Department of Statistics - Statistics Seminar – Fall 2012

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

9/10/2012

Nicolai Meinshausen
Professor of Statistics at Somerville College, Oxford University

Title: Regularization for large-scale regression in the physical sciences

Abstract: Many recent applications in the physical sciences generate large-scale datasets and modern regression techniques are routinely applied to these data for a wide range of purposes. Many of these approaches require a careful choice of a tuning parameter. Often though, simple qualitative and sign-constraints can be imposed on grounds of physical considerations. These constraints simplify the estimation problem as the tuning parameter becomes superfluous. We show the perhaps unexpected effectiveness of this approach for examples in climate science and beyond. Predictive accuracy is not compromised in general and we examine under which assumptions optimal convergence rates can be achieved.

9/17/2012

Prof. Uri Simonsohn, UPenn

Title: P-curve: a key to the file-drawer

Abstract: We propose using “p-curve,” the distribution of significant p-values behind any set of findings, to assess whether they are likely to replicate, and whether they have been “p-hacked” (analyses were chosen just to get statistical significance). We show that p-curve can be meaningfully analyzed when arising from small sets of p-values, say those present in a single paper. We illustrate its use comparing p-curves for a set of psychology experiments that we expected to have been p-hacked, and for a set we expected not to have been. P-curve can be used to aggregate the evidential value of disparate sets of findings, such as those by a given author, journal or institution. Finally, the p-curve of a set of published findings can be analyzed, ignoring all non-published ones, to obtain a 100% publication-bias-free effect-size estimate (though it is biased downwards in the presence of p-hacking).

9/24/2012

Speaker: Antonio Lijoi, University of Pavia

Title:
Gibbs-type priors and Bayesian nonparametric inference on species variety

Abstract:
Sampling from a population whose units belong to different types or species is a common feature in many applied areas. The inferential issues typically addressed in this framework include the evaluation of species richness, the design of sampling experiments and the estimation of rare species variety. Discrete random probabilities, acting as nonparametric priors, perfectly fit this framework and we rely on the flexible class of Gibbs-type priors. Given an observed sample of size n, focus will be on prediction of some key aspects of the outcome from an additional sample of size m. Indeed, we will address issues such as the estimation of: (i) a measure of species richness, (ii) the number of species that will be detected with a certain frequency in the enlarged sample of size n+m; (iii) the discovery probability; (iv) rare species variety. Finally, a concise illustration of asymptotic properties of these models will be sketched.

10/1/2012

Speaker: Jennifer Neville, Purdue University

Title:
How to learn from a single network: Statistical relational learning for social network domains

Abstract:
Machine learning researchers have focused on two distinct learning scenarios for structured graph and network data. In one scenario, the domain consists of a population of structured examples (e.g., chemical compounds) and we can reason about learning algorithms asymptotically, as the number of structured examples increases. In the other scenario, the domain consists of a single, potentially infinite-sized network (e.g., Facebook). In these "single network" domains, an increase in data corresponds to acquiring a larger portion of the underlying network.

Although statistical methods for relational learning have been successfully applied for social network classification tasks, the algorithms were initially developed based on an implicit assumption of an underlying population of networks---which does not hold for most social network datasets. Even when there are a set of network samples available for learning, they correspond to subnetworks drawn from the same underlying network and thus may be dependent. In this talk, I will present our recent efforts to outline a more formal foundation for single network learning and discuss how the analysis has informed the development of more accurate estimation, inference, and evaluation methods.

Bio:
Jennifer Neville is an assistant professor at Purdue University with a joint appointment in the

Departments of Computer Science and Statistics. She received her PhD from the University of Massachusetts Amherst in 2006. In 2008, she was chosen by IEEE as one of "AI's 10 to watch" and in 2012 she was awarded an NSF Career Award. She also received a DARPA IPTO Young Investigator Award in 2003 and was selected as a member of the DARPA Computer Science Study Group in 2007. Her research focuses on developing data mining and machine learning techniques for relational domains, including citation analysis, fraud detection, and social network analysis.

10/8/2012

Speaker: Kung-Sik Chan

Department of Statistics and Actuarial Science, University of Iowa

Title: Reduced rank stochastic regression with a sparse singular value decomposition

Abstract:

For a reduced rank multivariate stochastic regression model of rank r*, the regression coefficient matrix can be expressed as a sum of r* unit rank matrices each of which is proportional to the outer product of the left and right singular vectors. For improving predictive accuracy and facilitating interpretation, it is often desirable that these left and right singular vectors be sparse or enjoy some smoothness property. We propose a regularized reduced rank regression approach for solving this problem. Computation algorithms and regularization parameter selection methods are developed, and the properties of the new method are explored both theoretically and by simulation. In particular, the regularization method proposed can estimate the rank consistently and is shown to be selection consistent and asymptotically normal and to enjoy the oracle property. We illustrate the proposed method with high-dimensional bi-clustering lung-imaging data analysis.

