Department of Statistics - Statistics Seminar – Spring 2013

Statistics Seminars are on Mondays
Time: 12:00 - 1:00 PM
Location: Room 903, 1255 Amsterdam Avenue,
Tea and Coffee will be served before the seminar at 11:30 AM, Room 1025

1/28/2013

Speaker: Bin Yu
Departments of Statistics and EECS, University of California at Berkeley
www.stat.berkeley.edu/~binyu

Abstract
Reproducibility is imperative for any scientific discovery. More often than not, modern scientific findings rely on statistical analysis of high-dimensional data. At a minimum, reproducibility manifests itself in stability of statistical results relative to "reasonable" perturbations to data and to the model used. Jacknife, bootstrap, and cross-validation are based on perturbations to data, while robust statistics methods deal with perturbations to models.

In this talk, a case is made for the importance of stability in statistics. Firstly, we motivate the necessity of stability of interpretable encoding models for movie reconstruction from brain fMRI signals. Secondly, we find strong evidence in the literature to demonstrate the central role of stability in statis- tical inference. Thirdly, a smoothing parameter selector based on estimation stability (ES), ES-CV, is proposed for Lasso, in order to bring stability to bear on cross-validation (CV). ES-CV is then utilized in the encoding models to reduce the number of predictors by 60% with almost no loss (1.3%) of prediction performane across over 2,000 voxels. Last, a novel "stability" argument is seen to drive new results that shed light on the intriquing interactions between sample to sample varibility and heavier tail error distribution (e.g. double-exponential) in high dimensional regression models with p predictors and n independent samples. In particular, when $p/n \to \kappa \in (0.3, 1)$ and error is double-exponential, OLS is a better estimator than LAD.

2/4/2013

Speaker: Elizabeth Ogburn, Harvard University

Title: "Some challenges and results for causal and statistical inference with social network data"

Abstract:

Increasing interest in and availability of network data necessitates new methods for causal and statistical inference when observations are linked by network ties. My talk is motivated by the Health Outcomes, Progressive Entrepreneurship, and Networks (HopeNet) Study, which will collect three waves of complete social network data and implement clean water and microenterprise interventions in a small community in southwestern Uganda. Causal effects of interest include the effects of an individual's exposure to each intervention on his own outcome, and several different types of effects of an

individual's exposure on the outcomes of his social contacts. In order to clearly articulate these latter "interference" effects, I differentiate three different causal mechanisms that give rise to interference, defined as an effect of one individual's exposure on another's outcome, and briefly discuss new identification results for interference effects. I then turn to the problem of estimation when only a single network of non-independent observations is observed and the dependence among observations is informed by network topology. I explain why results on spatial-temporal dependence are not immediately applicable to this new setting and present some new methods for estimation in the presence of network dependence.

2/11/2013

Speaker: Tommy Wright, US Census Bureau

The Equivalence of Neyman Optimum Allocation for Sampling and
Equal Proportions for Apportioning the U. S. House of Representatives

Abstract
We present a surprising though obvious result that seems to have been unnoticed until now. In particular, we demonstrate the equivalence of two well-known problems -- the optimal allocation of the fixed overall sample size *n* among *L* strata under stratified random sampling and the optimal allocation of the 435 seats among the 50 states for apportionment of the U. S. House of Representatives following each decennial census. In spite of the strong similarity manifest in the statements of the two problems, they have not been linked and they have well-known but different solutions; one solution is not explicitly exact (Neyman allocation), and the other (equal proportions) is exact. We give explicit exact solutions for both and note that the solutions are equivalent. In fact, we conclude by showing that both problems are special cases of a general problem. The result is significant for stratified random sampling in that it explicitly shows how to minimize sampling error when estimating a total while keeping the final overall sample size fixed at *n*; this is usually not the case in practice with Neyman allocation where the resulting final overall sample size might be near *n + L* after rounding. An example reveals that controlled rounding with Neyman allocation does not always lead to the optimum allocation that minimizes variance. (During the last part of the talk, current general research topics/themes will be mentioned.)

