2013 Seminars

11/06/2013

The Markov Chinese Restaurant Process: A Non-parametric Bayesian Cluster Memory Model for Longitudinal Data

Robert Weiss, PhD

Professor, Department of Biostatistics, University of California, Los Angeles

Abstract: We develop a Dirichlet process mixture (DPM) model extension for regularly spaced longitudinal data. In longitudinal data, observations are both subject specific and a function of time. We account for both dependence between sampling densities across time and dependence in observations across time within the same subject. In the cluster memory Dirichlet process mixture (cmDPM) model, we use the inherent clustering properties of the DPM model to carry information from one time point to the next. Observations at baseline are modeled with a DPM. Cluster assignments at future time points depend on the previous assignment. Subjects may retain their cluster membership from the previous time point with nonzero probability. After baseline, given the previous time point, subjects are no longer exchangeable and their observed values depend on their previous clustering history. Clusters that are retained over time evolve through a time dependent process. There are several ways to look at the process including as a dynamic Markov Chinese Restaurant Process. We apply the cmDPM model to model annual tuberculosis (TB) incidence rates across 197 countries in the world from 1990-2010 and examine how the annual distribution of TB incidence rates has changed over time.

This is joint work with Yuda Zhu of Genentech.

10/08/2013

Varying index coefficient models for nonlinear interactions

Shujie Ma, PhD

Professor, University of California, Riverside

It has been a long history of utilizing interactions in regression analysis to investigate interactive effects of covariates on response variables. In this paper we aim to address two kinds of new challenges resulted from the inclusion of such high-order effects in the regression model for complex data. The first kind arises from a situation where interaction effects of individual covariates are weak but those of combined covariates are strong, and the other kind pertains to the presence of nonlinear interactive effects. Generalizing the single index coefficient regression model, we propose a new class of semiparametric models with varying index coefficients, which enables us to model and assess nonlinear interaction effects between grouped covariates on the response variable. As a result, most of the existing semiparametric regression models are special cases of our proposed models. We develop a numerically stable and computationally fast estimation procedure utilizing both profile least squares method and local fitting. We establish both estimation consistency and asymptotic normality for the proposed estimators of index coefficients as well as the oracle property for the nonparametric function estimator. In addition, a generalized likelihood ratio test is provided to test for the existence of interaction effects or the existence of nonlinear interaction effects. Our models and estimation methods are illustrated by both simulation studies and an analysis of body fat dataset.

10/04/2013

Permutation Tests 101

Joe Romano, PhD

Professor, Stanford University

09/09/2013

Real-Time Prediction in Clinical Trials: A Statistical History of REMATCH

Daniel F. Heitjan, PhD

Professor, Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

Randomized clinical trials often include one or more planned interim analyses, during which an external monitoring committee reviews the accumulated data and determines whether it is scientifically and ethically appropriate for the study to continue. With survival-time endpoints, it is often desirable to schedule the interim analyses at the times of occurrence of specified landmark events, such as the 50th event, the 100th event, and so on. Because the timing of such events is random, and the interim analyses impose considerable logistical burdens, it is worthwhile to predict the event times as accurately as possible. Prediction methods available prior to 2001 used data only from previous trials, which are often of questionable relevance to the trial for which one wishes to make predictions. With modern data management systems it is often feasible to use data from the trial itself to make these predictions, rendering them far more reliable. This talk will describe work that some colleagues and students and I have done in this area. I will set the methodologic development in the context of the trial that motivated our work: REMATCH, a randomized clinical trial of a heart assist device that ran from 1998 to 2001 and was considered one of the most rigorous and expensive device trials ever conducted.

06/19/2013

Statistical and Geographical Methods Exploring HIV-Risk along the Mexico-U.S. Border

Tommi Gaines, DrPH

Division of Global Public Health, Department of Medicine, UCSD

The Mexico-U.S. border region is home to an evolving HIV epidemic among vulnerable groups such as injection drug users and female sex workers. Features of one’s environment have been associated with individual health and therefore our objective is to highlight statistical and geographical techniques that examine HIV and risk-related behaviors. We describe the use of geographic information systems (GIS) data to map the location of sex work venues from epidemiologic studies conducted in Tijuana, Mexico and the application of statistical models to empirically assess the role of geography in shaping HIV and other sexually transmitted infections. We discuss the importance of combining statistical methods with GIS data to inform prevention and support services.

