Abstracts for the joint meeting of the

International Society for Business and Industrial Statistics
and the
Statistical Learning and Data Mining Section
of the
American Statistical Association

Abstracts are listed in the order in which the talk appears in the scientific program.

Monday Morning

Facing the Predictive Modeling Paradox: The Art of Making Do with the Data You Have, Not the Data You Wish You Could Get.

Claudia Perlich, Dstillery

Abstract: The most interesting paradox in predictive modeling is that in situations where a predictive model would be most useful, it is often the hardest to obtain adequate training data to build it. Between sample selection biases, non-stationary processes, and the cost/time constraints of running experiments to collect enough ’right’ data; the practical reality of building predictive models is the skillful embrace of second best solutions. Display advertising is one such application domain suffering from major data challenges: optimization for a campaign should consider thousands if not millions of parameters and should happen prior to the start of the campaign at which point nobody has even seen the ad. To make things ever harder, the outcomes towards which we are optimizing (e.g., post view purchases) are exceedingly rare if not entirely unobservable by the entity charged with running the campaign. The talk will survey a list of relevant modeling problems suffering some form of this predictive modeling paradox and provide a more in depth look at transfer learning as a potential solution in the case of display advertising.

Relation of Age to Human Brain DNA Methylation in a Cohort of Older Persons

Jingyun Yang, Rush University Medical Center

Abstract: DNA methylation plays a crucial role in gene expression regulation, cell differentiation and development. Previous studies have reported age-related methylation alteration in human brain across the lifespan. Using genome-wide DNA methylation data from 740 postmortem brains, we extend those studies by interrogating over 420,000 CpG sites across the genome in a cohort of old persons between the ages of 66 and 108, the ages at which many common chronic neurologic diseases become clinically manifest. We find that mean methylation level of interrogated CpGs follows a distinct bimodal distribution, and 29,004 CpGs (6.9%) exhibit little inter-individual difference in methylation. We observe 4,263 CpGs (1.0%) that are associated with age (p¡1.1910-7). These CpGs tag 2,823 genes across the genome, and are mainly located in CpG Islands (65.3%) and promoter regions (47.0%). The majority (85.7%) are hypo-methylated, i.e., mean methylation ¡0.5, and show positive association with age, i.e., older age is associated with higher level of methylation. We confirm many previously reported CpGs, and also identify novel genes harboring age-related CpGs. We find no age association with overall methylation for most chromatin states, except for inactive/poised promoter (p=5.9410-5). Age is associated with higher level of DNA methylation in the human brain, suggesting that methylation alteration may play a role in age-related brain diseases.

Tree-based Rare Variants Analyses

Heping Zhang, Yale University

Abstract: Since the development of next generation sequencing (NGS) technology, researchers have been extending their efforts on genome-wide association studies (GWAS) from common variants to rare variants to find the missing inheritance. Although various statistical methods have been proposed to analyze rare variants data, they generally face difficulties for complex disease models involving multiple genes. In this paper, we propose a tree-based method that adopts a non-parametric disease model and is capable of exploring gene-gene interactions. We found that our method outperforms the sequence kernel association test (SKAT) in most of our simulation scenarios, and by notable margins in some cases. By applying the tree-based method to the Study of Addiction: Genetics and Environment (SAGE) data, we successfully detected gene CTNNA2 and its 44 specific variants that increase the risk of alcoholism in women. This gene has not been detected in the SAGE data. Post hoc literature search also supports the role of CTNNA2 as a likely risk gene for alcohol addiction. This finding suggests that our tree-based method can be effective in dissecting genetic variants for complex diseases using rare variants data.

Computational Methods for Linear Mixed Effects Models in Large-Scale Genome-Wide Association Studies

Xiang Zhou, University of Chicago

Abstract: Linear mixed models have attracted considerable attention recently as powerful and effective tools to account for population stratification and relatedness in GWASs. However, existing methods for calculating likelihood ratio test (LRT) statistics are computationally impractical for even moderate-sized GWASs, and many studies have to rely on approximate LRT methods. To address this issue, I present novel computationally-efficient algorithms, which we refer to as genome-wide efficient mixed model association (GEMMA), for fitting both univariate and multivariate linear mixed models, and computing the LRT for SNP associations in GWASs. Our methods improve on existing approximate LRT methods in computation speed, power/correct control of type I error, and ability to deal with more than two phenotypes. I illustrate these features on real and simulated data.

Large-Scale Clustering Methods with Applications in Record Linkage

Sam Ventura, Carnegie Mellon University

Abstract: Deduplication is the process of linking records corresponding to unique entities within a single database. We frame deduplication as a clustering problem, where each observed record belongs to a cluster corresponding to some latent unique entity in the underlying population. Hierarchical clustering is an intuitive way to link groups of records, but pairwise distances are computationally expensive, and most deduplication problems have a large number of records, making it difficult to estimate a distance matrix. Our work focuses on improving the feasibility and computational scalability of clustering methods for large-scale deduplication problems. We define the distance between record-pairs to be a monotonically decreasing transformation of their pairwise matching probabilities, which are obtained via a supervised learning approach (given some set of pairwise match/non-match labels). In large-scale problems, building a single classifier can be computationally infeasible. We build an ensemble of classifiers on subsets of training data and estimate a distribution of pairwise matching probabilities for each pair of records. We then present a method for clustering records into unique entities by identifying the best approximation of the true distance between records (or probability of matching) given these distributions, called ”distribution linkage hierarchical clustering. Finally, we apply novel blocking techniques to reduce the required number of pairwise distance calculations. These clustering approaches can be used when we have very large datasets and/or unavailable or uncertain distances between observations. We apply our methodology to two deduplication problems: identifying unique inventors in the United States Patent and Trademark Office patent-inventor database (8+ million patents) and estimating the number of unique casualties resulting from the Syrian civil war conflict (150,000+ records of casualties).

Record linkage and other statistical models for quantifying conflict casualties in Syria

Beka Steorts, Carnegie Mellon University

Abstract: How do we know how many people have been killed in Syria? If violence is escalating or decreasing? The hard answer is we don't. But through careful application of machine learning and other statistical techniques, we can quantify what we do, and don't, know. In this talk, l present how the Human Rights Data Analysis Group uses random forests, multiple systems estimation, and various Python and R packages to estimate conflict casualties.

The Design and Implementation of a Record Linkage Pipeline

Patrick Ball, HRDAG

Abstract: We describe the evolution and design of a record linkage pipeline currently in use at Human Rights Data Analysis Group. Along the way we discuss a number of software design principles specific to statistical applications, for example auditability, testability, and automatic reporting. We highlight some recent developments in the python ecosystem, especially ipython, pandas, and sklearn, that have made these ideas practical.

On the Sensitivity of the Lasso to the Number of Predictor Variables

Cheryl Flynn, New York University

Abstract: The Lasso is a computationally efficient procedure that can produce sparse estimators when the number of predictors (p) is large. Oracle inequalities are commonly cited as theoretical justification for the Lasso; however, these inequalities are based on a deterministic choice of the regularization parameter. In practice, the regularization parameter is typically selected using a data-dependent procedure. We address this disconnect by studying the predictive performance of the Lasso when the regularization parameter is selected using a data-dependent procedure. Assuming orthonormal predictors and a sparse true model, we prove that the best possible predictive performance of the Lasso deteriorates as p increases with positive probability, and that this probability can be close to one in certain situations. We further demonstrate empirically that the deterioration in performance can be far worse than is commonly viewed by the literature.

Bayesian Reference Analysis for Exponential Power Regression Models

Marco Ferreira, University of Missouri-Columbia

Abstract: We develop Bayesian reference analyses for linear regression models when the errors follow an exponential power distribution. Specifically, we obtain explicit expressions for reference priors for all the six possible orderings of the model parameters. In addition, we show that associated with these six parameters orderings there are only two reference priors. Further, we show that both of these reference priors lead to proper posterior distributions. Furthermore, we show that the proposed reference Bayesian analyses compare favorably to an analysis based on a competing noninformative prior. Finally, we illustrate our Bayesian reference analysis for exponential power regression models with applications to two datasets. In the first application we analyze excess returns for a publicly traded company. In the second application we study the relationship between sold home videos versus profits at the box office.

Confluent Hypergeometric Mixture of g-priors in Generalized Linear Models

Yingbo Li, Clemson University

Abstract: This is a new Bayesian model selection and model averaging method for GLMs. It can be considered as a unified framework that has a lot of mixtures of g-priors in the literature as special cases and naturally extends them to be applicable to GLMs. Theoretical properties such as model selection and estimation consistency are also shown. The model has approximate marginal likelihood in closed-form so that enables quick search method for high dimensional data.

Complex Grouped Variable Selection

Xiaoli Gao, University of North Carolina at Greensboro

Abstract: Existing grouped variable selection methods rely heavily on the prior group information, thus may not be reliable if an incorrect group assignment is used. In this paper, we propose a family of novel shrinkage variable selection operators by controlling the k-th largest norm (KLAN). The proposed KLAN method is able to perform grouped variable selection naturally even though no prior group information is available. We also construct a group KLAN shrink-age operator using a composite of KLAN constraints. Neither ignoring nor relying completely on prior group information, the group KLAN method has flexibility of controlling the within group strength and therefore can reduce the damage caused by incorrect group information. Finally, we investigate an unbiased estimator of the degrees of freedom of (group) KLAN estimates. Some small sample simulation studies are performed to demonstrate the ad- vantage of both KLAN and group KLAN as compared to the LASSO and group LASSO, respectively.

Estimation in High-dimensional Spatial Conditional Autoregressive Model

Abdulkadir Hussein, University of Windsor

Abstract: In this talk we develop an array of shrinkage and absolute penalty estimators for the coefficients of the spatial conditional autoregressive model. The performance of the estimators will be assessed by using analytical results as well as through Monte Carlo simulations. The usefulness of the proposed estimators will be illustrated by using data sets on crime distribution and housing prices.

Big Data Analysis, Big Biases

S. Ejaz Ahmed, Brock University

Abstract: In high-dimensional data settings where number of variables is greater than observations, many penalized regularization approaches were studied for simultaneous variable selection and estimation. However, with the existence of covariates with weak effect, many existing variable selection methods may not distinguish covariates with weak signals and no signal. In this case, the prediction based on a selected submodel may not be highly efficient. In this talk, we propose a high-dimensional shrinkage estimation strategy to improve the prediction performance of a submodel. Such a high-dimensional shrinkage estimator (HDSE) is constructed by shrinking a weighted ridge estimator in the direction of a predefined candidate submodel. Under an asymptotic distributional quadratic risk criterion, its prediction performance is explored analytically. We show that the proposed HDSE performs better than the weighted ridge estimator. More importantly, it improves the prediction performance of any candidate submodel generated from most existing variable selection methods significantly. The relative performance of the proposed HDSE strategy is demonstrated by both simulation studies and the real data analysis.

Augmenting a Designed Experiment with New Runs to More Precisely Locate a Response Contour

Brad Jones, SAS

Abstract: After running an RSM experiment an investigator may have a contour line of predicted responses that match a targeted response. This contour line is a function of the control factors and may be subject to substantial uncertainty. In such cases it may be desirable to augment the original data with new data to more precisely locate that contour. In addition to improving the model predictions, however, the investigator may want the observed responses to be close to the target response so that the units produced in the experiment are not lost to production. This paper describes a new methodology for space filling design augmentation and applies it to the problem of approximating a response contour.

A Change Point Approach for Phase-I Analysis in Multivariate Profiles Monitoring and Diagnosis

Kamran Paynabar, Georgia tech

Abstract: Process monitoring and fault diagnosis using profile data remains an important and challenging problem in statistical process control (SPC). Although the analysis of profile data has been extensively studied in the SPC literature, the challenges associated with monitoring and diagnosis of multichannel (multiple) nonlinear profiles are yet to be addressed. Motivated by a real-data application in multi-operation forging processes, this paper develops a new modeling, monitoring and diagnosis framework for phase-I analysis of multichannel profiles. The proposed framework incorporates the multi-dimensional functional principal component analysis into change-point models. In this framework, the multichannel profiles are treated as multivariate functional observations and their low-dimensional projections on the principal components of profile data are used for monitoring and diagnosis. Simulation results show that the proposed approach has better performance in identifying change-points in various situations compared with some existing methods.

Statistical Process Control approaches for High-Density Dimensional Point Cloud Data-Sets

Lee Wells, Virginia Tech

Abstract: Statistical process control (SPC) methods have been extensively applied to monitor manufacturing processes to quickly detect and correct out-of-control conditions. As sensor and measurement technologies advance, there is a continual need to adapt and develop new SPC techniques to effectively and efficiently take advantage of these new data-sets. For instance, advanced multivariate and profile monitoring techniques have been developed to account for the increased dimensional data collected from technologies, such as coordinate measuring machines. Currently high-density dimensional (HDD) measurement technologies, such as 3D laser scanners, are being implemented in industry to rapidly collect point clouds consisting of millions of data points to represent an entire manufactured parts' surface. This gives HDD measurements a significant advantage over competing technologies that typically provide tens or hundreds of data points. Consequently, HDD data-sets have the potential to detect unexpected faults, i.e., faults that are not captured by measuring a small number of predefined dimensions of interest. However, in order for this potential to be realized SPC methods capable of handling these data-sets need to be developed. This presentation focuses on two recently developed SPC approaches for the use of HDD measurements. The first approach transforms HDD point clouds into Q-Q plots which are monitoring as linear profiles. The second approach transforms HDD point clouds into non-uniform rational basis spline (NURBS) surfaces. The model parameters for these NURBS surfaces are then monitored using surface monitoring techniques.

Interactive Learning for Claims Processing

Andrew Fano, Accenture

Classifiers built for real world systems are often not autonomous but part of a larger interactive system with an expert in the loop. Developing such systems effectively requires not only data, but a clear understanding of the broader business processes within which these experts work. We describe the development of a system that uses a classifier to detect payment errors. The intent of the system, however, was not simply to maximize the effectiveness of the classifier. Instead the intent was to make the overall detection and processing of these claims by an adjuster more efficient. The result is a system that balances the use of the expert to explore and refine the learned models, with the exploitation of those models. Finally this system served as the basis for a dissertation project that detailed a generalized approach to negotiating the tradeoffs between the experts cost, exploration (the potential future benefit in classifier performance from labeling cases likely to improve classifier performance) and exploitation (the potential benefit of labeling cases most likely to produce value). This collaboration between an academic department and an industry lab yielded work that is of interest to the field and value to companies that stand to benefit from the approaches developed by this field.

A Generalized Model of Advertising: Incorporating Electronic Word-of-Mouth into Advertising Model

Nicolas Glady, ESSEC Business School

Abstract: Today consumers decision journey has become a multi-channel and multi-stage process. Attempting to influence consumers brand consideration set, marketing managers are using an increasing variety of media vehicles for advertising. Such expansion is intensified by social web technologies and consumer-generated content that provide marketers with alternative advertising instruments.

The current research uses an extended version of an integrated utility-maximizing framework and investigates the effect of electronic word-of-mouth (eWOM) on category purchase incidence, brand choice and purchase quantity decisions. This study addresses eWOM as a special type of social media and contrasts it with traditional channels of awareness-raising. An empirical model based on Hierarchical Bayes methodology accounts for consumer heterogeneity and as a result, provides marketing managers with suggestions for multi-channel advertising strategies.

Large-scale Statistical Modeling for LinkedIn Advertising Platform

Liang Zhang, LinkedIn

Abstract: This talk gives a high-level overview of how large-scale statistical modeling practices are applied in LinkedIn advertising platform. Although the logistic regression model we use is well understood, we do face challenges with data collection, large-scale and dynamic model training, and scalability in real-time inference. To facilitate training with both large numbers of training examples and high dimensional covariates on commodity clustered hardware, we employ the Alternating Direction Method of Multipliers (ADMM). Because online advertising applications are much less static than classical presentations of response prediction, we employ a number of techniques that allows it to adapt in real time. The model we proposed can be divided into components with different re-training frequencies, allowing us to learn from changes in ad campaign performance frequently without incurring the cost of retraining larger, more stable sections of the model. Thompson sampling during online inference further helps by efficiently balancing exploration of new ads with exploitation of long running ones. Finally, we show via extensive offline experiments and online A/B tests that this system provides significant benefits to prediction accuracy and gains in revenue and click through rates.

