the dimension reduction in regression problems. The main objective is to estimate the Central Subspace without losing information for the conditional distribution of Y|X. The use of machine learning techniques in the sufficient dimension reduction framework was introduced in Li, Artemiou and Li (2011). These methods though in a regression setting create artificial inliers and outliers with heavy weight in estimating the Central subspace. New methodology to alleviate this problem is investigated.

Zhihua Su (University of Florida) 9:00-9:30am
Groupwise Envelope Models for Imaging Genetic Analysis

The aim of this paper is to study the association between brain volumes and candidate genes of the Alzheimer Disease. We develop a groupwise envelope model that allows for distinct regression coefficients and error structures for different groups. The envelope model introduced by Cook et al. (2010) is a new paradigm that improves estimation efficiency in multivariate linear regression. Theoretical properties of the proposed model are established. Numerical experiments as well as analysis on the Alzheimer dataset show the effectiveness of the model in efficient estimation.

Xin Zhang (Florida State University) 9:30-10:00am
Some Recent Advances in Envelope Methodology

Envelope methodology is essentially a form of targeted dimension reduction designed to increase efficiency in multivariate statistics. In some applications that increase can be equivalent to taking thousands of additional observations. In this talk we will discuss recent advances in envelope methodology, including improved algorithms and applications in neuroimaging data analysis.

New Sparse Methods for Regression and Classification

Mountain Laurel AB

(organized by Hao Helen Zhang; University of Arizona; chaired by Zhigen Zhao, Temple Univ.)

Ning Hao (University of Arizona) 8:30-9:00am
A rotate-and-solve procedure for high dimensional classification

Many high dimensional classification techniques have been proposed in the literature based on sparse linear discriminant analysis. To efficiently use them, sparsity of linear classifiers is a prerequisite. However, this might not be readily available in many applications, and rotations of data are required to create the needed sparsity. In this talk, we consider a family of rotations to create the required sparsity. The basic idea is to use the principal components of the sample covariance matrix of the pooled samples and its variants to rotate the data first and to then apply an existing high dimensional classifier. This rotate-and-solve procedure can be combined with any existing classifiers, and is robust against the sparsity level of the true model. The effectiveness of the proposed method is demonstrated by a number of simulated and real data examples.

Gen Li (Columbia University) 9:00-9:30am
Supervised Integrative Principal Component Analysis

It becomes increasingly common to have data from multiple sources for the same set of subjects in modern health science research. Integrative dimension reduction of multi-source data has the potential to leverage information in heterogeneous sources, and identify dominant patterns in data that facilitates interpretation and subsequent analyses. However, such methods are not well studied, and in particular, no existing method accounts for supervision effects from auxiliary covariates. In this talk, I will introduce a novel statistical framework for integrative dimension reduction of multi-source data with covariates adjustment. The method decomposes the total variation of multi-source data into the joint variation across sources and individual variation specific to each source. The framework is formulated as a hierarchical latent variable model where each latent variable easily incorporates covariates adjustment through a linear model or nonparametric models. We show that the model subsumes many recently developed dimension reduction methods as special cases. A computationally efficient algorithm is devised to fit the model. We apply the methods to two pediatric growth studies, where we discover interesting growth patterns and identify important covariates associated with growth.

Sijian Wang (University of Wisconsin Madison) 9:30-10:00am
Sparse additive index model for group variable selection

In this talk, we propose a regularized multiple-index model to integrate group structures to the model. It can not only identify important groups, but also select important predictors within selected groups. Furthermore, the proposed method has three good properties: 1) It is flexible to model the nonlinear association 2) It automatically considers the interactions among variables within the same group; 3) When the groups have overlaps, it may distinguish the effects of a predictor in all of groups it belongs to. We have studied the theoretical properties of the methods. The methods are demonstrated using simulation studies and analysis on a TCGA dataset.

Plenary Talk **10:30-11:30am**
(chaired by Eric Laber, NCSU) **Dogwood AB**

Michael R. Kosorok (UNC Chapel Hill)
The Evolution of Data Science and Statistics

The field of data science has been evolving in one way or another for over a hundred years and incorporates a number of disciplines including statistics, computer science, mathematics, library science, engineering, as well as many other domains. The current ascension of data science seems to have initiated around the beginning of the 21st century, with the term being attributed to a 2001 publication by statistician William Cleveland in the International Statistical Review. The field of statistics has been and continues to play an important role in this domain because of the inferential perspective it provides and the necessity of this perspective—in both designing studies which generate data and in drawing conclusions about data through analytics—to ensure suitable reproducibility and generalizability of results. Many other disciplines also bring tremendous value to the data science endeavor. For all of the disciplines involved, significant changes at the core of our various perspectives are happening. Moreover, data science is becoming a new, convergent discipline which is greater than the sum of its parts. These dramatic changes are both challenges and opportunities. As statisticians, computer scientists, mathematicians, and other contributing researchers, we will need to learn to better communicate with, learn about, work with, and respect one another. We will all need to think much more broadly and creatively than ever before. In this presentation, I will discuss several specific relevant research and educational endeavors currently underway, some potential future opportunities, and several perspectives about the path forward.

Parallel Sessions **1:00-2:30pm**

Data Integration and Meta Analysis **Dogwood AB**

(organized and chaired by Quefeng Li, UNC)

Haitao Chu (University of Minnesota) 1:00-1:30pm

Bayesian Hierarchical Models for Multiple Diagnostic Tests Meta-analysis

In studies evaluating the accuracy of diagnostic tests, three designs are commonly used: (1) the multiple tests design; (2) the randomized design; and (3) the non-comparative design. Existing methods on meta-analysis of diagnostic tests mainly considered the simple case when the reference test in all or none of studies can be considered as a gold standard test, and when all studies use either the randomized or the non-comparative design. Yet the proliferation of diagnostic instruments and diversity of study designs being used have boosted the demand to develop more general methods to combine studies with or without a gold standard test using different designs. In this talk, we discuss two related frameworks from the missing data perspective for meta-analysis simultaneously comparing multiple diagnostic tests. It accounts for the potential correlation among multiple diagnostic tests within a study and heterogeneity across studies. Our model is evaluated through simulations and illustrated using a real data analysis.