10/15/2012

Prof. Jianqing Fan, Princeton University
(with Ke Zeng and Jiashun Jin)

"Covariance Assisted Screening and Estimation"

Abstract: We propose a new procedure called the "Covariance Assisted Screening and Estimation" (CASE) to the sparse recovery when signals are both rare and weak . CASE first uses a linear filtering to reduce the original setting to a new regression model where the corresponding Gram (covariance) matrix is sparse. The new covariance matrix induces a sparse graph, which guides us to screen variables without visiting all the submodels. By interacting with the signal sparsity, the graph enables us to decompose the original problem into many separable small-size subproblems. Linear filtering also induces a so-called problem of information leakage, which can be overcome by a newly introduced patching technique. Together, these give rise to CASE, which is a two-stage Screen and Clean procedure, where we first identify candidates of these submodels by patching and screening, and then re-examine each candidate to remove false positives.

For any procedure for variable selection, we measure the performance by the minimax Hamming distance between the sign vectors of the estimator and the true parameter.
We show that in a broad class of situations where the Gram matrix is non-sparse but sparsifiable, CASE achieves the optimal rate of convergence. The results are successfully applied to a long-memory time series model and a change-point model.

Bio: Jianqing Fan is the Frederick L. Moore '18 Professor of Finance and a Professor of Statistics at Princeton University. He has received numerous awards including election as an Academician to Academia Sinica, a Guggenheim Fellowship, the Morningside Gold Medal in Applied Mathematics and the COPSS Presidents' Award in 2000.

10/22/2012

Speaker: Cynthia Rudin, Assistant Professor in the MIT Sloan School of Management

Title: Algorithms for Interpretable Machine Learning

Abstract:
It is extremely important in many application domains to have transparency in predictive modeling. Domain experts do not tend to prefer "black box" predictive model models. They would like to understand how predictions are made, and possibly, prefer models that emulate the way a human expert might make a decision, with a few important variables, and a clear convincing reason to make a particular prediction.

I will discuss recent work on interpretable predictive modeling with lists of rules. I will describe several approaches, including:

- an algorithm where not only the predictions, but the whole algorithm itself is interpretable to a human
- an algorithm based on Bayesian analysis
- an algorithm based on mixed-integer linear optimization

Collaborators are: Dimitris Bertsimas, Allison Chang, Ben Letham, Tyler McCormick, David Madigan, and Shawn Qian
Bio: Cynthia Rudin is an assistant professor at the MIT Sloan School of Management in the Operations Research and Statistics group. She works on machine learning and knowledge discovery problems relating to data-driven prioritization. Previously, Dr. Rudin was an associate research scientist at the Center for Computational Learning Systems at Columbia University, and prior to that, an NSF postdoctoral research fellow at NYU. She holds an undergraduate degree from the University at Buffalo, and received a PhD in applied and computational mathematics from Princeton University in 2004. She was given an NSF CAREER award in 2011, and is a finalist for the 2012 INFORMS Innovation in Analytics Award. Her work has been featured in articles appearing in IEEE Computer, Businessweek, ScienceNews, WIRED Science, U.S. News and World Report, Slashdot, Discovery Channel / Discovery News, CIO magazine, and Energy Daily, and very recently, on Boston Public Radio.

10/29/2012

Speaker: Adrian Raftery, University of Washington

Title: Bayesian Reconstruction of Past Populations for Developing and Developed Countries

Bayesian population reconstruction is a recently developed method for estimating past populations by age and sex, with fully probabilistic statements of uncertainty. It simultaneously estimates age-specific population counts, vital rates and net migration from fragmentary data while formally accounting for measurement error. As inputs, it takes initial bias-reduced estimates of age-specific population counts, vital rates and net migration. The output is a joint posterior probability distribution which yields fully probabilistic interval estimates for the inputs. It is designed for the kind of data commonly collected in modern demographic surveys and censuses and can be applied to countries with widely varying levels of data quality. We describe the method and demonstrate it with real data. This is joint work with Mark Wheldon, Patrick Gerland and Samuel Clark.

11/12/2012

Prof. Johan Segers, Universite Catholique de Louvain

"Markov Tail Chains"

Abstract: The extremes of a univariate Markov chain with regulary varying stationary marginal distribution are known to exhibit under general conditions a multiplicative random walk structure called the tail chain. In this paper, we extend this fact to Markov chains with multivariate regularly varying marginal distribution in Euclidean space. We analyze both the forward and the backward tail process and show that they mutually determine each other through a kind of adjoint relation. In a broader setting, it will be seen that even for non-Markovian underlying processes a Markovian forward tail chain always implies that the backward tail chain is Markovian as well. We analyze the resulting class of limiting processes in detail. An application of the theory yields the asymptotic distribution of the past and the future of the solution to a stochastic difference equation conditionally on the present value being large in absolute value.