Monday Afternoon

Factorization Machines for Predictive Modeling and Recommendation

Jorge Silva, SAS Institute

Abstract: Factorization Machines, introduced in 2012 by Rendle, are a generalization of matrix factorization which allows seamless introduction of features (aka ”feature engineering”) into the standard factor model. It is among the state-of-the-art , e.g., for recommender system applications. Moreover, they can be interpreted as a simplified version of polynomial regression/classification, thus allowing non-linear predictive modeling that scales linearly in computational complexity with both the number of observations and the number of features. While this linear scaling makes Factorization Machines attractive for big data applications, issues remain concerning the parallelization of stochastic gradient descent, which is the main method for optimizing the parameters. We present a methodology for distributed stochastic gradient descent applied to Factorization Machines, allowing the method to be efficiently implemented under SAS' High Performance Analytics platform.

Unstructured Algorithms to Discover Structured Patterns

Anjishnu Banerjee, Amazon

Abstract: Big data is the ubiquitous term among analytics and machine learning professionals, encompassing a wide array of algorithms, modeling techniques, tools, platforms and application areas. One common thread in many of these is how to reduce dimension for big data, to be used in subsequent prediction and/or inference. In this talk, we take a different viewpoint we show that explicit dimension reduction need not be necessary, one can get good inference/reduction by simply using random dimension reduction or random projection of the original data. We discuss the goodness of the resultant inference in light of theoretical results and empirical evidence. Along the way, we bridge techniques across a diverse array of domains, including, matrix algebra, genetics, economic time-series, machine learning.

Alternating Linearization for Structured Penalties

Minh Pham, SAMSI

Abstract: We adapt the alternating linearization method for proximal decomposition to regularization problems with structured penalties. The method is related to two well-known operator splitting methods, the Douglas-Rachford and the Peaceman-Rachford method, but it has descent properties with respect to the objective function. Its convergence mechanism is related to that of bundle methods of non-smooth optimization. A block coordinate descent method is developed to facilitate fast convergence. The framework is further extended to problems with structured non-convex penalties. We present several numerical studies involving synthetic data, cancer research data, brain imaging and image processing.

Bayesian Nonparametric Functional Models for High-dimensional Genomics Data

Veerabhadran Baladandayuthapani, University of Texas MD Anderson Cancer Center

Abstract: Due to rapid technological advances, various types of genomic, epigenomic, transcriptomic and proteomic data with different sizes, formats, and structures have become available. These experiments typically yield data consisting of high-resolution genetic changes of hundreds/thousands of markers across the whole chromosomal map. Modeling and inference in such studies is challenging, not only due to high dimensionality, but also due to presence of structured dependencies (e.g. serial and spatial correlations). Using genome continuum models as a general principle we present a class of Bayesian methods to model these genomic profiles using functional data analysis approaches. Our methods allow for simultaneous characterization of these high-dimensional functions using non-parametric basis functions, joint modeling of spatially correlated functional data and detection of local features in spatially heterogeneous functional data to answer several important biological questions. We illustrate our methodology by using several real and simulated datasets and propose methods to integrate various types of genomics data as well.

High-dimensional Joint Bayesian Variable and Covariance Selection: Applications in eQTL Analysis and Cancer Genomics

Anindya Bhadra, Purdue University

Abstract: We describe a Bayesian technique to (a) perform a sparse joint selection of significant predictor variables and significant inverse covariance matrix elements of the response variables in a high-dimensional linear Gaussian sparse seemingly unrelated regression (SSUR) setting and (b) perform an association analysis between the high-dimensional sets of predictors and responses in such a setting. To search the high-dimensional model space, where both the number of predictors and the number of possibly correlated responses can be larger than the sample size, we demonstrate that a marginalization-based collapsed Gibbs sampler, in combination with spike and slab type of priors, offers a computationally feasible and efficient solution. We demonstrate our method in an eQTL data set (SNPs as predictors and mRNA as responses) and in a glioblastoma data set (microRNA and copy number aberration as predictors and mRNA as responses). If time permits, we will also describe ongoing work on generalizations to non-linear, non-Gaussian models.

Bayesian Kernel-Based Modeling and Selection of Genetic Pathways and Genes for Cancer

Sounak Chakraborty, University of Missouri-Columbia

Abstract: Much attention has been given to the development of methods that utilize the large quantity of genetic information available in online databases. Recently a new philosophy emerged which considers the genetic pathways, which contain sets of genes, they have a combined effect on a disease. Under this new idea the goal is to identify the significant genetic pathways and the corresponding influential genes and their combined effect towards different diseases. In this article we propose a Bayesian kernel machine model for right censored survival data which incorporates existing information on pathways and gene networks in the analysis of DNA microarray data. Each pathway is modeled nonparametrically using a reproducing kernel Hilbert space. Mixture priors on the pathway indicator variables and the gene indicator variables are assigned. This helps us to model both linear and non-linear pathway effects, pinpoint the important pathways along with the active genes within each pathway. An efficient Markov Chain Monte Carlo (MCMC) algorithm is developed for our model. Simulation studies and a real data analysis, using, van't Veer et al. (2002) breast cancer microarray data, are used to illustrate the proposed method.

Leverage Score Perturbation

Thomas Wentworth, North Carolina State University

Abstract: Leverage scores were introduced in 1978 by Hoaglin and Welsch for outlier detection in statistical regression analysis. Starting about ten years ago, Mahoney et al. pioneered the use of leverage scores for importance based sampling in randomized algorithms for matrix computations. In this talk we present bounds for the sensitivity of leverage scores in terms of both matrix perturbations and principal angles. Our results show that the leverage scores of two matrices, A and B, are close if either the principle angles between them are small, or if A is well conditioned and the relative difference between A and B is small.

Dimensionality Reduction for large-scale Kernel Methods

Alex Gittens, eBay Research

Abstract: Kernel methods like kernel regression or the SVM remain among the most popular statistical/ML tools, but can be expensive to apply on massive datasets characterized by a large number of high-dimensional datapoints. Two complementary randomized approaches have emerged that significantly reduce the cost of applying kernel methods to such massive datasets: Nystrom methods, which form low-rank approximations to the kernel matrix, and random feature methods, which provide low-dimensional approximations of the feature map. This talk presents new results and reviews what is known about their performance (not much in the case of random features) and comments on the future of these methods.

Randomization, Block Splitting and Hybrid Parallelism for Scalable Kernel Methods

Vikas Sindwhani, IBM T.J. Watson Research Center

Abstract: The dramatic success of deep learning techniques on large-scale speech recognition, computer vision and natural language processing tasks emphasizes the significance and potential of marrying ”big models” with ”big data”. With this motivation, I consider the problem of massive-scale training of kernel-based models, via convex optimization, in distributed computing environments. I will describe a new solver for various statistical modeling tasks based on fast randomization techniques for approximating kernel functions via explicit approximate low-dimensional feature maps; while exploiting hybrid parallelism on a cluster of multicore machines.

Robust Speaker Identification Using Gaussian Mixture Models Based on Correlated MFCC-Derived Features

Amita Pal, Indian Statistical Institute

Abstract: Speaker identification using Gaussian Mixture Models (GMMs) based on Mel Frequency Cepstral Coefficients (MFCCs) as features, proposed by Reynolds (1995), is one of the most effective approaches available in the literature. The use of GMMs for modeling speaker identity is motivated by the interpretation that the Gaussian components represent some general speaker-dependent spectral shapes, and the capability of mixtures to model arbitrary densities. In this work, we have established empirically how the well-known principal component transformation, in conjunction with some robust estimation procedures, can be used to enhance significantly the performance of the MFCC-GMM speaker recognition systems, using the benchmark speech corpus NTIMIT.

A Hybrid Approach to Content-Based Image Retrieval

Smarajit Bose, Indian Statistical Institute

Abstract: In Content Based Image Retrieval (CBIR) Systems, features of the query image are matched with those of a candidate image in the database. It has been shown by several researchers in this area that segmentation of the images before matching improves retrieval precision for some image databases.

The features derived from the clusters obtained by segmentation of the query image are matched with those of the clusters from a given candidate image from the database. This approach is not uniformly effective for all types of image databases. In this work, an attempt has been made to combine the basic approach (without segmentation) with the segmentation approach, with the objective of improving retrieval precision. Relevance feedback has been used to boost the performance further. Results of extensive experiments have been presented to illustrate the effectiveness of the proposed approach. (Joint work with A. Pal, J. Mallick and S. Kumar)

Determining the Number of Clusters in a Dataset using ABC

Ilknur Kaynar Kabul, SAS Institute

Abstract: Determining the number of clusters in a dataset, k, is a fundamental problem in unsupervised learning. It is also an important business problem, e.g. in market segmentation. Existing approaches include the silhouette measure, the gap statistic and Dirichlet process clustering. For thirty years SAS procedures have included the option of using the Cubic Clustering Criterion (CCC) to estimate k. While CCC remains among the state-of-the-art, we propose a significant and original improvement to CCC, referred to herein as the Aligned Box Criterion (ABC). The proposed ABC leverages the scalability of SAS' High Performance platform and achieves higher accuracy in the determination of k.

Role of Expert Judgement in Analysing Large Complex Reliability Data Sets

Kevin Wilson, University of Strathclyde

Abstract: Increasing use of sensors and monitoring devices provide a plethora of data on the state of asset condition and performance. Such data can be modelled to support prognostics and inform decisions about future asset operation and maintenance. While data mining of such high dimensional, temporal data allows associations and trends to be detected, we argue that the answers to key engineering questions might not be contained in the empirical data only. Instead data should be augmented with other theoretical models and expert engineering judgment if we are to obtain satisfactory inference. Experts can be used to develop qualitative models explicating cause-effect relationships, thereby providing direction for empirical data analysis and improving efficacy in discriminating between true and spurious relationships. Quantifying subjective probability distributions can be cognitively challenging to experts, but analysis of empirical monitoring data can complement judgemental beliefs by effectively providing a way of testing hypotheses. We develop our ideas by drawing on industry cases relating to, for example, wind farms and vehicles. Viewing the examples from the goal of sound statistical inference, we discuss the value of expert judgement in a data mining context and outline future challenges.

Combining Information to Assess System Reliability

Alyson Wilson, North Carolina State University

Abstract: Reliability is an essential element of system suitability. However, in the current era of smaller budgets and increasing reliability requirements, it is challenging to demonstrate reliability requirements using a single test. This talk illustrates the benefits of using statistical models to combine multiple sources of information in the assessment of system reliability. In addition to demonstrating reliability, combining information from multiple tests can also be used to help with test planning. Although practitioners often use assuring and demonstrating synonymously, statisticians distinguish between reliability demonstration and reliability assurance testing. A traditional reliability demonstration test is essentially a classical hy-

pothesis test, which uses only the data from the current test to assess whether the reliability-related quantity of interest meets or exceeds a requirement. Many modern systems, such as communication devices, transportation systems and defense systems, are highly reliable and extremely complex. For these systems, reliability demonstration tests often require an impractical amount of testing. In response to this dilemma, we can consider an alternative reliability assurance test as one that uses additional supplementary data and information to reduce the required amount of testing. The additional data and information may include appropriate reliability models, earlier test results on the same or similar devices, expert judgment regarding performance, knowledge of the environmental conditions under which the devices are used, benchmark design information on similar devices, prior knowledge of possible failure modes, etc. Developing of this kind of reliability assurance testing strategy requires principled methods for combining information.

Searching for Powerful Supersaturated Designs

David Edwards, Virginia Commonwealth University

Abstract: An important property of any experimental design is its ability to detect active factors. For supersaturated designs, in which model parameters outnumber experimental runs, power is even more critical. In this talk, we consider several popular supersaturated design construction criteria in the literature, propose several of our own, and perform an extensive simulation study to evaluate these construction criteria in terms of power. We use two analysis methods- forward selection and the Dantzig selector- and find that although the latter clearly outperforms the former, most supersaturated design construction methods are indistinguishable in terms of power. This conclusion can be reassuring for practitioners as supersaturated designs can then be sensibly chosen based upon convenience. For instance, the Bayesian D-optimal supersaturated designs can be easily constructed in JMP and SAS for any run size and number of factors. On the other hand, software for constructing E(s^2)-optimal supersaturated designs is not as accessible.

Nested Cone Analysis Methods

Lingsong Zhang, Purdue University

Abstract: Driven by the analysis of constrained data objects, e.g. population of nonnegative data and population of diffusion tensor image data, we propose a novel statistical framework, nested cone analysis. The novel methods directly work with the constraints, and identify a sequence of approximations to the original data by different ranks. These methods provide a nested learning sequence, which properly handles both the complicated constraints and the goodness of fit. Simulations and applications are used to illustrate the usefulness of our methods.

Practical Exploration of Some Statistical and Computational Tools for Predictively Optimal Big Data Analytics

Ernest Fokoué, Rochester Institute of Technology

Abstract: Big data comes in various ways, types, shapes, forms and sizes. Indeed, almost all areas of science, technology, medicine, public health, economics, business, linguistics and social science are bombarded by ever increasing flows of data begging to analyzed efficiently and effectively. In this talk, we propose a rough idea of a possible taxonomy of big data, along with some of the most commonly used tools for handling each particular category of bigness. The dimensionality p of the input space and the sample size n are usually the main ingredients in the characterization of data bigness. The specific statistical machine learning technique used to handle a particular big data set will depend on which category it falls in within the bigness taxonomy. Large p small n data sets for instance require a different collection of tools from the large n small p variety. Among other tools, we discuss preprocessing, standardization, imputation, projection, regularization, penalization, compression, reduction, selection, kernelization, Hybridization, Parallelization, Aggregation, Randomization, Replication, sequentialization. We provide a comparison of the predictive performance of some of the most commonly used methods on a few data sets.

An Alternative to the Bayesian Hierarchical Model and its Application to Microarray Gene Expression Data

Zhaohui (Steve) Qin, Emory University

Abstract: Modern high throughput biotechnologies such as microarray produce massive amount of information for each sample assayed. However, in a typical high throughput experiment, only very limited amount of data are observed for each individual feature (like a probe on the microarray), thus the classical large p, small n problem. Bayesian hierarchical model, capable of borrowing strength across features within the same dataset, has been increasingly recognized as an effective tool in analyzing such data. In this work, we propose an alternative solution to the popular Bayesian hierarchical models. Through simulation and real data analysis, we showed that the proposed approach outperforms hierarchical model-based approaches. This is a collaboration with Ben Li, Qing He, Zhaonan Sun and Michael Yu Zhu.

Screening for Lung Cancer: Does it work, and Who Should Be Screened?

Marek Kimmel, Rice University

Abstract: The efficacy of computed tomography (CT) screening for lung cancer remains controversial despite the encouraging results from the National Lung Screening Trial. We first discuss how a single-arm CT screening data can be used to estimate the mortality reduction using a modeling-based approach to construct a control comparison arm. To estimate the potential lung cancer mortality reduction because of CT screening, a previously developed and validated model was applied to the screening trial to predict the number of lung cancer deaths in the absence of screening. By using age, gender, and smoking characteristics matching those of the trial participants, the model was used to simulate 5000 trials in the absence of CT screening to produce the expected number of lung cancer deaths along with 95% confidence intervals (95% CIs), while adjusting for healthy volunteer bias. The results of the study indicate that CT screening along with early stage treatment can reduce lung cancer-specific mortality. This mortality reduction is greatly influenced by the protocol of nodule follow-up and treatment, and the length of follow-up. Second, we discuss the selection of the right high-risk group for screening. Using computer modeling, we illustrate the tradeoffs resulting from the necessity of optimized allocation of limited means. This is the conundrum currently facing the policy-making bodies.

Predicting Occult Metastasis of Lung Cancer by Modeling the natural History and Detection

Olga Gorlova, Dartmouth School of Medicine

Abstract: The proportion of occult metastasis in clinically undetected non-small cell lung cancers will largely affect the outcome of screening. We developed a mathematical method for simulating the progression and detection of LC at an individual level that allows obtaining the characteristics of LC in a population at the time of diagnosis. LC data from Surveillance, Epidemiology and End Results (SEER) database were used to fit and validate the LC progression and detection model. Detection threshold of CT scan was assumed to determine undetectable metastases at the time of diagnosis. Based on the simulations, 9.3% and 26.0% of patients diagnosed with stage N0M0 respectively had nodal and distant metastases that were not discovered at the time of diagnosis. 56.9% of patients at stage N1M0 had hidden distant metastasis. A high proportion (over 80%) of hidden distant metastases occurred in these patients with a primary tumor smaller than 3 cm. An important observation is the existence of a turning point in the stage distribution of clinically undetected lung cancers, around the primary tumor size of approximately 2 cm. While among the tumors with sizes 1-2 cm, 78% are stage N0M0 tumors, among the tumors with sizes in the 2-3 cm range, only 37% are N0M0. This threshold can be translated into the window of opportunity for the curable disease, which has 0.75 year width. Since tumors detected by CT-screening are drawn from the clinically undetected pool, this window of opportunity is important for evaluation of screening strategies.

