BIRS Workshop:
Current Challenges in Statistical Learning
December 11-16, 2011
ABSTRACTS

Ejaz Ahmed, University of Windsor

Perspectives on Machine Bias Versus Human Bias

Penalized regression has been widely used in high-dimensional data analysis. In this talk, I consider estimation in generalized linear models when there are many potential predictor variables and some of them may not have influence on the response of interest. In the context of two competing models where one model includes all predictors and the other restricts variable coefficients to a candidate linear subspace based on prior knowledge, we investigate the relative performances of absolute penalty estimator (APE), shrinkage in the direction of the subspace, and candidate subspace restricted type estimators. We develop large sample theory for the shrinkage estimators including derivation of asymptotic bias and mean-squared error. The asymptotics and a Monte Carlo simulation study show that the shrinkage estimator overall performs best and in particular performs better than the APE when the dimension of the restricted parameter space is large. The estimation strategies considered in this talk are also applied on a real life data set for illustrative purpose.

David Banks, Duke University

Text Networks

There are many important examples of text networks: the Wikipedia, the Internet, political blogs. The confluence of new topic models, such as those based on nested Chinese restaurant processes, and new models for evolving networks, such as edge-weighted ERGMs with autoregressive dynamics, offer fresh hybrid strategies for machine learning in this domain.

Antonio Ciampi, McGill Univeristy

Patterns of Delirium: Latent Classes and Hidden Markov Chains as Modelling Tools

Delirium is a debilitating mental disorder frequently found in elderly populations. A psychometric instrument, the Delirium Index (DI), has been developed in order to assess the severity of Delirium, and is used in both a univariate and multivariate form. Using mixture of distributions for the univariate DI and latent class analysis with Hidden Markov Chains, we are developing an approach to pattern detection for Delirium aiming to identify states and course of the illness. The approach avoids 'hard' classifications but does identify potential and revisable diagnostic categories to which each patient can be assigned with a certain probaility. It can serve as an example to develop flexible, soft classification from data for any disease for which similar longitudinal data are available.

Bertrand Clarke, University of Miami

Clustering Stability: Impossibility and Possibility

In the first part of this talk we present a theorem that gives conditions under which high dimensional clustering is unstable. Specifically, for any fixed sample size, clustering becomes impossible (in a squared error sense) as the dimension increases unless the separation among the clusters is large enough in the sense that coordinatewise differences do not decrease too quickly with D, the dimension of the data points. We also show that clustering impossibility occurs with a theoretical rate of $\mathcal{O}(\sqrt{D})$.

In the second part of this talk we present a Bayesian method for assessing clustering stability. Roughly, the idea is to evaluate the probability that the distances between points and cluster centers can be re-ordered by random factors. The method seems to be consistent for choosing the number of clusters and we argue that it accurately reflects what we mean by the stability of a clustering. This is ongoing research and hence comments and discussion are particularly welcome.

Dean Eckles, Stanford University

Statistical causal inference for peer effects in online behavior

Peer effects can produce clustering of behavior in social networks, but so can homophily and common external causes.

For observational studies, adjustment and matching estimators for peer effects require often implausible assumptions, but it is only rarely possible to conduct appropriate field experiments to study peer influence. We describe research designs in which individuals are randomly encouraged to perform a focal behavior, which can subsequently influence their peers. Ubiquitous product optimization experiments on Internet services can be used for these analyses, presenting new opportunities and statistical challenges largely not addressed in the instrumental variables literature. We illustrate this approach with an analysis of peer effects in expressions of gratitude via Facebook on Thanksgiving Day 2010.

Bret Hanlon, University of Wisconsin

High Dimensional Variable Selection for Grouped Covariates with Applications in Cancer Genomics

In many scientific applications, parameters can be naturally grouped; for example, assayed genes can be grouped by biological pathways. This talk discusses variable selection methods which utilize group information to more effectively identify important variables. We study a class of variable selection procedures for parametric models via penalized likelihood, allowing for a general collection of likelihood functions and penalty functions. In particular, the penalty functions can be defined differently for each group. Under an asymptotic framework allowing both the number of parameters and the number of groups to diverge to infinity, we prove that the penalized likelihood estimators possess an oracle property. Our scientific motivation is to utilize pathway information to select a gene signature that is predictive for cancer patients' response to therapy. The goal is to use this predictive signature to make better informed decisions for treatment of new patients. Simulations and data analysis from a myeloma study illustrate the utility of the proposed methods.

