Abstracts

Talks

1. Tatiyana V. Apanasovich
Cornell University
tanya@orie.cornell.edu
Semiparametric Spatial Modeling of Binary Outcomes
In an experiment to understand colon carcinogenesis, all animals were exposed to a carcinogen while half the animals were also exposed to radiation. Spatially, we measured the existence of what are referred to as aberrant crypt foci (ACF), namely morphologically changed colonic crypts that are known to be precursors of colon cancer development. The biological question of interest is whether these binary responses occur at random through the colon: if not, this suggests that the effect of environmental exposures is localized in different regions.

Important features of the motivating example include a large number of repeated measures per subject and a small number of subjects. Many papers on longitudinal data address data sets of the opposite type where there are many subjects but relatively few repeated measures per subject. Our aim is to produce a theoretical framework for longitudinal and spatial data which is general enough to apply to our example as well as many others with a large number of outcomes per subject. Such examples typically exhibit several features: (1) nonlinearity may be seen so that a nonparametric model of the mean function, that is, the conditional expectation of the response given the covariates, is likely to be needed, and (2) correlation between repeated measures can be estimated sufficiently accurately that nonstationarity may be evident. Our model of the mean function is new and includes single-index models, additive models, and partially linear models as special cases. Though motivated by a spatial example, it is not restricted to longitudinal and spatial data.

2. John Aston
Institute of Statistical Science Academia Sinica
jaston@stat.sinica.edu.tw
Waiting Time Distributions for Runs and Patterns in Higher Order Markovian Sequences
In recent years, there has been significant research on the calculation of waiting time distributions for runs and patterns in sequences of elements from a Markov chain (first order Markovian sequence). These distributions include the waiting time until the first occurrence of a run of a specific length in a binary state space, until the r'th occurrence, or until the first occurrence of a simple or compound pattern in a more general finite state space. Here using the technique of Finite Markov Chain Imbedding (Fu and Koutras, 1994) methods to generate general waiting time distributions when the sequence is a higher order Markovian sequence will be established. Included distributions will be the extension of those above, and also the case of the r'th occurrence of a compound pattern in a general finite state space.

3. Antar Bandyopadhyay
Chalmers University of Technology
antar@math.chalmers.se
Recursive Distributional Equations : Application to Hard-Core Model on Random Graphs
In a variety of applied probability settings, from the study of Galton-Watson branching processes to mean-field and other type of statistical physics models, a central theme is to solve some fixed-point equation on an appropriate space of probabilities, such an equation is called a recursive distributional equation (RDE). Study of such equation has been one of my area of research in the last few years. In this talk I will briefly give some examples where such equations arise naturally. In particular I will discuss the hard-core model (random independent set) on Galton-Watson trees and sparse random graphs (e.g. Erdos-Renyi random graph). For the model on a Galton-Watson tree I will demonstrate a phase transition phenomenon which can be characterize by uniqueness of solution of a particular RDE. This will generalize an earlier work of Kelly (1985) on regular trees. Using this I will also provide some exact asymptotic behavior for the same model on sparse random graphs.

4. John Beam
University of Wisconsin Oshkosh
beam@uwosh.edu
Expectations for Coherent Probabilities: Defining the Integral
In the 1930's, around the same time that Kolmogorov developed the standard axioms for probability theory, Bruno de

Finetti proposed an alternative, more general, model. He interpreted a probability as an assignment of fair odds for a bet – an assignment of odds with respect to which it is impossible for a clever gambler to ensure himself a victory against an unwitting opponent. This model is consistent with that of Kolmogorov, but requires neither countable additivity of the measure nor any sort of structure on the domain. For perhaps two reasons, de Finetti's work has been followed by relatively few people: his mathematical style and terminology are nonstandard; the mathematical framework, in contrast to that of countably-additive measures, is largely undeveloped. (In particular, there has been no systematic development of the integral.) I will address the first issue by presenting de Finetti's definitions in a manner common to that of traditional measure theory. My primary aim is to address the second issue by introducing a theory of integration (expectations) for de Finetti's probabilities.

5. Swati Biswas
   The University of Texas-MD Anderson Cancer Center
   sbiswas@mdanderson.org
   Modeling Locus Heterogeneity in Linkage Analysis
   Heterogeneity is a ubiquitous feature of complex genetic traits, such as some cancers, asthma, diabetes, etc and poses a major difficulty in mapping disease-causing genes via linkage analysis. It refers to the situations when a disease is caused in some families by one gene while in other families it is caused by some other gene and/or non-hereditary factors. The currently used approach based on mixture likelihood uses a single mixing parameter to model the heterogeneity. However, in general, different types of families exhibit different heterogeneity levels, which a single heterogeneity parameter is unable to capture. To incorporate this variability, we propose a new approach, wherein each family has its own heterogeneity parameter. These parameters are nuisance parameters while the main parameter of interest is the location of the disease gene, if there is any. We model the problem in the Bayesian framework and implement it using the reversible jump Markov chain Monte Carlo methodology. We show that the proposed method is more powerful than the currently used approach in detecting linkage while the two approaches have comparable false positive rates. The proposed method is applied to a lung cancer dataset of Genetic Epidemiology of Lung Cancer Consortium.

6. Jose H. Blanchet
   Harvard University
   blanchet@fas.harvard.edu
   Approximations and Computational Algorithms in Stochastic Modeling
   In the performance analysis of complex stochastic systems one typically has to take advantage of analytic approximations and computational algorithms, such as efficient stochastic simulation. We shall illustrate some of these techniques in the context of problems motivated by risk insurance theory, inventory theory, finance and queuing theory.

7. Brian Caffo
   Johns Hopkins University
   bcaffo@jhsph.edu
   Statistical reconstruction algorithms in SPECT imaging
   Single photon emission computed tomography (SPECT) is a modern non-invasive functional imaging technique. In SPECT, harmless amounts of radioactive isotopes are introduced into a patient, such as via injection. Specialized cameras count emitted photons while being rotated around the subject. Computed tomography methods are used to reconstruct a three dimensional image given the collected photon counts. In gated cardiac SPECT, researchers synchronize the image acquisition with the cardiac cycle to estimate cardiac function. The well known EM algorithm is a popular alternative to the standard filtered back projection (FBP) algorithms used in clinical practice. The benefit of EM is the ease in which researchers can account for known physical processes that degrade image quality, such as attenuation, scatter, and depth dependent blurring. It has been shown that these factors not only degrade image quality via mathematical measures, but have a negative impact on clinical sensitivity and specificity. Though an improvement over the FBP algorithm, the (maximum likelihood) limit of the EM algorithm results in images of poor quality as the model ignores spatial and temporal correlation in the image sequence. This talk overviews the SPECT process, statistical reconstruction algorithms and their validation. Particular attention will be paid to Monte Carlo phantom studies via an anatomically accurate mathematical phantom and a physically accurate computational representation of the SPECT imaging process.

8. Gabriel Chandler
   Connecticut College
   gabriel.chandler@conncoll.edu
   Classification of Locally Stationary Time Series via the Excess Mass Functional
   Classification of locally stationary time series is a well-studied problem and has many applications in seismology, ecology, and many other fields. Most of what is found in the literature regarding this problem are distance-based approaches involving the time-varying spectrum. However, the use of distance based criteria requires that the observed series be aligned in time. Classification techniques based on measures of concentration of volatility are proposed. These techniques compare well with those existing in the literature under certain models, while not requiring alignment of the data. Applications to seismic data and animal footfall data will be given.

9. Pankaj K. Choudhary
   University of Texas at Dallas
   pankaj@utdallas.edu
   Assessment of Agreement Using Tolerance Intervals
   We will consider the problem of assessment of agreement between a test and a reference method of measurement of a continuous variable. The goal is to determine whether, for a subject, a measurement from the test method can be substituted for a measurement from the reference method without leading to any difference in the clinical interpretation of the measurement. This problem arises in medical applications where the reference method, sometimes called a gold-standard method, is expensive or invasive, and the test method provides a convenient alternative. We will describe a tolerance interval approach introduced by Lin (2000, Statist. Med., 19, 255-270) and Lin et al. (2002, J. Am. Stat. Assoc., 97, 257-270) and briefly discuss some of its extensions to handle non-constant mean or variance and repeated measurements. A real data example will also be presented.