Lung Cancer Growth Rate Estimates in the Context of a CT Screening Program

David Yankelevitz, Professor of Radiology at Mount Sinai Medical Center

Abstract: CT screening for lung cancer leads to early diagnosis of lung cancer in each round of screening. Broadly speaking, the types of cancers can be classified as those typically found in the baseline round of screening and those found in subsequent rounds. The types of cancers typically seen in the baseline round tend to be slower growing with greater variability in doubling times while those found in the annual rounds tend to be fast growing. Analogously, the cell types of the cancers found in the baseline rounds are more frequently of the adenocarcinoma subtype (more slow growing) while the cancers found in repeat rounds tend to mimic the cell type distribution found in clinical practice. An understanding of the doubling times of tumors found in the context of screening allows for development of a natural history model and can provide insight into such challenging issues such as overdiagnosis. The I-ELCAP database now has over 700 documented cases of lung cancer where we can calculate doubling times and compare this to the well-known doubling times associated with cancers found in clinical practice. Of particular interest is the subcategory of nonsolid appearing cancers where doubling times can exceed one thousand days and yet these cancers are detected at sizes sometimes exceeding 3 cm. This suggests a complex natural course as these tumors cannot be expected to reach this size with such long doubling times.

A Computationally Efficient Flexible Observed Factor Model with Separate Dynamics for the Factor Volatilities and Correlation Matrix

Sujit Ghosh, North Carolina State University

Abstract: Multivariate stochastic volatility (MSV) models have become important tools in financial econometrics, largely due to the successful utilization of Markov chain Monte Carlo (MCMC) methods. Recent developments in the MSV literature focus on dimension reduction via factor analysis, given that the complexity of computation and the difficulty in model interpretation drastically increase as the dimension of data increases. In order to obtain a flexible yet computationally efficient procedure, we propose a MSV structure with separate dynamics for the volatilities and the correlation matrix. The correlation matrix of the factors is allowed to be time varying, and its evolution is described by an inverse Wishart process. The proposed MCMC method is shown to have some major advantages compared to existing methods for similar models in the literature. For numerical illustrations, we compare the proposed model with other similar multivariate volatility models using Fama-French factors and portfolio weighted return data. The result shows that our model has better predictive performance and computationally more stable. This is joint work with Yu-Cheng Ku, Peter Bloomfield.

Efficient Wind Power Forecasting, Affected by Missing Values

Pilar Muñoz, Universitat Politcnica de Catalunya

Abstract: Often the time series related to wind power generation are affected by a large number of missing values, due, in general, to the poor quality of the sensors in the generators; whereby to obtain precise wind power forecasts is a difficult task if the wind power observations are not obtained regularly over time. The aim of this work is first of all, to impute missing values in wind generation series (Sorjamaa, 2010). Two algorithms are used to impute the missing data in these series, SOM (Self-Organizing Maps) proposed by Kohonen (1995) and EOF (Empirical Orthogonal Function) proposed by Beckers and Rixen (2003). The combination of these two algorithms allows imputing efficiently missing values in those series (Kolding and Stovring (2010). Once the complete wind generation series is obtained, precise prediction is possible.

Clustering Financial Time Series: A Polyspectral SLEX Approach

Priya Kohli, Connecticut College

Abstract: Clustering financial time series data is an important problem. The most widely used clustering methods are based on the assumption that the time series are linear and stationary, although financial time series rarely satisfy these assumptions. We focus on providing statistically and computationally efficient clustering schemes which can handle the challenges arising with nonlinear and/or nonstationary time series. Ombao et al. (2001) described spectral analysis of nonstationary linear time series using the Smooth Localized Complex Exponential (SLEX) library of complex- valued orthogonal transforms which are localized in time and frequency. Harvill et al. (2013) exploited properties of the bispectrum of stationary nonlinear time series to construct a clustering algorithm. We propose an extension for nonstationary, nonlinear time series through a Polyspectral SLEX (or PSLEX) approach. We illustrate our approach using simulated time series and a rich set of financial time series from several industries, with a goal to determining whether the series that move together correspond to companies belonging to the same economic sector. Joint work with Jane L. Harvill, Baylor University and Nalini Ravishanker, Uni- versity of Connecticut.

Dependent Probability Measures for Topic Modelling

Vinayak Yao, Duke University

Abstract: We propose a simple and general nonparametric framework to construct dependent random probability measures by transforming a single underlying Poisson process. When the Poisson process corresponds to a so-called Gamma subordinator, the result is a set of dependent DPs, each associated with a point in an index space, with the property that nearby DPs are more dependent. Our construction allows us to control both the marginal distribution of the probability measures, as well as the dependence across probability measures. We describe Markov chain Monte Carlo inference and consider an application to topic modeling through time of a corpus of documents.

Joint work with Yee Whye Teh.

Semiparametric Bernstein von-Mises theorem: Second Order Studies

Guang Cheng, Purdue University

Abstract: Semiparametric Bernstein von-Mises Theorem has been successfully developed by Bickel and Kleijn (2012) in a general setup among others. This talk mainly focuses on its second order extension with an attempt to figure out the influence of nonparametric Bayesian prior on the semiparametric inference.

Such results provide theoretical insights in constructing a general semiparametric setup, i.e., so-called semiparametric objective prior.

Tensor Factorizations and Sparse Log-linear Models

Anirban Bhattacharya, Texas A&M University

Abstract: Contingency table analysis routinely relies on log linear models, with latent structure analysis providing a common alternative. Latent structure models lead to a low rank tensor factorization of the probability mass function for multivariate categorical data, while log linear models achieve dimensionality reduction through sparsity. Little is known about the relationship between these notions of dimensionality reduction in the two paradigms. We derive several results relating the support of a log-linear model to the nonnegative rank of the associated probability tensor.

Big Data Analytics: An Integrated Business Solution in Wal-Mart

Rida Moustafa, Wal-Mart

Abstract: This talk will provide an overview of Business Analytics in corporate environment that bridges the gap between businesses and data. We will review how analytics may be used to facilitate the information flow from multiple data sources to businesses within a corporate environment. We will illustrate how analytics is applied to provide Preventative, Predictive, and Proactive solutions that could be customized for businesses to boost productivity in diverse industries. We will also show why multidisciplinary teams and diverse resources should be pulled together to enable functionalities and capabilities in analytics to pave the way for businesses to explore opportunities with manageable level of uncertainty.

A Bayesian Approach for Developing Climate Surfaces to Estimate Uncertainty in Daily Weather Interpolations

Rui Zhang, IBM T. J. Watson Research Center

Abstract: Weather conditions, i.e., temperature and humidity, are usually the most important predictors in various energy related forecasting models. However, the weather forecast provided is usually at very course resolutions. Especially, there can be different levels of uncertainty in the observations of the weather conditions. In this paper, we propose a framework to interpolate daily weather conditions data to high resolution surfaces and keep track of the uncertainties introduced by the interpolation. By providing the high-resolution climate metric surfaces and uncertainties, this work facilitates richer and more robust predictive modeling in building energy consumption forecast.

Statistical Modeling of Big and Complex Data in Industrial Applications

Jin Xia, General Electric Global Research Center

Abstract: With the advance of technology in recent years, large amounts of data are collected ubiquitously in academia, industry and government. They often contain complex information about machines, human behaviors, biological experiments, etc. These large and complex data sets create substantial challenges to computing and modeling in data analysis. In this talk, we aim to discuss what challenges big data bring to statistical analysis; promising computation through the Hadoop platform; and propose a statistical modeling framework called divide and recombine. Based on the highly scalable distributed computing environment provided by Hadoop, we can divide the data into subsets, model subsets in parallel, and recombine subset models to achieve statistical modeling on big and complex data sets.

Predictive Healthcare Analytics under Privacy Constraints

Joydeep Ghosh, University of Texas, Austin

Abstract: The move to electronic health records is producing a wealth of information, which has the potential of providing unprecedented insights into the cause, prevention, treatment and management of illnesses. Analyses of such data also promises numerous opportunities for much more effective and efficient delivery of healthcare. However (valid) privacy concerns and restrictions prevent unfettered access to such data. In this talk I will first provide a perspective on the privacy vs. utility trade-off in the context of healthcare analytics. I will then outline two approaches that we have recently and successfully taken that provide privacy-aware predictive modeling with little degradation in model quality despite restrictions on what can be shared or analyzed. The first approach focuses on extracting predictive value from data that has been aggregated at various levels due to privacy concerns, while the second introduces a novel, non-parametric sampler that can generate ”realistic but not real” data given a dataset that cannot be shared as is.

Analyzing Multiview Parliament Networks with Structured Matrix Factorization: Does Leadership Translate to the Twitter Universe?

Shawn Mankad, University of Maryland

Abstract: Do Twitter relations between policy makers predict their real-world interactions? We investigate this question by analyzing multiple Twitter networks that feature different link relations between the Members of Parliament (MPs) in the United Kingdom. Direct approaches that rely only on network statistics for identification of key MPs in the Twitter networks become overwhelmed by the high density of connections, and mixture of weighted and binary link relations. We develop and apply a matrix factorization technique for discovery of MPs that are important for content generation and subsequent transmission in the Twitter-universe. The technique allows the practitioner to emphasize nodes with context-specific local network structures by specifying network statistics that guide the solution. Relying only on link relations, we find that important MPs in Twitter networks are associated with real-world leadership positions, and that rankings from the proposed method are predictive of future media headlines.

Capital Asset Pricing using Horseshoe Prior

Rituparna Sen, Indian Statistical Institute, Chenai Centre

Abstract: A consequence of Fama’s Efficient Market Hypothesis (EMH) is that one cannot consistently achieve return in excess of average market return on risk-adjusted basis. However, since the late 1990 many empirical studies have shown that EMH is not necessarily true. If EMH is true then all the assets in market are always fairly priced, leading to the Capital Asset Pricing Model (CAPM). This turns out to be a testing of hypothesis problem, where null hypothesis is EMH is true or assets are fairly priced versus alternative hypothesis as EMH is false or assets are over/under-priced. As there are thousands of assets, this problem turns out to be a multiple testing problem. We develop a Bayesian multiple testing procedure with the Horseshoe prior for the CAPM. We present the back-testing (aka out-of-sample performance) of the method for 500 stocks that are considered in S&P 500 index for the period from 2008 to 2013.

Estimation in High-dimensional Vector Autoregressive Models

George Michailidis, University of Michigan

Abstract: Vector Autoregression (VAR) is a widely used method for learning complex interrelationship among the components of multiple time series. Over the years it has gained popularity in the fields of control theory, statistics, economics, finance, genetics and neuroscience. We consider the problem of estimating stable VAR models in a high-dimensional setting, where both the number of time series and the VAR order are allowed to grow with sample size. In addition to the "curse of dimensionality" introduced by a quadratically growing dimension of the parameter space, VAR estimation poses considerable challenges due to the temporal and cross-sectional dependence in the data. Under a sparsity assumption on the model transition matrices, we establish estimation and prediction consistency of 1-penalized least squares and likelihood based methods. Exploiting spectral properties of stationary VAR processes, we develop novel theoretical techniques that provide deeper insight into the effect of dependence on the convergence rates of the estimates. We study the impact of error correlations on the estimation problem and develop fast, parallelizable algorithms for penalized likelihood based VAR estimates.

Classification with Unstructured Predictors with an Application to Sentiment Analysis

Junhui Wang, City University of Hong Kong

Abstract: Unstructured data refers to information that lacks certain structures and cannot be organized in a predefined fashion. Unstructured data involve heavily on words, texts, graphs, objects or multimedia types of files that are difficult to process and analyze by traditional computational tools and statistical methods. In this talk, I will discuss ordinal classification with unstructured predictors and ordered class categories, where imprecise information concerning strengths between predictors is available for predicting the class labels. We integrate the imprecise predictor relations into linear relational constraints over classification function coefficients, where large margin ordinal classifiers are introduced, subject to quadratically many linear constraints. The proposed methods are implemented via a scalable quadratic programming algorithm based on sparse word representations. The advantage is demonstrated in a variety of simulated experiments as well as one large-scale sentiment analysis example on TripAdvisor.com customer reviews. If time permits, the asymptotic properties will also be discussed, which confirm that utilizing relationships among unstructured predictors can significantly improve prediction accuracy.

Estimation of a Directed Acyclic Gaussian Graph

Xiaotong Shen, University of Minnesota

Abstract: Directed acyclic graphs are widely used to describe, among interacting units, causal relations. Causal relations are estimated by reconstructing a directed acyclic graph's structure, presenting a great challenge When the unknown total ordering of a DAG needs to be estimated. In such a situation, it remains unclear if a graph's structure is reconstructable in the absence of an identifiable likelihood with regard to graphs, and in facing super-exponentially many candidate graphs in the number of nodes. In this talk, I will introduce a global approach to identify all estimable causal directions as well as model parameters jointly. This approach uses constrained maximum likelihood with nonconvex constraints reinforcing the non-loop requirement to yield an estimated directed acyclic graph, where super-exponentially many constraints characterize the major challenge. Computationally, we develop a reduction method that constructs a set of active constraints from the super-exponentially many constraints, which turn out to be of a cubic-polynomial order. This permits efficient computation. Theoretically, the proposed method has desirable the statistical properties with respect to reconstruction of identifiable directions of the true graph. Finally, some illustrative examples will be given. This work is joint with Y. Yuan, W. Pan and Z. Wang.

Tuesday Morning

Big Bayes

David Dunson, Duke University

Abstract: Bayesian methods have great promise in big data sets, but this promise has not been fully realized due to the lack of scalable computational methods. Usual MCMC and SMC algorithms bog down as the size of the data and number of parameters increase. For massive data sets, it has become routine to rely on penalized optimization approaches implemented on distributed computing systems. The most popular scalable approximation algorithms rely on variational Bayes, which lacks theoretical guarantees and badly under-estimates posterior covariance. Another problem with Bayesian inference is the lack of robustness; data contamination and corruption is particularly common in large data applications and cannot easily be dealt with using traditional methods. Motivated by a variety of interesting applications in computational advertising, ecology, finance and neurosciences, I provide a biased overview of recent developments in scaling up and robustifying Bayes in big and complex data settings. An emphasis will be on nonparametric Bayes modeling approaches, developing new models and computational algorithms tailored for big data settings.

A Posterior Predictive Approach to Process Conformance Optimization with Complex Multivariate Models

John Peterson, GlaxoSmithKline

Abstract: Quality improvement has been described succinctly as ”reduction in variation about a target”. This is intimately related to processes with high capability of meeting quality specifications. In particular for (stable and reproducible) processes characterized by multiple response types, we desire the multivariate distribution of response types to be tight around a vector quality target. Because of this, we want to optimize the probability of process conformance, not just mean responses. Since 2001, several researchers have investigated this quality capability approach to multiple-response process optimization, most recently Alshraideh and del Castillo (2013). A challenge for modeling multiple-response processes is that the ability to measure many different quality responses is growing much faster than the cost of executing experimental runs. In addition, difficulties such as jointly censored multivariate responses pose further modeling and analysis difficulties. As stated by Little (2013) in his JASA paper dedicated to the memory of George Box, “Make outcomes univariate (when it makes sense to do so).” Taking this advice, I modify a (univariate) desirability function due to Kim and Lin (2000) and use it as a basis for modeling a predictive distribution to optimize the probability of process conformance for complex multiple-response processes. I illustrate the methodology with two examples from pharmaceutical manufacturing.

Scan Statistics for Detection of Genome Structural Variation

Nancy R. Zhang, University of Pennsylvania

Abstract: Scan Statistics for Detection of Genome Structural Variation Structural variation, which includes deletion, insertion, and inversion of stretches of DNA, comprise an important class of genome variation in the human population, and are implicated in many diseases. High throughput paired end short read sequencing allows for genome-wide detection of a wide spectrum of structural variation. We develop a general model for this data, based on a Poisson random field, under which signals that are characteristic for each type of structural change can be modeled using a likelihood based framework. Scan Statistics

derived from the model integrate information from coverage, insert length, and other aspects of the data, and thus has improved sensitivity over methods that only utilize any single feature. We also describe how to control the false discovery rate for scan statistics of Poisson random fields, and illustrate our methods on 1000 genomes data. This is joint work with David Siegmund and Benjamin Yakir.