Xiaoming Huo, Georgia Institute of Technology

Data-driven Functional Estimation on Irregular Regions

Suppose input variable X_i and response y_i have the relation: $y_i = f(X_i) + \varepsilon_i$, where ε_i are i.i.d. noises. Furthermore, we assume that X_i's are 'adequately' sampled within a domain Ω and function $f(\cdot)$ is unknown. Estimating $f(\cdot)$ is the objective for many well-known parametric and nonparametric methods. The most influential existing approach follows the following framework: (1) assume that f belongs to a predetermined functional class $\mathcal{F}$; (2) Derive analytic description of the basis function of $\mathcal{F}$ in Ω; (3) Turn the functional estimation problem into a quadratic programming problem, for which analytical and numerical solutions are available. This approach runs into difficulty when the domain Ω is irregular, or nonstandard.

We have developed a strategy that can circumvent this difficulty. In particular, a method that is completely driven by data is invented. We show that nearly all good asymptotic properties of the existing state-of-the-art approaches are inherited by the data-driven approach. These properties include, e.g., optimal rate of convergence, asymptotic optimality. We use numerical examples to demonstrate better performance of the proposed method when the domain Ω is irregular.

This is joint work with Zhouwang Yang and Huizhi Xie.

Jiashun Jin, Carnegie Mellon University

New approach to spectral clustering

Consider a two-class clustering problem where we have n data vectors X_i, i = 1, 2, ..., n, from two possible classes, but the class labels Y_i are unknown to us and it is of interest to estimate them. We model each vector as p-dimensional Gaussian $N(Y_i\mu, I_p)$, where we assume the data vectors are centered and μ is the contrast mean vector, which is unknown to us but is presumably sparse.

We propose the following approach to clustering (a) for each feature (e.g. gene), we use Kolmogorov-Smirnov statistic to assess the importance of the feature, and rank all features in terms of p-values. (b) we then perform a feature selection, where we use the recent idea of Higher Criticism Thresholding (HCT) to decide how many features we should keep. (c) for the remaining features, we obtain the leading eigenvector of the Human matrix, (d). we apply simple thresholding to the leading eigenvector and obtain the labels. The method is tested on two gene microarray data.

We explain the rationale behind this procedure by (a) a careful study of the tail-behavior of Kolmogorov-Smirnov statistic, using results from change-point analysis, (b) a careful study of the asymptotic property of the leading eigenvector, using results from Random Matrix Theory (RMT), (c) exposing a surprising connection between the leading eigenvector and the recent Higher Criticism statistic.

Yongdai Kim, Seoul National University

On weak and strong oracle property

Penalized regression methods with noncovex penalties are considered. Even though various nonconvex penalties possess the oracle property, a problem of using nonconvex penalties is that there are multiple local minima, and we do not know which local minimum is the oracle estimator. A question is whether the problem of multiple local minima is a real practical problem. In this talk, I give a partial answer for the question. First, I will show that there are many bad local minima for certain noncnovex penalties possessing the oracle property. Then, I will give some conditions on the penalty to ensure not many bad local minima.

Yumi Kondo and Matias Salibian-Barrera, University of British Columbia

A robust and sparse K-means clustering algorithm

In many situations where the interest lies in identifying clusters one might expect that not all available variables carry information about these groups. Furthermore, data quality (e.g. outliers) might present a serious and hard-to-assess problem for some large and complex datasets. In this talk we show that a small proportion of atypical observations might have serious adverse effects on the solutions found by the sparse clustering algorithms of Witten and Tibshirani (2010). We propose a robustification of sparse K-means based on the trimmed K-means algorithm (Gordaliza, 1991a and 1991b). Our proposal is also able to handle datasets with missing values. The performance of the proposed robust sparse K-means is assessed in various simulation studies and one data analysis. Our simulation studies show that, when there are outliers in the data, the robust sparse K-means algorithm performs better than other competing methods both in terms of the selection of features and also the identified clusters. Finally, we illustrate our method on a microarray dataset where we are able to identify natural biological clusters. The robust sparse K-means algorithm has been implemented in the R package RSKC.