10. Samantha Cook
    Columbia University
    cook@stat.columbia.edu
    Validation of Software for Bayesian Models using Posterior Quantiles
    We present a simulation-based method designed to establish that software developed to fit a specific Bayesian model works properly, capitalizing on properties of Bayesian posterior distributions. The validation method involves repeatedly generating parameters and data from the model to be fit and then fitting the same model to these simulated data (i.e., generating a sample from the posterior distribution). For all scalar parameters, the quantile of the "true" parameter value with respect to its posterior distribution should follow a uniform distribution if the software is written correctly. Testing that the software works amounts to testing that these quantiles are uniformly distributed. We illustrate that the validation method finds errors in software when they exist and, moreover, the validation output can be informative about the nature and location of such errors.

11. Kimberly Drews
    Texas A & M University
    kdrews@stat.tamu.edu
    A Likelihood Based Approach to the Analysis of Coordinated Response Among Colonic Crypts
    The colon (large intestine) wall contains colonic crypts with cells lining the inside of each crypt. There has long been a conjecture that a coordinated response existed at the crypt level, i.e., that the biological responses in one crypt affect the biological responses in neighboring crypts. Our work examines this hypothesis. The biological response of interest is the amount of p27 present in the cell. p27 is a cyclin dependent kinase inhibitor which, at high levels, keeps the cell from progressing out of the G1 phase into the S phase and is believed to help promote apoptosis, a type of programmed cell death. The responses have a natural hierarchical structure which we posit follow a mixed model, which is difficult to analyze with traditional methods due to the fact that the set of distances between colonic crypts for each subject is unique. We propose the crypt correlations follow an AR(1) model or one of several Matern correlation models. We note that our interest ! lies in the lower level of the hierarchy rather than the top level. For data of this type we have developed a methodology allowing us to avoid fitting the entire hierarchical model thus permitting us to efficiently find maximum likelihood estimates for the value of the correlation. The method is illustrated using simulations and shown to perform well. The method is then applied to the data that inspired the development of this technique.

12. Yongchao Ge
    Mount Sinai School of Medicine
    Yongchao.Ge@mssm.edu
    An Upper Confidence Bound of the False Discovery Proportion
    Benjamini and Hochberg (1995) propose the false discovery rate (FDR) as an alternative criterion to the traditional family-wise error rate in multiple testing problems. However, most previous works on the FDR focus on the point estimation: the expectation of the false discovery proportion (FDP). Being motivated by the fixed rejection approach in multiple testing of Storey (2002), this paper proposes a simple procedure to construct an upper confidence bound of the FDP for a fixed rejection region, and also to construct an upper confidence bound of the total number of true null hypotheses. A procedure is given to construct an upper confidence band simultaneously valid for all rejection regions.

13. Kevin Gross
    North Carolina State University
    gross@stat.ncsu.edu
    Estimating abundances from count data for species with discrete generations
    Most methods of estimating wildlife abundance are based on mark-recapture methods. With fragile or threatened species, however, capture is not feasible and abundance estimates must rely on observational count data alone. When such species have discrete generations (as insects often do), a simple estimation question arises: how can time series of count data be used to estimate the total number of individuals in a single generation? Current methods for estimating single-generation abundances with count data resemble state space models, but treat the unobserved population dynamics as deterministic, only allowing for stochasticity in the counting process. By ignoring stochasticity in population dynamics, the precision of the abundance estimate is overestimated, and confidence intervals are too small. Here, we modify existing estimation methods to account for stochasticity in the population dynamics, and study the properties of the new abundance estimators. This work is motivated by the study of an endangered butterfly occuring in the North Carolina sandhills, and is collaborative work with Eric J. Kalendra, Brian R. Hudgens, and Nick M. Haddad.

14. Murali Haran
    The Pennsylvania State University
    mharan@stat.psu.edu
    Monte Carlo for spatial models: two issues and some relevant methodology
    Hierarchical models are increasingly used for spatial data arising from topics as diverse as epidemiology, environmental health and climatology. Markov chain Monte Carlo (MCMC) methods allow fully Bayesian analyses of hierarchical spatial models. However, as spatial models become more sophisticated and data sets grow in size, MCMC methods become less reliable. Worse yet, it also becomes more difficult to make decisions regarding how long to run MCMC algorithms (finding stopping criteria) and how to assess the accuracy of the MCMC-based estimates (calculating Monte Carlo standard errors). I will discuss general strategies used to help deal with such problems including (i) alternative sampling methods that produce independent and identically distributed samples and (ii) techniques for reliably assessing Monte Carlo standard error and providing sensible stopping criteria.

15. Amelia M. Haviland
    RAND
    haviland@rand.org
    Causal Inferences with Group Based Trajectory Models
    A central theme of research on human development and psychopathology is whether a therapeutic intervention or a turning point event, such as a family break-up, alters the trajectory of the behavior under study. This paper lays out and applies a method for using observational longitudinal data to make more confident causal inferences about the impact of such events on developmental trajectories. The method draws upon three distinct lines of research: Work on the use of finite mixture modeling to analyze developmental trajectories, work on propensity scores, and work on the application of the g-equation framework to causal inference. The essence of the method is to use the posterior probabilities of trajectory group membership from a finite mixture modeling framework to create balance on lagged outcomes and other covariates established prior to t for the purpose of inferring the impact of first-time treatment at t on the outcome of interest. The approach is demonstrated with an analysis of the impact of gang membership on violent delinquency based on data from a large longitudinal study conducted in Montreal. This is joint work with Daniel S. Nagin.

16. Chiu-Hsieh Hsu
University of Arizona
phsu@azcc.arizona.edu
Joint Modeling of Recurrence and Progression of Adenomas: A Latent Variable Approach
We treat the number of recurrent adenomatous polyps as a latent variable and then use a mixture distribution to model the number of observed recurrent adenomatous polyps. This approach is equivalent to zero-inflated Poisson regression, which is a method used to analyze count data with excess zeros. In a zero-inflated Poisson model, a count response variable is assumed to be distributed as a mixture of a Poisson distribution and a distribution with point mass of one at zero. In many cancer studies, patients often have variable follow-up. When the disease of interest is subject to late onset, ignoring the length of follow-up will underestimate the recurrence rate. In this paper, we modify zero-inflated Poisson regression through a weight function to incorporate the length of follow-up into analysis. The weight function can be estimated using profile likelihood approaches. We motivate, develop, and illustrate the methods described here with an example from a colon cancer study.

17. Mark Inlow
Rose-Hulman Institute of Technology
inlow@rose-hulman.edu
New Goodness-of-Fit/Goodness-of-Link Smooth Tests
Many goodness-of-fit/lack-of-fit procedures test a model by embedding it in a larger family of models and then comparing it with competing models from that family. Neyman (1937) proposed the first test test of this form, the "smooth test," so-called because the family of models is explicitly constructed to detect "smooth" departures of the data from the null model. Numerous smooth test generalizations have been developed for the composite case in which the hypothesized model is estimated from the data. We present new composite smooth tests (goodness-of-fit and goodness-of-link) which achieve greater power by ameliorating the under appreciated, deleterious effects of model estimation.

18. Woncheol Jang
Duke University
wjang@stat.duke.edu
Uniform Confidence Sets for Densities
In this talk, I will discuss recent work on constructing confidence sets for an unknown function in nonparametric density estimation problems. The goal is to construct a set – usually a ball in some space or, alternatively, bands – that provides (asymptotically) uniform coverage for the whole function. Inferences can then be generated by searching this set, possibly with added constraints from available side information.

One approach I'll describe extends results by Beran and Dumbgen (1998) to density estimation. We expand the density in an appropriate basis and we estimate the basis coefficients by using linear shrinkage methods. We then find the limiting distribution of an asymptotic pivot based on the quadratic loss function. Inverting this pivot yields a confidence ball for the density This is joint work with Larry Wasserman and Chris Genovese.

19. Jiashun Jin
Purdue University
jinj@stat.purdue.edu
Sparse Inference in Large Scale Multiple Comparisons and False Discovery Rate Thresholding
Control of the *False Discovery Rate* (FDR) is a recent innovation in multiple hypothesis testing, allowing the user to limit the fraction of rejected null hypotheses which correspond to false rejections (i.e. false discoveries). The FDR principle also can be used in multiparameter estimation problems to set thresholds for separating signal from noise when the signal is sparse. Success has been proven when the noise is Gaussian.

In this talk, we consider the application of FDR thresholding to sparse signals in general *additive* noise, in hopes of learning whether the good asymptotic properties of FDR thresholding as an estimation tool hold more broadly than just at the Gaussian model.