Shuangge Ma (Yale University) 1:30-2:00pm

Integrating multidimensional omics data for cancer prognosis

Prognosis is of essential interest in cancer research. Multiple types of omics measurements – including mRNA gene expression, methylation, copy number variation, SNP, and others – have been implicated in cancer prognosis. The analysis of multidimensional omics data is challenging because of the high data dimensionality and, more importantly, because of the interconnections between different units of the same type of measurement and between different types of omics measurements. In our study, we have developed novel regularization-based methods, effectively integrated multidimensional data, and constructed prognosis models. It is shown that integrating multidimensional data can lead to biological discoveries missed by the analysis of one-dimensional data and superior prognosis models.

Chi Song (Ohio State University) 2:00-2:30pm

A Bayesian Method for Transcriptomic Meta-analysis – Exploring the Homogeneity and Heterogeneity

Due to rapid development of high-throughput experimental techniques and fast dropping prices, many transcriptomic datasets have been generated and accumulated in the public domain. Meta-analysis combining multiple transcriptomic studies can increase statistical power to detect disease related biomarkers. In this presentation, I will introduce a Bayesian latent hierarchical model based on one-sided p-values from differential/association analysis to perform transcriptomic meta-analysis. The p-value based method is capable of combining data from different microarray and RNA-seq platforms, and the latent variables help quantify homogeneous and heterogeneous differential expression signals across studies. A tight clustering algorithm is applied to detected biomarkers to capture differential meta-patterns that are informative to guide further biological investigation. Simulations and two examples using a microarray dataset from metabolism related knockout mice and an RNA-seq dataset from HIV transgenic rats are used to demonstrate performance of the proposed method.

Flexible Learning Tools and Applications Azalea AB

(organized and chaired by Yuying Xie, MSU)

Wei Sun (Yahoo) 1:00-1:30pm

Provable Sparse Tensor Decomposition and Its Application to Personalized Recommendation

Tensor as a multi-dimensional generalization of matrix has received increasing attention in industry due to its success in personalized recommendation systems. Traditional recommendation systems are mainly based on the user-item matrix, whose entry denotes each user's preference for a particular item. To incorporate additional information into the analysis, such as the temporal behavior of users, we encounter a user-item-time tensor. Existing tensor decomposition methods for personalized recommendation are mostly established in the non-sparse regime where the decomposition components include all features. For high dimensional tensor-valued data, many features in the components essentially contain no information about the tensor structure, and thus there is a great need for a more appropriate method that can simultaneously perform tensor decomposition and select informative features. In this talk, I will discuss a new sparse tensor decomposition method that incorporates the sparsity of each decomposition component to the CP tensor decomposition. Specifically, the sparsity is achieved via an efficient truncation procedure to directly solve an L0 sparsity constraint. In theory, in spite of the non-convexity of the optimization problem, it is proven that an alternating updating algorithm attains an estimator whose rate of convergence significantly improves those shown in non-sparse decomposition methods. As a by-product, our method is also widely applicable to solve a broad family of high dimensional latent variable models, including high dimensional Gaussian mixtures and mixtures of sparse regression. I will show the advantages of our method in two real applications, click-through rate prediction for online advertising and high dimensional gene clustering.

Peng Wang (University of Cincinnati) 1:30-2:00pm

Selection by Partitioning the Solution Paths

The performances of penalized likelihood approaches profoundly depend on the selection of the tuning parameter; however there has not been a common agreement on the criterion for choosing the tuning parameter. Moreover, penalized likelihood estimation based on a single value of the tuning parameter would suffer from several drawbacks. This article introduces a novel approach for feature selection based on the whole solution paths rather than choosing one single tuning parameter, which significantly improves the selection accuracy. Moreover, it allows for feature selection using ridge or other strictly convex penalties. The key idea is to classify the variables as relevant or irrelevant at each tuning parameter and then select all the variables which have been classified as relevant at least once. We establish the theoretical properties of the method, which requires significantly weaker conditions compared to existing literature. We also illustrate the advantages of the proposed approach with simulation studies and a data example.

Tian Zheng (Columbia University) 2:00-2:30pm

Topic-adjusted visibility metric for scientific articles

Measuring the impact of scientific articles is important for evaluating the research output of individual scientists, academic institutions and journals. While citations are raw data for constructing impact measures, there exist biases and potential issues if factors affecting citation patterns are not properly accounted for. In this talk, I present a new model that aims to address the problem of field variation and introduce an article level metric useful for evaluating individual articles' topic-adjusted visibility. This measure derives from joint probabilistic modeling of the content in the articles and the citations amongst them using latent Dirichlet allocation (LDA) and the mixed membership stochastic blockmodel (MMSB). This proposed model provides a visibility metric for individual articles adjusted for field variation in citation rates, a structural understanding of citation behavior in different fields, and article recommendations which take into account article visibility and citation patterns. For this work, we also developed an efficient algorithm for model fitting using variational methods. To scale up to large networks, we developed an online variant using stochastic gradient methods and case-control likelihood approximation. Results from an application of our methods to the benchmark KDD Cup 2003 dataset with approximately 30,000 high energy physics papers will also be presented.