Resampling Approximations to the Distributions of Scan Statistics

Soumen N. Lahiri, North Carolina State University

Abstract: Scan statistics play an important role in many areas of science and technology in the context of analyzing the occurrence of observed clusters of events in time and space. Exact distributions of scan statistics are difficult to determine and often the associated computation becomes infeasible due to complexity of the problem. In this talk, we present a general approach based on resampling methods to derive approximations to the distributions of scan statistics. Under some regularity conditions, we establish asymptotic validity of the proposed method. Numerical results are presented to illustrate accuracy of the approximation in finite samples.

Distribution of Scan Statistics over Hidden States

Donald E.K. Martin, North Carolina State University

Abstract: In classification, it is sometimes important to determine whether a clustering of a particular category is statistically significant. Such clustering could indicate a change in the underlying process that should be noted. Examples include biosequences, where the clustering could be due to a biological function, and syndromic surveillance, where clustering could indicate a disease hotspot. Scan statistics are frequently used to detect clumping, while avoiding problems associated with multiple testing. We compute the distribution of scan statistics over hidden states of undirected and directed graphical models that are represented by conditional random fields and their corresponding factor graphs. The methods are relevant for graphs with a sparseness of edges that allows exact computation. Distributions are obtained by including matrix operators in messages of the sum-product algorithm so that a vector that indicates the statistics value is sequentially updated while computing associated probabilities.

Discovering Novel Anomalous Patterns in General Data

Edward McFowland III, Carnegie Mellon University

Abstract: We propose Discovering Novel Anomalous Patterns (DAP), a new method for continual and automated discovery of anomalous patterns in general datasets. Currently, general methods for anomalous pattern detection attempt to identify data patterns that are unexpected as compared to normal system behavior. We propose a novel approach for discovering data patterns that are unexpected given a profile of previously known, both normal and abnormal, system behavior. This enables the DAP algorithm to identify previously unknown data patterns, add these newly discovered patterns to the profile of known system behavior, and continue to discover novel (unknown) patterns. We evaluate the performance of DAP in two domains of computer system intrusion detection (network intrusion detection and masquerade detection), demonstrating that DAP can successfully discover and characterize relevant patterns for these two tasks. As compared to the current state of the art, DAP provides a substantially improved ability to discover novel patterns in massive multivariate datasets.

Methods and Models for Interpretable Linear Classification

Berk Ustun, MIT

Abstract: We present a comprehensive approach to create accurate and interpretable linear classification models using mixed-integer programming. Our approach can produce models that incorporate many interpretability-related qualities, and that strike a user-defined balance between accuracy and interpretability. We use our approach to train personalized scoring systems and M-of-N rule tables for applications in medicine, marketing and crime prediction. In addition, we describe methods to train interpretable models on large-scale datasets.

L1-Norm Prinicipal Component Analysis

Paul Brooks, Virginia Commonwealth University

Abstract: Principal component analysis (PCA) may be viewed in terms of optimization as finding a series of best-fit subspaces. Traditional PCA is based on using the L2 norm to measure distances of points to the fitted subspaces and can be sensitive to outlier observations. Several robust approaches based on the L1 norm have been proposed, including methods that estimate best-fitting L1-norm subspaces. In this talk, we review progress on the L1-norm best-fit hyperplane problem and the L1-norm best-fit line problem. We introduce methods for deriving solutions to these problems that can be used for robust principal component analysis.

A General Framework for Mixed Graphical Models

Yulia Baker, Rice University

Abstract: Markov Random Fields, or undirected graphical models are widely used to model high-dimensional multivariate data. Classical instances of these models, such as Gaussian Graphical and Ising Models, as well as recent extensions to graphical models specified by univariate exponential families, assume all variables arise from the same distribution. Complex data from high-throughput genomics and social networking for example, often contain a mixture of discrete, count, and continuous variables measured on the same set of samples. To model such heterogeneous or mixed data, we develop a novel class of mixed graphical models by specifying that each node-conditional distribution is a member of a possibly different univariate exponential family. Additionally, we show how this class of models can be further generalized using heterogenous conditional random fields to yield flexible classes of joint distributions over mixed variables. We study several instances of these models, and propose neighborhood selection estimators for recovering the underlying network structure. Simulations as well as an application to learning mixed genomic networks from next generation sequencing and mutation data demonstrate the versatility of our methods. Joint work with Pradeep Ravikumar, Eunho Yang, Yulia Baker and Zhandong Liu.

Properties of Optimizations Used in Penalized Gaussian Likelihood Inverse Covariance Matrix Estimation

Adam Rothman, University of Michigan

Abstract: We establish necessary and sufficient conditions for the existence of inverse covariance matrix estimates obtained by minimizing the negative Normal log-likelihood plus a weighted ridge or weighted L1 penalty. A new algorithm to solve this optimization with the weighted ridge penalty is developed and its convergence is established. This algorithm combines the majorize minimize principle with minorize minimize acceleration attempts. Numerical experiments show this algorithm is superior to its only competitor and that ridge penalization is useful within quadratic discriminant analysis.

Maximum Likelihood Network Estimates from Social Grouping Behavior

Yunpeng Zhao, George Mason University

Abstract: Within the field of network analysis, there is often an important distinction made between physical networks (i.e. highways, router systems, and electrical grids) and social networks (i.e. friendships, movie actors, and citations). In physical networks, the network topology is observable and analysis of the network properties directly informs the means by which the network functions. However, in social networks, the network topology is not explicit and must be inferred from the observed behavior. This effort is often complicated by the use of heuristic techniques for network inference which are not capable of reproducing the original behavior. In this presentation, the authors define a network based model to describe the social grouping behavior and present a maximum likelihood technique for inferring the network most likely to have produced the observed behavior.

Updating Sequential Probability Ratio Test for Real-Time Surveillance of Vaccine Safety

Tom McCurdy, Stanford University and Acumen, LLC

Abstract: Real-time data surveillance plays a vital role in the timely detection of adverse events associated with newly-introduced medical products, but delays in data acquisition create challenges in analyzing surveillance data. This study presents the testing methods and results of a real time surveillance of the safety of immunization by influenza vaccine as it was conducted for the 2008-09 flu season using the Medicare database. The analysis exploits the properties of sample structure and devises the Updating Sequential Probability Ratio Test (USPRT) that accounts for delays in data accrual encountered in the real-time monitoring of the safety of new vaccines. A total delay distribution is modeled as a convolution of the clinical delay distribution associated with the passage of time during at-risk windows, and the processing delay distribution associated with administrative processing delays in reporting health events in medical databases. We propose simulation methods to construct the distribution of the USPRT statistic and its critical values implied by alternative alpha-spending plans. The analysis demonstrates the power properties of the USPRT test and its superior performance over standard SPRT testing procedures in real-time surveillance settings. The empirical findings reveal that USPRT offers a timely and reliable approach for the identification of safety signals.

Clustering Mixed Data Subject to Measurement Error

Aliza Heching, IBM T.J. Watson Research Center

Abstract: In spite of the existence of a large number of clustering algorithms, clustering remains a difficult problem. We consider the problem of clustering mixed numeric and categorical data with measurement error. This problem is often encountered in practice, as very large datasets become increasingly common in a number of different domains and clustering algorithms must be applied to heterogeneous sets of error-prone variables. Examples include labor claiming records for resources deployed on service delivery projects, patient health records, and applications arising in economics.

In this talk, we will highlight the importance of considering measurement error in data clustering. Based on the results of a literature review, theoretical analysis, and Monte Carlo simulations, we will discuss and illustrate (1) the effect of measurement error on various geometric distances that are commonly used in clustering algorithms, (2) how to construct distance metrics that accurately combine continuous and categorical variables, (3) how to select optimal clustering methods, (4) the impact of measurement error models commonly encountered in industry, and (5) how to effectively cluster error-prone mixed data with replicate measurements.

We focus on hierarchical, partitioning, and model-based clustering approaches that are most commonly used in industry, and draw comparisons based on performance on mixed data, robustness to measurement error, and scalability. Specific characteristics considered include sample size, number of continuous/categorical variables, the distribution of the measurement error, severity of measurement error, and the number of underlying populations.

The Joint Analysis of Genomic and Pharmacological Data: A Novel Framework in Development

Ray Liu, Takeda International

Abstract: Drug target discovery involves a variety of complex data sources. Genome wide expression data, metabolomics data, and drug sensitivity profiles are examples of commonly used data. Traditional analysis methods consider each type of data one at a time, but pooling the data could reveal new information. Here we propose a new statistical method using tensor factorization and Bayes theorem for the joint modeling of various data sources. This model enables incorporation of prior knowledge on the associations between genes, drugs and pathways through the Bayesian sparse model set-up. The model has multiple usages, including the prediction of novel drug targets. Performance of this model is evaluated via various simulations.

Mining for Interactions Using Convex Optimization

Jacob Bien, Cornell University

Abstract: Searching for interactions between variables in a two-class setting is challenging in datasets where there are a large number of variables. A common approach is to use marginal (i.e., main effect) information to greatly reduce the number of variables and then search for interactions among the remaining variables. Such an approach imposes a very strong ”hierarchy” assumption that can backfire when it does not hold. We develop a testing procedure that incorporates the hierarchy assumption in a gentler way, allowing main effects to guide the search for interactions without blinding the procedure to strong interactions that lack strong main effects. Our test statistic is based on the solution path of a recently developed lasso-like procedure for building interaction models. The test statistic can be computed in closed-form and a procedure for controlling its false discovery rate is proposed.

Laplacian Shrinkage for Estimation of Inverse Covariance Matrices in Heterogeneous Populations

Ali Shojaie, University of Washington

Abstract: We introduce a general framework, using a Laplacian shrinkage penalty, for estimation of inverse covariance matrices from heterogeneous populations. This framework generalizes previously proposed methods for joint estimation of multiple graphical models to the analysis of observations from nonexchangeable populations with complex structures, including hierarchal relationships among subpopulations. We propose an efficient alternating direction method of multiplier (ADMM) algorithm for parameter estimation, and establish both variable selection and norm consistency of the estimator for distributions with exponential or polynomial tails. Finally, we discuss the selection of the Laplacian shrinkage penalty based on hierarchical clustering, in the settings where the true hierarchy is unknown, and discuss conditions under which this data driven choice results in consistent estimation of precision matrices. Extensive numerical studies and applications to gene expression data from subtypes of cancer with distinct clinical outcomes indicate the potential advantages of the proposed method over existing approaches.

Pathwise Calibrated Coordinate Descent Algorithm for Large-Scale Semiparametric Graph Estimation Problems: Nonconvexity with Theoretical Guarantees

Han Liu, Princeton University

Abstract: The pathwise coordinate descent strategy combined with the warm start and active set tricks is arguably one of the most popular optimization methods for estimating high dimensional graphical models. In particular, it is conceptually simple, easy to implement, and applicable to a wide range of convex and nonconvex problems. However, there is still a gap between its theoretical justification and practical success: for high dimensional convex problem, existing theory only shows sublinear rates of convergence; for nonconvex problems, almost no theory on the rates of convergence exists. To bridge this gap, we propose a new unified computational framework named PICASA (Pathwise Calibrated Active Shooting Algorithm) for pathwise coordinate optimization. Compared to the existing pathwise coordinate optimization strategies, the main difference of PICASA is that we exploit a proximal gradient pilot to identify an active set. Such a modification, though simple, has profound impact: with high probability, the PICASA method attains a global geometric rate of convergence to the oracle solution for optimizing a large family of convex and even nonconvex problems. Unlike most existing analysis which assumes all the computation can be carried out exactly without worrying about numerical precision, our theory explicitly counts the numerical computation accuracy and thus is more realistic. The PICASA framework is quite general can be combined with different coordinate descent optimization strategies, including cyclical descent, stochastic descent, and greedy descent. As an application, we apply this strategy on a family of nonconvex optimization problems motivated by estimating semiparametric graphical models. The PICASA method allows us to obtain new statistical recovery results on both parameter inference and graph estimation consistency, which do not exist in the existing literature. Thorough numerical results are also provided to back up our theoretical arguments.

Tuesday Afternoon

Component-Based Redundancy Path Modelling

Vincenzo Esposito Vinizi, ESSEC

Abstract: David Banks regrets that he was unable to find this abstract amongst his emails.

Statistical Modelling of Bidding Prices in Online Ad Position Auctions

Xiaoming Huo, Georgia Institute of Technology

Abstract: Ad position auctions are being held all the time in nearly all web search engines, and have become the major source of revenue in online advertising. We study statistical models of the bidding prices. Two approaches are explored: (1) a game theoretic approach that characterizes bidders behavior, and (2) a statistical generative approach, which aims at mimicking the fundamental mechanism underlying the bidding process. We compare/contrast these two approaches, and describe how auctioneer can take advantage of the obtained knowledge.

Interactions Between Machine Learning and Decision Making

Cynthia Rudin, MIT

Abstract: Predictions are often used within decision making problems (optimization problems) for creating policies about the future. In this talk I will discuss ways in which prediction and decision-making interact. In particular:

- Can prior knowledge about the outcome of the decision problem be used to create better predictions? I will quantify the answer in terms of generalization bounds.
- Can prior knowledge about the outcome of the decision problem be used to create better decisions from the predictions? I will provide a regularized learning method to do this.
- Can we learn how to make our decision robust? I will present a method for estimating the uncertainty set for robust optimization.

This is joint work with Theja Tulabandhula.

Model-free Variable Selection

Yuexiao Dong, Temple University

Abstract: Novel procedures for model-free variable selection are introduced under the sufficient dimension reduction paradigm. We remove the linear model assumption and extend the classical stepwise regression algorithm to handle nonlinear link functions. In the challenging setting with ultrahigh dimensional predictors, the newly proposed test statistics naturally leads to marginal utilities for screening, which serves as an initial reduction step by removing most irrelevant predictors. The algorithms can also be tailor made to either target all active predictors or only the active predictors for the regression mean function.

High-dimensional Ordinary Least-squares Projector for Screening Variables

Chenlei Leng, University of Warwick

Abstract: Variable selection is a challenging issue in many statistical applications when the number of predictors p far exceeds the number of observations n. In this ultra-high dimensional setting, Fan and Lv (2008) introduced the sure independence screening (SIS) procedure that can significantly reduce the dimensionality while preserving the true model with overwhelming probability, before a refined second stage analysis. However, the aforementioned sure screening property strongly relies on the assumption that the important variables in the model should have large marginal correlations with the response, which rarely holds in reality. Motivated by these concerns, we propose a novel and simple screening technique called the high-dimensional ordinary least-squares projector (HOLP) for high dimensional features. We show that HOLP possesses the sure screening property and gives consistent variable selection without the strong assumption, and has a low computational complexity. Simulation study shows that HOLP performs competitively compared to many other marginal correlation based methods including (iterative) SIS, forward regression and tilting. An application to a mammalian eye disease data illustrates the attractiveness of HOLP.
This is joint work with Xiangyu Wang.

Component Selection and Estimation for Functional Additive Models

Helen Zhang, University of Arizona

Abstract: Functional additive model provides a flexible yet simple framework for regressions involving functional predictors. The utilization of data-driven basis in an additive rather than linear structure naturally extends the classical functional linear model. However, the critical issue of selecting nonlinear additive components has been less studied. In this work, we propose a new regularization framework for joint component selection and estimation in the context of the Reproducing Kernel Hilbert Space. The proposed approach takes advantage of the functional principal components which greatly facilitates the implementation and the theoretical analysis. The selection and estimation are achieved by penalized least squares using a penalty which encourages the sparse structure of the additive components. Theoretical properties, such as the existence and the rate of convergence are investigated. The empirical performance is demonstrated through simulation studies and a real data application.

Linda Zhao, University of Pennsylvania

Abstract: We consider a high dimensional normal mean problem. Nonparametric Empirical Bayes solutions have been proposed and studied previously but mainly under the L2 loss. We extend the results and focus on the posterior inference for the unknown

High Dimensional Tests for Brain Networks with Desirable Resolutions

Jichun Xie, Temple University

Abstract: Large-scale resting-state fMRI studies have been conducted for patients with autism, and the existence of abnormalities in the functional connectivity between brain regions (containing more than one voxel) have been clearly demonstrated. Due to the ultra-high dimensionality of the data, current methods focusing on studying the connectivity pattern between voxels are often lack of power and computation-efficiency. In this talk, we introduce a new framework to identify gigantic network with desired resolution. We propose three procedures based on different network structures and testing criteria. The asymptotical null distributions of the test statistics are derived, together with its rate-optimality. Simulation results show that the tests are able to control type I error and familywise error rate, and yet very powerful. We apply our method to a resting-state fMRI study on autism. The analysis yields interesting insights about the mechanism of autism.