20. John Kern
Duquesne University
kern@mathcs.duq.edu
Bayesian Modeling Strategies for Longitudinal Frequency Data
The objective of this research is to develop a Bayesian model appropriate for frequency data collected regularly for

several individuals over an extended time period. We develop and implement competing discrete-data models that handle differently the time dependence inherent in longitudinal data. Motivated by a study investigating alternative treatments for relief of menopausal symptoms, we apply these models to actual study data in an effort to compare their effectiveness.

21. George Kordzakhia
University of California, Berkeley
kordzakh@stat.Berkeley.EDU
Stochastic spatial models of species competition and predator-prey interactions
We consider a class of multi-type spatial models where several types of particles (species) interact on the integer lattice Z^d ($d \geq 2$). One of the goals is to understand the limiting behavior of the occupied regions and conditions for the long-term coexistence of the distinct types of particles. The multi-type models can be classified roughly into two groups: competition models and predator-prey models. We give examples of models of each type, describe their properties and state some conjectures.

22. Michael Levine
Purdue University
mlevins@stat.purdue.edu
Variance Estimation in Nonprametric Regression – A Possible Apporach
Traditionally, the nonparametric regression research has been centered on the mean estimation problem when the variance is constant. Very often, however, homoscedasticity assumption is not quite a viable option. In a few applications, such as conditional variance function estimation in financial time series or immunoassay, the variance is a function of an observed argument and its estimate is needed to construct a confidence interval or prediction interval.

We consider the non-parametric regression model

$$y_i = g(x_i) + \sqrt{f(x_i)}\epsilon_i \tag{1}$$

for $i = 1, \ldots, n$ where the observations are ordered and $x_{i+1} - x_i = \frac{1}{n}$. We assume that both the mean function $g(x)$ and the variance function $f(x)$ belong to some smoothness class but are otherwise unknown. The object of interest is the variance while the mean function plays the nuisance parameter role.

We present a class of variance estimators that is based on smoothing transformed data. First, differences of observations of order r are defined as

$$\Delta_{r,i} = \sum_{j=0}^{r-1} d_j y_{j+i} \tag{2}$$

for a set of numbers $\{d_i\}$ such that $\sum_j d_j = 0$ and $\sum_j d_j^2 = 1$ and $i = 1, \ldots, n-r$. Then, the local polynomial smoother can be applied to these differences to obtain the estimated variance function at the point x_i. These estimators exhibit certain commonality with the more established kernel-based estimators of the mean function, such as Nadaraya-Watson or Gasser-Müller estimators. In particular, for p-times continuously differentiable variance function $f(x)$ the L_2-convergence rate for this class of estimators is n^{-l} where $l = \frac{2p}{2p+1}$. We derive exact expressions for the asymptotic risk and show that this rate of convergence is true for any finite $r > 0$. We show that our estimator class is asymptotically better in reducing the bias component of its L_2-risk then the competing class described in Fan and Yao (1998) while having the same asymptotic order of the variance component. We also demonstrate that when the ratio of the difference order to the sample size $\frac{r}{n}$ becomes large, the variance component of the L_2-risk of these estimators slowly decreases at the rate of $\frac{1}{r}$, achieving certain limiting value as $r \to \infty$. On the contrary, the bias component does not depend on the order of the differences used. On the basis of this result, some practical conclusions about the proper choice of r is made. Finally, the asymptotic minimaxity of our estimator class is also established.

To enable practical application of this estimator class, the bandwidth selection mechanism is needed. The plug-in type algorithm and crossvalidation-type algorithm for bandwidth selection are introduced and discussed. We conclude that the former results in the problem more complicated than the original variance estimation problem. The latter, on the contrary, seems to possess fairly good empirical properties that are demonstrated using simulated data. This is based on the joint work with Prof. Lawrence D. Brown.

23. Lexin Li
University of California, Davis
lexli@ucdavis.edu
Sufficient Dimension Reduction in High-dimensional Data
Given a large number of predictors in a regression, it is often desirable to reduce the dimensionality of the problem by replacing the original high-dimensional data with a low-dimensional space composed of a few key predictors or linear combinations of predictors. In this talk, I will first introduce the general framework of sufficient dimension reduction, which targets the reduction of dimension without losing any information on the conditional distribution of response given predictors, and without pre-specifying any parametric model. I will then briefly review some of my related methodological work within this framework.

24. Liang Li
Cleveland Clinic Foundation
lli@bio.ri.ccf.org
Some Measurement Error Models with Complicated Error Structures
I will describe two clinical problems where covariate measurement errors play an important role in regressions. Unlike many measurement error models in the literature, the regression models are pretty simple for these problems, while the error models are quite complicated. In the first problem, both the true and observed covariates are semi-continuous; in the second problem, the variance of the error depends on the unknown true covariate. I will discuss issues related to these problems and our proposed solutions.

25. Johan Lim
Texas A & M
johanlim@stat.tamu.edu
Function Estimation with Shape or Order Constraints
In this talk, two conventional optimization techniques, the constrained uniform approximation and the geometric programming, are introduced with applications to function estimation with shape or order constraints.

26. Anna Liu
University of Massachusetts Amherst
anna@math.umass.edu
Hypothesis Testing in Smoothing Spline Models
One of my research interests lies in model diagnostic. In this talk, I will present a smoothing spline based framework for testing the hypothesis of a generalized linear model, generalized additive model and more generally, mixed models. Within this framework, these seemingly complicated hypothesis are transformed to hypothesis on smoothing parameters, which allows both classical tests and innovative tests to be developed. I will show the comparative performances of several tests and their applications to real data.

27. Dacheng Liu
University of Rochester
dliu@bst.rochester.edu
Mixed-effects state space models
The rapid development of new biotechnologies allows us to deeply understand the biomedical dynamic systems in more details and at a cellular level. Many of the subject-specific biomedical systems can be described by a set of differential or difference equations which is similar to an engineering dynamic system. Motivated by HIV dynamic studies, we propose a class of mixed-effects state space models based on the longitudinal feature of the dynamic systems. Three estimation methods (global two-stage, Bayesian approach and MLE) for standard mixed-effects models are modified and investigated for estimating unknown parameters in the proposed mixed-effects state space models. Simulation results indicate that all the three methods perform well when the number of observations per subject is large. Finally, we apply the mixed-effects state space model to a data set from an AIDS clinical trial to illustrate the proposed methodologies.

28. Xueli Liu
University of Florida
xueli@stat.ufl.edu
Detecting Differentially Expressed Time Course Gene Expression Profiles
With the burgeoning filed of gene expression and microarrays, novel statistical methods are in great needs to analyze

such types of genomics data. Among these high-throughput data, time course gene expression profiles can reveal important dynamic features of cell activities. Yet not so much effort has been contributed to address the key question of detecting differentially expressed time course gene expression data. Furthermore, the experimental designs for time course gene expression data are often not consistent across subjects, e.g., varying sampling rates and the total number of time points for each subject sampled are often small. To address these questions, we present a statistical method for detecting statistical significance of time course gene expression data. The idea is to integrate a principal component analysis through conditional expectation method for sparse longitudinal data into a nonparametric bootstrap model framework. In doing so, we can define a significance measure for each gene expression profile. The method is applied to a developmental C. elegans microarray study. Our aim is to identify genes which are development-specific.

29. Wenbin Lu
    North Carolina State University
    wlu4@stat.ncsu.edu
    Marginal Regression of Multivariate Event Times Based on Linear Transformation Models
    Multivariate event time data are common in medical studies and have received much attention recently. In such data, each study subject may potentially experience several types of events or recurrences of the same type of event, or event times may be clustered. Marginal distributions are specified for the multivariate event times in multiple events and clustered events data, and for the gap times in recurrent events data, using the semiparametric linear transformation models while leaving the dependence structures for related events unspecified. We propose several estimating equations for simultaneous estimation of the regression parameters and the transformation function. It is shown that the resulting regression estimators are asymptotically normal, with variance-covariance matrix that has a closed form and can be consistently estimated by the usual plug-in method. Simulation studies show that the proposed approach is appropriate for practical use. An application to the well-known bladder cancer tumor recurrences data is also given to illustrate the methodology.

