Title: Active Diagnosis via AUC Maximization: An Efficient Approach for Multiple Fault Identification in Large Scale, Noisy Networks

Author: Gowtham Bellala Advisor: Clayton Scott
Dept: Electrical Engineering & Computer Science Year: 5th Candidacy Status: Yes

The problem of active diagnosis arises in several applications such as medical diagnosis, fault diagnosis in nuclear plants, power-delivery systems, communication networks, and computer networks. In these problems, the goal is to rapidly identify the binary states of a set of objects (e.g., faulty/working states of components) by sequentially selecting, and observing, (noisy) responses to binary valued queries. Current algorithms in this area rely on belief propagation for active query selection. These algorithms have an exponential time complexity, thus making them slow and even intractable in large networks. In this paper, we propose a novel rank-based greedy algorithm that sequentially chooses queries such that the area under the ROC curve (AUC) of the rank-based output is maximized. The AUC criterion allows us to make a simplifying assumption that significantly reduces the complexity of active query selection (from the current exponential complexity to polynomial complexity), with little or no compromise on the performance quality. We demonstrate the performance of the proposed algorithm on real world, large scale computer networks.

Title: Assessing Functional Connectivity in Neuronal Networks from Spike-Train Data: A New Class of Models and Likelihood-Based Inference

Author: Kohinoor Dasgupta Advisor: Vijay Nair, Stilian Stoev and Xuanlong Nguyen
Dept: Stat Year: 4th Candidacy Status: Yes

Recent advances in technology have allowed neuro-scientists to collect large amounts of electro-physiological data at fine time scales. An important class of such data is multi-neuronal spike trains - time sequences of firings of a group of neurons. Identifying the functional connectivity of the neurons from spike-train data has been a problem of considerable interest in recent years. This paper proposes a new class of models for characterizing the dependence and connectivity among the neurons over time. Likelihood-based
methods for estimating the underlying parameters including the connectivity matrix and base firing rates have also been developed. Asymptotic theory for the estimators is currently under investigation. Our inference methods are shown to be considerably less complex than other comparable methods in the literature and are illustrated on simulation and real data.

Title: Analysis of Periodontal Data using Circular Statistics

Author: Samopriyo Maitra Advisor: Dr. Thomas Braun
Dept: Biostatistics Year: 4th Candidacy Status: Yes

Periodontal disease is moderately prevalent in the United States and is a common cause of tooth loss in adults. The severity of periodontal disease is usually quantified based upon the magnitudes of several tooth-level clinical parameters, the most common of which is clinical attachment level (CAL). Visualizing the mouth as a circle and the teeth as points located on the circumference of the circle, we propose the use of circular statistical methods to determine the mean locations of diseased teeth. We assume the directions of diseased teeth, as determined by their tooth averaged CAL values, to be observations from a Von Mises distribution, the mean of which is a function of tooth-level and mouth-level covariates. Because multiple teeth from a subject are correlated, we use a bias-corrected generalized estimating equation approach (Mancl and DeRouen;2001) to obtain robust variance estimates for our parameter estimates. Via simulations of data motivated from an actual study of periodontal disease, we demonstrate that our methods have excellent performance in the moderately small sample sizes common to most periodontal studies.

Title: Threshold Estimation Using p-values

Author: Atul Mallik Advisor: Michael Woodroofe and Moulinath Banerjee
Dept: Stat Year: 3rd Candidacy Status: Yes

We seek to identify the threshold value at which a real valued function takes off from its baseline level, under regression and multiple dose-response setting. This is relevant to a broad range of problems, e.g., estimating the minimum effective dose level in certain dose-response models in pharmacology, detecting tidal disruptions in dwarf spheroidal galaxies by Milky way, determining high pollution zones in a densely inhabited region, advent of global warming etc. An important case involves the baseline set having the form \([0, d]\), the unknown \(d\) being the threshold. On this set, the function stays at its baseline value (minima or maxima), not necessarily known, and then takes off. The approach involves fitting stumps to p-values obtained from tests conducted at different points/bins under the hypothesis that the function is at its baseline level. This works well owing to the fact that the p-values exhibit a dichotomous behavior over the base-
line region and its complement. This problem has natural connections to change point estimation. The procedure is consistent under minimal conditions, involves at most one tuning parameter and is computationally easy to implement. It also attains the optimal rate of convergence under certain assumptions. The asymptotic distribution has been derived. Subsampling works for constructing confidence intervals for this procedure. This is a joint work with Moulinath Banerjee and Bodhisattva Sen.

Title: Synthetic Data for Small Geographic Areas

Author: Joseph W. Sakshaug Advisor: Trivellore Raghunathan
Dept: Survey Methodology Year: 4th Candidacy Status: Yes

Demand for small area estimates is growing among a variety of stakeholders who use these data to study issues affecting local communities. Statistical agencies regularly collect data from small geographic areas but are prevented from releasing small area identifiers due to disclosure concerns. Several disclosure control methods are used to disseminate micro-data, including summary tables, suppression of geographical details, and Research Data Centers, but none of these methods is ideal for meeting the growing demand for small area datasets. This research tests a new method for disseminating public-use micro-data that contains more geographical details than are currently being released. Specifically, the method replaces the observed micro-data with fully-synthetic, or imputed, micro-data generated from a posterior predictive distribution. A hierarchical Bayesian model is used to preserve the small area inferences and simulate the synthetic data. Confidentiality protection is enhanced because no actual values are released. The synthetic data is evaluated by comparing inferences obtained from the synthetic data with observed data from the 2005-2007 American Community Survey.

