2004 WNAR/IMS Abstracts

Monday, June 28th 8:30-10:15

I1 Invited Paper Session
Functional Data Analysis

Name: Hans-Georg Muller
Affiliation: University of California, Davis
Email: mueller@wald.ucdavis.edu
Title: *Functional Response Models*
Abstract: We review functional regression models and discuss in more detail the situation where the predictor is a vector or scalar such as a dose and the response is a random trajectory. Diagnostics for models that incorporate the influence of the predictor through the mean response function may be based on a "residual process". The residual process plays a role analogous to the classical residuals, and in well-fitting models does not depend on the covariates. In a case study, we analyze dose-response data with functional responses from an experiment on the age-specific reproduction of medflies. Daily egg-laying was recorded for a sample of 874 medflies in response to dietary dose provided to the flies. We study a product model for the mean response function and demonstrate how conditional residual processes can be incorporated to address possible lack-of-fit of the mean model. A useful criterion to evaluate functional response models is their ability to predict the functional response at a new dose. We quantify this notion by means of a conditional prediction error that is obtained through a leave-one-dose-out technique and is numerically efficient. This talk is based on joint work with Jeng-Min Chiou and Jane-Ling Wang.

Name: Naisyin Wang
Affiliation: Texas A&M University
Email: nwang@stat.tamu.edu
Title: *Accounting for Response Correlation in Varying-Coefficient Models*
Abstract: We consider a nonparametric varying coefficient model in a longitudinal data setting. An estimating procedure which accounts for the dependency of within subject responses is presented. The major gain is in the reduction of estimation variation. Asymptotic properties of this estimator are provided. The numerical efficacy of the proposed methodology is demonstrated through a small simulation study and the analysis of an Ophthalmology example.
Coauthor: Jeng-Min Chiou, National Health Research Institute, Taiwan

Name: Gareth James
Affiliation: University of Southern California
Email: gareth@usc.edu
Title: *Curve Synchronization with Applications to Functional Regression Models*
Abstract: Recently a number of standard statistical techniques, such as principal components analysis, clustering, linear discriminant analysis and generalized linear models, have been extended to functional data. However, functions or curves pose an additional problem over more standard data, namely curve alignment. It is often the case that variability is not only caused by differences in amplitude but also by warping of the time scale causing two curves to be out of alignment. Hence one can not meaningfully compare curves at the same point in time. Potentially both the warping function X and the amplitude function Z may contain important information yet the above mentioned approaches generally fail to incorporate the two separate components. In this talk I suggest an automated approach for aligning a set of curves or functions. This method involves defining the first K "moments" for any given curve Y. For example the first moment is a measure of the average value of Y in time while the second central moment measures the spread of Y. The observed curves are then

realigned so that the first moment is the same for every curve. This procedure is repeated for each of the first K moments subject to minimizing the variability in the warping functions X. I illustrate this approach on several data sets and end by showing how the warping and amplitude functions can be combined to produce superior results in clustering, classification and regression settings.

C1. Contributed Paper Session
Issues in Binary and Survival Data Analysis

Name: Stephen Shiboski
Affiliation: University of California, San Francisco
Email: steve@biostat.ucsf.edu
Title: *Issues in analysis of event times arising in case-control samples*
Abstract: We investigate issues in the use of data from case-control samples of rare disease outcomes to estimate properties of the distribution of an underlying event time. Because of the retrospective nature of these samples, all observed events are of the current status variety (i.e. are either left or right-censored). In such cases, interest may focus on models other than the conventional logistic regression model. Using simulations and analytic techniques, we investigate bias in estimates from alternative models arising from ignoring the sampling procedure. We also present applications of recently developed approaches to fitting generalized linear regression models to data from response-based samples (Scott and Wild, 1997), and discuss semiparametric generalizations. Finally, we provide sensitivity analysis approaches for situations where information on population-level information on cases and controls is unavailable. Results are illustrated using data from studies of HIV transmission.
Additional authors:
John Neuhaus, University of California, San Francisco
Eric Vittinghoff, University of California, San Francisco

