Department of Statistics - Statistics Seminar – Fall 2011

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

9/19/2011

Jane Wang, UC Davis

"Adjusting Covariate Effects in Functional Principal Component Analysis"

Classical principal component analysis (PCA) has been extended to functional data and termed functional PCA. In this talk, we explore approaches for functional PCA when covariate information is available. We start with the case of a single covariate and present two approaches. Next, we consider the multiple covariate case and employ a dimension reduction approach to collapse the covariate information to a few indices. The extension differs from current single index models in two ways: (i) it accommodates longitudinal data, both as responses and as covariates; and (ii) the time-dynamic effects of the single-index are reflected in the model. The proposed estimator for the index parameters is shown to be root-n consistent and asymptotically normally distributed. One advantage of the new approach is that the same bandwidth is used to estimate both the nonparametric mean function and the parameters in the single-index. The finite sample performance of the proposed procedure is studied through simulations and AIDS CD4 cell count data.

*The talk is based on joint work with Ciren Jiang, SAMSI

9/26/2011

Song Xi Chen, Iowa State University

The transition density of a diffusion process does not admit an explicit expression in general, which prevents the full maximum likelihood estimation (MLE) based on discretely observed sample paths. A\"it-Sahalia (1999, 2002) proposed asymptotic expansions to the {{transition}} densities of diffusion processes, which lead to an approximate maximum likelihood estimation (AMLE) for parameters.

Built on A\"it-Sahalia (2002, 2008)'s proposal and analysis on the AMLE, we establish the consistency and convergence rate of the AMLE, which reveal the roles played by the number of terms used in the asymptotic density expansions and the sampling interval between
successive observations. We find conditions under which the AMLE has the same asymptotic distribution as that of the full MLE. A first order approximation to the Fisher information matrix is proposed.

*The talk is based on joint work with Jinyuan Chang, Peking University

---

10/3/11

Yufeng Liu (UNC)

University of North Carolina at Chapel Hill

The Support Vector Machine (SVM) has been a popular margin-based technique for classification problems in both machine learning and statistics. It has weak distributional assumptions and great flexibility in dealing with high dimensional data. In this talk, I will present various aspects of the SVM as well as some of its recent developments. Issues including statistical properties of the SVM,multicategory SVM, as well as class probability estimation of the SVM will be discussed.

Special Statistics Seminar

Date: Tuesday, October 4
Time: 12:00 - 1:00
Location: Room 903

Wolfgang Polasek

Title: Does Globalization affect Regional Growth? Evidence
for NUTS-2 Regions in EU-27

Abstract: We analyze the influence of newly constructed globalization measures on regional growth for the EU-27 countries between 2001 and 2006. The spatial Chow-Lin procedure, a method constructed by the authors, was used to construct on a NUTS-2 level a complete regional data for exports, imports and FDI inward stocks, which serve as indicators for the influence of globalization, integration and technology transfers on European regions. The results suggest that most regions have significantly benefited from globalization measured by increasing trade openness and FDI. In a non-linear growth convergence model the growth elasticities for globalization and technology transfers decrease with increasing GDP per capita. Furthermore, the estimated elasticity for FDI decreases when the model includes a higher human capital premium for CEE countries and a small significant growth enhancing effect accrues from the structural funds expenditures in the EU.

10/10/2011

Professor Andrew Gelman, Columbia University

Title: Parameterization and Bayesian Modeling

Abstract: Progress in statistical computation often leads to advances in statistical modeling. For example, it is surprisingly common that an existing model is reparameterized, solely for computational purposes, but then this new configuration motivates a new family of models that is useful in applied statistics. One reason why this phenomenon may not have been noticed in statistics is that reparameterizations do not change the likelihood. In a Bayesian framework, however, a transformation of parameters typically suggests a new family of prior distributions. We discuss examples in censored and truncated data, mixture modeling, multivariate imputation, stochastic processes, and multilevel models.

10/17/2011

Tyler Vander, Harvard University

Title: Sensitivity analysis for contagion effects in social networks

Abstract: Analyses of social network data have suggested that obesity, smoking, happiness and loneliness all travel through social networks. Individuals exert "contagion effects" on one another through social ties and association. These analyses have come under critique because of the possibility that homophily from unmeasured factors may explain these statistical associations and because similar findings can be obtained when the same methodology is applied to height, acne and head-aches, for which the conclusion of contagion effects seems somewhat less plausible. In this talk we first review general sensitivity analysis results for causal effects in observational research. We then use sensitivity analysis techniques to assess the extent to which supposed contagion effects for obesity, smoking, happiness and loneliness might be explained away by homophily or confounding and the extent to which the critique using analysis of data on height, acne and head-aches is relevant. Sensitivity analyses suggest that contagion effects for obesity and smoking cessation are reasonably robust to possible latent homophily or environmental confounding; those for happiness and loneliness are somewhat less so. Supposed effects for height, acne and head-aches are all easily explained away by latent homophily and confounding. The methodology that has been employed in past studies for contagion effects in social networks, when used in conjunction with sensitivity analysis, may prove useful in establishing social influence for various behaviors and states. The sensitivity analysis approach can be used to address the critique of latent homophily as a possible explanation of associations interpreted as contagion effects.

