Department of Statistics - Statistics Seminar – Fall 2010

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

9/13/2010

Dr. Hadley Wickham, Rice University

"Removing the blindfold: visualising statistical models"

As the volume of data increases, so to does the complexity of our models. Visualisation is a powerful tool for both understanding how models work, and what they say about a particularly dataset. There are very many well-known techniques for visualising data, but far fewer for visualising models. In this talk I'll discuss three broad strategies for model visualisation: display the model in the data space; look all members of a collection; and explore the process of model fitting, not just the end result. I will demonstrate these techniques with two examples: neural networks, and ensembles of linear models.

9/20/2010

Dr. Jon A. Wellner, University of Washington

Title: Nonparametric estimation of log-concave densities

Abstract: I will discuss nonparametric estimation of log-concave densities in R^1 and R^d. In the case of R^1, I will present limit theory for the estimators at fixed points at which the population density has a non-zero second derivative and for the resulting natural mode estimator under a corresponding hypothesis. In the case of R^d with $d\ge 2$ will briefly discuss some recent progress and sketch a variety of open problems.

9/27/2010

Dr. Henrik Hult, Royal Institute of Technology, Sweden

Title: EFFICIENT CALCULATION OF RISK MEASURES BY IMPORTANCE SAMPLING - THE HEAVY TAILED CASE

Abstract: Computation of extreme quantiles and tail-based risk measures using standard Monte Carlo simulation can be inefficient. A method to speed up computations is provided by importance sampling. We show that importance sampling algorithms, designed for efficient tail probability estimation, can significantly improve Monte Carlo estimators of tail-based risk measures. In the heavy-tailed setting, when the random variable of interest has a regularly varying distribution, we provide sufficient conditions for the asymptotic relative error of importance sampling estimators of risk measures, such as Value-at-Risk and expected shortfall, to be small. The results are illustrated by some numerical examples.

10/4/2010

Dr. Hui Zou, University of Minnesota

Title: Non-concave Penalized Composite Likelihood Estimation of Sparse Ising Models

Abstract: The Ising model is a useful tool for studying complex interactions within a system. The estimation of such a model, however, is rather challenging especially in the presence of high dimensional parameters. In this work, we propose efficient procedures for learning a sparse Ising model based on a penalized composite likelihood with non-concave penalties. Non-concave penalized likelihood estimation has received a lot of attention in recent years. However, such an approach is computationally prohibitive under high dimensional Ising models. To overcome such difficulties, we extend the methodology and theory of non-concave penalized likelihood to penalized composite likelihood estimation. An efficient solution path algorithm is devised by using a new coordinate-minorization-ascent algorithm. Asymptotic oracle properties of the proposed estimator are established with NP-dimensionality. We demonstrate its finite sample performance via simulation studies and further illustrate our proposal by studying the Human Immunodeficiency Virus type 1 (HIV-1) protease structure based on data from the Stanford HIV Drug Resistance Database. This talk is based on a joint paper with Lingzhou Xue and Tianxi Cai.

10/11/2010

Dr. Eric Xing, Carnegie Mellon University

Title: Dynamic Network Analysis: Model, Algorithm, Theory, and Application

Abstract -- Across the sciences, a fundamental setting for representing and interpreting information about entities, the structure and organization of communities, and changes in these over time, is a stochastic network that is topologically rewiring and semantically evolving over time, or over a

genealogy. While there is a rich literature in modeling invariant networks, until recently, little has been done toward modeling the dynamic processes underlying rewiring networks, and on recovering such networks when they are not observable.

In this talk, I will present some recent developments in analyzing what we refer to as the dynamic tomography of evolving networks. I will first present a new class of statistical models known as dynamic exponential random graph models for evolving social networks, which offers both good statistical property and rich expressivity; then, I will present new sparse-coding algorithms for estimating the topological structures of latent evolving networks underlying nonstationary time-series or tree-series of nodal attributes, along with theoretical results on the asymptotic sparsistency of the proposed methods; finally, I will present a new Bayesian model for estimating and visualizing the trajectories of latent multi-functionality of nodal states in the evolving networks.

I will show some promising empirical results on recovering and analyzing the latent evolving social networks in the US Senate and the Enron corporation at a time resolution only limited by sample frequency. In all cases, our methods reveal interesting dynamic patterns in the networks.