Name: Susanne May
Affiliation: University of California, San Diego
Email: smay@ucsd.edu
Title: *Hosmer-Lemeshow type goodness-of-fit tests for the Cox proportional hazards model*
Abstract: The Cox proportional hazards (PH) model has been an extremely popular regression model in the analysis of survival data during the last decades. Even though a number of goodness-of-fit tests have been developed for the PH model, authors who utilize this model rarely compute these tests. One reason might be that only a few can be easily calculated in statistical software packages. We discuss goodness-of-fit tests for the Cox proportional hazards model, which are based on ideas similar to the Hosmer-Lemeshow goodness-of-fit test for logistic regression. All of these tests can be derived by adding group indicator variables to the model and testing the hypothesis that the coefficients of the group indicator variables are zero via the score test. We will call the tests that can be derived in this way the added variable tests. Care needs to be taken when implementing these tests since some of them require the use of time-dependent group indicator variables.
Authors: Susanne May and David W. Hosmer

Name: Reena Deutsch
Affiliation: University of California, San Diego
Email: rdeutsch@ucsd.edu
Title: *Testing significance of a cluster of multivariate items with binary outcomes: The tripartite T index vs. three common similarity measures*
Abstract: Numerous similarity measures have been developed to quantify resemblance between pairs of items containing a pattern of dichotomous responses. Results from three uncomplicated and widely used similarity measures, the Dice, Jaccard, and Simple Matching coefficients, are compared with the more complex tripartite T similarity index, recently proposed by Tulloss for use in mycology research. The measures are utilized for testing if items within a specified cluster are significantly more similar to each other than they are to items outside the cluster. A simple permutation test is applied using a measure of distinctness. Simulated data

generated for a range of prevalence levels and resemblance, plus data obtained from a neurocognitive testing battery, illustrate the methods. The tripartite T measure is comparable to the other indices for some scenarios, and essentially equivalent to the Dice coefficient overall when compared on the same data. Some limitations of the Tulloss algorithm were detected with the clinical data. Thus, the tripartite T index is not recommended to routinely replace the more traditional methods in all applications.
Co-author: Mariana Cherner, University of California, San Diego

Name: Holly Janes, Margaret Pepe
Affiliation: University of Washington
Email: hjanes@u.washington.edu
Title: *Adjusting for Covariate Effects in a Biomarker Study Using the Subject-Specific Threshold ROC Curve*
Abstract: In diagnostic and screening studies, it is often necessary to account for covariates which are associated with the biomarker of interest. For example, age is strongly associated with prostate-specific antigen (PSA), a biomarker for prostate cancer, and the discriminatory accuracy of PSA may also vary with age. We propose the subject-specific threshold ROC (SST-ROC) as a covariate-adjusted measure of the diagnostic accuracy of the biomarker. The SST-ROC is the ROC curve for a rule which uses covariate-specific thresholds to define "test-positive". It can also be interpreted as a weighted average of the covariate-specific sensitivities, holding the covariate-specific specificities constant. We motivate consideration of the SST-ROC, propose non parametric and semi-parametric estimators of it, and summarize asymptotic distribution theory along with its implications for efficient study design.

Name: Coen Bernaards
Affiliation: AMC Cancer Research Center
Email: bernaardsc@amc.org
Title: *Robustness of a multivariate normal approximation for imputation of incomplete binary data*
Abstract: Multiple imputation has become easier to perform with the advent of several software packages that provide imputations under a multivariate normal model, but imputation of missing binary data remains an important practical problem. Here, we explore three alternative methods for converting a multivariate normal imputed value into a binary imputed value: (1) simple rounding of the imputed value to the nearer of 0 or 1, (2) a Bernoulli draw based on a coin flip where an imputed value between 0 and 1 is treated as the probability of drawing a 1, and (3) an adaptive rounding scheme where the cutoff value for determining whether to round to 0 or 1 is based on a normal approximation to the binomial distribution. We perform simulation studies on a data set of 206,802 respondents to the California Healthy Kids Survey, where complete cases are viewed as a finite population and incomplete cases provide a basis for imposing missing-data patterns. Frequently, we found satisfactory bias and coverage properties, suggesting that approaches such as these that are based on statistical approximations are preferable in applied research than either avoiding settings where missing data occur or relying on complete-case analyses.
Co-authors: Thomas R. Belin, University of California, Los Angeles; Joseph L. Schafer, The Pennsylvania State University

Monday, June 28th 10:30-12:15

I2 Invited Paper Session
Interval Censored Data: Theory and Methods.