30. Zeng-Hua Lu
    University of South Australia
    Zen.Lu@unisa.edu.au
    A Mixture Model of Heterogeneous Covariates with an Application to the Censored Dependent Variable
    Mixture regression models have been used to model structural parameter instability in the regression analysis. But it has often been assumed that a same set of covariates exert different effects on the response variable in each mixture component; the heterogeneity of explanatory mechanism arises from the parameter heterogeneity with homogeneous regressors. The effects of the heterogeneity in terms of the explanatory factors across regression regimes have so far largely been ignored. This paper examines the heterogeneity of explanatory mechanism, which may arise from the parameter heterogeneity effect, and / or the covariates heterogeneity effect. We illustrate our model with a particular interest to the censored response variable. Issues concerning the statistical inference of the resulting model are studied. First, the strong consistency of our estimator under the nonidentifiabilty condition, where all regressors in a mixture component are mistakenly included, is investigated. Second, because both the number of mixture components and component covariates are assumed to be unobservable, the number of candidate models can be enormous. We suggest a model selection procedure, which adds on little more computational burden and is based on combining the information of the estimator of regression coefficients and sample value of model selection criterion function, such as BIC. Our simulation studies confirm the suggested method works well. An empirical study of female labor supply is provided.

31. Yajun Mei
    Fred Hutchinson Cancer Research Center
    ymei@fhcrc.org
    Change-point problems and information fusion
    The problem of quickest change detection, or the change-point problem, has a variety of applications including industrial quality control, reliability, (bio)surveillance, and security systems. By monitoring data streams which are generated from a process, we are interested in quickly detecting malfunctioning once the process goes out control, while keeping false alarms as infrequent as possible when the process is in control. The classical or centralized version of this problem, where all observations are available at a single, central location, is a well-developed area. In this talk, motivated by information fusion and its applications in mobile and wireless communication and surveillance systems, we generalize this problem to the decentralized version where the information available is distributed across a set of sensors. Each sensor receives a sequence of observations, and sends a sequence of sensor messages to a central processor, called the fusion center, which makes a final decision when observation are stopped. In order to reduce the

communication costs, it is required that the sensor messages belong to a finite alphabet. In the decentralized version, the goal is to detect the change as soon as possible over all possible protocols for generating sensor messages and over all possible decision rules at the fusion center, under a restriction on the frequency of false alarms. We will present a general asymptotic theory, and provide procedures that are asymptotically optimal and easy to implement.

32. Ayrin Molefe
University of Central Arkansas
calachan@yahoo.com
An Extension of the Neyman-Johnson Technique to Binary Regression
In the comparative study of two treatments based on a normally distributed outcome where a confounding factor (covariate) is present, the analysis of covariance (ANCOVA) is a standard tool used in many fields of application. When the usual parallel-line assumption in the ANCOVA is untenable, the Johnson-Neyman technique provides a useful alternative which can identify a region of covariate values for which there are significant treatment differences. When the response is dichotomous rather than continuous, logistic regression is used analogously to ANCOVA to control for confounding. As such, logit analysis likewise relies on the assumption of parallelism or no treatment-covariate interaction. We propose an extension of the Johnson- Neyman technique which allows comparison of non-parallel logistic regression lines. We illustrate the method using a fictitious and a real data.

33. Samantha Bates Prins
Virginia Tech
sbates@vt.edu
Scaling by Reference Conditions for Ecological Assessment
Reference sites provide important information about the range of biological, physical and chemical measurements. Using reference information to scale data from other sites is useful for evaluating the status of sites and establishing impairment. A common approach is to use all the data on the reference conditions to standardize measurements for a new site. Rather than using all available reference sites to scale the observed value of a particular metric at a test site, I will present an alternative approach that uses only the k closest (in terms of selected predictors) reference sites. Using a set of data from the Mid-Atlantic Highlands, I will show that the nearest neighbor method improved on the ability of the regression approach to classify test sites correctly without affecting the ability to predict reference sites. I will also present an overview of my interest in Bayesian uncertainty assessment for multicompartment deterministic simulation models.

34. Jing Qiu
University of Missouri
qiujing@missouri.edu
Sharp Simultaneous Intervals for the Means of Selected Populations with Application to Microarray Data Analysis
Simultaneous inference is a challenge when the number of populations, N, or the dimensionality is large. In some situations, including microarray experiments, the scientists are only interested in the K populations with parameters (such as means) that have the most extreme estimates. In these situations, can we construct simultaneous intervals for the means corresponding to these K selected populations? The answer is yes, as demonstrated here, and the approach allows us to cut down the dimensionality of the problem from N to K. The naive simultaneous intervals for the K means (applied directly without taking into account the selection) have low coverage probabilities. We take an Empirical Bayes approach (or an approach based on the mixed effect model) and we construct simultaneous intervals with good coverage probabilities. For N=10,000 and K=100, typical for microarray data, the lengths of our intervals could be 82% shorter than those of the Bonferroni's N-dimensional simultaneous intervals and 77% shorter than those of the naive K-dimensional simultaneous intervals. This is joint work with J.T. Gene Hwang.

35. Arni SR Srinivasa Rao
University of Guelph
arnirao@uoguelph.ca
Limit theorems approach in epidemic reporting and virus dynamics
Rate of convergence of independent sequence of real variables can be well explained using limit theoretic approach. In this work, we define a relation between reported and actual disease cases as a sequence of differences and explain its convergence properties through a function called an efficiency function. Three limit theorems were developed that explain the error of reporting of disease cases in epidemiology. A new methodology is developed using limit theory of branching process, that given a cohort of virus at time T, then what is the probability that this cohort of viruses will extinct at time Tn (for Tn¿¿T). This method is successfully demonstrated to explain the probabilities of extinction

of human immunodeficiency virus under the regulation of therapy. Application of Cauchy's convergence principle is explored to explain the probability convergence of virus concentrations.

36. Richard Samworth
Cambridge University
R.J.Samworth@statslab.cam.ac.uk
First order properties of k-nearest neighbor and bagged nearest-neighbor classifiers
Suppose we have two random samples $X_1, ..., X_m$ and $Y_1, ..., Y_n$, and wish to classify a new observation z as coming from one or other of the two populations. A traditional and appealing nonparametric classifier is the k-nearest neighbor rule, which assigns z to the X population if at least half of the nearest k observations in the combined sample come from the X population.

We show how to choose k optimally, in the sense of minimizing the asymptotic error rate, or risk. Moreover, we also study a natural competitor called the bagged nearest neighbor classifier, which behaves asymptotically like a weighted k-nearest neighbor classifier, with geometrically decreasing weights on successively further observations from z. Bagging involves combining the results of classifications based on resamples of the original data, and the optimal choice of resample size is derived. Simple consequences of these results are the optimal rates of convergence to zero of the regret, i.e. the difference in asymptotic risk between the classifier in question and the optimal Bayes classifier.

Finally, we examine the ratio of the asymptotic regrets of the two classifiers when both tuning parameters are chosen optimally. Interestingly, this ratio does not depend on the underlying distributions, instead only on the dimension of the data. In particular, the bagged nearest-neighbor classifier always improves asymptotically on the k-nearest neighbor rule, provided that the dimension of the data is at least three.

37. Damla Senturk
The Pennsylvania State University
dsenturk@stat.psu.edu
Covariate Adjusted Regression
We introduce covariate-adjusted regression for situations where both predictors and response in a regression model are not directly observable, but are contaminated with a multiplicative factor that is determined by the value of an unknown function of an observable covariate. We demonstrate how the regression coefficients can be estimated by establishing a connection to varying-coefficient regression. The proposed covariate adjustment method is illustrated with an analysis of the regression of plasma fibrinogen concentration as response on serum transferring level as predictor for 69 haemodialysis patients. In this example, both response and predictor are thought to be influenced in a multiplicative fashion by body mass index. A bootstrap hypothesis test enables us to test the significance of the regression parameters. We establish consistency and convergence rates of the parameter estimators for this new covariate-adjusted regression model. Simulation studies demonstrate the efficacy of the proposed method.

38. Claude Messan Setodji
RAND
setodji@rand.org
Multivariate variable reduction and applications
In this computer age where there is an increasing thirst of understanding things around us, our society has been investing massively in the collection and processing of data of all kinds. Most observed phenomena of interest are usually characterized by high-dimensional variables but the majority of classical statistical methods are not designed to cope with the increasingly high dimensionality of the problems. As the number of variables one has to deal with increases, the luxury of graphical representation, pattern recognition in the data, statistical inferences and predictions become harder to do. In this talk, we will present the method of variable reduction without any loss of information and without any model assumption that deals with multivariate response problems. Applications to real life problems will be discussed.