False Discovery Control under General Dependence

Xu Han, Temple University

Abstract: Multiple hypothesis testing is a fundamental problem in high dimensional inference, with wide applications in many scientific fields. In genome-wide association studies, tens of thousands of hypotheses are tested simultaneously to find if any genes are associated with some traits. In practice, these tests are correlated. False discovery control under arbitrary covariance dependence is a very challenging and important open problem in the modern research. In this talk, we extend our principal factor approximation approach (PFA) that was designed for the known covariance matrix case to a more general situation where the covariance dependence is unknown. In practice, this unknown dependence has to be estimated first, and the estimation accuracy can greatly affect the convergence of FDP or even violate its consistency. We will give conditions on the dependence structures and estimation procedures such that the estimate of FDP is consistent. Such dependence structures include sparse covariance matrices and strong dependence matrices, which encompass most practical situations. The finite sample performance of our procedure is critically evaluated by various simulation studies. Our approach is further illustrated by some real data in genome-wide association studies.

Consistency of Co-Clustering Exchangeable Graph Data

David Choi, Carnegie Mellon University

Abstract: We analyze the problem of partitioning a 0-1 array or bipartite graph into subgroups (also known as co-clustering), under a relatively mild assumption that the data is generated by a general nonparametric process. This problem can be thought of as co-clustering under model misspecification; we show that the additional error due to misspecification can be bounded by $O(n^{(} - 1/4))$. Our result suggests that under certain sparsity regimes, community detection algorithms may be robust to modeling assumptions, and that their usage is analogous to the usage of histograms in exploratory data analysis.

Fast Hierarchical Modeling for Recommender Systems

Patrick Perry, New York University

Abstract: In the context of a recommender system, a hierarchical model allows for user-specific tastes while simultaneously borrowing estimation strength across all users. Unfortunately, existing likelihood-based methods for fitting hierarchical models have high computational demands, and these demands have limited their adoption in large-scale prediction tasks. We propose a moment-based method for fitting a hierarchical model, which has its roots in a method originally introduced by Cochran in 1937. The method trades statistical efficiency for computational efficiency. It gives consistent parameter estimates, competitive prediction error performance, and dramatic computational improvements.

Contextualized Spectral Network Analysis

Norbert Binkiewicz, University of Wisconsin-Madison

Abstract: Many modern data sets involve relationships between interacting units that can be intuitively represented in the form of a graph or a network. Many graphs contain underlying structure that is informative to the nature of the process that generated the graph. One type of underlying structure is clusters (aka communities, modularities, blocks). Spectral clustering is a popular and computationally efficient method of discovering such communities. In numerous applications, graph data is also accompanied by covariate measurements on each node. We develop a novel approach called covariate assisted spectral clustering (CASC), which can utilize node specific covariates to help uncover latent communities in a graph. We give theoretical results for CASC under the Stochastic Block Model, including a bound on the mis-clustering rate. In simulations, we demonstrate that under most conditions CASC yields superior results relative to either a modified canonical correlation analysis or vanilla spectral clustering that ignores the node covariates. We then apply CASC to brain graphs derived from DTI data, using the node locations as covariates. The results demonstrate that CASC gives more spatially coherent clusters than spectral clustering and has the potential of yielding neurologically meaningful clusters.

Analyzing Neurological Disorders Using Functional and Structural Brain Imagining Data

Brian Caffo, Johns Hopkins University

Abstract: In this talk we overview methodology for predicting clinical outcomes, and especially neurological disorders, using functional and structural brain imaging data. We focus on resting state functional connectivity data via fMRI as well as structural imaging data via T1 MRI and diffusion weighted MRI. We consider these modalities and variety of methods for feature extraction, prediction and analysis. We apply the methodology to developmental disorders, particularly attention deficit hyperactivity, cognitive impairment and Alzheimers disease and multiple sclerosis.

Multiscale Weighted Principal Component Analysis for High-Dimensional Data on Graphs

Hongtu Zhu, University of North Carolina-Chapel Hill

Abstract: The aim of this paper is to develop a multiscale weighted principal component analysis (MWPCA) method for the use of high dimensional data on graphs $\sigma = \{G, E\}$ defined by their domain of vertexes G and edge system E to predict a low-dimensional outcome variable. MWPCA has three key features: being adaptive, being hierarchical, and being supervised. In MWPCA, we introduce two sets of weights including importance score weights for the selection of individual features at each node of G and spatial weights for the incorporation of the neighboring pattern of E on the graph $\sigma = \{G, E\}$. We develop a three-stage algorithm to adaptively determine weights and compute principal components for MWPCA. We systematically investigate the theoretical properties of MWPCA under some mild conditions. We demonstrate the utility of our methods through simulations and a case study on Alzheimer's disease neuroimaging initiative data.

A Dynamic Directional Model for Effective Brain Connectivity using Electrocorticographic (ECoG) Time Series

Tingting Zhang, University of Virginia

Abstract: We introduce a dynamic directional model (DDM) for studying brain effective connectivity based on intracranial electrocorticographic (ECoG) time series. The DDM consists of two parts: a set of differential equations describing neuronal activity of brain components (state equations), and observation equations linking the underlying neuronal states to observed data. When applied to functional MRI or EEG data, DDMs usually have complex formulations and thus can accommodate only a few regions, due to limitations in spatial and/or temporal resolution of these imaging modalities. In contrast, we formulate our model in the context of ECoG data. The combined high temporal and spatial resolution of ECoG data result in a much simpler DDM, allowing investigation of complex connections between many regions. To identify functionally-segregated sub-networks, a form of biologically economical brain networks, we propose the Potts model for the DDM parameters. The neuronal states of brain components are represented by cubic spline bases and the parameters are estimated by minimizing a log-likelihood criterion that combines the state and observation equations. The Potts model is converted to the Potts penalty in the penalized regression approach to achieve sparsity in parameter estimation, for which a fast iterative algorithm is developed. An L1 penalty is also considered for comparison. The methods are applied to an auditory ECoG data set.

Weighted principal support vector machines for sufficient dimension reduction in binary classification

Seung Jun Shin, University of Texas/MD Anderson Cancer Center

Abstract: David Banks regrets that he was unable to find this abstract amongst the email.

Talent Analytics to Predict Employee Job Roles and Skill Sets using Diverse Data Sources

Kush R. Varshney, IBM Thomas J. Watson Research Center

Abstract: David Banks regrets that he was unable to find this abstract amongst the email.

Sequential Neutral Zone Classifiers

Daniel Jeske, University of California-Riverside

Abstract: David Banks regrets that he was unable to find this abstract amongst the email.

Group Regularized Estimation Under Strong Hierarchy

Yiyuan She, Florida State University

Abstract: In many high-dimensional models involving interaction effects, statisticians usually favor variable selection obeying certain logical hierarchical constraints. The talk focuses on strong hierarchy which means that the existence of an interaction term implies that both associated main effects must be present. Although lately the hierarchical lasso has been proposed, the existing computational algorithms converge quite slow and cannot meet the challenge of big data. Moreover, the literature of finite-sample studies of such estimators is extremely scarce, largely due to the difficulty that multiple sparsity-promoting penalties are enforced on the same subject. The talk investigates a new type of estimators based on group multi-regularization for hierarchical variable selection. Some nonasymptotic results are presented, together with the minimax optimal rate. A general-purpose algorithm is developed with a theoretical guarantee of strict iterate convergence. It is extremely scalable and meets the needs of big data computation. This is joint work with He Jiang.

Shrinkage Priors in High Dimensions

Debdeep Pati, Florida State University

Abstract: Shrinkage priors are routinely used as alternative to point-mass mixture priors for sparse modeling in high-dimensional applications. The question of statistical optimality in such settings is under-

studied in a Bayesian framework. We provide theoretical understanding of such Bayesian procedures in terms of two key phenomena: prior concentration around sparse vectors and posterior compressibility. We demonstrate that a large class of commonly used shrinkage priors lead to sub-optimal procedures in high-dimensional settings. As a remedy, we propose a novel shrinkage prior that leads to optimal posterior concentration. A novel sampling algorithm for our proposed prior is devised and illustrations are provided through simulation examples and an image-denoising application. Extension to massive covariance matrix estimation is discussed.

Bayesian Analysis for Partially Identified Models

Yuan Liao, University of Maryland, College Park

Abstract: Bayesian partially identified models have received a growing attention in recent years in econometrics, due to their broad applications in empirical studies. We contribute to this literature by proposing a novel (two-sided) Bayesian credible set (BCS) for the identified set which is easy to compute. We show that, while the BCS and frequentist confidence set (FCS) for the partially identified parameter are asymptotically different, our BCS for the identified set is asymptotically a FCS. The proposed BCS is constructed based on the support function of the identified set. The Bernstein-von Mises theorem for the posterior distribution of the support function is proved. Our Bayesian procedure is semi-parametric and places a nonparametric prior on the unknown likelihood function but also works when the likelihood is known up to a finite dimensional parameter. Thus, our approach only requires a set of moment conditions, so that it is robust to the risk of likelihood misspecification, and still possesses a pure Bayesian interpretation. A frequentist validation of our procedure is also provided. Finally, the proposed method is illustrated in a financial asset pricing problem.

Geometric Ergodicity of Gibbs Samplers for Bayesian Mixed Models

Jorge Román, Vanderbilt University

Abstract: Due to advances in Markov chain Monte Carlo (MCMC) methods, the use of Bayesian statistical models in the applied sciences has increased dramatically over the last decade. MCMC methods allow for the estimation of intractable quantities associated with complex posterior distributions. However, in a large number of applications, MCMC-based estimates are reported without a valid measure of their quality. This is largely due to the fact that establishing central limit theorems (CLTs), which allow for the computation of valid asymptotic standard errors, is typically challenging. In this talk, we consider several Bayesian versions of linear mixed models and discuss recent results that provide simple sufficient conditions for the geometric ergodicity of the associated Gibbs samplers. These results are important from a practical standpoint because they provide a simple way of proving the existence of CLTs that allow for the computation of valid asymptotic standard errors for the estimates computed using the Gibbs sampler.

Estimates and Standard Errors for Ratios of Normalizing Constants from Multiple Markov Chains

Aixin Tan, University of Iowa

Abstract: In the classical biased sampling problem, we have k densities $\pi_1, \ldots, \pi_k$, each known up to a normalizing constant, i.e. for $l = 1, \ldots, k, \pi_l = \nu_l/m_l$, where ν_l is a known function and m_l is an unknown constant. For each l, we have an iid sample from π_l, and the problem is to estimate the ratios m_l/m_s for all l and all s. This problem arises frequently in several situations in both frequentist and Bayesian inference. An estimate of the ratios was developed and studied by Vardi and his co-workers over two decades ago, and since then there has been much subsequent work on this problem from many different perspectives. In spite of this, there are no rigorous results in the literature on how to estimate the standard error of the estimate. In this paper we present a class of estimates of the ratios of normalizing constants that are appropriate for the case where the samples are not iid sequences, but are Markov chains. We also develop an approach based on regenerative simulation for obtaining standard errors for the estimates of ratios of normalizing constants. These standard error estimates are valid for both the iid case and the Markov chain case.

Efficient estimation and prediction for spatial generalized linear mixed models

Vivekananda Roy, Iowa State University

Abstract: Markov chain Monte Carlo (MCMC) methods are often used for parameter estimation and prediction in model-based geostatistics. Due to complexity of the spatial models, often some ad-hoc methods like discretization are used for MCMC sampling from the corresponding posterior distributions. Here we consider a principled approach based on MCMC and efficient importance sampling methods for parameter estimation and model selection in spatial generalized linear mixed models. We introduce a spatial robit model for binomial data and illustrate our methodology in the context of this model. The techniques that we use here can also be applied to other types of geostatistical models. We present results from some simulation studies and a real data application involving prediction of a map of disease severity for a particular plant root disease.

Using Targeted Maximum Likelihood Estimation To Estimate The Impact Of Online Advertising

Ori Stitelman, Dstillery

Abstract: This talk will examine different approaches for estimating the causal effect of display advertising on browser post-view conversion (i.e. visiting the site after viewing the ad rather than clicking on the ad to get to the site). The effectiveness of online display ads beyond simple click-through evaluation is not well established in the literature. Are the high conversion rates seen for subsets of browsers the result of choosing to display ads to a group that has a naturally higher tendency to convert, or does the advertisement itself cause an additional lift? How does showing an ad to different segments of the population affect their tendencies to take a specific action, or convert? An approach for assessing the effect of display advertising on customer conversion that does not require the cumbersome and expensive setup of a controlled experiment, but rather uses the observed events in a regular campaign setting. The advantages of using Targeted Maximum Likelihood Estimation (TMLE) to efficiently estimate the impact of advertising will be highlighted. The presented approach can be applied to many additional types of causal questions in display advertising.

An SMS Text Classification System for UNICEF Uganda

Rick Lawrence, IBM T.J. Watson Research Center

Abstract: U-report is an open-source SMS platform operated by UNICEF Uganda, designed to give community members a voice on issues that impact them. Data received by the system are either SMS responses to a poll conducted by UNICEF or unsolicited reports of problems occurring anywhere within Uganda. There are currently over 240,000 U-report participants, and they send up to 10,000 unsolicited text messages a week. The objective of the program in Uganda is to understand the data in real-time, and to insure that critical issues identified in the messages are addressed by the appropriate department in

UNICEF in a timely manner. This talk describes an automated message-understanding and routing system deployed by IBM Research at UNICEF. We discuss a dual-supervision learning technique to leverage human-generated labels on both features and text examples, and conclude with a discussion of the societal impact that U-report is already driving in Uganda.

Building probabilistic classification predictions from aggregated ground truth

Melinda Han, Columbia University

Abstract: We explore methods for building probabilistic classification models using training data where the ground truth is only available in aggregate, rather than as individually labeled examples. We outline various methods for converting probabilistic class labels into hard class membership labels and discuss their implications on model performance. We show that model evaluation and selection can be reliably carried out with access to only aggregate or probabilistic ground truth examples. Finally, we share the results from our own application of these techniques to the problem of demographic prediction in online advertising.

Detecting Complex Rare Categories: Theory and Applications

Jingrui He, Steven Institute of Technology

Abstract: Complex rare categories exist in many real-world problems. Take insider threat detection from various social contexts as an example. While the target malicious insiders may only be a very small portion of the entire population (i.e., rarity), each person can be characterized by rich features, such as social friendship, emails, instant messages, etc (i.e., feature heterogeneity). Moreover, different types of insiders, though correlated, may exhibit different statistical characteristics (i.e., task heterogeneity). For such problems, how can we quickly identify an example from a new rare category? How can we leverage both feature heterogeneity and task heterogeneity to help detect more examples from the same rare category?

In this talk, I will present our recent work on addressing these problems. In the first part, I will answer the question of how to detect the rare examples with the help of a labeling oracle. In the second part, I will present a graph-based classification method taking into consideration both feature heterogeneity and task heterogeneity. I will also talk about how these techniques can be used in other applications such as semiconductor manufacturing.

A Statistical-Physical Approach for Air Quality Forecasting

Youngdeok Hwang, IBM Watson

Abstract: Physical processes, such as advection and diffusion, can transport atmospheric pollutants from the location at which they are generated, resulting in adverse impacts which can be distant from the pollution source. A physical model to incorporate the change of fluid field is a crucial tool in building an accurate forecasting model; but this requires instantaneous calibration of the physical model. In this paper, we propose a method to build a statistical forecasting model that couples the physical knowledge and observed data. The method obtains a regression fit for each of the spatial locations by solving a convex program, while simultaneously finding the input configurations for the physical model. The proposed approach is demonstrated through a real application.

Bayesian Robust Analysis for Large Vector Auto Regression Model

Ban Kawas, IBM Watson

Abstract: The number of parameters in the Vector Auto Regression Model (VAR) is typically very large, hence it is complicated to investigate the dynamic relations between the time series. Due to the large space of possible structures in VAR, the widely used Markov chain Monte Carlo (MCMC) techniques do not provide a practical solution for the general models considered. We propose a Bayesian robust framework for large VAR problems through compressing the computational expensive loading matrices while accommodating uncertainty in the subspace dimension and the noise from the measurement process. Practical performance relative to competitors is illustrated in simulations and real data applications.