Name: Nicholas Jewell
Affiliation: University of California, Berkeley
Email: jewell@stat.berkeley.edu
Title: *Nonparametric Estimation from Bivariate and Case-Control Current Status Data*

Abstract: Researchers working with survival data are by now adept at handling issues associated with incomplete data, particular those associated with various forms of censoring. An extreme form of interval censoring, known as current status observation, refers to situations where the only available information on a survival random variable T is whether or not T exceeds a random independent monitoring time C. This talk will introduce some recent developments and applications of current status to allow for (i) response-based sampling, and (ii) observation of bivariate survival times. Our remarks will be focused on nonparametric techniques. Some examples where these methods have been applied will be used for illustration.

Name: Moulinath Banerjee
Affiliation: University of Michigan
Email: moulib@umich.edu
Title: *Likelihood based inference for interval censored data: some recent developments and future directions*
Abstract: This talk will feature some recent developments in likelihood based nonparametric and semiparametric inference for interval censored (with emphasis on current status) data. I will talk about the universality of the limit distribution of the likelihood ratio statistic for testing the value of the survival distribution at one or multiple points in the simple Current Status Model (without covariates) and demonstrate the superiority of the likelihood ratio based method of constructing confidence sets for the survival distribution to other methods. I will also present analogous results for likelihood ratio inference in the Cox Model with Current Status Data, where the problem is to estimate the baseline hazard function. Natural connections between Current Status Data and Binary Regression Models under certain montonicity assumptions and how these may be exploited to study the latter will also be illustrated. If time permits, I will discuss some tentative results along similar lines for general (mixed--case) interval censoring models and panel count data.

Name: Shuangge Ma
Affiliation: University of Wisconsin-Madison
Email: shuangge@stat.wisc.edu
Title: *Penalized Log-Likelihood Estimation for Partly Linear Transformation Models with Current Status Data*
Abstract: Current status data arises when a continuous response is reduced to an indicator of whether the response is greater or less than a random threshold value. We consider partly linear transformation models applied to current status data. Partly linear transformation models are flexible semiparametric regression models where a continuous outcome U, conditional on covariates Z and W, is modeled as H(U)=beta*Z+h(W)+e. Here H is an unknown non-decreasing transformation, h is a an unknown smooth function, and e has a known distribution F with support R. It is shown that the penalized MLE for the regression parameter beta is asymptotically normal and efficient and converges at the parametric rate, although the penalized MLE for the transformation function H and nonparametric regression effect h are only n^1/3 consistent. Inference for the regression parameter based on the weighted bootstrap is investigated. We also study computational issues and demonstrate the proposed methodology with a simulation study, which shows satisfactory numerical results. The transformation models and partly linear regression terms, coupled with new estimation and inference techniques,provide flexible alternatives to the Cox model for current status data analysis.
Co-author: Michael R. Kosorok, University of Wisconsin-Madison

W1 Invited Paper Session
Mortality Displacement: Inference from Air Pollution Time Series

Name: Steven Roberts
Affiliation: Australian National University
Email: steven.roberts@anu.edu.au
Title: *Properties of Mortality Displacement Estimates in the Context of Frail Population Models*
Abstract: Numerous time series studies have investigated the association between daily mortality and daily ambient particulate air pollution concentrations. The consensus from these studies is that increases in ambient particulate air pollution concentrations are associated with increases in daily mortality. However, it may be

that increases in the particulate air pollution concentration only hasten the deaths of individuals in a small frail subset of the population who would have died shortly in the absence of particulate air pollution. This hypothesis has been termed mortality displacement or harvesting. Distributed lag coefficients have been proposed as a method to explore mortality displacement. We use simulation studies to investigate properties of distributed lag coefficients in frail population models. The properties of distributed lag models are compared to explicit estimation of frail population longevity, together with effect estimates for particulate air pollution.