Title: A Reassessment of Consensus Clustering for Class Discovery and Cluster Stability

Author: Yasin Senbabaoglu Advisor: Jun Li and George Michailidis
Dept: Bioinformatics Year: 4th Candidacy Status: Yes

Consensus clustering (C-C) is an unsupervised class discovery method highly popular for defining sample subtypes using high-dimensional data. "Consensus" rate between two samples measures how frequently they are grouped together in multiple clustering runs under perturbation. The matrix of pairwise consensus rates has been used for visualization, clustering, defining optimal number of clusters and assessing cluster stability. However, statistical properties of C-C, especially for the null situation of uniform or unimodal data, have not been examined in detail. We evaluated the performance of C-C in (1) a real dataset previously reported to contain four subtypes by using C-C (gene expression data for TCGA’s glioblastoma multiforme tumors) and (2) a series of sim-
ulated data drawn from a homogeneous population (i.e., no structure). We found that many unsupervised partitioning methods were able to divide homogeneous data into pre-specified numbers of clusters, and C-C was able to show apparent stability of such chance partitioning of random data. Summary statistics of C-C for the real dataset were often within the null distribution. These results suggest that C-C needs to be applied with caution as it is prone to over-interpretation: finding apparent structure when there is none, or showing cluster stability when it is subtle. If samples are not well-separated, it is necessary to objectively evaluate the evidence of clustering, and if appropriate, consider alternative models such as partial membership models or continuous distributions. C-C is a powerful tool for minimizing false negatives in the presence of genuine structure, but in exploratory phase of many studies can lead to false positives if not accompanied by a suitably formed null distribution.

Title: Multi-sensor image fusion using local dimension

Author: Kumar Sricharan Advisor: Alfred O. Hero III
Dept: Electrical Engineering & Computer Science Year: 5th Candidacy Status: Yes

Advances in technology has resulted in acquisition and subsequent fusion of data from multiple sensors of possibly different modalities. Fusing data acquired from different sensors occurs near the front end of applications and therefore becomes a critical bottleneck. In this work, we are specifically concerned with fusing hyper-spectral images of the earth’s surface, which can then be used for quantitative characterization of surface features. We take the following novel approach to fusion based on dimension estimation: In a (M) channel hyper-spectral response of a region, the effective ‘intrinsic’ dimension (m) of the response will typically be much smaller than (M), i.e., the response will lie on a smaller (m)-dimensional sub-manifold with m<<M. Furthermore, we conjecture that different material on the earth’s surface has different reflectance properties. Therefore, regions with a wider range of hyper-spectral responses should have a higher ‘intrinsic’ dimension, as opposed to regions with a narrow range of responses. The local dimension of each region can therefore be used to distinguish between different types of material on the surface, for e.g., water and land masses. To this end, we propose a local dimension estimator based on weighted k-nearest neighbor graphs for high-dimensional data and provide confidence intervals on the local dimension. The fused local dimension image can subsequently be used to perform image segmentation. Furthermore, the confidence intervals can be used to set thresholds for desired false alarm rates in anomaly detection applications. We show specific examples of our method on the AVIRIS (Airborne Visible Infra-Red Imaging Spectrometer) data set.

Title: Two-stage semi-parametric Bayesian model-based inference for the finite population total in probability proportional to size samples
We study estimation and inference on data arising from finite population surveys with unequal probabilities of selection. We focus on the problem of estimating the population total in probability proportional to size (PPS) sampling. In PPS sampling, the sizes for non-sampled units are not required for the usual Horvitz-Thompson or Hajek estimates, and this information is rarely included in public use data files. However, Zheng and Little (JOS 2003) show that including the sizes of the non-sampled units as predictors in a spline model can result in improved estimates of the finite population total, without strong parametric assumptions. Assuming, external information on the mean of the size variables in the population, we use a two-stage procedure to estimate the finite population total of continues survey variables. In first stage, we predict the non-sampled sizes using a Bayesian Bootstrap (BB) model with weakly informative prior distributions, at the second stage, we use a Bayesian penalized spline model to predict the outcome for the non-selected units based on the predicted sizes. We introduce a Bayesian method for variance estimation inspired by decomposing the posterior predictive variance of the proposed estimator into components corresponding to estimating the non-sampled sizes and spline prediction model with known sizes. Our numerical studies indicate that in addition to providing improved point estimates, the proposed estimator offers a reliable inferential framework for interval estimation. We find that, in populations with homoscedastic errors or with limited curvature, our method is fairly robust to the distributional assumption on the errors, as well as the underlying assumption of the BB, that all of the sizes in the population are captured in the sample. The performance of the method is evaluated on two data sets from the US census.

II. Poster Presentations

Title: Random Matrix Theory Approach to Classification and Clustering

Author: Nicholas Asendorf Advisor: Raj Rao Nadakuditi
Dept: Electrical Engineering & Computer Science Year: 1st Candidacy Status: No

Classification and clustering algorithms are used to linearly separate data believed to belong to multiple classes. Using kernel functions, we may extend these algorithms to separate classes with nonlinear dependencies. We consider a high-dimensional setting containing noisy low-dimensional signals where (linear or kernel) PCA or SVD is used as the first step in the classification or clustering algorithm. In settings with a low signal-to-noise ratio (SNR), the performance of these algorithms degrades due to the increased overlap of the training clusters as well as the errors in the estimation of the true signal subspace. We use recent results from random matrix theory to develop a data-driven technique for optimally approximating the true signal subspace given a noisy estimate.
Using this improved signal subspace estimate in clustering and classification algorithms, we show an improvement in classification ability. Extensions to the kernel setting are discussed.