New Sparse Learning Techniques **Mountain Laurel AB**
(organized by Yin Xia, UNC; chaired by Jingxiang Chen UNC)

Botao Hao (Purdue University) 1:00-1:30pm
Simultaneous Clustering and Estimation of Multiple Graphical Models

We consider high dimensional estimation of multiple graphical models arising from heterogeneous observations. An appealing feature of our methodology is to learn clustering structure while estimating graphical models. This is achieved via a high dimensional EM algorithm. Meanwhile, a joint graphical lasso penalty is imposed to extract a common structure shared across all clusters. In theory, a non-asymptotic estimation error bound is derived for understanding the trade-off between statistical accuracy and computational complexity in the regularized Gaussian mixture models.

Aaron Molstad (University of Minnesota) 1:30-2:00pm
Indirect multivariate response linear regression

We propose a new class of estimators of the multivariate response linear regression coefficient matrix that exploits the assumption that the response and predictors have a joint multivariate Normal distribution. This allows us to indirectly estimate the regression coefficient matrix through shrinkage estimation of the parameters of the inverse regression, or the conditional distribution of the predictors given the responses. We establish a convergence rate bound for estimators in our class and we study two examples. The first example estimator exploits an assumption that the inverse regression's coefficient matrix is sparse. The second example estimator exploits an assumption that the inverse regression's coefficient matrix is rank deficient. These estimators do not require the popular assumption that the forward regression coefficient matrix is sparse or has small Frobenius norm. Using simulation studies, we show that our example estimators outperform relevant competitors for some data generating models.

Xue Wang (Penn. State University) 2:00-2:30pm
Folded Concave Penalized Nonconvex Learning via a Modern Optimization Lens

We consider non-convex learning with folded-concave penalty, an important statistical problem aimed for high-dimensional data analysis. Due to the presence of a folded concave penalty, the problem formulation involves both non-smoothness and non-convexity. We show that local solutions satisfying a second order necessary condition (SONC) entail good estimation accuracy with a probability guarantee. We discuss (1) a novel potential reduction method (PR) and (2) a new mixed integer linear programming-based global optimization (MIPGO) scheme. This talk is based on joint work with Hongcheng Liu, Runze Li and Yinyu Ye.

Parallel Sessions **3:00-4:30pm**

Data Science and Networks: Methodology and Applications **Dogwood AB**
(organized and chaired by Shankar Bhamidi, UNC)

Bailey Fosdick (Colorado State University) 3:00-3:30pm
Multiresolution models for networks

Many social networks exhibit global sparsity and local density. That is, overall the propensity for interaction between any two randomly selected actors is infinitesimal, but for any given individual there is massive variability in the propensity to interact with others in the network. In this talk, we propose a class of multiresolution statistical network models that account for such variability by mixing models that represent structure in different levels of the graph. As a byproduct of this approach, we are able to characterize and compare structure within the network at various scales. We discuss properties of these models and demonstrate their utility using simulation and data on social network interactions from Karnataka, India.

Anand Vidhyashankar (George Mason University) 3:30-4:00pm
Implicit Networks in High Dimensional Problems

In a variety of contemporary applications, especially those involving big-data, it is becoming a common practice to use high-dimensional regression models for data analysis. While such methods yield important information concerning associations between a response and a set of features, they fail to capture the global characteristics of the feature set. To address some of these limitations, we introduce the concept of supervised implicit networks and investigate the theoretical properties of various network wide metrics (NWM). Specifically, we provide an assessment of variability in the statistical estimates of NWM and discuss their use in the context of data analysis. Finally, we apply these methods to develop supervised clustering algorithms and use it to cluster genotypes associated with a phenotype. We illustrate our methods using a recent data set in the study of bipolar disorder. This is joint work with my doctoral student Brandon Park.

James Wilson (University of San Francisco) 4:00-4:30pm
A Significance-based Community Extraction Method for Multilayer Networks

Multilayer networks are a useful way to capture and model multiple, binary relationships among a fixed group of objects. While community detection has proven to be a useful exploratory technique for the analysis of single-layer networks, little work has been done in the development of methods suitable for multilayer networks. In this talk I will introduce an extraction procedure, called Multilayer Extraction, that identifies densely connected vertex-layer sets in multilayer networks. Multilayer Extraction makes use of a significance-based score that quantifies the connectivity of an observed vertex-layer by comparison with a fixed-degree random graph model. I will investigate the use of Multilayer Extraction on two real networks including a European Airlines transportation network as well as a collaboration network. I will also describe large graph consistency properties of Multilayer Extraction under the multilayer stochastic block model. This is joint work with Andrew B. Nobel, and Shankar Bhamidi.

Industrial Applications Azalea AB

(organized by Cynthia Rudin, MIT; chaired by J. Paul Brooks, Virginia Commonwealth Univ.)

Veena Mendiratta (Bell Labs, Alcatel-Lucent) 3:00-3:30pm
Anomaly Detection in Wireless Networks using Mobile Phone Data

Communications traffic on wireless networks generates large volumes of metadata on a continuous basis across the various servers involved in the communication session. The data fields include event type, event duration and error codes in the form of logs. Since these networks are engineered for high reliability, the data is predominantly normal with only a small proportion of the data being anomalous. It is, however, important to detect these anomalies when they occur because such anomalies are indicators of vulnerabilities in the network. In this work we will present the use of neural network based Kohonen Self Organizing Maps (SOM) and visual analytics for network anomaly detection and analysis using data from a 4G wireless network.

Matthew Lanham (Virginia Tech) 3:30-4:00pm
A Framework for Combining Statistical & Business KPIs for Low-Turn Product Forecasts

This study provides a framework for retailer's using binary classification techniques to model low-turn products to gauge the propensity that a product will sell within a certain time horizon. We show how a firm performing predictive modeling should consider the statistical performance measures in conjunction to the business performance measures that the predictive model is designed to support. Model assessment statistics (e.g. AUC, precision, recall) are important metrics that gauge how well a model will predict future observations, but we have discovered that using them in isolation is insufficient when deciding which model performs optimally with regard to the business. Modeling the propensity that a product will sell in a particular store using several binary classification techniques, we capture their traditional assessment statistics and point out which model would likely be chosen. We then show that a better solution would be to build a decision model that selects the best forecasts using both traditional assessment statistics and business performance in collaboration. Finally, we provide results from a large scale simulation showing the performance among the various predictive models and our solution with regard to generating a retailer's assortment stocking decision.