Wonyul Lee and Yufeng Liu, University of North Carolina at Chapel Hill

Joint statistical modeling of multiple high dimensional data

With the abundance of high dimensional data, shrinkage techniques are very popular for simultaneous variable selection and estimation. In this talk, I will present some new shrinkage techniques for joint analysis of multiple high dimensional data. Applications on cancer gene expression data and microRNA data will be presented.

This is joint work with Wonyul Lee.

Hongzhe Li, University of Pennsylvania

Robust Detection and Identification of Sparse Segments in Ultra-High Dimensional Data Analysis

Copy number variants (CNVs) are alternations of DNA of a genome that results in the cell having a less or more than two copies of segments of the DNA. CNVs correspond to relatively large regions of the genome, ranging from about one kilobase to several megabases, that are deleted or duplicated. Motivated by CNV analysis based on next generation sequencing data, we consider the problem of detecting and identifying sparse short segments hidden in a long linear sequence of data with an unspecified noise distribution. We propose a computationally efficient method that provides a robust and near-optimal solution for segment identification over a wide range of noise distributions. We theoretically quantify the conditions for detecting the segment signals and show that the method near-optimally estimates the signal segments whenever it is possible to detect their existence. Simulation studies are carried out to demonstrate the efficiency of the method under different noise distributions. We present results from a CNV analysis of a HapMap Yoruban sample to further illustrate the theory and the methods.

George Michailidis, University of Michigan

Network Granger Causality with Inherent Grouping Structure

The problem of estimating high-dimensional network structures arises naturally in the analyses of many physical, biological and socio-economic systems. Examples include stock price fluctuations in financial markets and gene regulatory networks in biology. We aim to learn the structure of the network over time employing the framework of Granger causal models under the assumptions of sparsity of its edges and inherent grouping structure among its nodes. We introduce a truncated penalty variant of Group Lasso to discover the Granger causal interactions among the nodes of the network. Asymptotic results on the consistency of the new estimation procedure are developed. The performance of the proposed methodology is assessed through an extensive set of simulation studies and comparisons with existing techniques.

Alejandro Murua and Thierry Chekouo Tekougang, Université de Montréal

The penalized plaid biclustering model

Very few statistical models for biclustering have been proposed in the literature. Instead, most of the research has focused on algorithms to find biclusters. The models underlying them have not received much attention. Hence, very little is known about the adequacy and limitations of the models and the efficiency of the algorithms. In this work we shed light on the actual statistical models behind the algorithms. This allows us to generalize most of the known popular biclustering techniques, and to justify, and many times improve on, the algorithms used to find the biclusters. It turns out that most of the known techniques have a hidden Bayesian flavor. Therefore, we proposed a Bayesian framework to model biclustering. We propose a measure of biclustering complexity (overlapping) trough a penalized plaid model, and present a modified DIC criterion to choose the appropriate number of biclusters, a problem that has not been adequately addressed yet. We show some applications of these ideas to gene expression data.

Art Owen, Stanford University

Bootstrapping r-fold Tensor Data

The famous Netflix data is a sparsely sampled table with rows for customers and columns for movies (or vice versa). Both movies and rows are naturally modeled as random effects. Bootstrapping such data is problematic: no proper bootstrap can exist, according to a theorem of Peter McCullagh.Resampling rows and columns independently is effective though slightly conservative.

Computerized data gathering frequently produces data sets with three-way or even higher order data tables. We present a bootstrap for such tensor valued data. Our version uses independent weights instead of multinomial ones. It remains mildly conservative. Poisson weights are close to the original bootstrap, but binary weights have computational and statistical advantages. Under certain conditions a single bootstrap replicate suffices to give a variance estimate. We apply our method to show that Facebook users in the UK make longer comments than those in the US when using their phone. The opposite pattern holds for comments made via computer.

This work is joint with Dean Eckles, Stanford University.

Zhaohui Qin, Emory University

Inference of correlated hidden Markov models with application to genome-wide studies

Hidden Markov models (HMM) have been widely used to analyze large-scale genomic data because of its ability to incorporate spatial correlation, in areas such as genome-wide mapping of binding sites for regulatory proteins. With advances in sequencing technology, it is now common to analyze data for multiple proteins jointly, but existing algorithms analyze data for each protein separately. However, extending the HMM framework to a multivariate form is not trivial because hidden state space increases quickly with the number of experiments and it is therefore of interest to develop a powerful yet computationally tractable algorithm. Here we present a fast inferential method for correlated hidden Markov models for multiple sequential data (series). Instead of jointly inferring hidden states for all series, we adjust the transition kernel of each series with the hidden state configuration in other correlated series and keep the hidden state inference marginal in each. Through simulation, we show that the new scheme achieves sensitivity comparable to the fully coupled HMM fit at a computational cost as low as fitting an independent HMM for each series separately. The method was applied to the analysis of histone modification data in mouse.