Population Size Estimation with Inactive Lists: Hierarchical mixture models and Missing Data with Application to Armed Conflict Data

Shira Mitchell, Harvard University

Abstract: Since 1964, tens of thousands of people have died in Colombia's armed conflict. Underreporting the level of violence obscures the true nature of the conflict, precluding development of effective solutions. We develop hierarchical log-linear capture-recapture models to estimate the number of armed conflict killings that occurred in Casanare, Colombia in the years 1998-2007. Lack of data the early years motivates the use of hierarchical models that borrow strength across time. We investigate two methods to handle groups actively collecting data in different but overlapping time-periods. One fills in the inactive periods, treating counts in those years as missing data. Another does not, instead incorporating the inactivity into the model. We compare these, as well as hierarchical versus unpooled models. A simulation study shows that the Bayesian hierarchical models have shorter confidence interval width, with similar or better coverage than the unpooled models. They give useful intervals for the number of killings in the early years, where there are less data, so we can look at trends across time that guide political analysis of the conflict.

High-Resolution Spatiotemporal Estimates of Undocumented Killings in a Conflict Zone

Kristian Lum, Consultant

Abstract: A common focus in the statistical analysis of human rights data is the estimation of the number of undocumented rights violations that took place in a particular geographic location during a specified period of time. Multiple systems estimation (MSE) a method that utilizes overlaps in multiple lists of victims to estimate the number of victims who appear on none of the lists is generally employed for this purpose. It is often difficult to obtain reliable estimates of the number of violations at a granular geographic and temporal level because of small sample sizes. To achieve further disaggregation, we develop a Bayesian method for spatial variable selection that borrows information across regions in estimating the region-specific dependence structure among the lists. Our method is applied to the estimation of the number of undocumented killings in the state of Casanare, Colombia during a seven-year period of civil war. We produce estimates for 19 separate districts within Casanare for each year, while arriving at an aggregate estimate of undocumented killings that is very similar to those from previous studies conducted with this dataset.

Record linkage and capture-recapture models for quantifying conflict casualties in Syria

Megan Price, HRDAG

Abstract: How do we know how many people have been killed in Syria? If violence is escalating or decreasing? The hard answer is we don't. But through careful application of machine learning and other statistical techniques, we can quantify what we do, and don't, know. In this talk Megan will present how the Human Rights Data Analysis Group uses random forests, multiple systems estimation, and various Python and R packages to estimate conflict casualties.

Optimal Stability for the Nearest Neighbor Classifier

Wei Sun, Purdue University

Abstract: The nearest neighbor classifier is very popular in machine learning community due to its simplicity. In this talk, we penalize the weighted nearest neighbor classifier for automatically achieving the optimal trade-off between the classification accuracy and stability. The resulting classifier is shown to possess the minimax optimal stability by slightly sacrificing the classification accuracy. Our simulation results further demonstrate a significant improvement of stability over the optimal weighted nearest neighbor classifier (Samworth, 2012). This is joint work with Xingye Qiao and Guang Cheng.

Multiclass Distance Weighted Discrimination

Hanwen Huang, University of Georgia

Abstract: DWD is a powerful tool for solving binary classification problems which has been shown to improve upon SVM in high dimensional situations. We extend the binary DWD to the multiclass DWD. In addition to some well-known extensions which simply combine several binary DWD classifiers, we propose a global multiclass DWD (MDWD) which finds a single classifier that simultaneously considers all classes. Our theoretical results show that MDWD is Fisher consistent, even in the particularly challenging case when there is no dominating class (i.e., maximal class conditional probability is less than 1/2). The performances of different multiclass DWD methods are assessed through simulation and real data studies.

Regularized Outcome Weighted Subgroup Identification for Differential Treatment Effects

Sijian Wang, University of Wisconsin-Madison

Abstract: To facilitate comparative treatment selection when there is substantial heterogeneity of treatment effectiveness, it is important to identify subgroups that exhibit differential treatment effects. Existing approaches model outcomes directly and then define subgroups according to treatment and covariates interaction. However outcomes are affected by both the covariate-treatment interactions and covariate main effects. Consequently mis-specification of the main effects interferes with the covariate-treatment interaction estimation thus impedes valid predictive variable identification. We propose a method that approximates a target function whose value directly reflects correct treatment assignment for patients. This can disconnect the covariate main effects from the covariate-treatment interactions. The function uses patient outcomes as weights instead as modelling targets. Consequently, our method can deal with binary, continuous, time-to-event, and possibly contaminated outcomes in the same fashion. We first focus on identifying only directional estimates from linear rules that characterize important subgroups. We further consider estimation of differential comparative treatment effects for identified subgroups. We demonstrate the advantages of our method in simulation studies and in an analysis of two real data sets. This is joint work with Yaoyao Xu, Quefeng Li, Menggang Yu, Yingqi Zhao and Jun Shao.

Signal Detection in Massive Data with Applications in Genome-wide Genetic and Genomic Association Studies

Xihong Lin, Harvard University

Abstract: The genetic and genomic era provides an unprecedented promise of understanding genetic underpinnings of complex diseases or traits. Massive array based genome-wide SNP data, next generation sequencing data, as well as different types of omics data have been rapidly collected. These massive genetic and genomic data present statisticians with many exciting opportunities as well as challenges in data analysis and result interpretation, e.g., how to develop effective strategies for signal detection using massive genetic and genomic data when signals are weak and sparse. Many variable selection methods have been developed for analysis of high-dimensional data in the statistical literature. However limited work has been done on statistical inference for massive data. In this talk, I will discuss hypothesis testing for analysis of high-dimensional data motivated by gene, pathway/network based analysis in genome-wide association studies using arrays and sequencing data. I will focus on signal detection when signals are weak and sparse, which is the case in genetic and genomic association studies. I will discuss hypothesis testing for signal detection using penalized likelihood based methods and aggregated marginal test statistics based methods. The results are illustrated using data from genome-wide (sequencing) association studies.

Sharing Information with Behaviorally Similar Stores for Better Retail Forecasting

Ozden Gur Ali, Koc University

Abstract: Medium term (up to a year) multi-period sales forecasts are important inputs to retail operations, planning and budgeting. Retail chains have hundreds of stores and many formats serving different customer segments with many product categories. This challenging complexity offers an opportunity to generate insights and to increase the prediction accuracy at all aggregation levels. The forecasts need to reflect the seasonality and holiday effects, as well as the impact of the marketing plans. Further, there are many unspecified factors that manifest themselves in the sales trends of particular segments, which are difficult to distinguish from noise at the disaggregate level. Our approach first uses segment specific panel regressions with seasonality and marketing variables. Next, the residual time series are extrapolated into the future with lead-time specific regression models using features constructed from the residual time series. The final forecast is constructed by combining the regression forecasts, that take seasonality and marketing plans into account, with the extrapolated residual series. Working with the extensive dataset of the largest Turkish retailer, we show that the sophisticated residual structure that we develop outperforms the panel regression models with AR error structure: The farther out the prediction, the more the improvement, regardless of the aggregation level. Further, sharing information from stores that are similar based on business hierarchy provides even better prediction accuracy, compared with using only the focal store-categorys residual time series. But the best performance is obtained when the similarity for information sharing is established based on behavioral similarity rather than business hierarchy. To quantify behavioral similarity we use the Dynamic Time Warping measure which is capable of identifying similarities in time series with varying speeds (lags), and cluster the store-categories with the k-medoid method.

Experimental Designs and Estimation for Online Display Advertising Attribution in Marketplaces

Ram Akella, University of California-Berkeley

Abstract: Online advertising's importance as a marketing channel is due to its presumed ability to attribute conversions to an on-going campaign. Current industry practice is to run a randomized experiment to measure the ad creative effect (using a placebo); this is used to predict future performance of the ad. However, these attribution methods ignore other campaign components, including user selection (or targeting), in marketplaces with competitor effects, such as RTB (Real-Time-Bidding). We propose a novel experimental design to estimate both the campaign attribution and the ad creative impact in marketplaces. Our motivation is to find the campaign attribution under current conditions, as opposed to predicting future performance, and to demonstrate cost-effectiveness for continuous attribution. We estimate the campaign effects for users exposed to the ad using the Potential Outcomes Causal Model and the Principal Stratification framework. This analysis enables assessing user targeting, based on: 1) the probability of being influenced by the ad, and 2) the probability of selecting influenceable users. We analyze the effects of 3 independent campaigns from a major ad network with 20M+ users for each campaign. We demonstrate that the well-accepted practice of targeting users with highest conversion probability does not necessarily improve campaign attribution.

Unsupervised Consensus Analysis for On-line Review and Questionnaire Data

Stephen L. France

Abstract: We describe a set of Cultural Consensus Theory (CCT) models for analyzing review and questionnaire data. The basic single culture/cluster model can be used to estimate user competencies, user biases, and aggregate review scores. The model is unsupervised and only utilizes the input review scores. A maximum likelihood approach is used to estimate the model. We expand existing work by developing a clusterwise multi-culture continuous CCT model, for which we use the acronym CONSCLUS (CONSensus CLUStering). The original single culture CCT model is a special one-cluster case of CONSCLUS. We show that when all user competencies are equal, CONSCLUS is equivalent to k-means clustering. CONSCLUS is estimated using an alternating least squares variant of the algorithm for k-means clustering, which we denote as CCT-Means. CONSCLUS is a partitioning clustering technique. We describe extensions to CONSCLUS to incorporate fuzzy clustering and overlapping clustering.

We run a series of simulation experiments using generated data with random error. We test both the single cluster and multiple cluster cases. These experiments show that CONSCLUS is able to recover aggregate rating values and latent cluster assignments better than a range of other aggregation methods. The performance increase over the other aggregation methods is particularly strong when the users have varying competencies. We give an illustrative example using the Movielens dataset. We give a set of recommendations for the practical implementation of CONSCLUS on real world data and show how the user competencies can be used to gain insight into these data that cannot be gained from simple partitioning clustering.

An Asynchronous Scalable Distributed Expectation-Maximization Algorithm For Massive Data: The DEM Algorithm

Sanvesh Srivastava, Duke University

Abstract: Massive data with complex latent structures have become common independent of discipline. The computer architectures to store these data are also rapidly evolving. Classical iterative statistical algorithms, such as Expectation-Maximization (EM), for fitting models with latent structures are practically infeasible for these data due to two main reasons: massive size of the data and the large number of parameters required to model the complex dependencies in the data. These two limitations are relaxed by the Distributed Iterative Statistical Computing (DISC) framework presented in this work for implementing

iterative statistical algorithms by taking advantage of widely available computing power, such as cluster of computers. Using EM as a concrete example of an iterative algorithm, DISC extends and scales it for massive data as DISC-EM (DEM). We analyze the convergence properties of the sequence of parameter estimates generated by DEM and show that DEM retains the attractive properties of EM: monotone ascent of the log likelihood at each iteration and stability of iterations. DEM can also be easily implemented in cluster and grid computing environments using R package disc and existing EM implementations. To illustrate its application, we use DEM for estimating the effect of movie genres on their ratings in a movie ratings data.

Sparse Robust Graphical Models

Myung Hee Lee, Colorado State University

Abstract: A graphical model is a way of inferring conditional relationships among multiple variables. When the variables follow multivariate normal distribution, we can fit Gaussian Graphical Models (GGMs) and identify the conditional independence relationship by the zero entries of the precision matrix. However, when the variables do not follow Gaussian, the conditional independence can no longer be inferred from the precision matrix. We propose a graphical model that is robust to the distributional assumption, and we do this via applying a set of sparse quantile regression models. The conditional quantile probabilities of one variable as function of the rest is shown to bear sufficient information on the conditional dependence between variables under appropriate assumption. We demonstrate the advantages of our approach using simulation study under various scenarios and then we apply our method to an interesting real biological dataset, where considerable amount of the dataset is contaminated, illustrating the advantage of the proposed method in a real setting.

Incremental Response Modeling Based on Novelty Detection via One-Class Support Vector Machine

Taiyeong Lee, SAS Institute

Abstract: In direct marketing campaign, using conventional predictive model often leads to wasting marketing expenses because the model tries to predict all the people who are likely to make a purchase. Among the predicted purchasers, there is a customer group who would buy regardless of the marketing action such as promotion coupons. Incremental response model builds a predictive model which looks for only the customers who are likely to buy or respond positively to marketing campaigns when they are targeted but are not likely to buy if they are not targeted. One popular methodology for incremental response modeling, as known as the difference score model, is a simple method to model the incremental effect through building two response models separately from treatment and control groups. Another method is a tree based method which uses the incremental effect measure as a split criterion. We propose a new incremental response model based on novelty detection via one-class support vector machine. The detected abnormal events are considered as the incremental responses and we show how to train and validate the model between treatment and control group data. Simulation studies and a real data set analysis are illustrated.

Calibration of Complex Computer Models: An Application in Cardiac Cell Modeling

Huan Yan, Georgia Institute of Technology

Abstract: This paper studies the calibration of a complex computer model that describes the Potassium currents in a Cardiac cell. The computer model is expensive to evaluate and contains 24 unknown calibration parameters, which makes the problem very challenging for the traditional methods of calibration

using kriging. We propose physics-driven strategies for the approximation and calibration of this large-complex model. Another difficulty with this problem is the presence of large cell-to-cell variation, which is modeled through random effects. We propose approximate Bayesian methods for the identification and estimation of the parameters in this complex nonlinear mixed-effects statistical model.

Analysis of Computer Experiments with Qualitative and Quantitative Factors

Chunfang Devon Lin, Queen's University

Abstract: Computer experiments with qualitative and quantitative factors occur frequently in various applications in science and engineering. Analysis of such experiments is not yet completely resolved. In this work, we propose a flexible modeling approach to build predictive models. The approach makes use of a flexible function to capture the correlation among qualitative and quantitative factors. Several examples are given to demonstrate significant improvement in prediction.

Tapered Correlation Matrix Preconditioning for Fast Approximate Kriging

Lulu Kang, Illinois Institute of Technology

Abstract: Kriging has been very popular for analyzing computer experiments due to its accurate prediction. But it also needs extensive computation due to the matrix inversion involved in the predictor. We propose using the preconditioning approach to approximate the matrix inversion in the kriging predictor. How to construct the preconditioning matrix is crucial to the accuracy and efficiency of the approximate kriging. Here we suggest using the tapered correlation matrix as the preconditioning matrix.

Statisticial Inference Using Agent-Based Models

Daniel Heard, Duke University

Abstract: Agent-based models (ABMs) are computational models used to simulate the behaviors, actions and interactions of agents within a system. The individual automonous agents each have their own set of assigned attributes and rules, which determine their behavior within the ABM system, allowing us to observe how the behaviors of the individual agents impact the system as a whole and if any emergent structure develops within the system.

I begin by presenting some background and theory related to ABMs, including procedures for model validation, assessing model equivalence and measuring model complexity.

I then discuss two approaches for performing likelihood-free inference involving ABMs: Gaussian Process emulators and Approximate Bayesian Computation. I conclude by demonstrating the approaches for inference in two applications.

myEpi. Modeling a 'digital self

Georgiy Bobashev, RTI

Abstract: I will discuss novel approaches to the analysis of intensive data collected within a single individual. While technological advantages of social media and mobile technology open new opportunities in collecting intensive behavioral and drug-using data within one individual, traditional epidemiological mindset demands generalizability to a larger population. I will argue that a single individual could be viewed as an entire population of drug using and behavioral events. I will show how traditional epidemiological methods (regression analysis, Markov models, and agent-based models) that are usually applied

to populations of humans, could be applicable to a single individual and thus used for self-monitoring, forecasting and controls of health events.

Classification of Dirichlet Observations applied to Industrial Problems

Roelof Coetzer, SASOL and University of the Free State

Abstract: In industry the response(s) from many processes are functions of mixture or composition variables i.e., that is a set of proportions that add up to one. In many of these processes the responses themselves are also observed as compositional data. For example, the amount and the composition of gas produced from a coal gasification facility depend crucially on the properties and the size distribution of the coal being used in the process, and both inputs are observed as compositional data. In this talk we apply the Dirichlet distribution for the compositional inputs and present a new classification scheme for the important responses. The approach presented is a linear partitioning of the Dirichlet simplexes that can also be extended to high dimensional cases. We will use two very different examples from industrial applications to illustrate the concepts.