Roshanak Nateghi (Purdue University) 4:00-4:30pm
Predicting Observed GRACE Satellite Groundwater Storage Trends Using Data in 81 Countries

Groundwater is an essential component of global access to fresh water. Groundwater availability has an impact on political stability, health, economic growth and well-being. Identifying trends in groundwater storage using GRACE satellite observations is a promising new line of research. We have conducted extensive data analytics to identify the best predictors of the observed groundwater trends using agricultural, climate, demographic, land-use and economic data in 81 countries. Our study sheds light on the main contributors to groundwater stress across the globe.

Modern Statistical Learning Methods for Big Data Mountain Laurel AB

(organized by Yingying Fan USC; chaired by Siliang Gong UNC)

Will Fithian (UC Berkeley) 3:00-3:30pm
Local Case-Control Sampling: Efficient subsampling in imbalanced data sets

For classification problems with significant class imbalance, subsampling can reduce computational costs at the price of inflated variance in estimating model parameters. We propose a method for subsampling efficiently for logistic regression by adjusting the class balance locally in feature space via an accept-reject scheme. Our method generalizes standard case-control sampling, using a pilot estimate to preferentially select examples whose responses are conditionally rare given their features. The biased subsampling is corrected by a post-hoc analytic adjustment to the parameters. The method is simple and requires one parallelizable scan over the full data set. Standard case-control sampling is inconsistent under model misspecification for the population risk-minimizing coefficients theta*. By contrast, our estimator is consistent for theta* provided that the pilot estimate is. Moreover, under correct specification and with a consistent, independent pilot estimate, our estimator has exactly twice the asymptotic variance of the full-sample MLE - even if the selected subsample comprises a miniscule fraction of the full data set, as happens when the original data are severely imbalanced. The factor of two improves to 1 + 1/c if we multiply the acceptance probabilities by c>1 (and weight points with acceptance probability greater than 1), taking roughly (1+c)/2 times as many data points into the subsample. Experiments on simulated and real data show that our method can substantially outperform standard case-control subsampling. This is joint work with Trevor Hastie.

Kun Chen (University of Connecticut) 3:30-4:00pm
Sequential Estimation in Sparse Factor Regression

Multivariate regression models of large scales are increasingly required and formulated in various fields. A sparse singular value decomposition of the regression component matrix is appealing for achieving dimension reduction and facilitating model interpretation. However, how to recover such a composition of sparse and low-rank structures remains a challenging problem. By exploring the connections between factor analysis and reduced-rank regression, we formulate the problem as a sparse factor regression and develop an efficient sequential estimation procedure. At each sequential step, a latent factor is constructed as a sparse linear combination of the observed predictors, for predicting the responses after adjusting for the effects of the previously found latent factors. Each sequential step reduces to a regularized unit-rank regression; when exact orthogonality among the sparse factors is desirable, it can be conveniently achieved through linear constrained optimization. The ideas of coordinate descent and Bregman iterative methods are utilized to ensure fast computation and algorithmic convergence even in the presence of missing data. Theoretically, we show that the sequential estimators enjoy the oracle properties for recovering the underlying sparse factor structure. The efficacy of the proposed approach is demonstrated by simulation studies and two real applications in genetics.

Yuekai Sun (UC Berkeley) 4:00-4:30pm
Feature distributed sparse regression

Most existing approaches to distributed sparse regression assume the data is partitioned by samples. However, for high-dimensional data, it is more natural to partition the data by features. We propose an approach to distributed sparse regression when the data is partitioned by features rather than samples. Evaluating our estimator requires only a single round of communication, making it amenable to modern distributed computing frameworks.

Poster Session

June 7, 4:30-6:30pm
Atrium Center

1. **Anthony S. Abrantes** (UNC)
Classifying EEG data for working memory load

The objective of this research was to identify a satisfactory classification method to predict the working-memory load of physicians when performing basic working memory tasks and simulated medical scenarios. Data was collected using a Nicolet nEEG V32 amplifier. Electrodes were placed using the 10-20 international system on Fp1, Fp2, F3, F4, T3, T4, Cz, O1, O2 with reference and ground electrodes at FCz/A1&A2 and CPz respectively. Data processing was accomplished by subtracting averaged A1/A2 reference signal from the remaining nine neural signals. Data was filtered using 4th order Butterworth band-pass filter with cutoff frequencies of 0.3Hz and 250Hz. Independent Component Analysis (ICA) was performed to remove temporal muscle activity from signals. Extreme Value Rejection (EVR) was performed on epoched data (1.4 sec event-locked trials for baseline data; 1sec epochs for simulated medical scenarios data) with rejection threshold at 3 times the largest standard deviation across all electrode sites. Frequency content extraction was accomplished using the Morlet wavelet for 0.5 to 50 Hz at 0.5 Hz increments. The data from the 3 vs. 6 dots experiment was used to develop and test the 4 classification models (LASSO regression, support vector machines (SVM), nearest shrunken centroids (NSC), and iterated supervised principal components (ISPC)) to predict a working-memory state when physicians would be more likely to have optimal performance (e.g., memorizing and correctly recalling 3 dots) vs. sub-optimal performance (memorizing and not correctly recalling the 6 dots). The naïve misclassification (predicting more common 3 dots state) rate was 19.74%. LASSO and SVM did outperform this threshold with misclassification rates of 18.10% and 12.21% respectively. Both classification models had relatively high specificity; LASSO: 97.2%; SVM was 99.8% correct) but relatively low sensitivity; LASSO: 20.7%; SVM: 39.6%). The LASSO model classified that physicians were on average ≈83% of time in the optimal performance state in each scenario. Results from the basic working-memory tasks suggest the LASSO model is superior to others, although there is room for improvement in model sensitivity.