10/24/2011

Mark Tygert, NYU

Title:
Chi-square and classical exact tests often wildly misreport
significance; the remedy lies in computers

Abstract:

If a discrete probability distribution in a model being tested for goodness-of-fit is not close to uniform, then forming the Pearson chi-square statistic can involve division by nearly zero. This often leads to serious trouble in practice -- even in the absence of round-off errors -- as the talk will illustrate via numerous examples.

Fortunately, with the now widespread availability of computers, avoiding all the trouble is simple and easy: without the problematic division by nearly zero, the actual values taken by goodness-of-fit statistics are not humanly interpretable, but black-box computer programs can rapidly calculate their precise significance.
http://arxiv.org/abs/1108.4126
(joint work with Will Perkins and Rachel Ward)

10/31/2011

Date: Monday, October 31, 2011

Mokshay Madima, Yale University

Title: A Shannon-McMillan-Breiman theorem for log-concave measures and applications

Abstract: Our primary goal is to describe a strong quantitative form of the Shannon-McMillan-Breiman theorem for log-concave probability measures on linear spaces, even in the absence of stationarity. The main technical result is a concentration of measure inequality for the ``information content'' of certain random vectors. We will also briefly discuss implications. In particular, by combining this concentration result with ideas from information theory and convex geometry, we obtain a reverse entropy power inequality for convex measures that generalizes the reverse Brunn-Minkowski inequality of V. Milman. Along the way, we also develop some Gaussian comparison inequalities for the entropy of log-concave probability measures, and discuss an information-theoretic formulation of Bourgain's hyperplane conjecture. This is joint work with Sergey Bobkov (Minnesota).

11/14/11

Prof. Tong Zhang (Rutgers University)

Title: Structured Sparsity: theory and practice

Abstract: The notion of sparsity is central to modern high dimensional statistical data analysis. While the standard sparsity concept has been well studied in the past decade, current developments focus on its extensions to more complex problems. A large class of such problems can be referred to as structured sparsity.

The first part of the talk presents a general theory of structured sparsity, focusing on convex relaxation methods. A simple example of convex formulation for structured sparsity is group-Lasso. Although this method was proposed a number of years ago, it had been regarded as an engineering trick without theoretical justification until very recently. As an introduction to the structured sparsity idea, I will first present our recent theoretical results for group Lasso. I will then extend group Lasso to a much larger class of convex structured sparsity formulations, and present a very general theoretical framework that is capable of analyzing these formulations. Some consequences of this general theory will be demonstrated.

In the second part of the talk, I will show how to employ the concept of structured sparsity to improve boosted decision trees with general loss functions. The most influential statistical procedure for learning three ensemble under general loss function is Friedman's gradient boosting method. Although this method has led to many successful industrial applications, it suffers from several theoretical and practical drawbacks. By employing ideas from structured sparsity, we are able to design a regularized greedy forest procedure to address these issues. The resulting method constructs tree ensembles more effectively than gradient boosting, and achieves better performance on most datasets we have tested on.

Joint work with Cunhui Zhang and Rie Johnson

11/21/2011

Ejaz Ahmed (University of Windsor)

Perspectives on Machine Bias versus Human Bias: Generalized Linear Models

In this talk, I consider a mosaic of estimation strategies in generalized linear models when there are many potential predictor variables and some of them may not have influence on the response of interest. In the context of two competing models where one model includes all predictors and the other restricts variable coefficients to a candidate linear subspace based on prior knowledge, we investigate the relative performances of absolute penalty estimator (APE), shrinkage in the direction of the subspace, and candidate subspace restricted type estimators. We develop large sample theory for the shrinkage estimators including derivation of asymptotic bias and mean-squared error. The asymptotic and a Monte Carlo simulation study show that the shrinkage estimator overall performs best and in particular performs better than the APE when the dimension of the restricted parameter space is large. The estimation strategies considered in this talk are also applied on a real life data set for illustrative purpose.