David Stenning, University of California, Irvine

Automatic Classification of Sunspot Groups Using SOHO/MDI Magnetogram and White-Light Images

The morphological classification of sunspot groups, based on the complexity of magnetic flux polarity in associated active regions, is predictive of both their future evolution and of explosive events higher in the solar atmosphere. Currently, active region identification and classification is done manually by experts. This process is both laborious and prone to inconsistencies stemming from the subjective nature of the classification schemes. In addition, manual classification is unfeasible for the massive high cadencedatasets being generated by new instruments such as NASA's Solar Dynamics Observatory. Using mathematical morphology, we extract numerical summaries from magnetogram and white-light images of sunspot groups that are relevant to their classification. These features are then used in an automated classification scheme based on supervised learning techniques. Furthermore, since the discrete grouping of active region complexity is artificial, unsupervised learning techniques are being explored to develop a classification scheme based on a continuum of classes. Our ultimate goal is to capture the evolutionary patterns of sunspot groups that are useful for predicting volatile solar events, such as solar flares and coronal mass ejections.

David van Dyk, Imperial College London

Statistical Learning Challenges in Astronomy and Solar Physics

In recent years, technological advances have dramatically increased the quality and quantity of data available to astronomers. Newly launched or soon-to-be launched space-based telescopes are tailored to data-collection challenges associated with specific scientific goals. These instruments provide massive new surveys resulting in new catalogs containing terabytes of data, high resolution spectrography and imaging across the electromagnetic spectrum, and incredibly detailed movies of dynamic and explosive processes in the solar atmosphere. The spectrum of new instruments is helping scientists make impressive strides in our understanding of the physical universe, but at the same time generating massive data-analytic and data-mining challenges for scientists who study the resulting data. In this talk I will introduce discuss the statistical learning challenges inherent in data streams that are both massive and complex.

Marina Vannucci, Rice University

Bayesian Models for Variable Selection that Incorporate Biological Information

In this talk I will review Bayesian methods for variable selection that use spike and slab priors. Specific interest will be towards high-dimensional data. Linear and nonlinear models will be considered, with continuous, categorical and survival responses. Applications will be to genomics data from DNA microarray studies. The analysis of the high-dimensional data generated by such studies often challenges standard statistical methods. Models and algorithms are quite flexible and allow us to incorporate additional information, such as data substructure and/or knowledge on gene functions and on relationships among genes.

Joe Verducci, The Ohio State University

Long tau paths to detect heterogeneity of association in large populations

Given a large sample from a potentially heterogeneous population, the problem is to discover a subpopulation over which a pairs of continuous variables are monotonely associated. Reordering the sample to minimize the number of inversions appearing in the beginning of the sequence has been shown to be

effective in small samples. Strategies for practical extension to large samples draw on asymptotical distributional results for Longest Increasing Subsequences and Runs in random permuations. An extension of the beta family of distributions to include the truncated normal turns out to be useful for this problem.

Peng Wang, Bowling Green State University

Conditional Inference Functions for Mixed-Effects Models with Unspecified Random-Effects Distribution

In longitudinal studies, mixed-effects models are important for addressing subject-specific effects. However, most existing approaches assume a normal distribution for the random effects, and this could affect the bias and efficiency of the fixed-effects estimator. Even in cases where the estimation of the fixed effects is robust with a misspecified distribution of the random effects, the estimation of the random effects could be invalid. We propose a new approach to estimate fixed and random effects using conditional quadratic inference functions. The new approach does not require the specification of likelihood functions or a normality assumption for random effects. It can also accommodate serial correlation between observations within the same cluster, in addition to mixed-effects modeling. Other advantages include not requiring the estimation of the unknown variance components associated with the random effects, or the nuisance parameters associated with the working correlations. Real data examples and simulations are used to compare the new approach with the penalized quasi-likelihood approach, and SAS GLIMMIX and nonlinear mixed effects model (NLMIXED) procedures.