2. **Xuan Bi** (UIUC)
A Group-Specific Recommender System

Recently there has been a growing demand to develop efficient recommender systems, which track users' preferences and recommend items of interest. We propose a group-specific method which utilizes dependency information from similar users and items under the matrix factorization framework. The new approach is effective for the "cold-start" problem, where majority responses in the testing set are obtained from new users whose preferences are not available. Another novelty is that we incorporate non-random missingness information through clustering, based on the numbers of ratings from each user and variables associated with missing patterns. Computationally, we propose a scalable algorithm that embeds backfitting into alternating least squares. This avoids large matrices operation and big memory storage. Our simulation studies and MovieLens data analysis both indicate that the proposed method improves prediction accuracy significantly compared to existing competitive approaches.

3. **Kelly Bodwin** (UNC)
Coherent Itemset Mining

It is often of interest to find associations between variables based on observed binary data. This problem has been previously studied under the heading of "frequent itemset mining" or "association rule mining". However, these classical approaches break down in the presence of non-identically distributed samples, which we argue is a common structure in real datasets of interest. In this paper, we propose a new model that accounts for differences in sample behavior while maintaining a common underlying variable dependence structure. We then introduce an algorithm that makes use of this model to identify groups of associated variables, which we refer to as coherent itemsets. The Coherent Itemset Mining (CIM) algorithm relies on an iterative update procedure that adaptively selects variable sets based on statistical testing principles. It is designed to run efficiently for high dimensional data. The CIM algorithm is tested on a variety of simulations as well as real datasets in genetics and text analysis.

4. **Frederick Campbell** (Rice University)
Within Group Variable Selection through the Exclusive Lasso

Many data sets consist of variables with an inherent group structure. The problem of group selection has been well studied, but in this paper, we seek to do the opposite: our goal is to select at least one variable from each group in the context of predictive regression modeling. This problem is NP-hard, but we propose the tightest convex relaxation: a composite penalty that is a combination of the one and two norms. Our so-called Exclusive Lasso method performs structured variable selection by ensuring that at least one variable is selected from each group. We study our method's statistical properties and develop computationally scalable algorithms for fitting the Exclusive Lasso. We study the effectiveness of our method via simulations as well as using NMR spectroscopy data. Here, we use the Exclusive Lasso to select the appropriate chemical shift from a dictionary of possible chemical shifts for each molecule in the biological sample.

5. **Jingxiang Chen** (UNC)
Estimating Individualized Treatment Rules for Ordinal Treatments

Precision medicine is an emerging scientific topic for disease treatment and prevention by taking into account individual patient characteristics. It is an important direction for clinical trial research and many statistical methods have been proposed recently. One of the primary goals in precision medicine is to obtain an optimal individual treatment rule (ITR) which can help make decisions on treatment selection according to each patient's specific characteristics. Recently, outcome weighted learning (OWL) has been proposed to estimate such an optimal ITR in a binary treatment setting by maximizing the expected clinical outcome. However, for the ordinal treatment settings such as dose finding, it is unclear how to use OWL. Furthermore, OWL requires transformation of the clinical outcome when the outcome has negative values. In this paper, we propose a new technique for estimating ITR with ordinal treatments. In particular, we propose a data duplication technique with a piecewise convex loss function. We establish Fisher consistency for the resulting estimated ITR under certain conditions, and also obtain the convergence and risk bound properties. Simulated examples and two applications to an irritable bowel problem and a type-2 diabetes mellitus clinical trial demonstrate the highly competitive performance of the proposed method compared to several existing ones.

6. **Glen Colopy** (University of Oxford)
Personalized Patient Monitoring with Gaussian Processes using Fast Adaptive Kernel Selection

The step-down unit (SDU) is a high-acuity hospital environment, to which patients may be sent after discharge from the intensive care unit (ICU). About 1- in-7 patients will deteriorate in the SDU and require emergency readmission to the ICU. Upon readmission, these patients experience significantly higher lengths of stay and mortality risks. Gaussian Process Regression (GPR) models are proposed as a flexible, principled, probabilistic method to address the clinical need to monitor continuously patient time-series of vital signs acquired in the SDU. The proposed GPR models focus on the robust forecasting of patient heart rate time series and on the early detection of patient deterioration. Results suggest that GPR-based heart rate monitoring provides superior advanced warning of deterioration compared to the current clinical practice of rules-based thresholding, as well as the current state-of-the-art kernel density method, which requires 4 additional vital sign features. The results show the value of principled Bayesian non-parametric methods to capture the underlying generative process of patient physiology, and use this information for significant clinical benefit. Predictive accuracy of the GPR model can be improved by an astute choice in the covariance function. As further patient data is acquired, the complexity of the kernel function better reflect the patient's specific physiological dynamics. The number of kernels considered in real-time is limited by the O(n3) computations of fitting the GP. To overcome this computational burden, the GP is approximated as a Kalman Filter, which requires O(n) computations. This fast approximation allows for the rapid comparison of many covariance functions of varying complexity. An optimal kernel is then selected according to maximize expected forecast accuracy, or minimize probability of extremely poor performance.