Title: **Imputing Data Via Quasi-Likelihood Regression on a Latent Variable**

**Author:** Jonggyu Baek  
**Advisor:** Dr. Brisa N. Sanchez  
**Dept:** Biostatistics  
**Year:** 3rd  
**Candidacy Status:** No

Missing data arise in many applications, and many methods to impute continuous variables exist. However, methods to impute data from continuous distributions with finite support are less common. In the application motivating our work, a continuous variable representing the proportion of unfit children within gender and race subgroups of school children in the State of California was missing from many observations. Although a Beta regression with random effects could be one way to be used to impute missing values, imputed values may be biased because they tend to be shrunk towards the estimated parameters from the assumed distribution. Instead, we employed a mixture of Beta and Uniform distributions. For a given observation, the probability of arising from either the Beta or Uniform distribution was estimated using auxiliary predictors. That is, in a classic mixture problem the probability that observations are from a given distribution is computed using the observed value of the variable (e.g., proportion unfit). In our case, the value is missing and we wish to impute it. However, by using our proposed quasi-likelihood regression method, where the unobserved mixture weight is regressed on mixture predictors within an E-M algorithm, observations with missing values can be assigned to an appropriate target distribution, and then imputing procedure for missing values can be applied.

Title: **Isotonic Regression with Short and Long-Range Dependent Errors**

**Author:** Pramita Bagchi  
**Advisor:** Moulinath Banerjee and Stilian Stoev  
**Dept:** Statistics  
**Year:** 1st  
**Candidacy Status:** No

Let \( x(t) t \in [0,1] \) be a monotone non-decreasing function. We consider the isotonic regression problem of estimating \( x(t) \) from the data \( y(t_{i,n}) = x(t_{i,n}) + \epsilon_i, \ i = 1, \cdots, n, \) where \( t \) is different from the given design points \( t_{i,n} := \frac{i}{n} \in [0,1]. \) This is an important problem arising in many applications such as climate studies, economics, current status data in biostatistics, among many others. When the errors \( \epsilon_i \)'s are independent and Gaussian, the well-known Pooled Adjacent Vector Algorithm provides the maximum likelihood estimate (MLE) of \( x(t). \) Its asymptotic properties are well understood. In many applications, however, the errors \( \epsilon_i \)'s are dependent. Motivated by the work of Banerjee & Wellner (2001), our goal is to study the pseudo-LRS. This statistic does not involve the slope of \( x, \) which is difficult to estimate and arises in the asymptotic
distribution of MLE. We present some simulation results, which show that the pseudo-LRS with suitably modified scaling will have the same asymptotic distribution in the case when the $\epsilon_i$’s are weakly (short-range) dependent. The situation when the $\epsilon_i$’s are long-range dependent is completely different. In this case, the pseudo-LRS diverges because $\sum_{k=1}^{\infty} \text{Cov}(\epsilon_k, \epsilon_0) = \infty$. The ratio of the pseudo-LRS and the Residual Sum of Squares, however, is likely to have a non-degenerate limit distribution, as confirmed by our simulations. We conclude with some theoretical intuition behind these preliminary results. We expect that the proposed statistics to find many applications as they do not assume independence on the errors and do not involve estimating the scaling factor for LRS.

Title: Discovering Graphical Granger Causality in Sparse High-dimensional Networks with Inherent Grouping Structure

Author: Sumanta Basu  Advisor: George Michailidis  Dept: Statistics  Year: 2nd  Candidacy Status: Yes

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

Title: Statistical analysis of Raman-spectroscopic images, with applications to clinical assessment of bone structure parameters

Author: Ruth Cassidy  Advisor: Liza Levina and Kerby Shedden  Dept: Stat  Year: 3rd  Candidacy Status: Yes

Raman-spectroscopic imaging is a non-invasive technique used to study the chemical composition of a given material or living system. Images can be used for either pixel- or subject-level prediction. Analysis of Raman-spectroscopic images can be challenging due to the high-dimensional nature of spectra and the large amount of noise. Existing dimension reduction methods for Raman spectra often assume an idealized linear superposition model, which may be inadequate in practice. Moreover, many methods ignore the spatial structure and/or the functional nature of the data. We propose a method
that models the salient features of spectral peaks while accounting for spatial information via the use of regularization. The method is demonstrated on an application to bone tissue classification.

**Title: Testing for Serial Homogeneity in the Correlation of a Longitudinally Measured Biomarker and Disease**

**Author:** Su Chen  **Advisor:** Thomas Braun  
**Dept:** Biostatistics  **Year:** 2nd  **Candidacy Status:** Yes

Our methods are motivated by studies that longitudinally collect values of a continuous biomarker with the goal of predicting for future disease occurrence, which is quantified by a single continuous outcome. The association of the biomarker at each time point with future disease is summarized with a Pearson’s correlation coefficient, from which we derive an across-time pooled correlation coefficient. Our methods derive a test for assessing whether or not the serial correlation of the biomarker with disease is homogeneous, thereby allowing for computation of the pooled correlation coefficient. We derive the asymptotic covariance matrix of the time-specific correlations, based on which a Wald test for homogeneity is derived. We also derive a Wald test based on Fisher’s $Z$-transformation of the correlation coefficients. We also present findings from a variety of simulation settings that demonstrate both tests have nominal size and sufficient power.

