Department of Statistics - Statistics Seminar – Spring 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

2/15/2010

Dr. Mark Handcock, University of California, Los Angeles

"Modeling networks when data is missing or sampled"

Network models are widely used to represent relational information among interacting units and the structural implications of these relations. Recently, social network studies have focused a great deal of attention on random graph models of networks whose nodes represent individual social actors and whose edges represent a specified relationship between the actors.

Most inference for social network models assumes that the presence or absence of all possible links is observed, that the information is completely reliable, and that there are no measurement (e.g. recording) errors. This is clearly not true in practice, as much network data is collected though sample surveys. In addition even if a census of a population is attempted, individuals and links between individuals are missed (i.e., do not appear in the recorded data).

In this paper we develop the conceptual and computational theory for inference based on sampled network information. We first review forms of network sampling designs used in practice. We consider inference from the likelihood framework, and develop a typology of network data that reflects their treatment within this frame. We then develop inference for social network models based on information from adaptive network designs.

We motivate and illustrate these ideas by analyzing the effect of link-tracing sampling designs on a collaboration network.

This is joint work with Krista J. Gile, Nuffield College, Oxford.

3/22/2010

Dr. Michael Sobel, Columbia University

"Does Marriage Boost Men's Wages? Identification of Treatment Regression Models for Longitudinal Data Effects from Fixed and Random Effects "

Abstract

Social scientists have generated a large and inconclusive literature on the effect(s) of marriage on men's wages. Researchers have hypothesized that the wage premium enjoyed by married men may reflect both a tendency for more productive men to marry and an effect of marriage on productivity. To sort out these explanations, researchers used effects regression models for panel data to adjust for unobserved time invariant confounders, interpreting coefficients or the time varying marriage variables as effects. But they inadvertently omit important time varying confounders. Further, they do not define the effects they purport to estimate, and as there are various effects of possible interest, this leads to misinterpretation and to potentially inappropriate policy recommendations. The same problems also arise in many other literatures where x and random effects regressions are used to adjust for unobserved variables. To clarify these issues, I build on recent statistical work on causal inference with longitudinal data. A basis set of treatment effects is defined and also used to define derived effects.

Causal fixed and random effects regression models are defined and the treatment effects are reexpressed in terms of these models. Ignorability conditions under which the parameters of the regression models and causal models are identical are given. Even when these hold, a number of interesting and important treatment effects are typically not identified.

3/29/2010

Dr. Marten Wegkamp, Florida State University

"Classification with a reject option"

We study a modification of the binary classification problem that allows for withholding a decision. This third option is called the reject option, and is introduced to reduce the misclassification rate and to avoid ambiguous examples. We incorporate this reject option in both plug-in classification rules and classifiers that are based on empirical risk minimization.

We show that the performance of the new plug-in classification rules depends on (a) the accuracy of the conditional class probability estimates and (b) the behavior of the conditional class probabilities themselves near values that depend on the cost of withholding a decision.

Next, we consider empirical risk minimization classifiers based on a convex surrogate loss and a threshold level that automatically reject ambiguous examples. We derive a necessary and sufficient condition on the loss function to be classification calibrated. A loss function is classification calibrated if the minimizer of the risk corresponding to this loss function yields the optimal rule. We show that the rates of convergence of the empirical risk minimizers crucially depend on the behavior of the conditional class probabilities near values that depend on the cost of withholding a decision.

Finally, we discuss in detail the support vector machines that use a generalized hinge loss. We show that the usual hinge loss cannot incorporate the reject option. We propose a generalization of the hinge loss and derive oracle inequalities for the resulting empirical minimization rule that is penalized by a lasso type penalty to guarantee sparse solutions.

We illustrate our theoretical findings with simulations.

4/5/2010

Dr. David Seigmund, Stanford University

"BIC Applied to Model Selection of a Large Number of Change-points"

In a previous paper (Biometrics, 2006, pp. 22-32) we derived a Bayes Information Criterion (BIC) for determining the number of change-points in a sequence of independent observations when the number m of change-points is assumed to remain bounded as the number of observations increases. Here we generalize that result to include multiple alignedsequences with intervals of simultaneous change that occur in a fraction of the sequences and a total number of of change-points m that can increase with the sample size; and we include in the criterion terms that increase at rate m. Stochastic terms that enter into the new criterion involve integrals and maxima of two-sided Brownian motion with negative drift. Examples involve segmenting aligned DNA sequences according to copy number variations that occur at the same position in a fraction of the sequences.

This is joint research with N. Zhang.

4/12/2010

Speaker: Stuart Geman, Brown University

Title: Orchestrating Computation in Hierarchical Models

Abstract:
The vision problem is hard. In fact, under any reasonable formulation it is provably NP-hard. So are many other problems that nature handles routinely, yet remain out of reach of artificial systems, e.g. protein folding and the planning and coordination of muscle action. How does nature compute? One thing these problems have in common is their combinatorial structure. Focusing on vision, and image analysis in particular, I will propose a probabilistic combinatorial formulation of the image analysis problem, and examine optimal recognition performance from the Neymann-Pearson point of view. Through an asymptotic analysis, and basic large-deviation theory, I will argue that essentially optimal performance can be attained through a computationally feasible sequential decision analysis. I will illustrate with some recognition experiments, and discuss possible connections to some problems in biology.

4/26/2010

Speaker: Dr. Michael L. Littman from Rutgers University

Bayesian Models in Reinforcement Learning

Reinforcement learning (RL) is a subfield of Artificial Intelligence concerned with decision makers that adapt their behavior given utility-based feedback. While temporal difference methods dominate the field, there are a growing number of researchers examining Bayesian statistics as a way of learning about unfamiliar environments while modeling their own uncertainty and addressing the classic exploration/exploitation dilemma. I will introduce the RL problem and provide some background into existing Bayesian methods and the extensions underway in our lab at Rutgers.

5/3/2010

Dr. Nan Laird, Harvard University

"Testing for Gene and Environmental Interactions with Family-Based Designs"

It is valuable to have robust gene-environment interaction tests that can utilize a variety of family structures in an efficient way. Our objective is to develop powerful tests that can combine trio data with parental genotypes and discordant sibships when parents' genotypes are missing. With trios alone, methods based on the Conditional on Parental Genotype (CPG) likelihood are very efficient; with discordant sibships, methods based on conditional logistic regression are more powerful than the CPG approach. We propose a hybrid approach to utilize different family structures in the same analysis in an efficient way. An example involving interactions between SNPS in theSerpine2 gene and smoking on the risk of COPD is given.