7. **Derek Feng** (Yale University)
ABtree: An Algorithm for Subgroup-Based Treatment Assignment

Given two possible treatments, there may exist subgroups who benefit greater from one treatment than the other. This problem is relevant to the field of marketing, where treatments may correspond to different ways of selling a product. It is similarly relevant to the field of public policy, where treatments may correspond to specific government programs. And finally, personalized medicine is a field wholly devoted to understanding which sub-groups of individuals will benefit from particular medical treatments. We present a computationally fast tree-based method, ABtree, for treatment effect differentiation. Unlike other methods, ABtree specifically produces decision rules for optimal treatment assignment on a per-individual basis. The treatment choices are selected for maximizing the overall occurrence of a desired binary outcome, conditional on a set of covariates. In this poster, we present the methodology on tree growth and pruning, and show performance results when applied to simulated data as well as real data.

8. **Siliang Gong** (UNC)
Testing-based Variable Selection for High-dimensional Linear Models

Variable selection plays a fundamental role in high-dimensional data analysis. Various methods have been developed for variable selection in recent years. Some well-known examples include forward stepwise regression (FSR), least angle regression (LARS), and many more. These methods typically have a sequential nature in the sense that variables are added into the model one-by-one. For sequential selection procedures, it is crucial to find a stopping criterion, which controls the model complexity. One of the most commonly used techniques for model evaluation in practice is cross-validation (CV). Despite its popularity, CV has two major drawbacks: expensive computational cost and lack of statistical interpretation. To overcome these drawbacks, we introduce a flexible and efficient testing-based variable selection approach that could be incorporated with any sequential selection procedure. At each step of the selection, we test the overall signal in the remaining inactive variables using the maximal absolute partial correlation among the inactive variables with the response conditioning on active variables. Furthermore, we develop a stopping criterion using the stepwise p-value. Numerical studies show that the proposed method delivers very competitive performance in terms of both variable selection accuracy and computational complexity compared to CV.

9. **Nhat Ho** (University of Michigan)
Singularity Structures and Parameter Estimation in Finite Mixtures of Skew Normal Distributions

Understanding singularity structures of the Fisher information matrix in finite mixture models has been a challenging problem since this matrix is singular and has very low rank around specific values of parameters. In this paper, we propose a general way to study these singularity structures under the specific setting of finite mixtures: skew normal mixtures. These models have become increasingly popular in recent years due to their flexibility in modeling asymmetric data. However, they appear to contain various kinds of singularities under both the exact-fitted setting, i.e the setting when the number of mixing components is known, or the over-fitted setting, i.e the setting where the number of mixing components is bounded but unknown. These singularities happen not only in the vicinity of symmetry but also in the setting of homologous sets, a new phenomenon due to the complex interaction among the parameters of the mixing measure. Apart from these singularities, skew normal density also has the non-linear partial differential equation structure. It leads to two interesting ways of characterizing the singularity levels of the Fisher information matrix. One way is based on the solvability of inhomogeneous system of polynomial equations while the another way is based on the combining strength phenomenon among multiple homogeneous and inhomogeneous systems of polynomial equations. The rich spectrum of the singularity structures consequently leads to various intricate degrees of parameter estimation under skew normal mixtures.

10. **Wenhao Hu** (NCSU)
Assessing Tuning Parameter Selection Variability in Penalized Regression

Penalized regression methods that perform simultaneous model selection and estimation are now ubiquitous in statistical modeling. The use of such methods seems almost unavoidable as manual inspection of all possible models quickly becomes intractable when there are more than a handful of predictors. However, such automated methods may fail to incorporate domain-knowledge, exploratory analyses, or other factors that might guide a more interactive model-building approach. A hybrid approach is to use penalized regression to identify a set of candidate models and then to use interactive model-building to examine this candidate set more closely. To identify a set of candidate models, we derive estimators of the probability that each model along the solution path will minimize some model selection criteria, e.g., AIC, BIC, etc., conditional on the observed solution path; models with a high selection probability are considered for further examination. Thus, the proposed methodology attempts to strike a balance between algorithmic modeling approaches that are computationally efficient but fail to incorporate expert knowledge and interactive modeling approaches that are labor intensive but informed by experience, intuition, and domain knowledge.

11. **Meilei Jiang** (UNC)
A Novel Method for Identifying Community Subtypes in the Sparse Microbiome of the Infected Lower Lung

A novel data object, Tree Weighted Abundance (TWA), was developed to do clustering, which combines the phylogenetic tree with relative abundance of microbiome. Based on a simulation study that models the sparseness of our data, we show that k-means to TWA is likely to perform better than two standard methods from the literature. Three statistically distinct subtypes of subjects have been discovered through the application of k-means clustering to TWA.

12. **Dehan Kong** (UNC)
High-dimensional Matrix Linear Regression Model

We develop a high-dimensional matrix linear regression model (HMLRM) to correlate matrix responses with high-dimensional scalar covariates when coefficient matrices have low-rank structures. We propose a fast and efficient screening procedure based on the spectral norm to deal with the case that the dimension of scalar covariates is ultra-high. We develop an efficient estimation procedure based on the nuclear norm regularization, which explicitly borrows the matrix structure of coefficient matrices. We systematically investigate various theoretical properties of our estimators, including estimation consistency, rank consistency, and the sure independence screening property under HMLRM. We examine the finite-sample performance of our methods using simulations and a large-scale imaging genetic dataset collected by the Alzheimer's Disease Neuroimaging Initiative study.

13. **Vered Madar** (SAMSI)
New and Simpler Algebraic Framework for Generating Correlated Random Variables

We will offer a surprising, novel algebraic representation to the well known Cholesky factor of a semi -positive-definite correlation matrix. As an useful application we shall offer a simple algorithm for the generation of realistic (semi-positive-definite) correlation structures, and illustrate how to simulate correlated random normal variables for variety of correlation structures.