39. Aleksandra B. Slavkovic
The Pennsylvania State University
sesa@stat.psu.edu
Statistical Disclosure Limitation Beyond the Margins
Statistical disclosure limitation applies statistical tools to the problem of limiting releases of sensitive information about individuals and groups that are part of statistical databases while allowing for proper statistical inference. Within this context, Dobra and Fienberg (2000, 2002) have employed Markov bases, in connection with decomposable

and graphical log-linear models given a set of margins, to establish bounds and distributions for the cell entries in contingency tables. In this talk, we present a framework for finding the bounds and distribution when given an arbitrary collection of marginals and conditionals. We extend the results of Arnold et al. (1999) on the uniqueness of discrete distributions and describe new results on bounds for cell entries in k-way tables estimated via optimization methods such as linear and integer programming. We give a complete characterization of the two way table problem and discuss extensions to multi-way tables including relationships to directed acyclic graphical models. We use tools from algebraic geometry to represent the tables of counts and describe the locus (T) of all possible tables under the given constraints. Markov bases needed to construct a connected Markov chain over T are described. These bases can be used to induce probability distributions over the space of possible tables via Markov Chain Monte Carlo sampling. This research presents new theoretical links between disclosure limitation, statistical theory and computational algebraic geometry and practical implications for confidentiality and statistical disclosure limitation.

40. George Sirbu
Bentley College
GSIRBU@bentley.edu
Optimizing Adaptive Design with Covariates
Many important real world experiments, like clinical trials in medicine, are designed using sampling procedures that are fixed before any data is collected. The ability to change the design once some data becomes available is known as adaptive sampling and can result in more efficient decision making. The potential benefits of adaptive allocation for clinical trials were recognized quite early. Implementation of these application schemes might ease the ethical problem involved in trials on human subjects. My work concentrates on inference for covariate-adaptive and response-adaptive randomization procedures with emphasis on the consistency for the inference and the optimality properties of the design.

41. Brian Smith
The University of Iowa
brian-j-smith@uiowa.edu
Statistical Issues in the Study of Residential Radon
Environmental radon is a radioactive gas that originates from uranium found in rocks and soil. It is present to some extent in all dry-land surface air. The decay products of radon emit alpha particles which are potentially harmful to lung tissue. Studies of radon have given rise to several rich datasets and opportunities for statistical research. I will touch on some of the statistical issues related to the study of radon, including spatial modeling, quantification of exposure, and estimation of disease risk.

42. Russell Stocker
Mississippi State University
rstocker@math.msstate.edu
Some Results Concerning A General Class of Parametric Models for Recurrent Event Data
A general class of models for recurrent event data is considered under a fully parametric specification. An estimation scheme is given and the resulting standardized estimators are shown to be Gaussian under certain regularity conditions. The regularity conditions pertain to calendar time and may be difficult to verify in practice. To facilitate their verification it becomes useful to introduce a gap time formulation. This formulation is given and the transformed gap time regularity conditions are discussed. The model is used to analyze a data set pertaining to the hydraulic subsystems of "load-haul-dump" machines used in mining.

43. Elizabeth Stuart
Mathematica
EStuart@Mathematica-Mpr.com
Matching with multiple control groups and adjusting for differences between the groups
When estimating causal effects using observational data, it is desirable to replicate a randomized experiment as closely as possible by obtaining treated and control groups with similar distributions of observed covariates. This goal often can be achieved by choosing well-matched samples of the original treated and control groups, thus reducing bias due to these covariates. However, sometimes the originally selected control units cannot provide adequate matches for the treated units. In these cases, it may be desirable to obtain matched controls from multiple control groups. Multiple control groups have been used to test for hidden biases in causal inference (Rosenbaum 2002); however, little work has been done on their use in matching or adjustment for these biases. In addition, there may be concern regarding systematic differences between the control groups. Here, we present a method that uses matches from multiple control

groups and adjusts for potentially unobserved differences between the groups in the analysis of the outcome. The methods are applied to data from the evaluation of a school dropout prevention program.

44. Jeffrey Thompson
North Carolina State University
thompson@stat.ncsu.edu
Estimation of Generalized Simple Measurement Error Models with Instrumental Variables

Measurement error (ME) models are used in situations where at least one independent variable in the model is imprecisely measured. Having at least one independent variable measured with error leads to an unidentified model and a bias in the naive estimate of the effect of the variable that is measured with error. One way to correct these problems is through the use of an instrumental variable (IV). An IV is one that is correlated with the unknown, or latent, true variable, but uncorrelated with the measurement error of the unknown truth and the model error. An IV provides the identifying information in our method of estimating the parameters for generalized simple measurement error (GSME) models. The GSME model is developed and it is shown how many well studied ME models with one predictor can fit into its framework. Included in these are linear, generalized linear, nonlinear, multinomial, multivariate regression, and other ME models. The GSME model, by design, can handle situation for continuous, discrete, and categorical observable, or manifest, variables. We provide theorems that give conditions under which the GSME model is identified. The initial step in our estimation method is to "categorize" all continuous and discrete variables. Categorical variables remain unchanged. Assuming conditional independence given the latent variable, the joint distribution of the categorized manifest variables and any that were already categorical is the product of the conditional cell probabilities and conditional distributions of the categorized continuous and discrete manifest variables summing over the categorical values of the latent variable. Maximum likelihood estimates of the joint categorical distribution are used to solve nonlinear equations for the parameters of interest which enter through the conditional probabilities. Estimated generalized nonlinear least squares is used to solve the equations for the parameters of interest. We show that our estimators have favorable asymptotic properties and develop methods of inference for them. We show how many commonly studied ME model problems fit into the general framework developed and how they can be solved using our method.

45. Heather Turner
The University of Warwick
Heather.Turner@warwick.ac.uk
Clustering Microarray Data
The advent of microarray technology nearly a decade ago stimulated considerable research into the statistical analysis of data from microarray experiments. The potential of clustering to identify biologically meaningful patterns was quickly realized and several specialized clustering methods have been introduced since then.

This talk gives an overview of such methods, discussing the motivating features of microarray data and the different forms of clustering that have been developed in response. The talk will conclude with some challenges for future research.

46. Antai Wang
Georgetown University
aw94@georgetown.edu
Parameter Estimation in Bivariate Copula Models
Many models have been proposed for multivariate failure-time data (T_1, T_2) arising in reliability and other applications. A bivariate survivor function $S(t_1, t_2)$ is said to be generated by an archimedean copula if it can be expressed in the form $S(t_1, t_2) = p[q\{S_1(t_1)\} + q\{S_2(t_2)\}]$ for some convex, decreasing function q defined on $(0, 1]$. Here p is the inverse function of q. Usually, p is specified as some function of an unknown parameter θ. Given a sample from $S(t_1, t_2)$, the distribution function of $V = S(T_1, T_2)$, called the Kendall distribution, can be expressed simply in terms of q. We use the score function from the log-likelihood of the V's to estimate θ. Although the V's are unknown, they can be estimated empirically. Interestingly, our estimates based on the empirical V's are much more precise than the estimates based on the true and unknown V's. We also investigate an alternative procedure based on iteratively estimating the V's using the assumed copula structure. We discuss the asymptotic theory for both methods and present some illustrative examples. This is joint work with David Oakes.

47. Haonan Wang
Colorado State University
wanghn@stat.colostate.edu
Object oriented data analysis: sets of trees
Object Oriented Data Analysis is the statistical analysis of populations of complex objects. In the special case of Functional Data Analysis, these data objects are curves, where standard Euclidean approaches, such as principal components analysis, have been very successful. Recent developments in medical image analysis motivate the statistical analysis of populations of more complex data objects which are elements of mildly non-Euclidean spaces, such as Lie Groups and Symmetric Spaces, or of strongly non-Euclidean spaces, such as spaces of tree-structured data objects. These new contexts for Object Oriented Data Analysis create several potentially large new interfaces between mathematics and statistics. This point is illustrated through the careful development of a novel mathematical framework for statistical analysis of populations of tree structured objects. This is joint work with J. S. Marron

48. Jin Wang
Northern Arizona University
Jin.Wang@NAU.EDU
On Peakedness, Kurtosis, and Tailweight
Kurtosis is much more difficult to characterize and interpret than location, spread, and skewness, due to rather sophisticated linkage with peakedness and tail weight, and with asymmetry if present. The meaning of kurtosis has been being a topic of considerable debate. This talk will introduce discussion of historical antecedents on kurtosis, recent and current developments of multivariate kurtosis measures, and some general perspectives of kurtosis. Particular focus will be on interpretation of kurtosis, connections of kurtosis with peakedness and tail weight, and applications of kurtosis. Some comparative analysis of various kurtosis measures along with comments might be made.