Joint work with K. Doksum and S. Hossain

11/28/2011

Marvin Zelen (Harvard University)

"EARLY DETECTION OF DISEASE AND STOCHASTIC MODELS"

The early detection of disease presents opportunities for using existing technologies to significantly improve patient benefit. The possibility of diagnosing a chronic disease early, while it is asymptomatic, may result in diagnosing the disease in an earlier stage leading to better prognosis. Many cancers, diabetes, tuberculosis, cardiovascular disease, HIV related diseases, etc. may have better prognosis when combined with an effective treatment. However gathering scientific evidence to demonstrate benefit has proved to be difficult. There are current controversies about the benefit of; (1) screening women to diagnose breast cancer using mammography and (2) using prostate specific antigen (PSA) to diagnose prostate cancer. Clinical trials have been arduous to carry out, because of the need to have large numbers of subjects, long follow-up periods and problems of non-compliance. Implementing public health early detection programs have proved to be costly and not based on analytic considerations. Many of these difficulties are a result of not understanding the early disease detection process and the disease natural histories. One way to approach these problems is to model the early detection process. This seminar will discuss stochastic models for the early detection of disease. The seminar will discuss : current controversies on screening for breast and prostate cancers, length biased sampling and its implications, randomized trials, optimal scheduling of examinations (time based schedules vs risk based schedules) and the over diagnosis of disease.. These issues cannot be addressed by clinical trials because such trials are unethical or are impractical.

12/5/2011

Marcel Nutz (Columbia Mathematics)

Duality and Superreplication under Model Uncertainty

We consider the problem of superreplication under Knightian uncertainty in a discrete-time financial market. In the absence of a reference probability measure, we develop a duality theory which is based on a locally convex vector space and allows to treat measurable quantities without further topological restrictions. We obtain the existence of an optimal strategy and a duality relation between (non-equivalent) martingale measures and superreplicable claims. The continuum hypothesis plays an important role in our approach, which is also related to certain ideas from robust statistics.

12/12/2011

Melanie Wall (Columbia Biostatistics)

TITLE: Clustering Variables Clustering Individuals - Examining factor mixture models and a new structured latent class method.

ABSTRACT: The latent class model (LCM) is a model based clustering method typically used for clustering individuals into k distinct classes (clusters) based on p ordered categorical variables. The basic assumption of LCM is that all p variables are conditionally independent given class (cluster) membership. Factor mixture models (FMM) extend the traditional factor analysis model for p continuous or ordered categorical variables such that the q (q<p) underlying continuous latent factors are assumed to come from a k component mixture. Unlike LCM which directly clusters individuals based on all p variables, the FMM reduces the dimension from p observed variables to q latent factors and then clusters individual on the q continuous latent factors. Motivated by the FMM, a structured LCM is considered in this talk that partitions the p variables into smaller groups of variables that are indicative of particular categorical aspects of the underlying clusters.

Two examples applying these methods will be presented throughout: one which explores potential clusters of adolescents' weight related behaviors and environments and the other which explores potential clusters of behavioral disturbances reported by caregivers of Alzheimer's patients.

BIO: Dr. Wall is Professor of Biostatistics in Psychiatry at Columbia University. She received a Ph.D. from the Department of Statistics at Iowa State University in 1998 and was on the biostatistics faculty in the School of Public Health at the University of Minnesota for 12 years before coming to Columbia in 2010. Here research includes latent variable modeling (e.g. factor analysis, item response theory, latent class models, structural equation modeling), spatial data modeling (e.g. disease mapping), and longitudinal data analysis including the class of longitudinal models commonly called growth curve mixture models. She have extensive experience working with epidemiological observational data related to behavioral psycho-social public health and psychiatry.

***Date: Thursday, December 15**
**Time: 1:00
***Location: 903 SSW

Simon Tavaré

DAMTP and Oncology, University of Cambridge
Molecular and Computational Biology, University of Southern California

Life without likelihoods

Abstract: Approximate Bayesian Computation (ABC) arose in response to the difficulty of finding posterior distributions determined by intractable likelihoods. The method exploits the fact that while likelihoods may be impossible to compute in complex probability models, it is often easy to simulate observations from them. ABC in its simplest form proceeds as follows: (i) simulate a parameter from the prior; (ii) simulate observations from the model with this parameter; (iii) accept the parameter if the simulated observations are close enough to the observed data. The magic, and the source of potential disasters, is in step (iii). This talk will outline what we know (and don't!) about ABC. The ideas will be illustrated with an example from stem cell biology and another concerning the evolution of primates.
(Biological expertise is not required!)