**Title: Media Coverage of Polling Methodology in the 2010 Alaskan Senate General Election**

**Author:** Benjamin Duffey  **Advisor:** Michael Traugott  
**Dept:** Survey Methodology  **Year:** 2nd, Masters  **Candidacy Status:** No

In multi-candidate, state-level elections, the public relies largely on media coverage of polling data to inform their voting strategy. In this paper, I use the 2010 Alaskan Senate Race to analyze the nature of poll coverage in elections, and assess the potential impact on voting strategy. First, I establish a basic notion of poll quality using methodological criteria such as field period, collection mode, and sample size, and then identify their relationship with stability and accuracy of estimates. Finally, I provide a content analysis of methods coverage from both local and national media sources, and then discuss potential effects on voting behavior.

**Title: Assessing the Causal Effect of Treatment Dosages in the Presence of Self-Selection**

**Author:** Xin Gao  **Advisor:** Dr. Michael R. Elliott
To make drug therapy as effective as possible, patients are often put on an escalating dosing schedule. But patients may choose to take a lower dose because of side effects. Thus even in a randomized trial, the dose level received is a post-randomization variable, and comparison with the control group may no longer have a causal interpretation. Hence we use the potential outcomes framework to define pre-randomization "principal strata" from the joint distribution of doses selected under control and treatment arms, with the goal of estimating the effect of treatment within the subgroups of the population who will select a given set of dose levels. We utilize Bayesian framework with Markov chain Monte Carlo algorithm for analysis of a randomized clinical study on painful bladder syndrome, and compare the analysis results from our proposed model with traditional analysis method. Simulation results show that the estimates of interests in our proposed causal model have correct coverage.

Title: Adjusted Prostate Cancer Mortality Rates Under Mis-attributed Cause of Death

Author: Jinkyung Ha Advisor: Dr. Alexander Tsodikov
Dept: Biostatistics Year: 4th Candidacy Status: Yes

One of possible explanations for the recent trend in prostate cancer mortality rates is that a certain number of men who die of other causes may be mislabeled as dying from prostate cancer. To assess this hypothesis, we consider the problem of estimating mortality rates where other cause of death can be mis-attributed to prostate cancer. We first propose a mortality model which is a convolution of cause-specific survival and distributions for variables measured at diagnosis. Under proportional hazards assumption, we propose survival estimates adjusted for attribution bias, which are used to correct inflated mortality rates. With a variety of mis-attribution models, a sensitivity analysis is performed to assess the effect of attribution bias on the recent trend in mortality rates. Data used in this article is obtained from the Surveillance, Epidemiology, and End Results Program.

Title: Series Estimators with a Penalization

Author: Ashley Holland Advisor: Matias Cattaneo and Virginia Young
Dept: Mathematics Year: 5th Candidacy Status: Yes

Nonparametric econometrics seeks to find the expected value of a response variable given values of a covariate set, without assuming a form for the regression function. Penalized spline estimation, a type of nonparametric estimation, uses splines (a type of series) and a criterion function that penalizes for an increased roughness of the estimate. This
framework encompasses both regression and smoothing splines as it makes no assumptions about the number of basis functions with respect to the number of data points. The current literature for penalized splines does not provide rates of convergence for all possible sequences of basis functions and values of the smoothing parameter in the multivariate context, nor does it present asymptotic normality (even in the univariate case). Our research allows for various sequences of basis functions and values of the smoothing parameter, while still providing rates of convergence and asymptotic normality. It also determines how the smoothing parameter and the number of basis functions should vary with the number of data points in order to minimize the estimates rate of convergence. Thus, it can guide practitioners in the best method of spline estimation and provide assurance of convergence and normality for any chosen sequence. This work unifies current results for regression and smoothing splines into one theory, including the results for univariate penalized splines with a restriction on the number of basis functions by Claeskens et. al. (2008), regression splines by Newey (1997), and smoothing splines by Cox (1984).

Title: Model selection for the number of distinct genetic signals

Author: Ming Chi Hsu Advisor: Kerby Shedden and Ji Zhu
Dept: Stat Year: 4th Candidacy Status: Yes

For several common human diseases, genome-wide association studies (GWAS) have resulted in strong evidence that genetic risk factors exist in multiple distinct regions of the genome. Within each such region, one or more specific genetic variants may confer risk, but efficiently summarizing the information contained in multiple dependent variants is challenging. To date, only a fraction of known genetic heritability for common diseases has been explained by multigenic models. The primary goal of our analysis is to devise a means for assessing whether one of the genetic variants dominates the others. In this case, we say that there is only one independent signal in the data. If two or more genetic variants values have approximately the same size, we say that there are multiple distinct signals in the data. Here we consider the likelihood based approach under two hypotheses and consider their predictive performance in a realistic setting for genetic data. We propose that estimates of the relative predictive performances on the same data set can be used to make inferences about the underlying structure of genetic effects.