49. Patrick J. Wolfe
Harvard University
patrick@deas.harvard.edu
A Bayesian Approach to Imputation of Missing Data Values in Audio Time Series
Audio time series such as digital speech and music recordings may suffer from degradations that result in missing data values, for instance in the case of a damaged vinyl disc in need of restoration or as a result of packet loss in voice-over-IP transmission. As shown in previous work, the structures of such series lend themselves to natural prior specifications in terms of time-frequency behavior, via a decomposition according to the principles of Gabor analysis over finite cyclic groups. Here such a method is applied to address the inherently ill-posed problem of interpolating over repeated short gaps in audio signals. Bayesian models for time-frequency coefficients are postulated based on the idea of a Gabor regression, in which a signal is represented as a superposition of translated, modulated versions of a window function exhibiting good time-frequency concentration. Prior structures suitable for typical audio time series are used in conjunction with stochastic computation via Markov chain Monte Carlo methods to impute missing data values; qualities of the resultant reconstructions are shown to be in keeping with those of the time series under consideration.

50. Jing Wu
Purdue University
jingwu@stat.purdue.edu
Improving the Specificity of Gene Prediction Using Genomic Homology
In the area of computational gene prediction, existing tools routinely generate large volumes of predicted exons (*putative exons*). One common limitation of these tools is the relatively low specificity. To address this issue, a statistical approach is developed that largely improves the prediction specificity. The key idea is to the implement the evolutionary conservation principle to the putative exons. By first exploiting homology between genomes of two related species, a probability model for the evolutionary conservation pattern across different genomes is developed. Second, probability model for the dependency between adjacent codons/triplets is added to differentiate exons and random sequences. Last, based on these models, the log odds ratio is developed to classify putative exons into the group of coding exons and the group of non-coding regions. The approach substantially improves the specificity of the existing methods without much loss in the sensitivity. When tested on pre-aligned human-mouse exons where the putative exons are predicted by GENSCAN and TWINSCAN, we are able to improve the exon specificity by 73% and 32% respectively, while the loss of the sensitivity $\leq 1\%$.

51. Gideon Zamba
The University of Iowa
GZamba@mail.public-health.uiowa.edu
Quality Control Techniques for Disease Monitoring: An example in the Area of Syndromic Surveillance
Quality control faces major issues, one of which is the detection of change in statistical process with unknown parameters. The unknown or partially known parameter context is seen not only in industrial setting but also in disease monitoring such as influenza control. Developing control technique for statistical process with unknown parameters however is a complex problem in which only few observations are available for use in phase I stage. These problems are more difficult in public health arena when, in disease monitoring, the pattern depicted by a disease may involve some unpredictability of its time of emergence. We propose the unknown parameter change-point control methodology and sequential Bayesian control techniques to tackle this problem.

52. Donglin Zeng
University of North Carolina
dzeng@bios.unc.edu
General Transformation Hazard Models in Survival Analysis
The proportional hazards model and the proportional odds model have been commonly used in analyzing survival data. However, these models are restrictive and invalid in modeling complex survival data, for example, crossed survival curves. We study some general classes of transformation hazard models, which includes the proportional hazards and proportional odds models as two special cases. A variety of these models are applied to model independent failure times, clustered failure times and recurrent event times. Nonparametric maximum likelihood estimation is proposed to derive the estimators for both regression parameters and other nuisance parameters. Some efficient algorithms are described to calculate the maximum likelihood estimators. The approaches are also applied to some commonly used medical data in survival literature.

53. Junni Zhang
Peking University
zjn@gsm.pku.edu.cn
Causal Inference, Sequential Monte Carlo and Clustering
After graduation, my research has been concentrated on causal inference, sequential Monte Carlo (SMC) and clustering.

In research on causal inference, I'm mostly interested in making inferences about causal effects when some outcomes are "truncated by death." Together with my collaborators, we have developed a principal stratification approach and showed that this approach can help researchers extract more detailed information from the data compared with traditional methods.

My research on SMC has been focused on developing new variants of SMC methods that can improve statistical inference. Together with my collaborators, we have developed a new set of SMC methods called the Independent Particle Filters, which can better deal with stochastic dynamic systems in which the current observation provides significant information about the current state whereas the state dynamics is weak. We have also borrowed ideas from both SMC and Markov Chain Monte Carlo to develop lookahead and piloting strategies for variable selection.

In research on clustering, I'm interested in clustering in some non-conventional settings, which I called relation-based clustering and hybrid clustering. In relation-based clustering, for each object, various attributes and some response variables are observed, and we can cluster the objects using heterogeneity in the relation between the response variables and the attributes. In hybrid clustering, we can cluster the objects using heterogeneity in both the distribution of attributes, and the relation between the response variables and the attributes.

54. Hongtu Zhu
Columbia University and New York State Psychiatric Institute
zhuh@childpsych.columbia.edu
Latent Variable Models and NeuroInformatics
In this talk, I will give an overview of my research topics in both latent variable model (LV) and neuroinformatics. My primary interests in the LV model include new estimation procedures, testing procedure, model diagnostic tools, and their applications. My current interests in the neuroinformatics include statistical analyses of MRI image, fMRI image, diffusion tensor image, and developing statistical models for analyzing data from different MRI modalities. Specific examples will be discussed throughout the talk.

Posters

1. Ruta Bajorunaite
   Marquette University
   ruta@mscs.mu.edu
   Unadjusted Methods for Comparing Cumulative Incidence Curves
   Problems involving competing risks are common in biomedical and engineering applications. In such problems there are several possible causes of failure. Occurrence of one risk causes the subject to fail and precludes occurrence of other events. Competing risks are often summarized by the cause specific hazard rate. The second basic competing risks summary measure is the cumulative incidence function representing the probability of having experienced failure from the specific cause in the setting where competing risks are acknowledged to exist. It is often of interest to compare probabilities of a particular event between different groups. We consider methods for comparing probabilities of a specific event between two groups in the presence of competing risks. Attention will be focused on various hypothesis testing methods to directly compare cumulative incidence functions. We will present simulation results to estimate type I error rates and power under different scenarios. Recommendations regarding which hypothesis tests are most appropriate and perform best in the competing risks setting will be provided.

2. Joseph Beyene
   University of Toronto
   joseph@utstat.toronto.edu
   Statistical approaches for high throughput data integration
   Data integration and synthesis is becoming increasingly important. Technologies constantly change leaving behind a trail of data with different forms, shapes and sizes. Statistical and computational methodologies are therefore critical for extracting the most out of these related but not identical sources of data. We extended traditional meta-analytic effect size models to combine information from different gene expression datasets using quality scores. We illustrate and assess our approach using real as well as simulated data sets.

3. Ming Dai
   University of North Carolina at Charlotte
   mdai@email.uncc.edu
   Smoothing Spline Models with Stationary Time Series Errors
   We consider smoothing splines with stationary time series errors. We study the penalized weighted least squared estimate for the regression function, derive the generalized cross validation (GCV) for selecting the smoothing parameter, and show that the estimate is consistent. For short-range dependent errors, the average of the mean squared errors(AMSE) of the smoothing spline estimate has the same convergence rate as with uncorrelated errors; but with long-range dependent errors, it achieves higher convergence rate than the conventional kernel estimate. And for short-range dependent errors, we propose a nonparametric estimation procedure for the covariance structure. We show that the estimates for both the regression function and the covariances are consistent. The method is illustrated using a numerical example and a real data example. This is joint work with Wensheng Guo.

4. Giles Hooker
   McGill University
   giles.hoooker@mcgill.ca
   Inference in Dynamical Systems
   Differential Equations are an increasingly popular and powerful tool in modeling real world phenomena. They have been used extensively in models of population dynamics, neural processes and in chemical engineering. Increasingly, they are applied to noisy data and in situations where parameters are not known or are poorly estimated.

   I present a new approach to the problem of parameter estimation for systems that are described by non-linear differential equations. The methodology is based on using a spline approximation using the differential operator from the equation as a smoothness penalty. Parameters within the penalty are then chosen to optimize the agreement between the resulting spline and the data. I will discuss some properties of the technique and diagnostic tools for analyzing model miss-specification.