High-Dimensional Variable Screening under Unobserved Causal Factors

John Daye, University of Arizona

Abstract: The presence of unobserved causal e effects is a common problem in genetic association studies. For example, unknown variables may arise from unmeasured clinical risks or environment exposures. In practice, genetic association studies rarely consider all potentially causal factors, which can dramatically impact statistical inference by introducing additional perturbations. Dealing with unknown factors is a statistically challenging problem, as knowledge on entire variables is missing. In this talk, we introduce a novel variable screening procedure based on mixture models to deal with some common situations when unknown causal factors are potentially present. Results suggest that the new procedure is robust towards the effects of unknown causal factors for variable screening and ranking of high-dimensional genomic data.

Functional Feature Construction for Personalized Treatment Regimes

Eric Laber, North Carolina State University

Abstract: Evidence-based personalized medicine formalizes treatment selection as a decision rule that maps up-to-date patient information into the space of possible treatments. Available patient information may include static features such race, gender, family history, genetic and genomic information, as well as longitudinal information including the emergence of comorbidities, waxing and waning of symptoms, side-effect burden, and adherence. Dynamic information measured at multiple time points before treatment assignment should be included as input to the decision rule. However, subject longitudinal measurements are typically sparse, irregularly spaced, noisy, and vary in number across subjects. Existing estimators for decision rules require equal information be measured on each subject and thus standard practice is to summarize longitudinal subject information into a scalar ad hoc summary during data pre-processing. This reduction of the longitudinal information to a scalar feature precedes estimation of a decision rule and is therefore not informed by subject outcomes, treatments, or covariates. We propose a data-driven method for constructing maximally prescriptive yet interpretable features that can be used with standard methods for estimating optimal decision rules including both regression-based estimators. In our proposed framework we treat the subject longitudinal information as a realization of a stochastic process at discrete time points, and observed with error. The estimated feature is expressible as the integral of the subject longitudinal process against a smooth coefficient function; the estimated coefficient function therefore describes the optimal weighting of subject-specific longitudinal information which is potentially informative for clinical practice.

Random KNN Classification with Variable Selection

E. James Harner, West Virginia University

Abstract: Random KNN (RKNN) is a novel generalization of traditional nearest-neighbor modeling. Random KNN consists of an ensemble of base k-nearest neighbor models, each constructed from a random subset of the input variables. A collection of r such base classifiers is combined to build the final Random KNN classifier. Since the base classifiers can be computed independently of one another, the overall computation is embarrassingly parallel. Random KNN can be used to select important features using the RKNN-FS algorithm. RKNN-FS is an innovative feature selection procedure for small n, large p problems. Empirical results on microarray data sets with thousands of variables and relatively few samples show that RKNN-FS is an effective feature selection approach for high-dimensional data. RKNN is similar to Random Forests (RF) in terms of classification accuracy without feature selection. However, RKNN provides much better classification accuracy than RF when each method incorporates a feature-selection step. RKNN is significantly more stable and robust than Random Forests for feature selection when the input data are noisy and/or unbalanced. Further, RKNN-FS is much faster than the Random Forests feature selection method (RF-FS), especially for large scale problems involving thousands of variables and/or multiple classes. Random KNN and feature selection algorithms are implemented in an R package rknn, which supports both classification and regression. The time complexity of the algorithm, including feature selection, is O(rkpn log n), assuming the number of variables randomly selected in a base classifier is m = log p. This choice of m, in contrast to p p, reduces the time complexity from exponential time to linear time. However, it is important to choose r sufficiently large to ensure adequate variable coverage. By paralleling the code in rknn, the time can be reduced linearly depending on the number of cores or compute nodes. We will show how to apply the Random KNN method via the parallelized rknn package to high-dimensional genomic data.

Model Selection in High-Dimensional Misspecified Models

Yang Feng, Columbia University

Abstract: Model selection is vital to high-dimensional modeling in selecting the best set of covariates among a sequence of candidate models. Most existing work assumes implicitly that the model under study is correctly specified or of fixed dimensions. Both model misspecification and high dimensionality are, however, common in real applications. In this paper, we investigate two classical Bayesian and Kullback-Leibler divergence principles of model selection in the setting of high-dimensional misspecified models. Asymptotic expansions of these model selection principles in high dimensions reveal that the effect of model misspecification is crucial and should be taken into account, leading to the generalized BIC and generalized AIC. With a natural choice of prior probabilities, we suggest the generalized BIC with prior probability (GBIC$_p$) which involves a logarithmic factor of the dimensionality in penalizing model complexity. We further establish the consistency of the covariance contrast matrix estimator in the general setting. Our results and new method are also supported by numerical studies.

Weak Signal Detection in High-Dimensional Data Analysis

Jessie Jeng, North Carolina State University

Abstract: This paper considers a frequently observed phenomenon in real-data applications that strong signals often stand out easily whereas weak signals can be indistinguishable from the noise. This natural phenomenon is especially relevant with high-dimensional data as signals are more easily obscured by the large amount of noise. Contemporary studies in signal detection mainly focus on relatively strong signals. Very few works have been developed to provide a rigorous characterization of the weak signals that can be indistinguishable from the noise. This paper seeks to facilitate the identification of weak signals by introducing a new approach that delineates the ranges of strong and weak signals separately. We start with theoretical developments to investigate the formations of signal, noise, and indistinguishable subsets. A novel Two-Level Thresholding (TLT) procedure is, then, proposed to identify the three subsets with statistical accuracy. Theoretical and simulation studies demonstrate that, in addition to controlling false positives, the proposed method can efficiently control false negatives at a desirable level. As a result, relatively weak signals that cannot be claimed significant due to multiplicity correction can be efficiently kept for follow-up study. The proposed procedure is further evaluated and compared to existing methods under various model settings. We apply TLT in a real-data analysis on detecting genomic variants with varying signal intensities.

Tests for High-dimensional Nonparametric Functions

Pingshou Zhong, Michigan State University

Abstract: We propose a test statistic for testing high-dimensional nonparametric functions in a repducing kernel Hilbert space generated by a positive definite kernel. The asymptotic distributions of the test statistic are derived under the null hypothesis and a series of local alternative hypotheses in a “large p, small n” setup. Simulation studies and a real data set analysis are used to demonstrate the proposed method.

Wednesday Afternoon

Distance-weighted Support Vector Machine

Xingye Qiao, SUNY Binghampton

Abstract: David Banks regrets that he was unable to find this abstract amongst his email.

Efficient Distributed Topic Modeling with Provable Guarantees

Mohammad Rohban, RIT and Boston University

Abstract: Topic modeling for large-scale distributed web-collections requires distributed techniques that account for both computational and communication costs. We consider topic modeling under the separability assumption and develop novel computationally efficient methods that provably achieve the statistical performance of the state-of-the-art centralized approaches while requiring insignificant communication between the distributed document collections. We achieve trade-offs between communication and computation without actually transmitting the documents. Our scheme is based on exploiting the geometry of normalized word-word co-occurrence matrix and viewing each row of this matrix as a vector in a high-dimensional space. We relate the solid angle subtended by extreme points of the convex hull of these vectors to topic identities and construct distributed schemes to identify topics.

Two Sample Inference on Populations of Graphical Models: Applications to Multi-Subject Functional Brain Connectivity

Manjari Narayan, Rice University

Abstract: Gaussian Graphical Models (GGM) are popularly used in neuroimaging studies based on fMRI, EEG or MEG to estimate functional connectivity, or relationships between remote brain regions. In multi-subject studies, scientists seek to identify the functional brain connections that are different between two groups of subjects, i.e. connections present in a diseased group but absent in controls or vice versa. This amounts to conducting two-sample large scale inference over network edges post graph selection, a novel problem we call Population Post Graph Selection Inference. Current approaches to this problem include estimating a network for each subject, and then assuming the subject networks are fixed, conducting two-sample inference for each edge. These approaches, however, fail to account for the variability associated with estimating each subject's graph, thus resulting in high numbers of false positives and low statistical power. By using Resampling and Random penalization to estimate the post graph selection variability, and the proper Random Effects test statistics, we introduce a new procedure, termed R3, that solves these problems. Through simulation studies we show that R3 offers major improvements over current approaches in terms of error control and statistical power. We provide a demonstration of our method by identifying functional connections present or absent in autistic subjects based on the ABIDE multi-subject fMRI study.

Learning with Hierarchical Deep Models

Ruslan Salakhutdinov, University of Toronto

Abstract: We introduce HD (or Hierarchical-Deep) models, a new compositional learning architecture that integrates deep learning models with structured hierarchical Bayesian (HB) models. Specifically, we show how we can learn a hierarchical Dirichlet process (HDP) prior over the activities of the top-level features in a deep Boltzmann machine (DBM). This compound HDP-DBM model learns to learn novel concepts from very few training example by learning low-level generic features, high-level features that capture correlations among low-level features, and a category hierarchy for sharing priors over the high-level features that are typical of different kinds of concepts. We present efficient learning and inference algorithms for the HDP-DBM model and show that it is able to learn new concepts from very few examples on CIFAR-100 object recognition, handwritten character recognition, and human motion capture datasets.

Sum-Product Networks: A New Deep Architecture

Hoifung Poon, Microsoft Research

Abstract: The key limiting factor in graphical model inference and learning is the complexity of the partition function. We thus ask the question: what are general conditions under which the partition function is tractable? The answer leads to a new kind of deep architecture, which we call sum-product networks (SPNs). SPNs are directed acyclic graphs with variables as leaves, sums and products as internal nodes, and weighted edges. We show that if an SPN is complete and consistent it represents the partition function and all marginals of some graphical model, and give semantics to its nodes. Essentially all tractable graphical models can be cast as SPNs, but SPNs are also strictly more general. We then present algorithms for generative and discriminative learning with SPNs, as well as algorithms for learning the structure of an SPN. Experiments show that inference and learning with SPNs can be both faster and more accurate than with standard deep networks. Finally, we will review open problems and exciting future directions for SPN learning and applications.

Deep Learning with Hierarchical Convolutional Factor Analysis

Lawrence Carin, Duke University

Abstract: Unsupervised multi-layered (“deep”) models are considered for imagery. The model is represented using a hierarchical convolutional factor-analysis construction, with sparse factor loadings and scores. The computation of layer-dependent model parameters is implemented within a Bayesian setting, employing a Gibbs sampler and variational Bayesian (VB) analysis, that explicitly exploit the convolutional nature of the expansion. In order to address large-scale and streaming data, an online version of VB is also developed. The number of dictionary elements at each layer is inferred from the data, based on a beta-Bernoulli implementation of the Indian buffet process. Example results are presented for several image-processing applications, with comparisons to related models in the literature.

Sparse and Smooth Principal Components Analysis

Genevera Allen, Rice University

Abstract: Regularized principal components analysis (PCA), especially Sparse PCA and Functional PCA, has become widely used for dimension reduction in high-dimensional settings. Many examples of massive data, however, may benefit from estimating both sparse and smooth factors. These include neuroimaging data where there are discrete brain regions of activation (sparsity) but these regions tend to be smooth spatially. Here, we introduce an optimization framework that can encourage both sparsity and smoothness of the row and/or column PCA factors. This framework generalizes many of the existing approaches to Sparse PCA, discrete versions of Functional PCA and two-way Sparse PCA and Functional PCA, as these are all special cases of our method. In particular, our method permits flexible combinations of sparsity and

smoothness that lead to improvements in feature selection and signal recovery as well as more interpretable PCA factors. We demonstrate our method on simulated data and a neuroimaging example on EEG data. This work provides a unied framework for regularized PCA that can form the foundation for a cohesive approach to regularization in high-dimensional multivariate analysis.

Analysis of Agreement between Two Long Ranked Lists

Srinath Sampath, Hamilton Capital Management

Abstract: The problem of determining the endpoint of agreement between two rankings of a long list of objects is addressed by modifying the method of estimation described in Fligner and Verducci (1988)s multistage model for rankings, from maximum likelihood of conditional agreement over a sample of rankings to a locally smoothed estimator of stage-wise agreement. Simulations show that this moving average maximum likelihood estimator (MAMLE) performs very well under several conditions. The MAMLE method is applied to a database of popular names for newborns in the United States and insights into trends as well as differences in naming conventions between male and female infants are uncovered. The technique is also applied as a stopping rule to augment the tau-path algorithm of Yu, Verducci and Blower (2011), in an analysis of associations between gene expression and compound potency in data from the National Cancer Institute, and of the length of these associations before agreement degenerates into noise. Fligner, M. A., and Verducci, J. S. (1988), Multistage Ranking Models, Journal of the American Statistical Association, 83(403), 892901.

Combining nonparametric inferences using data depth and confidence distribution

Dungang Liu, Yale University

Abstract: For the purpose of combining inferences from several nonparametric studies for a common hypothesis, we develop a new methodology using the concepts of data depth and confidence distribution. A confidence distribution (CD) is a sample-dependent distribution function that can be used to estimate parameters of interest. It is a purely frequentist concept yet can be viewed as a distribution estimator of the parameter of interest. Examples of CDs include Efrons bootstrap distribution and Frasers significance function (also referred to as p-value function). In recent years, the concept of CD has attracted renewed interest and has shown high potential to be an effective tool in statistical inference. In this project, we use the concept of CD, coupled with data depth, to develop a new approach for combining the test results from several independent studies for a common multivariate nonparametric hypothesis. Specifically, in each study, we apply data depth and bootstraps to obtain a p-value function for the common hypothesis. The p-value functions are then combined under the framework of combining confidence distributions. This approach has several advantages. First, it allows us to resample directly from the empirical distribution, rather than from the estimated population distribution satisfying the null constraints. Second, it enables us to obtain test results directly without having to construct an explicit test statistic and then establish or approximate its sampling distribution. The proposed method provides a valid inference approach for a broad class of testing problems involving multiple studies where the parameters of interest can be either finite or infinite dimensional. The method will be illustrated using simulations and flight data from the Federal Aviation Administration (FAA).

Inferring Latent Structure in Unsolicited Network Data

Tyler McCormick, University of Washington

Abstract: David Banks regrets that he was unable to find this abstract amongst his email.

Estimating Undirected Graphs Under Weak Assumptions

Mladen Kolar, University of Chicago

Abstract: David Banks regrets that he was unable to find this abstract amongst his email.

Relaxing Conditional Independence Assumptions in Data Fusion

Bailey Fosdick, SAMSI and University of Colorado

Abstract: Survey practitioners often try to combine information across multiple surveys by performing data fusion, whereby separate survey data sets are merged to form a single data file with no missing entries. The individual data sets usually contain information on disjoint sets of respondents and have distinct, but overlapping, variable sets. The primary task is the imputation of the variables not included in each of the surveys. Frequently numerous pairs of variables are never observed simultaneously in the data. In these cases, it is standard to use an imputation method that assumes conditional independence between the variables given the variables common to all surveys. In this talk, we consider a situation where auxiliary information, referred to as glue, is available on the dependence between variables not observed concurrently. We discuss different types of glue that may be obtained and compare their utilities when using a latent class mixture model to perform imputations on data from HarperCollins Publishers.

On the Practical Use of Training Surrogates for Machine Learning in Online Display Advertising

Brian Dalessandro, Dstillery

Abstract: Display advertising within real-time bidding systems is the ideal setting for the application of machine learning systems. Despite this perfect match, many practical challenges exist that can severely limit the efficacy of modeling systems used to optimize online display campaigns. In particular, training data can be described as very high dimensional, yet sparse, with extremely rare outcomes. Thus, exploiting the high dimensionality can easily lead to massive over fitting and poor performance. In this talk, we'll discuss how Dstillery uses techniques from transfer learning to boost performance in settings that are plagued by severe sparsity. Additionally, we'll discuss how this is coupled with state-of-the-art methods for learning in high dimensional, big data settings to achieve a robust and scalable machine learning system.

Bayesian Analysis of Weighted Networks

James Johndrow, Duke University

Abstract: The analysis of networks is of increasing interest in many ap- plication areas. In networks encoded by a weighted graph, the ob- served data are generally count or other non-negative matrices. In high-dimensional settings, the data and underlying graph are often sparse or approximately sparse, with most edges receiving weights that are nearly or exactly zero. We propose a Bayesian nonpara- metric approach to the analysis of sparse weighted networks. The model induces a low-rank factorization of the non-negative matrix parameter for count data. Dynamics and covariate information are incorporated in the class weights, inducing a conditional matrix fac- torization. Theoretical and empirical results on model support and properties are provided. We propose a simple and efficient Gibbs sampler for computation, and illustrate performance in simulation examples as well as on a real large-scale dataset of traffic between two groups of websites.