Title: Semi-parametric Regression Inference for Age-Stage at Diagnosis Relationship in Cancer Studies

Author: Chen Hu Advisor: Dr. Alexander Tsodikov
Dept: Biostatistics Year: 4th Candidacy Status: Yes

In cancer studies, it is of great interest to understand the relationship between dis-
ease progression (from non-metastasis to metastasis) and time-to-diagnosis, as well as what factors affect such relationship. In many observational studies and cancer registry datasets, age and stage at cancer diagnosis are often jointly observed but contain only partial information of interest due to cross-sectional data structure. Joint modeling such composite response of survival and binary outcome may provide more efficient results than using either component alone. We address this problem through a semi-parametric regression model for stage-specific cancer incidence. Constructed through a series of Cox relative risk models with time-dependent covariates, our model can be represented as a transformation model induced by a complex non-proportional frailty. We propose an estimating equation based estimation approach and a nonparametric maximum likelihood estimation approach for such transformation models. The estimation procedure and asymptotic variance of proposed estimator can be easily implemented and obtained numerically. The methodology is illustrated by Monte Carlo simulation studies and real prostate cancer incidence data from the Surveillance, Epidemiology and End Results (SEER) program.

Title: On the extension of Isomap based on the generalized BHH theorem

Author: Sung Jin Hwang Advisor: Alfred Hero III Dept: Electrical Engineering & Computer Science Year: 6th Candidacy Status: Yes

The generalized Beardwood-Halton-Hammersley (BHH) theorem states the relation between power-weighted graphs on the sample points and convergence of certain types of quasi-additive Euclidean functionals. We observe that such convergence values have geometric interpretation and it points to the deformations of the measure of the space where the probability distribution is supported. Such deformations based on the BHH theorem essentially lead to probability density deformations and we will show that they have some favorable properties to be useful in statistical learning. We apply this idea of deformation to the manifold learning problems. In particular, we are interested in the geodesic distances of the space through the sample points under such deformations. There is a natural and direct application to Isomap algorithm, and we propose an extension $p$-isomap based on power-weighted graphs. We will also try to establish its connections to anisotropic diffusion map.

Title: Transfer Learning for Automatic Gating of Flow Cytometry Data

Author: Gyemin Lee Advisor: Clayton Scott Dept: Electrical Engineering & Computer Science Year: 6th Candidacy Status: Yes

Flow cytometry is a technique for rapidly quantifying physical and chemical properties of large numbers of cells. In clinical applications, flow cytometry data must be manually "gated" to identify cell populations of interest. Because multiple, iterative gates are
often required to identify and characterize these populations, several researchers have investigated statistical methods for automating this process, most of them falling under the framework of unsupervised learning and mixture model fitting. We view the problem as one of transfer learning, which can leverage existing datasets previously gated by experts, while accounting for biological variation, to automatically gate a new flow cytometry dataset. We illustrate our proposed method by automatically gating lymphocytes from peripheral blood samples.

Title: Using Sequential, Multiple Assignment, Randomized Trial (SMART) Designs to Build Individualized Treatment Sequences

Author: Huitian Lei Advisor: Susan Murphy
Dept: Stat Year: 1st Candidacy Status: No

Long term management of chronic disease calls for the use of adaptive interventions, in which interventions are chosen based on individuals characteristic and are adjusted repeatedly over time. Sequential Multiple Assignment Randomized Trials (SMART) is used in evaluating adaptive interventions and in finding the optimal one. In this review, we provide an overview of SMART alongside with comparison to alternative designs. Selected examples of SMART studies are described and contrasted. Data analysis methods are illustrated using an extended example involving a SMART trial studying behavioral and pharmacological interventions for alcohol dependent individuals. Secondary analysis which aims at greater individualization is also discussed.

Title: Surrogate tuning in imbalanced learning

Author: Tzu-Yu Liu Advisor: Clayton Scott
Dept: Electrical Engineering & Computer Science Year: 4th Candidacy Status: Yes

Classification on imbalanced datasets is an important problem, in which the class of interest is outnumbered by the majority class. At the algorithmic level, uneven cost, uneven margin have been studied and shown to improve the performance. However, these results are usually based on comparing sensitivity, specificity, g-mean, AUC, and neither of them is directly related to the topic of being classification calibrated. We compare these different approaches that deal with imbalanced datasets, and discuss their weighted risks that relate to classification calibration.

Title: A Policy Search Method for Estimating Treatment Policies

Author: Xi Lu Advisor: Susan Murphy
Dept: Stat Year: 2nd Candidacy Status: No
A treatment policy or dynamic treatment regime is a sequence of decision rules. At each stage a decision rule inputs patient history and outputs a treatment. The value of a treatment policy is the expected outcome when the policy is used to assign treatment. Data from sequential, multiple assignment, randomized trials can be used to estimate an optimal treatment policy. One approach is to parameterize the policy value (Robins et al. 2008); this may result in bias if the model is misspecified. Alternately the value of any specific policy can be estimated non-parametrically; however this method may have high variance. We propose a new method in which each stages treatment effect or blip is parameterized. These treatment effects are easily interpretable to scientists and thus more meaningfully parameterized than the policy value. To estimate the parameters we utilize a telescoping sum representation of the policy value and uses ideas from missing data theory. We illustrate the proposed method with data from the ExTENd trial, a recently completed alcohol dependence study.