14. **Christian Mueller** (Simons Foundation)
Generalized Stability Approach for Regularized Graphical Models

Selecting regularization parameters in penalized high-dimensional graphical models in a principled, data-driven, and computationally efficient manner continues to be one of the key challenges in high-dimensional statistics. One state-of-the-art model selection scheme is the Stability Approach to Regularization Selection (StARS) which determines the smallest regularization parameter that results in a graph that is sparse and simultaneously stable under random subsampling of the data. We here show that modeling the StARS stability criterion (the sum of variances of Bernoulli indicators across N subsamples) as Poisson-Binomial distribution allows us to derive a lower bound on the regularization path that results in identical StARS selection at greatly reduced computational cost for typical graphical model inference. Furthermore, we generalize the StARS criterion from single edge to induced subgraph (graphlet) stability and show superior graph recovery performance independent of the underlying graph topology with clear interpretability. We believe that these insights, along with our efficient parallel software utilities, make Gaussian graphical model inference and selection a routine task on standard multi-core computers at unprecedented speed.

15. **John Palowitch** (UNC)
The Continuous Configuration Model: A Null for Community Detection on Weighted Networks

Community detection is the process of grouping strongly connected nodes in a network. Many community detection methods for un-weighted networks have a theoretical basis in a null model, which provides an interpretation of resulting communities in terms of statistical significance. In this work, we introduce a null for edge-weighted networks called the continuous configuration model. We prove a Central Limit Theorem for sums of edge weights under the model, and propose a community extraction method called CCME which applies this result to an iterative multiple testing framework. To benchmark the method, we provide a simulation framework that incorporates the continuous configuration model as a way to plant null or “background” nodes in weighted networks with communities. We show CCME to be competitive with existing methods in accurately identifying both disjoint and overlapping communities, while being particularly effective in ignoring background nodes when they exist. We present two real-world data sets with potential background nodes and analyze them with CCME, yielding results that correspond to known features of the data. A pre-print of this work is available at http://arxiv.org/abs/1601.05630v2.

16. **So-Young Park** (NCSU)
Functional data analysis for quantile regression modeling, with application to feed intake of lactating sows

We propose a unifying modeling framework for quantile regression when covariates are of various types, vector and functional. The proposed approach first estimates the conditional distribution of the response under a generalized penalized regression modeling framework for a functional response and then obtains the estimated quantile by inverting it. The method allows a flexible model structure and is computationally efficient as it involves a single step penalization in estimation. We validate its performance through extensive simulation studies. The method is motivated by and applied to the sow data, where the primary interest is to understand how dynamic change of temperature in the farrowing rooms within a day (functional) is associated to low quantiles of feed intake of lactating sows while accounting for other sow-specific information (vector).

17. **David Pritchard** (UNC)
Composite Quantile-based Classifiers

Supervised dichotomous-outcome classification has received a large amount of attention in the literature. Numerous applications include spam detection, image recognition, and disease classification, among many others. Many well-known methods have been proposed such as linear discrimination, k-nearest neighbor, logistic regression, classification trees, and support vector machine. Yet despite this attention, new developments are still needed for high-dimensional and complex data. In this project, we propose a new efficient composite quantile-based classifier which captures the class differences for each covariate among different within-class quantiles. Our new method covers the median-based classifiers method by Hall and the quantile-based classifiers method by Hennig and Viroli as special cases. Our numerical and theoretical studies demonstrate the competitive performance of the proposed new classifier.

18. **Mauricio Sadinle Garcia-Ruiz** (Duke and NISS)
Set-Valued Multiclass Classifiers: Lowest Ambiguity with Bounded Error Levels

We introduce a framework for multiclass classification where the classifiers are allowed to output a set of plausible labels rather than a single label. These set-valued classifiers guarantee user-defined levels of coverage or confidence (the probability that the true label is contained in the set) while minimizing the ambiguity (the expected size of the output). First we derive oracle classifiers assuming the distribution is known. The oracle classifiers are obtained from level sets of the functions that define the conditional probability of each class. We then develop estimators with good asymptotic and finite sample properties. The proposed classifiers build on and refine many existing single-label classifiers. The optimal classifier can sometimes output the empty set. We provide two solutions to fix this issue that are suitable for various practical needs.

19. **Andrey Skripnikov** (University of Florida)
Estimation of Multi-Granger Network Causal Models

Network Granger causality focuses on estimating Granger causal effects from multivariate time series and it can be operationalized through Vector Autoregressive Models (VAR). The latter represent a popular class of time series models that has been widely used in applied econometrics and finance and more recently in biomedical applications. In this work, we discuss joint estimation and model selection issues of multiple Granger causal networks. We present a modeling framework for the setting where the same variables are measured on different entities (e.g. same set of economic activity variables for related countries). The framework involves the introduction of appropriate structural penalties on the transition matrices of the respective VAR models that link the underlying network Granger models and use of factor modeling for error covariance estimation. ADMM algorithm is presented for implementation of joint optimization procedure and the model is evaluated on both synthetic and real data.

20. **Samuel Ventura** (CMU)
PREDS: Prediction with Ensembles using Distribution Summaries

In classification, it is common to use an ensemble of models to predict an outcome of interest. For example, a random forest is an ensemble of decision trees built using bootstrap samples of the observations and using randomized subsets of the covariates at each split in each of the T trees in the ensemble. In a two-class problem, random forests aggregate the predictions of each underlying tree using a “majority vote" scheme, assigning the class with the majority vote amongst the underlying trees as the predicted class of the random forest. However, valuable information can be discarded in this process, including the individual predicted probabilities from each of the T underlying classifiers, which we show to often have multimodal or heavily skewed distributions. We introduce new prediction approaches that extract and incorporate information from these distributions of tree probabilities for the two-class problem. More specifically, we calculate distributional summary statistics to be used as covariates in a secondary classification model that makes more accurate predictions. We assess these approaches on a large record linkage dataset of death records from the Syrian Civil War conflict (Price et al, 2014) as well as several commonly used classification datasets. We compare to existing ensemble prediction approaches, such as stacking (Wolpert, 1992) and demonstrate that our approach lowers prediction error rates in most applications.