06/05/2013

On Hypothesis Testing and Interval Estimation for Monotone Dose-Response Means with a Control Mean

Lin Liu, PhD

Division of Biostatistics and Bioinformatics, Department of Family Medicine and Public Health, UCSD

In dose-response studies, one of the most important issues is the identification of minimum effective dose (MED), where the MED is defined as the lowest dose such that the mean response is better than the mean response of a zero-dose control by a clinically significant difference. Dose-response curves are sometimes

monotonic in nature. A union-intersection type of likelihood ratio test is proposed. One-sided lower confidence bounds can be inverted from the test to detect the differences between the dose-response means and a control mean. The evaluation of the lower confidence bounds is a concave programming problem subject to homogeneous linear inequality constraints. An efficient computing algorithm is proposed. A real data example from a dose-response study is used to illustrate the method.

05/08/2013

Statistical and Bioinformatics Challenges in Systems Biology Research for Influenza Infection

Jaroslaw Harezlak, PhD

Assistant Professor, Department of Biostatistics, Fairbanks School of Public Health and School of Medicine, Indiana University, Indianapolis, IN

Collection of functional data has vastly grown in the past decade, including functional data collected longitudinally. For example, in the HIV Neuroimaging Consortium (HIVNC) study, metabolite spectra were obtained using magnetic resonance spectroscopy (MRS) from multiple brain regions at a number of study time points. Analysis of such data usually follows a two-step procedure: (1) metabolite concentration extraction and (2) association study of extracted features and outcome of interest.

Our approach does not rely on this frequently unreliable feature extraction. Instead, it incorporates prior scientific knowledge to estimate regression function associating the whole functional profile with the outcome without explicitly extracting the feature characteristics. Specifically, we propose a method for functional linear model estimation using partially empirical eigenvectors for regression (PEER) in the longitudinal data setting. Our method allows the regression function to vary across both time and space. We derive the estimator's statistical properties and discuss their connections to the generalized singular value decomposition (GSVD). The results of the simulation studies and an application to the analysis of HIV patients' neurocognitive impairment as a function of the metabolite profiles are presented.

Joint work with Madan G. Kundu and Timothy W. Randolph

05/01/2013

Statistical and Bioinformatics Challenges in Systems Biology Research for Influenza Infection

Hulin Wu, PhD

Dean’s Professor, Department of Biostatistics and Computational Biology, Director, Center for Integrative Bioinformatics and Experimental Mathematics, University of Rochester School of Medicine and Dentistry

Many systems in engineering and physics can be represented by differential equations, which can be derived from well-established physics laws and theories. However, currently no laws or theories exist to deduce exact quantitative relationships and interactions among the huge amount of elements at different levels in a biological system. It is unclear whether the biological systems follow a mathematical representation such as differential equations, similar to that for a man-made physics or engineering system. Fortunately, recent advances in cutting-edge biomedical technologies allow us to generate intensive high-throughput data to gain insights into biological systems. It is badly needed to develop statistical methods and bioinformatics approaches to test whether a biological system follows a mathematical representation based on experimental data so that quantitative predictions can be made for biomedical interventions in a biological system. In this talk, I will present and discuss how to construct data-driven differential equations (ODE) to describe biological systems, in particular for dynamic gene regulatory network systems. We propose to combine the high-dimensional variable selection approaches and ODE model estimation methods to construct the high-dimensional ODE

models based on experimental data. We apply the proposed approaches to study how our immune system responds to influenza infections and vaccination based on the time course high-throughput experimental data.

04/15/2013

Predicting Health Care Costs of Individual Patients

Andrew Zhou, PhD

Professor, Department of Biostatistics, University of Washington Director, Research Career Scientist, Biostatistics Unit, VA Puget Sound Health Care System

The rising cost of health care is one of the most important problems facing the United States. Accurately predicting such costs is an important first step in addressing this problem. However, due to some special distributional features of health care costs, including high skewness, presence of excessive zero values, and heteroscedasticity, it is difficult to obtain an accurate prediction of future health care costs of patients.

In this talk, I will describe some new models for using covariates to predict the future health care costs of patients. These new models include: (1) a parametric heteroscedastic transformation model, (2) a semi-parametric two-part heteroscedastic transformation model, (3) a quantile regression model, (4) a non-parametric heteroscedastic transformation regression model, and (4) a semi-parametric two-part mixed-effects heteroscedastic transformation model.