2/18/2013

Speaker: Joseph Pickrell, Harvard Medical School

Title: Statistical models for functional and evolutionary analysis of human genetic variation

Abstract:
Over the past 100,000 years, humans have expanded across the globe and adapted to a wide rage of environments. With the explosion of genomic technologies, we can now hope to answer long-standing questions about the history of the human expansion and the molecular causes of human variation and adaptation. In this talk, I will discuss two questions that are now tractable using large-scale genetic data. First, I will present a statistical model that uses genetic data to learn

about the history of population splits and mixtures within a species. Second, I will discuss methods for identifying genetic variants that influence the cellular processes of transcription and splicing.

Date: *Thursday, March 7, 2013
Time:12:00 - 1:00 PM
Location: Room 903, 1255 Amsterdam Avenue
Tea and Coffee will be served before the seminar at 11:30 AM, Room 1025

Speaker: Yimin Xiao, Michigan State University

"Anisotropic Gaussian Random Fields: Geometry and Excursion Probability"

Many space-time models and solutions to stochastic partial differential equations are anisotropic random fields. In this talk we present some recent results on geometric properties and extreme value theory of anisotropic Gaussian random fields.

3/11/2013

Speaker: Ethan Anderes, UC Davis

Title: Decomposing CMB lensing power with simulation

Abstract:
Over the past year, data from two ground based telescopes, ACT and SPT, have resulted in the first direct measurement of the weak lensing power spectrum solely from cosmic microwave background (CMB) measurements. In the coming years, the data from Planck and upcoming experiments ACTpol and SPTpol will begin probing this lensing at much greater resolution. The state-of-the-art estimator of weak lensing, the quadratic estimator developed by Hu and Okomoto, works in part through a delicate cancelation of terms in a Taylor expansion of the lensing effect on the CMB. In this talk we present a new simulation based approach for exploring the nature of this cancelation and the resulting higher order bias they create. Along with new FFT algorithms our method can easily analyze all polarization estimators along with some recently proposed modifications used to mitigate higher order bias. We illustrate our simulation methodology by uncovering some surprising behavior of the quadratic estimate.

3/25/2013

Speaker: Peter Hoff from University of Washington

Title: Rank likelihood for copula estimation: application, method and theory.

Abstract: Often of primary interest in the analysis of multivariate data are the associations among the variables rather than their univariate marginal distributions. In such cases, a copula model is a natural framework for data analysis. In this talk, we review the "extended rank likelihood", a type of marginal likelihood that only depends on the copula parameters, and not on the univariate

marginal distributions. As such, inference with the extended rank likelihood does not require plug-in estimation of or prior specification for these potentially high-dimensional nuisance parameters.

However, using a marginal likelihood or some other rank-based procedure could potentially lead to inefficient inference because such approaches do not use the complete information in the data. We explore this possibility with a comparison of the asymptotic information bounds for rank-based estimators to those of estimators based on the complete data. Our results suggest the possibility of semiparametric efficient rank-based estimators. We also show that the commonly-used rank based approach of ``plugging-in'' the empirical marginal distributions is not semiparametric efficient.

***Friday 3/29/2013**

Richard Samworth

Title: Variable selection with error control: Another look at Stability Selection

Abstract: Stability selection was recently introduced by Meinshausen and Buhlmann as a very general technique designed to improve the performance of a variable selection algorithm. It is based on aggregating the results of applying a selection procedure to subsamples of the data. We introduce a variant, called complementary pairs stability selection, and derive bounds both on the expected number of variables included by complementary pairs stability selection that have low selection probability under the original procedure, and on the expected number of high selection probability variables that are excluded. These results require no (e.g. exchangeability) assumptions on the underlying model or on the quality of the original selection procedure. Under reasonable shape restrictions, the bounds can be further tightened, yielding improved error control, and therefore increasing the applicability of the methodology.