Title: Smooth Plaid Models: A Dynamic Clustering Algorithm with Application to Electronic Financial Markets

Author: Shawn Mankad  Advisor: George Michailidis and Andrei Kirilenko
Dept: Stat  Year: 3rd  Candidacy Status: Yes

Not so long ago most securities were traded in face-to-face markets. Now trading in most securities is increasingly done electronically by computer algorithms. Did this technological change also change the strategies employed by the traders? Or, did computer algorithms simply automate what traders used to do face-to-face? To answer these questions, we design and employ a dynamic online bi-clustering method based on plaid models. Specifically, we introduce a regularization framework that smooths the model parameters over time, thus allowing us to identify persistent groups (traders) and the critical features that separate them consistently over time. We apply our method to cluster traders in the E-mini S&P 500 stock index futures contract. We employ audit-trail, transaction-level data (with identities of traders) for all transactions in the September 2009 E-mini S&P 500 futures contract during the month of August 2009. On any given day, 10-12 thousand traders engage in up to 1 million transactions. Yet, we find that traders can be consistently separated into five groups that we characterize as high frequency traders, market makers, opportunistic traders, fundamental traders, and small traders. We argue that the emergence of high frequency trading strategies is the direct consequence of technological changes in financial markets.

Title: Reduced Rank Ridge Regression and its Kernel Extensions

Author: Ashin Mukherjee  Advisor: Ji Zhu and Naisyin Wang
Dept: Stat  Year: 3rd  Candidacy Status: Yes
In multivariate linear regression, it is often assumed that the response matrix is intrinsically of lower rank. This could be due to the correlation structure among the prediction variables or the coefficient matrix being lower rank. To accommodate both, we propose a reduced rank ridge regression for multivariate linear regression. Specifically, we combine the ridge penalty with the reduced rank constraint on the coefficient matrix to come up with a computationally straightforward algorithm. Numerical studies indicate that the proposed method consistently outperforms relevant competitors. A novel extension of the proposed method to the reproducing kernel Hilbert space (RKHS) set-up is also developed.

Title: Adaptive Search for Sparse Moving Targets under Resource Constraints

Author: Gregory Newstadt Advisor: Alfred Hero III
Dept: Electrical Engineering and Computer Science Year: 4th Candidacy Status: Yes

Previous work with adaptive search for sparse static targets under resource constraints has produced two-stage allocation policies with desirable properties for asymptotic SNR, such as convergence to the true region of interest (ROI) and optimal gains over an exhaustive search. This work investigates the problem of extending the allocation policies to T ≥ 2 stages, with particular emphasis on cases where the SNR for any particular stage is considerably less than the asymptotic level. Furthermore, this work provides a new formulation that can account for non-static targets, including a transition model for target locations and a population model to account for targets that leave or enter the scene. Under this formulation, a dynamic adaptive resource allocation policy (D-ARAP) is proposed with a heuristic derivation. It is shown through Monte Carlo simulation that this policy provides significant gains over an exhaustive search in both static and dynamic target cases. Moreover, the comparison is made to the optimal solution of partially observable Markov decision processes (POMDPs).

Title: In Defense of Randomization: A Subjectivist Bayesian Approach

Author: Raphael Nishimura Advisor: Jim Lepkowski
Dept: Survey Methodology Year: 1st Candidacy Status: No

In research situations usually approached by Decision Theory, it is only considered one researcher who collects a sample and makes a decision based on it. It can be shown that randomization of the sample does not improve the utility of the obtained results. Nevertheless, we present situations in which this approach is not satisfactory. First, we present a case in which randomization can be an important tool in order to achieve
agreement between people with different opinions. Next, we present another situation in which there are two agents: the researcher - a person who collects the sample; and the decision-maker - a person who makes decisions based on the sample collected. We show that problems emerge when the decision-maker allows the researcher to arbitrarily choose a sample. We also show that the decision-maker maximizes his expected utility requiring that the sample is collected randomly.

Title: Vitamin C Improves Short-term Memory Performance Among Adults: A Randomized, Placebo-Controlled, Double-Blind Clinical Trial

Author: Rachel Petrak Advisor: Brenda Gillespie
Dept: Health Behavior & Health Education Year: 2nd Candidacy Status: No

Vitamin C is crucial for proper neuro-cognitive development and as a co-factor in a variety of biological functions. Clinical studies have found special populations of children to show improvements in attention and mood with vitamin C supplementation (Kidd, 2000). Among older adults, vitamin C can contribute to improved memory (Gale et al, 1996), and reduced risk of Alzheimer’s disease (Fotuhi et al, 2008). In a general adult population, there has been very little controlled research on the use of vitamin C to improve short-term memory performance. To address this gap, we conducted a randomized, placebo-controlled, double-blind clinical trial to test the efficacy of vitamin C in improving short-term memory performance among an adult population. This research was designed and conducted as part of a graduate course in Biostatistics. 14 participants were recruited from the course, and assigned randomly to vitamin C supplementation (250 mg daily, oral) or placebo. At baseline and once weekly for 2 additional weeks, participants completed a brief survey of dietary intake of vitamin C, sleep, mood and study compliance (when appropriate), and a 10 minute memory test (24 trials of a reverse digit recall task). Participants were included in analyses using an intent-to-treat approach, and Maximum Likelihood Estimation was used to look for treatment effects in participants performance on the memory test. By week 2 of the trial, participants who were in the vitamin C group had significantly better digit-string recall than participants in the placebo group, (p= 0.0063, Chi-sq = 0.3359). This effect was not observed at week 1. Although further research is needed to look at the role of vitamin C in improving performance on a wider variety of tasks, these data provide preliminary evidence for the efficacy of vitamin C supplementation on improving short-term memory performance among a general adult population.