Short bio of the speaker:

Dr. Eric Xing is an associate professor in the School of Computer Science at Carnegie Mellon University. His principal research interests lie in the development of machine learning and statistical methodology; especially for solving problems involving automated learning, reasoning, and decision-making in high-dimensional and dynamic possible worlds; and for building quantitative models and predictive understandings of biological systems. Professor Xing received a Ph.D. in Molecular Biology from Rutgers University, and another Ph.D. in Computer Science from UC Berkeley. His current work involves, 1) foundations of statistical learning, including theory and algorithms for estimating time/space varying-coefficient models, sparse structured input/output models, and nonparametric Bayesian models; 2) computational and statistical analysis of gene regulation, genetic variation, and disease associations; and 3) application of statistical learning in social networks, data mining, vision. Professor Xing has published over 100 peer-reviewed papers; he is an action editor of the Machine Learning Journal, an associate editor of the Annals of Applied Statistics, and the PLoS Journal of Computational Biology. He is a recipient of the NSF Career Award, the Alfred P. Sloan Research Fellowship in Computer Science, and the United States Air Force Young Investigator Award.

Special Statistics Seminar
Wednesday, October 20th
1:00-2:00 p.m.
Room 963 Schermerhorn

Speaker: Cosma Rohilla Shalizi, CMU

Title: "What happens when a Bayesian can't handle the truth?"

Abstract:
Much is now known about the consistency of Bayesian updating on infinite-dimensional parameter spaces with independent or Markovian data. Necessary conditions for consistency include the prior putting enough weight on the correct neighborhoods of the data-generating

distribution; various sufficient conditions further restrict the prior in ways analogous to capacity control in frequentist nonparametrics. The asymptotics of Bayesian updating with mis-specified models or priors, or non-Markovian data, are far less well explored. Here I establish sufficient conditions for posterior convergence when all hypotheses are wrong, and the data have complex dependencies. The main dynamical assumption is the asymptotic equipartition (Shannon-McMillan-Breiman) property of information theory. This, along with Egorov's Theorem on uniform convergence, lets me build a sieve-like structure for the prior. The main statistical assumption, also a form of capacity control, concerns the compatibility of the prior and the data-generating process, controlling the fluctuations in the log-likelihood when averaged over the sieve-like sets. In addition to posterior convergence, I derive a kind of large deviations principle for the posterior measure, extending in some cases to rates of convergence, and discuss the advantages of predicting using a combination of models known to be wrong. An appendix sketches connections between these results and the replicator dynamics of evolutionary theory.
http://arxiv.org/abs/0901.1342

10/18/2010

Dr. David Dunson, Duke University

"High-dimensional categorical data analysis via simplex factor models"

Gaussian latent factor models are routinely used for modeling of dependence in continuous, binary and ordered categorical data. For unordered categorical variables, Euclidean latent factor models lead to challenging computation and overly complex modeling structures. With this motivation, we propose a novel class of simplex factor models. In the single factor case, the model treats the different categorical outcomes as independentwith unknown marginals. The model can characterize highly flexible dependence structures parsimoniously with few factors, and as factors are added, any multivariate categorical data distribution can be accuratelyapproximated. Using a Bayesian approach for computation and inferences, a highly efficient MCMC algorithm is proposed that scales well with increasing dimension, with an adaptive Gibbs step enabling selection of the number of factors. Relationships with mixed membership models and tensor decompositions are described, and we evaluate the approach through simulation examples and applications to modeling dependence and classification from nucleotide sequences. The framework is natural forsparsely characterizing higher order interactions in categorical predictors, and can be directly applied to nonparametric modeling of data having mixed measurement scales including not only categorical, continuousand count variables but also images, text and curves.

Joint work with Anirban Bhattacharya

10/25/2010

Dr. Yali Amit, University of Chicago

"Generative Models for Scene Annotation"

The goal of Computer Vision is the automatic annotation of scenes containing multiple occluded objects as well as noise and clutter. Recent work has focusedon two main tasks. The first is the classification among object classes in segmented images containing onlyone object and the second is the detection of a particularobject class in a large image. Both tasks have been primarily addressed using discriminative learning.It is not clear however how these methods can extendto deal with the recognition of multiple object classes in images containing a number of objects in a wide range of configurations. I will present an approach which starts from simple statistical models for individual objects. With these models the important notion of invariance can be clearly formulated. Furthermore the individual object models can be composed to define models for object configurations. Decisions are likelihood based and do not depend on pretrained decision boundaries. I will briefly discuss some computational strategies forcomputing the scene annotation, show some applications,and describe some major difficulties we face in making further progress.

Speaker: Surya Ganguli
http://keck.ucsf.edu/~surya/

Title: Dynamical compressed sensing and short-term sequence memory in neuronal networks.