Junhui Wang, University of Illinois at Chicago

On positive definite estimation of covariance and precision matrices

In recent years, estimation of the covariance and precision matrices has attracted growing attention of both statistics and computer science communities. It has close connection with Gaussian graphical models in that the Gaussian distribution is fully characterized by its first two moments and its dependence and conditional dependence structure is fully determined by the covariance and precision matrices. In this talk, a generic gradient descent algorithm will be present, which is applicable to estimation of both covariance and precision matrices. It can assure positive definiteness of the estimated matrices, attain sparseness when appropriately stopped, and is computationally efficient. Numerical examples will be provided to demonstrate the effectiveness of the proposed estimation scheme, and convergence properties will also be discussed.

Steven Wang, York University

Nominal Association and Dimensionality Reduction for Categorical Variable

When response variables are nominal and populations are cross-classified with respect to multiple polytomies, questions often arise about the degree of association of the responses with explanatory variables. When populations are known, we introduce a nominal association vector and matrix to evaluate the dependence of a response variable with an explanatory variable. These measures provide detailed evaluations of nominal associations at both local and global levels. We also propose a general class of global association measures which embraces the well known association measure by Goodman-Kruskal (1954). The hierarchy of equivalence relations defined by the association vector and matrix are also shown. We also prove that a dimensionality reduction for high dimensional categorical data is theoretically possible by using the proposed association measures.

Yichao Wu, North Carolina State University

Adaptively Weighted Large Margin Classifiers

Large margin classifiers have been shown to be very useful in many applications. The Support Vector Machine is a canonical example of large margin classifiers. Despite their flexibility and ability in handling high dimensional data, many large margin classifiers have serious drawbacks when the data are noisy, especially when there are outliers in the data. In this paper, we propose a new weighted large margin classification technique. The weights are chosen adaptively with data. The proposed classifiers are shown to be robust to outliers and thus are able to produce more accurate classification results.

This is joint work with Yufeng Liu.

S. Stanley Young, NISS

Assessing variable importance in environmental observational studies

Often authors do not address the relative importance of variables under consideration, choosing instead to concentrate on specific claims they are making. Geographic variation in the possible effects of air pollution may go unrecognized. Yet good policy decisions would seem to require knowing the relative importance of variables, not just their statistical significance. Of course, authors are expected to use statistical methods and write papers as effectively as possible to support their specific claims. Authors will use statistical methods to adjust out other factors without highlighting their relative importance. Often data used is not available so the reader is in the position of having to take the authors at their word. We obtained data pertinent to an important policy question: how does air quality relate to all cause mortality across the US. We use three methods of determining variable importance to show how predictor variables can be put into a context useful for policy decisions. Using both regression and recursive partitioning, we are able to confirm a spatial interaction of an air quality variable, PM2.5, known in the literature. We also determine the relative importance of this variable in the context of other variables. Knowing that effects are specific

to geographic regions and the relative importance of predictor variables within regions should be useful to policy decision-makers.

This is joint work with Jessie Q. Xia

Ruben Zamar, UBC

Robust Classification

I'll present a classification method which is robust not only to outliers in the training and also in the test data. We achieve that by using an ensemble of robust classifiers based on mixture models. We apply our method to a large SNP genotype dataset.

This is joint work with Mohua Podder, Will Welch and Scott Tebbutt.

Yunzhang Zhu, University of Minnesota

Maximum likelihood estimation over multiple undirected graphs

Graphical models are useful to analyze and visualize conditional independence relationships between interacting units, in addition to their structural implications. In dynamic network analysis, a structural change may be a result of certain events. To estimate dependency structures between adjacent matrices we study maximum likelihood estimation over multiple Gaussian graphical models. Of particular interest is the grouping structure over these matrices as well as sparseness over and within matrices, which is characterized in terms of homogeneous subgroups of elements and zero-elements of the matrices. A nonconvex method is proposed to seek sparse representations within each matrix and identify subgroups over all matrices. Theoretically, a non-asymptotic error bound for exact recovery for sparseness and clusters is derived. This leads to consistent reconstruction of sparseness and cluster structures over matrices simultaneously, permitting the number of unknown parameters to be in exponential of the sample size. Simulation studies suggest that the method enjoys the benefit of clustering and feature selection at the same time, and compares favorably against its convex counterpart in the accuracy of selection and predictive performance.