4/1/2013

Runze Li, The Pennsylvania State University at University Park

Title: Feature Screening for Ultrahigh Dimensional Data

This talk is concerned with screening features in ultrahigh dimensional data analysis, which has become increasingly important in diverse scientific fields. I will first introduce a sure independence screening procedure based on the distance correlation (DC-SIS, for short). The DC-SIS can be implemented as easily as the sure independence screening procedure based on the Pearson correlation (SIS, for short) proposed by Fan and Lv (2008). However, the DC-SIS can significantly improve the SIS. Fan and Lv (2008) established the sure screening property for the SIS based on linear models, but the sure screening property is valid for the DC-SIS under more general settings including linear models. Furthermore, the implementation of the DC-SIS does not require model specification (e.g., linear model or generalized linear model) for responses or predictors. This is a very appealing property in ultrahigh dimensional data analysis. Moreover, the DC-SIS can be used directly to screen grouped predictor variables and for multivariate response variables. We establish the sure screening property for the DC-SIS, and conduct simulations to examine its finite sample performance. Numerical comparison indicates that the DC-SIS

performs much better than the SIS in various models. We also illustrate the DC-SIS through a real data example. If time is permitted, I will introduce some newly developed model free screening procedure for categorical high-dimensional data.

4/8/2013

Speaker: Larry Carin, Duke University

Title: Multichannel Electrophysiological Spike Sorting via Joint Dictionary Learning& Mixture Modeling

Abstract: We propose a construction for joint feature learning and clustering of multichannel extracellular electrophysiological data across multiple recording periods for action potential detection and discrimination ("spike sorting"). Our construction improves over the previous state-of-the art principally in four ways. First, via sharing information across channels, we can better distinguish between single-unit spikes and artifacts. Second, our proposed "focused mixture model" (FMM) elegantly deals with units appearing, disappearing, or reappearing over multiple recording days, an important consideration for any chronic experiment. Third, by jointly learning features and clusters, we improve performance over previous attempts that proceeded via a two-stage learning process. Fourth, by directly modeling spike rate, we improve detection of sparsely spiking neurons. Moreover, our Bayesian construction seamlessly handles missing data. We present state-of-the-art performance without requiring manually tuning of many hyper-parameters on both a public dataset with partial ground truth and a new experimental dataset.

4/15/2013

Speaker: Peter Hall, University of Melbourne

Title: NONPARAMETRIC REGRESSION WITH HOMOGENEOUS GROUP TESTING DATA

Abstract: In this talk we introduce new nonparametric predictors for homogeneous pooled data in the context of group testing for rare abnormalities, and show that they achieve optimal rates of convergence. In particular, when the level of pooling is moderate then, despite the cost savings, the method enjoys the same convergence rate as in the case of no pooling. In the setting of ``over-pooling'' the convergence rate differs from that of an optimal estimator by no more than a logarithmic factor. Our approach improves on the random-pooling nonparametric predictor, which is currently the only nonparametric method available, unless there is no pooling, in which case the two approaches are identical.

4/29/2013

Speaker: Thomas Mikosch (University of Copenhagen, visiting at Columbia University)

Title: Precise large deviation probabilities for random walks with stationary heavy tailed steps

Abstract: We study precise large deviation probabilities in the spirit of A.V. and S.V. Nagaev; see A.V. Nagaev (1969) in Theory Probab. Appl., 14: 51--64 and 193--208, and S.V. Nagaev (1979) in Ann. Probab., 7, 745--789. They studied random walks of iid steps with a regularly varying right tail and

showed that the right tail of the random walk at a given time is equivalent to the tail of the maximum step up to this time. In this talk, analogs are provided for random walks generated from a strictly stationary step sequence. The dependence structure is rather general, but excludes long range dependence. In particular, analogs of Nagaev's theorem can be derived for Markov chains and return models for speculative prices (GARCH, stochastic volatility model). The general framework for these results is regular variation of the finite-dimensional distributions of the step sequence. The proofs use ideas of Adam Jakubowski developed for the central limit theory with infinite variance stable limits. Precise large deviations can be used, for example, to derive precise bounds for ruin probabilities for such random walks. For linear regularly varying processes this approach was chosen in T. Mikosch and G. Samorodnitsky (2000) in Ann. Appl. Probab., 10, 1025--1064, and for solutions to affine stochastic difference equations in D. Buraczewski, E. Damek, T. Mikosch and J. Zienkiewic (2011) (2011). The results of this talk generalize the mentioned papers and also give insight how large deviations occur in a random walk with dependent heavy tail steps.

This is joint work with Olivier Wintenberger (Paris Dauphine).

5/6/2013

Tingting Zhang, Department of Statistics at University of Virginia