Data mining for onomastic analysis

Olivier Coppet, Global Data Excellence

Abstract: Unlike in the US, ethnicity is not considered as a legal descriptive attribute of an employee in the French administration, whatever the organization is in the public or private sector. This measure, though aiming to protect individuals from racial and ethnic discriminations, does not prevent them from such based inequities, actually. In fact, with omitting ethnic or gender description, for example, organizations made it difficult to track underlying injustices. Onomastics can tackle that issue by performing extensive statistical analysis (especially association rules) on employees names and surnames. Such tools are able to associate each pair (name, surname) to an ethnic group therefore enabling correlation analysis with other employee descriptive attributes within the database.

Dimension Reduction through LqSVM

Andreas Artemiou, Cardiff University

Abstract: Sufficient Dimension Reduction and Support Vector Machine have been used separately for handling high dimensional datasets. Recently Li, Artemiou and Li (2011) combined these two areas to achieve linear and nonlinear sufficient dimension reduction in a common framework. In this talk we will show that under specific circumstances, their algorithm might not have unique solution and we propose the use of LqSVM (q¿1) which ensures the uniqueness of the solution. The benefits of the new algorithm in performance are demonstrated through simulation results and real data analysis. Finally, we will also discuss some further advances on the use of Support Vector Machines in the sufficient dimension reduction framework.

Variable Selection in Multi-index Models

Peng Zeng, Auburn University

Abstract: Multi-index models are semi-parametric regression models that generalize linear regression by assuming that the response depends on several linear combinations of predictors via an unknown link function. Because of the weak assumptions on the link function, they are commonly used to facilitate dimension reduction in high-dimensional data analysis. In this talk, we propose a novel penalization method for simultaneous variable selection and parameter estimation in multi-index models, which naturally extends the lasso method for linear regression. The objective function enjoys a transformation invariance property, which enables an efficient computational algorithm. For multi-index models, the proposed method can automatically perform row-wise variable selection, which is a special case of group-wise variable selection. The properties of the estimates and the solution paths are investigated. The excellent performance is demonstrated using simulation examples and real examples. This is a joint work with Yu Zhu at Purdue University.

Tensor Dimension Reduction

Wenxuan Zhong, University of Georgia

Abstract: In this talk, I will introduce a sufficient dimension reduction method for semi-parametric regression with tensor predictors. The method introduced here generalized the popular vector-based sufficient dimension reduction methods to regression with tensor predictors. A tensor dimension reduction model is proposed assuming that a response depends on some low dimensional representation of tensor predictors through an unspecified link function. A sequential iterative dimension reduction algorithm that can effectively utilizes the tensor structure is proposed to estimate the parameters. Empirical and theoretical studies will be discussed to demonstrate empirical and asymptotic behavior of the proposed method.

Fantope Projection and Selection

Vince Vu, Ohio State

Abstract: Sparse PCA is a relatively new technique for simultaneous dimension reduction and variable selection in high-dimensional data analysis. However, the computation is challenging and most of the past decade's proposals are based on intractable optimization problems and heuristic algorithms for their solution. This talk will present some recent developments in convex relaxation of sparse PCA with a special focus on subspace estimation. The results include: a new method based on the convex hull of rank-k projection matrices (the Fantope) that can be solved efficiently by alternating direction method of multipliers, and some statistical properties of this new method such as subspace convergence and variable selection consistency. The results hold for general covariance models, do not require rank assumptions, and they can be applied to a wide array of settings beyond PCA. Even when the ”truth” is not sparse and nothing is assumed beyond independence, the method retains an optimality property and can be interpreted in an agnostic manner.

Based on collaborations with J. Lei (CMU), J. Cho (UWisc), and K. Rohe (UWisc)

Continuum discriminant directions and their high dimensional asymptotic properties

Sungkyu Jung, University of Pittsburgh

Abstract: We investigate a two-group classification direction which embraces several well-known methods. A generalized criterion for linear discrimination is proposed as an extension from Fisher's idea, which leads to a sequence of discriminant directions. These directions and their limits include the directions of linear discriminant, mean difference and the first principal component as well as ridge discriminant directions. We also show that an extension of this method leads to a supervised principal component analysis where the group information is blended in the computation of principal components. In the high dimension low sample size data situation, the maximum data piling direction is better suited for Fisher's criterion than the linear discriminant direction with a generalized inverse. In such a case, the method embraces the maximal data piling direction, instead of the linear discriminant direction. A high dimensional asymptotic investigation reveals the conditions under which the maximum data piling direction is preferable for classification than the mean difference. Simulation studies and real data analysis demonstrate the advantages of this new classification framework.

Efficient Dimension Reduction for a Group of Images

Dong Wang, University of North Carolina-Chapel Hill

Abstract: David Banks regrets that he was unable to find this abstract amongst his emails.

Personalized Dose Finding Using Outcome Weighted Learning

Guanhua Chen, University of North Carolina-Chapel Hill

Abstract: There is growing recognition that it is important to consider individual level heterogeneity when searching for an optimal treatment dose. The optimal individualized treatment rule (ITR) for dosing should maximize the expected clinical outcome averaged across the population of interest. In this talk, we describe a randomized trial design which allows the candidate dose levels to be continuous. To find the optimal ITR under such a design, we propose an outcome weighted learning method which directly maximizes the expected clinical outcome. This method converts the individualized dose selection problem into penalized weighted regression with a truncated 1 loss. A difference of convex functions (D.C.) algorithm is adopted to efficiently solve the associated non-convex optimization problem. The consistency and convergence rate of the estimated ITR are derived and moderate sample performance is evaluated via simulation studies. We demonstrate that the proposed method outperforms alternative approaches. We illustrate the method using data from a clinical trial for Warfarin (an anti-thrombotic drug) dosing.

Set-valued Dynamic Treatment Regimes for Competing Outcomes

Daniel Lizotte, University of Waterloo

Abstract: Dynamic treatment regimes operationalize the clinical decision process as a sequence of functions, one for each clinical decision, where each function maps up-to-date patient information to a single recommended treatment. Current methods for estimating optimal dynamic treatment regimes, for example Q-learning, require the specification of a single outcome by which the ‘goodness’ of competing dynamic treatment regimes is measured. However, this is an over-simplification of the goal of clinical decision making, which aims to balance several potentially competing outcomes, e.g., symptom relief and side-effect burden. When there are competing outcomes and patients do not know or cannot communicate their preferences, formation of a single composite outcome that correctly balances the competing outcomes is not possible. This problem also occurs when patient preferences evolve over time. We propose a method for constructing dynamic treatment regimes that accommodates competing outcomes by recommending sets of treatments at each decision point. Formally, we construct a sequence of set-valued functions that take as input up-to-date patient information and give as output a recommended subset of the possible treatments. For a given patient history, the recommended set of treatments contains all treatments that produce non-inferior outcome vectors. Constructing these set-valued functions requires solving a non-trivial enumeration problem. We offer an exact enumeration algorithm by recasting the problem as a linear mixed integer program. The proposed methods are illustrated using data from the CATIE schizophrenia study.

Tree Based Methods for Optimal Treatment Allocation

Yingqi Zhao, University of Wisconsin

Abstract: Individualized treatment rules recommend treatments on the basis of individual patient characteristics. With their promise of better patient outcomes as well as lower cost and patient burden, individualized treatment rules are increasingly being sought for use by clinical and intervention scientists. If a treatment rule learned from data is to be used to inform clinical practice or provide scientific insight, it is crucial that it be interpretable; clinicians may be unwilling to implement models they do not understand and black-box models may not be useful for guiding future research. The hallmark of interpretable prediction models is a decision tree. We propose a method for estimating the optimal individualized treatment rule from the class of rules that are representable as a decision tree. The class of rules we consider is thereby interpretable yet expressive. A novel feature of this problem, say compared with regression or classification, is that the data are not supervised as the optimal treatment allocation for each patient is not known and must be estimated from the data. The proposed method applies for both categorical and continuous treatments and performs favorably to competing methods in simulated experiments. We illustrate the proposed method using data from study of major depressive disorder.

Active Completion of Coherent Low-rank Matrices and Tensors

Aarti Singh, Carnegie Mellon University

Abstract: The ability to learn large-scale matrices from few observed entries is important in myriad applications including imputing latencies between hosts in a communication network, expression levels of genes under various drugs, and user ratings for movies. In many of these applications, one can employ judicious feedback-driven choice of which entries to observe to minimize network traffic, experimental cost or user effort needed. I will describe how feedback-driven active queries can be used to learn large low-rank matrices and tensors sequentially. Compared to random queries, active queries result in 1) lower sample complexity, 2) lower computational complexity, 3) lower memory complexity and 4) ability to handle matrices with coherent rows/columns that might arise due to anomalous hosts, genes or users.

Adaptive Compressive Sensing of Signals Exhibiting Tree-structured Sparsity Fundamental Limits, Implications, and Applications

Jarvis Haupt, University of Minnesota

Abstract: Recent breakthrough results in compressive sensing (CS) have established that many high dimensional signals can be accurately recovered from a relatively small number of non-adaptive linear observations, provided that the signals possess a sparse representation in some basis. Subsequent efforts have shown that the performance of CS can be improved by exploiting additional structure in the locations of the nonzero signal coefficients during inference, or by utilizing some form of sequential adaptive measurement focusing during the sensing process. In this talk I will discuss a powerful fusion of these notions.

Specifically, I will present results of our recent work which analyzes a class of adaptive sensing strategies that are specifically tailored to structured sparse signals characterized by nonzero components whose locations exhibit tree-structured dependencies. We establish fundamental performance limits for the task of support recovery of such tree-sparse signals from noisy measurements, in settings where measurements may be obtained either non-adaptively (using a randomized Gaussian measurement strategy motivated by initial CS investigations) or by any adaptive sensing strategy, and we show that a simple adaptive sensing strategy is nearly minimax optimal for these tasks. Our analysis reveals a surprising fact that the sufficient conditions under which our simple support recovery procedure succeeds do not depend at all on the ambient signal dimension. This is in stark contrast to the best possible adaptive or non-adaptive sensing strategies that leverage adaptivity or structure in isolation, each of which require that the signal energy grow in proportion to the signal dimension in order to guarantee support estimation consistency. We briefly discuss implications of our results in the context of a stylized imaging application.

Minimax Analysis of Active Learning

Steve Hanneke, Free Spirit

Abstract: In this talk, I will discuss recent work characterizing the minimax optimal label complexities of active learning with VC classes. In particular, it turns out the minimax label complexity of active learning under Mammen-Tsybakov noise is always smaller than the minimax sample complexity of passive learning, and in many cases is smaller than the previously-published upper bounds for active learning. I will also present new active learning strategies that nearly achieve this optimal performance.

Le Song, Georgia Institute of Technology

Abstract: David Banks regrets that he was unable to find this abstract amongst his emails.

Sivaranman Balakrishnan, UC Berkeley

Abstract: David Banks regrets that he was unable to find this abstract amongst his email.

Text Mining in a Blog Network

David Banks, Duke University

Abstract: Recently, there has been progress both in dynamic network models and topic modeling. Since the political blogosphere represents a network of documents, we seek to use network models to improve topic discovery and topic discovery to improve network modeling on the corpus of all posts to the top 1,509 political blogs in the U.S. in 2012.

Asymptotic Equivalence of Regularization Methods in Thresholded Parameter Space

Yingying Fan, University of Southern California

Abstract: High-dimensional data analysis has motivated a spectrum of regularization methods for variable selection and sparse modeling, with two popular classes of convex ones and concave ones. A long debate has been on whether one class dominates the other, an important question both in theory and to practitioners. In this paper, we characterize the asymptotic equivalence of regularization methods, with general penalty functions, in a thresholded parameter space under the generalized linear model setting, where the dimensionality can grow up to exponentially with the sample size. To assess their performance, we establish the oracle inequalities, as in Bickel, Ritov and Tsybakov (2009), of the global minimizer for these methods under various prediction and variable selection losses. These results reveal an interesting phase transition phenomenon. For polynomially growing dimensionality, the L_1-regularization method of Lasso and concave methods are asymptotically equivalent, having the same convergence rates in the oracle inequalities. For exponentially growing dimensionality, concave methods are asymptotically equivalent but have faster convergence rates than the Lasso. We also establish a stronger property of the oracle risk inequalities of the regularization methods, as well as the sampling properties of computable solutions. Our new theoretical results are illustrated and justified by simulation and real data examples.

Network Based Discriminant Analysis with Applications to fMRI

Xi Lou, Brown University

Abstract: Network structures are believed to be fundamental for brain functioning, and changes in brain networks may lead to mental disorders and behavior changes. Motivated by this theory, we study the problem of simultaneous classification and network estimation from high dimensional data with class labels. Using the Gaussian graphical models framework, it becomes popular to study joint estimation of networks from multiple labels. In many disease studies, however, disease classification is only a simplification of a wide spectrum of disorders with varying degrees. Our approach thus focuses on inferring sparse graphical model structures that contain discriminant power. It is based on a convex optimization formulation that simultaneously computes a discriminant function and its input network structures. In particular, the use of a discriminant function allows assigning varying weights for observations with the same label, and such weights are derived from the estimated networks. To compute this estimator, we further develop an iterative algorithm based on alternating direction method of multipliers. Numeric merits of our proposal are illustrated using simulated data and a resting-state functional MRI study on Parkinson's disease.

Online Statistical Learning Algorithms

Hua Zhou, North Carolina State University

Abstract: Online estimation is one of the most intriguing problems of big data analysis. This talk presents adaptations of three dominant approaches for maximum likelihood estimation in statistics - Newton's method, Fisher's scoring method, and minorization-maximization (MM) algorithm - to the online scheme.

Credit Risk Assessment: A Random Forest Based Approach

Imad Bou-Hamad, American University of Beirut

Abstract: Cerchiello and Guidici (2013) proposed a Bayesian approach to improve the ordinal variable selection in credit rating assessment. However, no comparison has been done with other advanced methods that could deliver higher performance and superiority. In this paper we propose an increasingly used data mining approach in an attempt of identifying important internal or external variables in evaluating credit risk. The means applied is a random forest based method. Two different types of decision trees are used to feed into the integration of our recommended model, Classification and Regression Trees (CART) and Conditional Inference Trees (CIT). Random forests, assembled using CART or CIT, provide similar variable importance measures ranking the predictors with respect to their association with the response. Compared to Bayesian method proposed by Cerchiello and Giudici (2013), random forest approach was superior and produced higher performance in predicting default when it has been applied to a European credit risk database.

Visualization and Measuring of Dynamic Customer Satisfaction: a Banking Case

Caterina Liberati, University of Milano-Bicocca

Abstract: Customer Satisfaction for banking services is, arguably, a construct that develops and changes over time for a number of different endogenous and exogenous factors (modification of products, transparency of banking transactions and financial services, customer relationships, changes of market conditions an so on). Measuring change requires a longitudinal perspective, and most of the times such perspective is missing in the market research. This paper aims to analyse the customer evaluation evolution of the banking services, in order to catch differences among the clusters and time lags through a dynamic factorial model (STATIS). Moreover, while existing techniques compare temporal trajectories using dissimilarity measures, an additional innovative aspect of this work lies in the proposal of an new index can summarize some characteristics of the trajectories such as distance covered, the shape and direction. A real case of study on customers of an Italian bank is presented. Results based on a sample of 27000 instances per 3 waves, obtained via a questionnaire framed according to SERVQUAL model, reveal the effectiveness of such an approach.

Comparison of models for time series forecasts of tourism in Argentina

Federico Armentano, Universidad Nacional de Rosario

Abstract: Tourism contribute significantly to the economic growth of many countries and regions. Given the rise of tourism demand in recent decades, accurate predictions of future trends in tourism demand are vital. The forecast has an important role in the planning process of the tourism industry.This undoubtedly has generated a considerable degree of interest in academics and professionals who wish to understand this phenomenon and future trends. The models and forecasts of tourism have been important areas of research in the last five decades.. The overall objective of this paper is to analyze time series tourism in Argentina using different models to determine which ones are the best for forecasting, both being specific and accurate forecasts intervals. We will compare the performance of different time series models such as models: ARIMA state space innovations and Naive method. Comparing the performance of different models relating especially to forecast their behavior, using measures of forecast accuracy.