21. **Kristopher Williams** (University of Texas at San Antonio)
Unsupervised Outlier Detection Using Robust Locally Weighted Density Estimation

Most modern outlier detection methods use a local neighborhood as a reference set when constructing outlier scores. Among such methods, density-based techniques have proven to be a powerful tool for detecting outliers in the unsupervised setting. Despite this, many density-based methods fail to utilize useful information into the density estimate and subsequent outlier score. For one, most density-based methods fail to incorporate the local covariance structure in the density estimate primarily due to the inefficiencies associated with computing and storing the local covariance matrices. Secondly, how can the information gained from computing the outlier score itself be used to improve detection? Rarely, if ever, do outlier detection methods directly address this question. In this paper, a new density-based outlier detection method is presented using a local parametric density with robust location and scatter parameter estimates. The proposed density estimator provides a more flexible way to incorporate the local covariance structure compared to nonparametric techniques such as kernel density estimation. To further improve detection, a simple updating procedure is introduced to re-estimate location and scatter using the initial density estimates with the goal to improve overall detection. While the updating procedure does not guarantee improved results for any parameter setting, it does show that improvements can easily be obtained using a small number of iterations over a wide range of parameter settings. Experimental results using both simulated and real data sets demonstrate these improvements, and show significant increases in performance compared to other density-based methods.

22. **Zidian Xie** (University of Rochester)
Building a Predictive Model for Type 2 Diabetes Using Machine Learning

As one of the most prevalent chronic diseases in the US, diabetes, especially type 2 diabetes, has enormous financial burden on US economy. Since type 2 diabetes is predictable, improving predictive models to identify subjects at high risk for diabetes could help facilitate early intervention and reduce medical cost. However, due to its causal complexity, the prediction of type 2 diabetes has not been very successful. While a few risk factors, such as body-mass index (BMI), age, and sex, have been identified for type 2 diabetes, many risk factors remain to be identified. Using the 2014 Behavioral Risk Factor Surveillance System (BRFSS) dataset, we have applied a number of machine learning techniques to predict type 2 diabetes. Using Support Vector Machine (SVM), Decision Tree, Random Forest and Gaussian Naive Bayes classifiers, we have significantly improved the detection rate for type 2 diabetes from 37% (by previous studies) to 60%. Furthermore, while confirming previously reported risk factors, we have identified a few new potential risk factors related to type 2 diabetes, such as sleeping time.

23. **Guan Yu** (UNC)
Sparse Regression for Block-missing Multi-modality Data

In modern scientific research, many data are collected from multiple modalities (sources or types). Since different modalities could provide complementary information, sparse regression methods using multi-modality data could deliver better prediction performance. However, one special challenge for using multi-modality data is related to missing data. In practice, the observations of a certain modality can be missing completely, i.e., a complete block of the data is missing. In this paper, we propose a new two-step sparse regression method for block-missing multi-modality data. In the first step, we estimate the covariance matrix of the predictors using either the well-conditioned estimator or the robust Kendall’s tau estimator. Rather than deleting samples with missing data or imputing the missing observations, the proposed method makes use of all available information. In the second step, based on the estimated covariance matrix, a Lasso-type estimator is used to deliver a sparse estimate of the regression coefficients in the linear regression model. The effectiveness of the proposed method is demonstrated by theoretical studies, simulated examples, and a real data example from the Alzheimer’s Disease Neuroimaging Initiative. The comparison between the proposed method and some existing methods also indicates that our method has promising performance.

24. **Zheqi Zhang** (UNC)
Efficient Gauss-Newton-type Algorithms for Low-Rank Matrix Optimization

Low-rank matrix approximation has numerous applications in machine learning and statistics such as linear model reduction and matrix completion. In this work, we deal with this problem in a general setting by reformulating it into a nonconvex optimization problem with a rank constraint. We develop a generic Gauss-Newton framework to solve this problem, which has global convergence guarantee to its critical point. As a special instance, we customize our framework to handle the symmetric case, which covers a prominent application in semidefinite least-squares problems. We test our algorithms on various practical problems including matrix completion, image processing and quantum tomography using both synthetic and real data sets. The numerical results from these problems demonstrate the advantages of our Gauss-Newton scheme in speed and precision over some existing algorithms.

25. **Yi Zhao** (Brown University)
Pathway Lasso: Estimate and Select Sparse Mediation Pathways with High Dimensional Mediators

In many scientific studies, it becomes increasingly important to delineate the causal pathways through a large number of mediators, such as genetic and brain mediators. Structural equation modeling (SEM) is a popular technique to estimate the pathway effects, commonly expressed as products of coefficients. However, it becomes unstable to fit such models with high dimensional mediators, especially for a general setting where all the mediators are causally dependent but the exact causal relationships between them are unknown. This paper proposes a sparse mediation model using a regularized SEM approach, where sparsity here means that a small number of mediators have nonzero mediation effects between a treatment and an outcome. To address the model selection challenge, we innovate by introducing a new penalty called Pathway Lasso. This penalty function is a convex relaxation of the non-convex product function, and it enables a computationally tractable optimization criterion to estimate and select many pathway effects simultaneously. We develop a fast ADMM-type algorithm to compute the model parameters, and we show that the iterative updates can be expressed in closed form. On both simulated data and a real fMRI dataset, the proposed approach yields higher pathway selection accuracy and lower estimation bias than other competing methods.

Conference on Statistical Learning and Data Science

Department of Statistics and Operations Research

Department of Biostatistics

University of North Carolina at Chapel Hill

Chapel Hill, NC

Designed by

Siliang Gong • Yufeng Liu • Kai Zhang