Title: On the extension of Isomap based on the generalized BHH theorem

Author: Arnau Tibau Puig Advisor: Alfred Hero III
Dept: Electrical Engineering & Computer Science Year: 3rd Candidacy Status: Yes
We consider the problem of estimating a signal subspace from noisy and misaligned observations. This problem arises in the context of multivariate time series analysis, where sampling times cannot be synchronized across different observations, due to technical or physical limitations. The most popular approach for estimating a signal subspace is called Principal Component Analysis (PCA). PCA estimates the signal from the eigendecomposition of the empirical covariance matrix of the data, where it is assumed that the population covariance is a low-rank perturbation of the (usually isotropic) noise covariance matrix. In this work we first study the effects of misalignment on the performance of the PCA estimators, both in the small ($p/n \to 0$) and the large-dimensional setting ($p/n \to c > 0$). Specifically, we show that (i) misalignments introduce bias in the PCA estimator and (ii) they have a negative effect in the phase transition point from which estimation of the signal from the eigenvectors of the observed covariance is possible. In the second part of this work we consider two algorithms to jointly estimate the signal subspace and the misalignment effects. Both algorithms provide approximate solutions to the Misaligned PCA (Mis-PCA) problem, which is a combinatorial optimization problem that arises as a maximum-likelihood estimation under a misaligned data model. We study the performance of both algorithms in simulations and apply our methodology to the study of immune system responses of infected individuals from longitudinal gene expression data.

**Title: Using Displacement Effects to Study Returns to Education**

**Author:** Adam Sales  
**Advisor:** Ben Hansen  
**Dept:** Stat  
**Year:** 4th  
**Candidacy Status:** No

In an observational study, let each subject, indexed by $i$, be characterized by $(x_i, y_{i0}, y_{i1}, Z)$, where $x_i$ and $u_i$ is a vector of covariates, $y_{i0}$ and $y_{i1}$ are potential outcomes, following Neyman (1923) and Rubin (1974), and $Z \in \{0, 1\}$ is an indicator of treatment status. For a given percentile $p$ of $y_0$, let $y_{k0}$ be an observed potential response such that $F_{Y_0}(y_{k0}) = p$, so $y_{k0}$ is the $p^{th}$ percentile of the distribution of the subjects’ potential outcomes under the control condition. Subject $i$ experiences a “displacement” (Rosenbaum 2001) around $p$ if $y_{i0} < y_{k0} < y_{i1}$, so treatment moves subject $i$ from below $y_{k,0}$ to above. Rosenbaum (2001) provides a method to estimate a $(1 - \alpha)$ confidence interval for $A_0 = \#\{i : Z_i = 1 and y_{i,0} < y_{k,0} < y_{i,1}\}$ for a given $p$ when subjects are stratified into matched pairs. We apply Rosenbaum’s general method to a particular problem in labor economics and education research. In 1947, the United Kingdom increased the minimum age at which one could leave formal schooling from 14 to 15 years. Oreopoulos (2006) used this “regression discontinuity” design to estimate the local average treatment effect of an extra year of compulsory education on wages. In this paper, we analyze Oreopoulos’ data using the displacement effects framework. We believe that this has several advantages over the traditional regression-based method Oreopoulos used, which we will explicate. In so doing, we expand Rosenbaum’s method to arbitrarily-structured strata, and show how to apply it in a regression-discontinuity setting. We also attempt
to shed light on the question of whether women tend to experience similar returns to education as their male counterparts.

Title: Asymptotics for Current Status Data Under Varying Observation Time Sparsity

Author: Runlong Tang Advisor: Moulinath Banerjee and George Michailidis
Dept: Stat Year: 6th Candidacy Status: Yes

In this paper, we consider the isotonic regression estimation and hypothesis testing for a survival distribution function at a point in the current status model with an equally-spaced grid distribution for the covariate. The grid resolution is specified as $cn^{-\gamma}$ with a scaler $c > 0$ and an order of the grid resolution $\gamma > 0$. The asymptotic results belong to three cases according to the value of $\gamma$. The case with $\gamma = \frac{1}{3}$ is the boundary case, whose limit distributions converge weakly to those of the other two cases with $\gamma \in (0, \frac{1}{3})$ and $\gamma \in (\frac{1}{3}, \infty)$ as $c$ goes to $\infty$ and $0$, respectively. This discovery claims that the counterparts of normal and chi-square distributions for isotonic regression problems are Chernoff and the so-called $D$ distributions. Further, we propose from the discovery an adaptive procedure to make statistical inference without the estimation or specification of $\gamma$, which is of practical interest. Finally, a simulation study on a practical version of the adaptive procedure is provided.

This is a joint work with professors Moulinath Banerjee and Michael R. Kosorok.

Title: Non-Linear De-trending In Panel Models To Estimate Macroeconomic Effects on Mortality

Author: Zhen Wang Advisor: Edward Ionides
Dept: Stat Year: 4th Candidacy Status: Yes

Previous investigations of the effect of macroeconomic conditions on mortality have often used fixed-effect panel models applied to mortality rates and unemployment rates in levels (or log-transformed). We show in the present investigation that this type of panel models often generates special patterns of the residuals and therefore violation of the regression model assumptions, that probably bias the effect estimates. Using panel regression models applied to differenced or Hodrick-Prescott-detrended data the high serial correlation of the observations is thoroughly removed and the patterns of the residuals disappear.