03/08/2013

Topics in Biostatistics: Trial Design (n=20, p=2), Prognostic Modeling (n=3,000, p=20), and Genomic Data Analysis (n=2, p=3,000)

Karen Messer, PhD

Professor, Family Medicine and Public Health, Director, UC San Diego Moores Cancer Center Biostatistics/Bioinformatics shared resource

As a biostatistician, one aims to support high-quality inference from experimental or observational data across a wide variety of scientific settings. To this sometimes bewildering array, the discipline of statistics brings a unifying set of tools and objectives which can help sort out what one knows with high confidence, with low confidence, and most especially, not at all. Although the approaches to sound inference may differ with the number of subjects (n- big or small) and the number of variables (p- small or big), the principles of control of Type I error, modeling sources of bias and variation, and quantifying the limits of statistical power provide a helpful framework for a variety of problems. In this talk, I will give examples of approaches to statistical inference from three areas of my work in cancer biostatistics: early phase trial design (small n, small p), prognostic modeling for survival (big n, medium p), and analysis of next generation sequencing data (small n, big p). In the first two topics, some recent approaches to older problems will be presented and in the third, traditional tools will be applied to modern data.

03/06/2013

The breakage fusion bridge, Chromothripsis and other exotic structural variations: combinatorics and cancer genomics

Vineet Bafna, PhD

Professor in the Department of Computer Science at UCSD and in the Bioinformatics PhD program. His research area is Bioinformatics, with a focus on Genomics and Proteomics.

Cancer genomes are marked by genomic instability and massive rearrangements. Recently, many exotic mechanisms have been proposed as mechanistic explanations for these rearrangements. For example, the breakage-fusion-bridge (BFB) mechanism, proposed over seven decades ago, has seen renewed interest as a source of genomic variability and gene amplification in cancer. Here, we formally model and analyze the BFB mechanism, the first rigorous formulation of the mechanism. Using this model, we show that BFB can achieve a surprisingly broad range of amplification patterns, and describe efficient combinatorial algorithms to characterize patterns consistent with BFB. An extensive analysis of simulated, cell-line, and primary tumor data reveals the existence of BFB. Our results also suggest that BFB may be hard to detect under heterogeneity and polyploidy.

As a second example, the model of chromothripsis--extensive shattering followed by regrouping of small parts of a chromosome-- has been proposed to explain the extensive rearrangements seen in some tumors. Time remaining, we will critique this model using 3 different lines of evidence.

(joint work with Shay Zakov, and Marcus Kinsella).

02/06/2013

Designing and monitoring clinical trials with survival endpoints: statistical issues, proposals, and opportunities

Daniel Gillen, PhD

Associate Professor, Department of Statistics, University of California, Irvine

Researchers frequently elect to evaluate new therapies on the basis of patient survival. For example, clinicians might consider five-year survival when investigating drugs developed for use in childhood cancer, or 28-day survival when investigating the treatment of sepsis in patients suffering traumatic injury. Both of these examples focus on patient responses over a fixed period of time. However, for ethical reasons it is common for data to be periodically analyzed for early indications of efficacy, futility, or harm. In the case of censored survival data, inference is typically based upon a semiparametric model assuming a time-invariant treatment effect and standard group sequential methodology is used to generate multiple criteria for guiding the decision of whether a trial should be stopped early given the observed data. However, it is often the case that a given treatment might have a delayed effect within individuals or that the effect of treatment might dissipate over time. Special issues arise in such settings, mostly due to the dependence of results on the censoring distribution observed in the trial. In this talk, we discuss general issues associated with the sequential testing of a survival endpoint. Specific attention is given to the uncertainty of future observations under a potentially time-varying treatment effect. In this case we propose a method of imputation of future treatment effects based on random walks, which assumes minimally informative Bayesian prior distributions on the smoothness of survival of each comparison group. Imputation of future survival differences is carried out using standard Bayesian predictive distributions, thereby allowing for quantification of uncertainty in future treatment differences.

01/13/2013

Quantitative challenges in advancing the HIV prevention research agenda

Victor DeGruttola, Sc.D.

Professor and Chair, Department of Biostatistics, Harvard School of Public Health

The UC San Diego Center for AIDS Research and AIDS Research Institute are pleased to present Victor DeGruttola, Sc.D.. Dr. DeGruttola will discuss the quantitative challenges in advancing the HIV prevention research agenda.