*Joint work with Anja Janssen (University of Hamburg)

11/19/2012

Speaker: Xiao-Hua Andrew Zhou, Ph.D
Professor, Department of Biostatistics, University of Washington
Director, Biostatistics Unit, Department of Veterans Affairs Seattle Medical Center

Title:Identifiability and Estimation of Causal Effects of a New Treatment on Patient□s Outcomes Truncated by Deaths

Abstract: In comparative effectiveness studies, we are interested in estimating the causal effect of a new treatment relative to a standard one on patient□s outcomes, such as the health related quality of life (HRQOL), after a certain time period of taking the treatment. In these studies, some patients may die before their outcomes can be measured, and hence, their measures are not well defined.

For example, in a randomized trial comparing a new drug with a traditional drug, we are interested in estimating the causal effect of the new drug relative to the old drug on the health related quality of life (HRQOL) after a certain time period of taking the treatment. But some of the patiens in the study may die before their outcomes are measured. One main issue with estimation of such the causal effect is parameter identifiability. We first show that the causal effect of interest is not identifiable non-parametrically under the commonly made regularity conditions in the causal inference literature. We then introduce a concept of using one additional baseline covariate associated with principal strata to make the causal effect identifiable. After we derive the sufficient conditions for identifiability of the causal effect, we then propose a non-parametric method for estimating the causal effect of interest. Our simulation studies show the proposed estimation methods work well in finite-sample sizes. Finally, we apply our approach to a data set from Southwest Oncology Group (SWOG) clinical trial on the effectiveness of the treatment of docetaxel and estramustine (DE) with mitoxantrone and prednisone (MP) in patients with metastatic, androgen-independent prostate cancer. This is a joint with Peng Din, Zhi Geng, and Wei Yan.

11/26/2012

Speaker: Prof. John Patrick Cunningham, Washington University

"Statistical analyses of populations of neurons"

Abstract

The remarkable computational processing of the human nervous system remains a great scientific mystery. In recent years, the field of neuroscience has been dramatically expanding the quantity and complexity of its data acquisition. Interrogating this data, both for basic science and for medical applications, requires new classes of statistical methods that can exploit this changing data paradigm. In this talk I will discuss a few examples of such methods. In particular, I will focus on novel methods that analyze the activity of populations of neurons to study internally-driven dynamical activity. This pursuit will require rethinking one of our most classical and simple methods - principal component analysis. I will describe a linear dimensionality reduction technique that chooses projections based on rotational dynamical activity, which will reveal surprisingly lawful structure in recorded neural data. I will discuss future directions of this work, both for studying other forms of temporal structure and for considering dimensionality reduction more generally. This future work will suggest broader implications for applied statistical and biomedical sciences.

12/3/2012

Speaker: Yufeng Liu, University of North Carolina, Chapel Hill

Title: Statistical Significance of Clustering for High Dimensional Data

Abstract: Clustering methods provide a powerful tool for the exploratory analysis of high dimensional datasets, such as gene expression microarray data. A fundamental statistical issue in clustering is which clusters are "really there," as opposed to being artifacts of the natural sampling variation. In this talk, I will present Statistical Significance of Clustering (SigClust) as a cluster evaluation tool. In particular, we define a cluster as data coming from a single Gaussian distribution and formulate the problem of assessing statistical significance of clustering as a testing procedure. Under this hypothesis testing framework, the cornerstone of our SigClust analysis is accurate estimation of those eigenvalues of the covariance matrix of the null multivariate Gaussian distribution. A likelihood based soft thresholding approach is proposed for the estimation of the covariance matrix eigenvalues. Our theoretical work and simulation studies show that our proposed SigClust procedure works remarkably well. Applications to some cancer microarray data examples demonstrate the usefulness of SigClust

---

12/10/2012

Speaker: Jonathan Goodman (NYU)

Title: Time stepping methods and accuracy measures for SDE

Abstract:
This talk discusses two topics in the numerical solution of stochastic differential equations. The first is a new analysis of time stepping methods for SDE. We present an accuracy measure that we call microscopic total variation, MTV. This is a purely statistical measure, as with weak error, but it also measures accuracy of paths, as with strong error. The analysis allows us to suggest new time stepping methods that are simpler than the Milstein method but have the same accuracy in the MTV sense. The second is a Monte Carlo method for estimating the derivatives of expected hitting times (and functions of hitting times) with respect to perturbations in the hitting set. This method constructs a random variable with the properties that (i) the work to make a sample is on the order of the work to make a single SDE path, (ii) the bias (difference between the expectation value and the true derivative) is on the order of the step size, (iii) the variance is bounded as the time step and the bias go to zero. The method is the obvious method, slightly re-interpreted. This is joint work with Vidal Alcala (second part), and Peter Glynn and Jose Antonio Perez (first part).