5. Rima Izem
Harvard University
izem@stat.harvard.edu
Analyzing Nonlinear Modes of Variation in Functional Data, New Look at Genetic Trade-offs
Scientists in an increasing number of fields, including biology, medicine, and chemistry, collect samples of curves or images of common shape. Although these data are discrete, the processes generating them are continuous. Analyzing variation in these samples, with the aim of making inferences about the general population from which the sample is drawn, is often the main statistical interest. Functional Data Analysis (FDA) methods use the underlying continuity of the data to analyze the variation. However, usual FDA methods, such as Principal Components Analysis, are only effective in analyzing linear variations, and do not always produce interpretable results. In this talk, a general model for curves or images of common shape is considered. We present a new method for analyzing variation under this model. Our method achieves two important goals. The first goal is to decompose the variation in the data into predetermined and interpretable directions of interest, and these could be linear or non-linear. The second goal is to quantify each direction by a newly defined ratio of sums of squares, to allow for a comparison of the contributions to the total variation. The new ratio of sums of squares quantifies a non-linear direction by taking into account the curvature of the space of variation. We discuss, in the general case, consistency of our estimates of variation, using mathematical tools from differential geometry and shape statistics. We successfully applied our method to two different examples of reaction norm curves in Biology. Our analysis shows that non-linear components are dominant. Moreover, our decomposition allows biologists to compare the prevalence of different genetic tradeoffs in a population and to quantify the effect of selection on evolution.

6. Shane T. Jensen
University of Pennsylvania
stjensen@wharton.upenn.edu
Hierarchical Bayesian Models for Combining Heterogenous Biological Data
The computational approaches that are used to identify clusters of co-regulated genes have traditionally used information either from expression data, sequence features or genome-wide location analysis of DNA-binding regulators. Although those approaches have been proven useful, their power is inherently limited by the fact that each data resource provides only partial information: expression data provides only functional or indirect evidence, whereas binding data or sequence features only provide physical location information. Recent efforts on integrating these data types have drawbacks, such as arbitrary parameter cutoffs or little systematic modeling.

We discuss a set of Bayesian hierarchical models that integrate heterogeneous information including expression data, ChIP binding data and sequence features in a principled and intuitive fashion. We have applied our model to data from 500 experiments, and we present several general validation analyses which indicate that our predicted clusters of co-regulated genes are biologically relevant. Our general approach of Bayesian modeling for integrating heterogeneous biological data to discover regulatory networks will be applied in other organisms and should be useful in additional problems in molecular biology. This is joint work with Guang Chen in the Department of Bioengineering and Christian J. Stoeckert in the Department of Genetics at the University of Pennsylvania.

7. Baha-Eldin Khaledi
Razi University
bkhaledi@hotmail.com
Stochastic Comparisons of of Order Statistics from Heterogeneous Random Variables
Let $\{x_{(1)} \leq \ldots \leq x_{(n)}\}$ denote the increasing arrangement of the components of a vector $\mathbf{x} = (x_1, \ldots, x_n)$. A vector $\mathbf{x}$ is said to majorize another vector $\mathbf{y}$ (written $\mathbf{x} \overset{m}{\succeq} \mathbf{y}$) if $\sum_{i=1}^{j} x_{(i)} \leq \sum_{i=1}^{j} y_{(i)}$ for $j = 1, \ldots, n-1$ and $\sum_{i=1}^{n} x_{(i)} = \sum_{i=1}^{n} y_{(i)}$. A vector $\mathbf{x}$ in $\mathbb{R}^{+n}$ is said to be *p-larger* than another vector $\mathbf{y}$ also in $\mathbb{R}^{+n}$ (written $\mathbf{x} \overset{p}{\succeq} \mathbf{y}$) if $\prod_{i=1}^{j} x_{(i)} \leq \prod_{i=1}^{j} y_{(i)}, j = 1, \ldots, n$. This paper is a survey of recent results on the stochastic properties of order statistics associated with independent random variables $X_1, \ldots, X_n$ when X_i, $i = 1, ..., n$, belongs to particular scale family of the form $F(\lambda_i x)$. It is of interest to investigate the effect on the survival function, the hazard rate function and other characteristics of the time to failure of a system consisting of such components when we switch the vector $(\lambda_1, ..., \lambda_n)$ to another vector say $(\lambda_1^*, ..., \lambda_n^*)$ according to majorization as well as p-larger order.

8. Erning Li
   Texas A & M University
   eli@stat.tamu.edu
   Likelihood and Pseudo-likelihood Methods for Semiparametric Joint Models for a Primary Endpoint and Longitudinal Data
   Inference on the association between a primary endpoint and features of longitudinal profiles of a continuous response is of central interest in medical and public health research. Joint models that represent the association through shared dependence of the primary and longitudinal data on random effects are increasingly popular; however, existing inferential methods may be inefficient or sensitive to assumptions on the random effects distribution. We consider a semiparametric joint model that makes only mild assumptions on this distribution and develop likelihood-based inference on the association and distribution, which offers improved performance relative to existing methods that are insensitive to the true random effects distribution. Moreover, the estimated distribution can reveal interesting population features, as we demonstrate for a study of the association between longitudinal hormone levels and bone status in peri-menopausal women.

9. Mengling Liu
   New York University School of Medicine
   mengling.liu@med.nyu.edu
   Mapping Quantitative Trait Loci with Time-to-Event Data from a Population of Mixed Susceptibility
   When quantitative trait loci (QTL) with a time-to-event phenotype are mapped, the latent population heterogeneous susceptibility produces new challenges. The distribution of the time-to-event phenotype usually has a spike because of the existence of a nonsusceptible sub-population. If we simply ignore the heterogeneous susceptibility or inappropriately handle the nonsusceptible subjects, we may be unable to detect the true genetic effects and may also find spurious significance at locations with low genotypic information. In this article, we propose a parametric mixture cure model for interval mapping of the QTL which can characterize the genetic effects on the susceptibility and/or the distribution of event times for susceptible subjects. A likelihood ratio based testing procedure is proposed with the genome-wide significance level obtained by a resampling method. The performance of proposed method and the importance of considering the heterogeneous susceptibility are demonstrated by simulation studies and an application to a survival data from an experiment on mice infected with Listeria monocytogenes.

10. Yufeng Liu
    University of North Carolina
    yfliu@email.unc.edu
    Optimizing Psi-Learning via Mixed Integer Programming
    Classification is one of the most useful statistical tools. Among many different classification methods, margin-based techniques have become very popular recently and are generally expected to yield good performance. As a new margin-based classifier, psi-learning shows great potentials with high accuracy. However, the optimization of psi-learning involves non-convex minimization which is very challenging to implement. In this research, we convert the optimization of psi-learning into a mixed integer programming (MIP) problem. This enables us to utilize the state-of-art algorithm of MIP in the field of operations research to solve psi-learning. Moreover, the new proposed algorithm provides a connection between the support vector machine (SVM) and psi-learning. We also exam the variable selection property of 1-norm psi-learning and compare it with SVM.

11. Shuangge Ma
    University of Washington
    shuangge@u.washington.edu
    Additive Risk Models for Survival Data with High Dimensional Covariates
    As a useful alternative to Cox's proportional hazard model, the additive risk model assumes that the hazard function is the sum of the baseline hazard function and the regression function of covariates. This study is concerned with estimation and prediction for the additive risk models with right censored survival data, especially when the dimension of the covariates is comparable to or larger than the sample size. Principal component regression is proposed to give unique and numerically stable estimators. Asymptotic properties of the proposed estimators, component selection based on the weighted bootstrap, and model evaluation techniques are discussed.This approach is illustrated with analysis of the PBC clinical data and the DLBCL genomic data. It is shown that this methodology is numerically stable and effective in dimension reduction, while still being able to provide satisfactory prediction and classification results.