Title: Observational strategies associated with increased accuracy of interviewer observations: Evidence from the National Survey of Family Growth
Managers of surveys conducted using in-person (or face-to-face) interviews commonly request that interviewers collect observations on key features of selected households, given that interviewers are the eyes and ears of the survey organization out in the field. This is done in an effort to obtain as much auxiliary information on all sample units as possible, given trends of declining response rates in surveys of all formats worldwide. Unfortunately, the error properties of these types of observations and the implications of these error properties for the subsequent effectiveness of a variety of nonresponse adjustments have received only minimal research focus to date. In addition, no studies have investigated the strategies being used by interviewers in the field when they are making these observations. This research examines the associations of observational strategies used by field interviewers collecting face-to-face interviews from a large area probability sample with the accuracy of observations collected by those interviewers. A qualitative analysis shows that certain strategies are in fact associated with increased accuracy of the observations, and recommendations for future practice are presented.

Title: Tracking Communities in Dynamic Social Networks

Author: Kevin S. Xu Advisor: Alfred Hero III
Dept: Electrical Engineering & Computer Science Year: 4th Candidacy Status: Yes

The study of communities in social networks has attracted considerable interest from many disciplines. Most studies have focused on static networks, and in doing so, have neglected the temporal dynamics of the networks and communities. This paper considers the problem of tracking communities over time in dynamic social networks. We propose a method for community tracking using an evolutionary clustering framework that detects communities at each time step using an adaptively weighted combination of current and historical data. We apply the method to reveal the temporal evolution of communities in two real data sets.

Title: Finite population density estimation in unequal probability surveys

Author: Sahar Zanganeh Advisor: Roderick Little and Robert Keener
Dept: Stat Year: 4th Candidacy Status: Yes

We study the problem of estimating the finite population distribution function of a continuous survey response variable in probability proportional to size (PPS) sampling. In PPS sampling, the sizes for non-sampled units are not required for the usual Horvitz-Thompson or Hajek estimates, and this information is rarely included in public use data files. Previous studies have shown that incorporating information on the sizes of the
non-sampled units through semi-parametric models can result in improved estimates. We use Bayesian nonparametric mixture modeling with Dirichlet Process priors to estimate the density of size and survey response variables in situations where auxiliary information is only available on the sampled units. Estimation of medians, quantiles and other nonlinear statistics would be facilitated by having a good estimate of the density function and Bayesian methodology provides us with a unified approach for performing reliable inference. We compare our method to existing design-based alternatives and demonstrate its performance in estimating finite population quantiles.

Title: Differential effects of measurement error in outcome prediction

Author: Juan Zhang Advisor: Kerby Shedden
Dept: Stat Year: 4th Candidacy Status: Yes

Often regression models are defined in terms of independent variables (covariates) that are not measured perfectly, or for some reason are not directly observable. In such situations, we focus on the performance of prediction models using covariates measured with error. Obviously, predictability tends to decline with increasing measurement error magnitude. But in more detail, is the amount of decline strongly dependent on the structure of the outcome generating distribution $P(Y | X)$? What attributes of $P(Y, X)$ might increase the negative effects of measurement error on predictability is our main interest.

Title: A Pseudo-Bayesian Shrinkage Approach to Regression With Missing Covariates

Author: Nanhua Zhang Advisor: Dr. Roderick Little and Dr. Mike Elliott
Dept: Biostatistics Year: 4th Candidacy Status: Yes

We consider the regression of outcome $Y$ on regressors $W$ and $Z$ with some values of $W$ missing, when our main interest is the effect of $Z$ on $Y$, controlling for $W$. Three common approaches to regression with missing covariates are (a) complete-case analysis (CC), which discards the incomplete cases, and (b) ignorable likelihood methods, which base inference on the likelihood based on the observed data, assuming the missing data are missing at random (Rubin, 1976), and (c) non-ignorable modeling, which posits a joint distribution of the variables and missing data indicators. Another approach that has not received much attention is to drop the regressor variables containing missing values from the regression modeling (DV, for drop variables). DV is most useful when either (I) the regression coefficient of $W$ is zero or (II) $W$ and $Z$ are uncorrelated. We propose a pseudo Bayesian approach for regression with missing covariates, which compromises between the CC and DV estimates. We illustrate favorable properties of the method by simulation, and apply the proposed method to a liver cancer study. Extensions of the method to more than one missing covariates and to generalized linear models are also
Detecting communities within networks has attracted wide interest in several different fields including computer science, social science and biology. The stochastic block model provides a statistical approach to model networks with multiple communities. However, fitting the stochastic block model with real-world networks often yield poor community detection results, for it ignores the variation among nodes within a community. To resolve this issue, Newman (2010) proposed a degree-corrected block model which allows flexibility among vertex degrees within the same community. In this presentation, we establish a general theory for checking asymptotic consistency of community labels estimated by any criterion under the assumption of the degree-corrected block model. As examples, we obtain consistency conditions for several commonly used criteria under both uncorrected and corrected block models.

Since the seminal introduction of propensity score by Rosenbaum and Rubin (1983), propensity-score-based methods, including matching, stratification and weighting, have been widely used for drawing a causal inference in the behavioral and social sciences. However, the primary limitation of propensity score lies in the impossibility of capturing unobserved individual and contextual confounders. In this paper, we survey the performance of propensity-score-based methods under different types of observed and unobserved selection. First, we discuss population heterogeneity in causal inference, selection bias, the ignorability assumption, and propensity-score-based methods. Then, we investigate the implication of propensity-score-based methods in settings where the ignorability assumption partially or completely breaks down. Finally, with simulated data under different assumptions, we compare propensity-score-based methods and marginal-treatment-effect-based methods in the estimation and inference of causal parameters.