903 SSW
Tuesday, November 2, 2010
12:00

Compressed sensing (CS) is an important recent advance that shows how to reconstruct sparse high dimensional signals from surprisingly small numbers of random measurements. However, the nonlinear nature of the reconstruction process poses a challenge to understanding the performance of CS. We employ techniques from statistical physics to compute the typical behavior of CS as a function of the signal sparsity and measurement density. Our computations reveal surprising and useful regularities in the nature of errors made by CS. We then extend the framework of CS to a dynamical setting by showing that certain neuronal networks can essentially perform compressed sensing of sparse temporal sequences in an on-line fashion. The lifetime of memory traces in such networks can actually exceed the number of neurons in the network times the single neuron time constant. This result circumvents prior no-go theorems which place more stringent limits on the memory capacity of neuronal networks for gaussian input sequences. More generally, we analytically compute the decay of memory traces in neuronal networks as a function of network size, connectivity, and input statistics, and our calculations reveal an intimate relationship between the statistics of temporal input sequences, and the optimal network connectivity required to remember them.

11/8/2010

Dr. Susan Murphy, University of Michigan

Title: Inference for Dynamic Treatment Regimes

Abstract:
Dynamic treatment regimes (or treatment policies) are used to operationalize multi-stage decision making in the medical field. Common approaches to constructing dynamic treatment regimes from data, such as Q-Learning, employ non-smooth functionals of the data. The non-smoothness leads to non-regular asymptotics under common data generating models. As a result methods that ignore the non-regularity have poor performance in small samples. In this talk, we present a bootstrap based method for constructing asymptotically valid confidence sets. This method is adaptive in the sense that it provides exact coverage when the true underlying generative model leads to regular asymptotics and is conservative otherwise. We discuss how a variety of modern statistical procedures exhibit this nonregularity and provide interesting statistical challenges in developing measures of confidence.

(this is a joint seminar with Columbia Biostatistics & NYSPI)

11/22/2010

Terry Speed, University of California, Berkeley

"Removing unwanted variation from microarray data"
(joint work with Johann Gagnon-Bartsch)

Abstract
Principal components have been used a lot with microarray data to exhibit unwanted variation, and to some extent to remove such variation.Such efforts are also called normalization. In this talk I will review some of these methods, and related work, and propose a simple but apparently novel variant which works a lot of the time,though not always.
Part of my story will be a discussion of how one tells whether one is helping or hurting the analysis by adjusting.

11/29/2010

Dr. David L. Weakliem, University of Connecticut

"Measuring Political Ideology in the American States"

In recent years, there has been much discussion of political differences among the states. This research reports on an effort to obtain state-level estimates of political ideology. We selected twenty-one opinion questions that had been asked with reasonable frequency on national surveys from the 1970s to the early 21^st century. By combining a large number of surveys, it is possible to obtain reasonably large samples for even states with small populations. If the questions are treated as indicators of a latent variable, we can obtain estimates of average ideology at the state level.

After examining the general state estimates, we consider differences by educational level. State differences in ideology show somewhat different patterns among college graduates and non-graduates, and are generally larger among college graduates. Finally, we find a substantial association between the measure of state ideology and ratings of state policy. However, the ideological position of college graduates in a state has a stronger association with state policy than the ideological position of all respondents, suggesting that the opinions of less educated people have little influence on state policy.

(Joint work with Casey Borch, Department of Sociology, University of Alabama--Birmingham)

12/6/2010

Harrison Zhou, Yale University

Title: Optimal Estimation of Large Covariance Matrices

Covariance matrices play a central role in statistical analysis. A large collection of fundamental statistical methods require the estimation of covariance matrices. With the emergence of high dimensional data from modern technologies, estimating large scale covariance matrices as well as their inverse is becoming a crucial problem in many fields. In this talk we give some theories to unveil the precision to which (inverse) covariance matrices can be estimated and to develop general methodologies for optimal estimation of the (inverse) covariance matrices under various settings.

12/13/2010

Monday, December 13, 2010

Professor Andrew Gelman, Department of Statistics and Department of Political Science, Columbia University

"Creating structured and flexible models: some open problems"

A challenge in statistics is to construct models that are structured enough to be able to learn from data but not be so strong as to overwhelm the data. We introduce the concept of "weakly informative priors" which contain important information but less than may be available for the given problem at hand. We also discuss some related problems in developing general models for interactions. We consider how these ideas apply to in several social science applications.