12. Anandamayee Majumdar
Arizona State University
ananda@math.la.asu.edu
Hierarchical Spatial Modeling of Multiple Soil Nutrients and Carbon in Heterogenous Land-Use Patches of the Phoenix Metropolitan Area
Modeling the multivariate spatial distribution of soil carbon and nutrients has been a challenge for ecosystem ecologists. There is a need for explanatory models, which give insight into the process of socio-economic and biophysical controls on soil spatial variability within and among land-use types. We propose a hierarchical Bayesian modeling specification, an approach that takes into account the spatial covariates as well as the inter-dependent nature of the different multiple soil nutrient and carbon pools. We develop the model to explain variability in soil nutrient and carbon pools in the for the Central Arizona Phoenix Metropolitan region where land-use has changed considerably over the years due to socio-economic among other factors in the region. Comparison of how these land-use changes affect the soil nutrients provides insight as to how socio-economics influence changes in ecology. Our model included 13 geomorphic, ecologic, and socio-economic independent variables that were used to predict soil total Nitrogen, organic Carbon, inorganic Carbon, and extractable Phosphorous. Using five levels of hierarchy we fit a suitable spatial hierarchical model. Using a Bayesian imputation strategy we generate appropriate covariate values used for predictions at new locations where some of the covariate information is unavailable. We compare prediction results from standard models and show that our model is richer and so is the interpretation. To the best of our knowledge this is the first work that applies hierarchical Bayesian modeling techniques and imputation strategies to study multivariate soil nutrient and carbon concentrations. We conclude a discussion of our findings and indication of the broader ecological applicability of our modeling style.

13. Jie Peng
University of California Davis
jie@wald.ucdavis.edu
Genome Scans With Gene-Covariate Interaction
Genetic control of a complex trait is widely thought to involve a number of loci, which may interact with one another and/or with environmental covariates. Standard genome scanning methods usually ignore these possibilities, presumably because they involve larger, more complex models and/or because of difficulties in formulating a suitable model. In this paper, genetic models for gene-covariate interaction are described. Methods of linkage analysis that utilize special features of these models and the corresponding score statistics are derived. Their power is compared with that of simple genome scans that ignore these features, and substantial gains in power are observed when the gene-covariate interaction is strong. Quantitative trait mapping and affected sibpair mapping are discussed. For the latter case, a simpler statistic is proposed that has similar performance to the score statistic, but does not require the estimation of nuisance parameters. Since the nuisance parameters are not estimable solely from affected sibpair data, this statistic is much easier to apply in practice. Similarities with linkage analysis of models for longitudinal data and multivariate phenotypes are also briefly discussed. Approximations for the genome-wide p-value and power are derived under the framework of local alternatives.

14. Qin Shao
The University of Toledo
qshao@UTNet.UToledo.Edu
Mixture Periodic Autoregressive Time Series Models
Periodic autoregressive (PAR) models have been widely used to model periodic time series. However, the major drawback of PAR models is assuming that the distributions are normal; therefore, PAR models cannot be applied to fit periodic time series of which density functions exhibit several departures from Gaussian distributions. Mixture periodic autoregressive models are introduced to fit periodic time series with asymmetric or multimodal distributions. The stationary conditions of such series are derived, the asymptotic property of maximum likelihood estimators is obtained, and the application of EM algorithm is discussed. The new model class is illustrated by analyzing the particulate matter concentrations in Cleveland, OH.

15. Samiran Sinha
Texas A&M University
sinha@stat.tamu.edu
Bayesian Regression Splines for Measurement Error in Matched Case-Control Studies
We propose a Bayesian method for estimating parameters of a prospective logistic regression model for matched case-

control studies when one or more covariates are subject to measurement error. Typically primary dataset consists of a number of matched sets and each matched set contains one case (diseased subject) and number of controls(non-diseased subject). Corresponding to each subject, a fallible surrogate W of the true exposure variable X, and some other covariates which are assumed to be measured without error, are recorded. A substudy contains information on the both X and W for some subjects. To handle measurement error, the unknown regression function of W on X is modeled with a penalized regression spline, which is a piecewise polynomial function with a penalty term to avoid overfitting. We adopt a fully Bayesian approach and estimate the model parameters through a joint likelihood of the primary study and its substudy. The Bayesian technique seems to be very powerful in handling measurement error and in selecting the smoothing parameter adaptively. The simulation study shows that the method efficiently detects nonlinearity present in the regression of W on X, and estimates the disease-exposure association parameters with higher precision than the method which assumes a linear regression of W on X. Finally, we analyze a real data on matched case-control study of colon cancer. This is joint work with Bani K. Mallick, and Raymond J. Carroll.

16. Liping Tong
University of Washington
tong@stat.washington.edu
Multilocus Lod Scores in Large Pedigrees: A New Approach to Combine Exact and Approximate Calculations
To detect the positions of the disease loci, a bunch of LOD scores needs to be calculated within a (several) pedigree(s) for a given set of markers. The exact LOD score calculations are often impossible when the size of the pedigree and the number of markers are both large. In this case, Markov Chain Monte Carlo (MCMC) approach is able to provide an approximation. However, the mixing performance, within reasonable amount of time, is always a key issue in these MCMC methods. We propose a new approach, which divides a large pedigree into several parts by conditioning on parental haplotypes. We perform exact calculation for the offspring parts where more data are often available, and combine this information to sample the hidden variables for the parental parts. We also improve the parental sampling part using a mixture of several conditional Hidden Markov Chains across loci or meiosis. Our approach is not only more efficient for large pedigree(s) with large number of markers, but also very useful for a looped pedigree, in which case most current methods can not give satisfactory results. This is a joint work with Elizabeth Thompson.

17. Chi-hong Tseng
New York University School of Medicine
ch.tseng@med.nyu.edu
Nonparametric estimation of a survival function with two-stage design studies
The two stage design is popular in epidemiology studies and clinical trials due to its cost effectiveness. Typically the first stage sample contains cheaper and possibly biased information, while the second stage validation sample consists of a subset of the subjects from the first stage with accurate and complete information. In this paper we study estimation of a survival function with right censored survival data from a two stage design. A nonparametric estimator is derived by combining data from both stages. We also study its large sample properties and derive pointwise and simultaneous confidence intervals for the survival function. The proposed estimator effectively corrects the potential bias in the first stage sample. It also reduces the variance of the Kaplan-Meier estimator solely based on the second stage validation sample. In addition, it has a better tail behavior than the second stage Kaplan-Meier estimate as observed in our simulation. A simulation study was conducted to study the small sample performance of the developed methods. We finally illustrate our methods on a real data from a medical device postmarket surveillance study.

18. Amanda Wang
University of Virginia
xw5a@cms.mail.virginia.edu
A scale-based approach to finding effective dimensionality in manifold learning
The discovering of low dimensional manifolds in high dimensional data is the main goal of manifold learning. We propose a new approach to identify the effective dimension (intrinsic dimension) of low-dimensional manifolds that summarize underlying structure in the data. The scale space viewpoint is the key to our approach enabling us to meet the challenge of noisy data. Our approach finds the effective dimensionality of the data over all scale without any prior knowledge. We also show that our approach has better performance compared with other methods especially in the presence of relatively large noise.

19. Haiyan Wang
Kansas State University
hwang@stat.ksu.edu
Hypothesis Testing for Heteroscedastic Functional Data
Models for analyzing data involving repeated measurements within a subject or stratum include ordinary and generalized linear and nonlinear mixed-effects models, and the fully nonparametric marginal model. These approaches are mainly suitable when the number of within stratum measurements is relatively small. Time series models, smoothing spline models, varying coefficient models, can also be used for functional or curve data where the number of within stratum measurements is large. However, these impose modeling assumptions that restrict full generality. Here we consider the fully nonparametric marginal model with unspecified covariance structure in the context of functional data, and present procedures for evaluating the effect of several crossed factors on the curve, as well as their interactions with time. The asymptotics, which rely on the large number of measurements per curve and not on large group sizes, hold under the general assumption of α-mixing and do not require the measurements to be continuous or homoscedastic. However, such asymptotics require strong moment conditions. A competing set of rank procedures is developed which require no moment conditions. Simulation results show that the (mid-)rank procedures provide both robustness and increased power away from the normal distribution. Two real data sets are analyzed. This is a joint work with Michael Akritas.

20. Pei Wang
Fred Hutchinson Cancer Research Center
pwang@fhcrc.org
Jointly learning from CGH and expression microarrays
DNA copy number alterations are key genetic events in the development and progression of human cancers. Array-based Comparative Genomic Hybridization (aCGH) is a new technique to map genome-wide alterations in DNA copy number at sub-megabase resolutions. Expression profiling and aCGH can be performed on the same microarray platform, which provides paired measurements of RNA expression level and DNA copy number of each gene/clone in a tumor sample.

The genome instability creates somatic genomic aberrations, but only a portion of these events (oncogenes) are involved in tumor progression. On the other hand, RNA expression levels are expected to change not only in oncogenes, but also in the down-stream genes regulated by oncogenes. Therefore, in order to understand the tumor progression as well as to correctly identify the oncogenes, it is advantageous to jointly learn from these two types of data, which are complementing each other. I will discuss some statistics methods addressing this problem.