Name: Leah Welty
Affiliation: Johns Hopkins University
Email: lwelty@jhsph.edu
Title: *Flexible distributed lag models for quantifying the effects of weather and air pollution on daily mortality.*
Abstract: Distributed lag models, time series models that include lags of exposure variables as covariates, are powerful tools for quantifying the cumulative effects of environmental exposures on a health outcome. Distributed lag models have been used extensively in economics since being popularized by Almon in 1965, and have more recently been used for quantifying the acute health effects of weather and air pollution. The degree of mortality displacement for a particular exposure and population may be estimated by comparing distributed lag models that ignore remote lags with richer models that include them. Standard methods for fitting distributed lag models include penalized splines and polynomial distributed lags, but these fail to take full advantage of prior information about the shape of the distributed lag function for environmental exposures. In this talk, we derive alternate distributed lag models of two types that more appropriately inform our understanding of mortality displacement. We consider linear and quadratic distributed lag models, the latter allowing for interactions in exposure levels at different lags. We propose semi-parametric methods, which more fully account for prior information of the lag structure, to estimate these models, and we contrast the results with simpler distributed lag parameterizations. We apply our models to weather, air pollution, and mortality data from 1987-2000 for 100 US cities.
Additional author: Scott Zeger, Johns Hopkins University, Department of Biostatistics

Monday, June 28 1:45-3:30

WI1 Invited Paper Session
Applications of Multivariate Survival Analysis in Genetic Epidemiology

Name: Hongzhe Li
Affiliation: University of California, Davis
Email: hli@ucdavis.edu
Title: *The additive genetics gamma frailty model: estimation and inferences*
Abstract: The additive genetic gamma frailty model has been proposed for multipoint genetic linkage analysis (Li and Zhong, Biostatistics, 2002). In this model, the genetic frailties are constructed based on the inheritance vectors. In order to compute the retrospective likelihood function, one needs to compute the multipoint inheritance distribution, which could be quite computationally intenstive. However, the probabilty distributions of pair-wise IBD sharing are easy to calculate. We show that the induced multivariate survival model from the additive genetic gamma frailty model is a special form of the Copula model, and propose a two stage procedure for estimating the model parameters. We also describe a psudolikelihood ratio test for genetic linkage. Performance of the estimation and testing procedure is evaluated by simulation studies. The issue of ascertainment correction will also be briefly discussed.

Name: Li Hsu
Affiliation: Fred Hutchinson Cancer Research Center
Email: lih@fhcrc.org
Title: *On estimation of marginal hazards functions from the case-control family studies*

Abstract: Estimating marginal hazard function from the correlated failure time data arisingfrom case-control family studies is complicated by non-cohort study design andrisk heterogeneity due to unmeasured, shared risk factors among the family members. Accounting for both factors in this talk, I will present a two-stageestimation procedure. The first stage is to estimate the dependence parameter inthe distribution for the risk heterogeneity without obtaining the marginal distribution first or simultaneously. Assuming that the dependence parameter isknown, the second stage one can estimate the marginal hazard function byiterating between estimation of the risk heterogeneity (frailty) for each family and maximization of the partial likelihood function with an offset to account for the riskheterogeneity. I will also present some practical considerations on themisspecification of frailty distribution. The simulation study as well as an analysis of a case-control family study of early onset breast cancer will be presented.

Name: Dave Glidden
Affiliation: University of California, San Francisco
Email: dave@biostat.ucsf.edu
Title: *General copula-checking strategies: applications to family studies*
Abstract: This paper develops a strategy for checking the bivariate dependence structure of clustered failure time data. The approach is based on calculating residuals thatcompare the observed number of failures to the expected based on an assumeddependence structure. The method uses the cumulative sums of these residuals as the basis for model assessment. Cross-ratio functions form the basis for theapproach, which can be used to evaluate a wide variety of models. In this talk, Idescribe the methods and consider it^Òs application to family disease studies.

Name: Jason Fine
Affiliation: University of Wisconsin, Madison
Email: fine@biostat.wisc.edu
Title: *Functional Association Models for Multivariate Survival Processes*
Abstract: We consider multivariate temporal processes that are continuously observed within overlapping time windows. The intended application is censored multistate and multivariate survival settings, where point processes are continuouly observed. This data differs from other functional data, like longitudinal data, which are discretely observed at irregular times. Functional mean and association regression models are studied for the point processes, with completely unspecified time-varying coefficients. The continuous observation scheme is exploited: the coefficients may be estimated nonparametrically by extending GEE to continuously observed data. The estimators automatically converge at the parametric rate, without smoothing, unlike with discretely observed data. Uniform consistency and weak convergence is established with empirical process techniques. Existing functional approaches to survival processes utilize intensity models, which require smoothing and depend critically on the choice of smoothing parameters, similarly to discretely observed data. The nonparametric estimators yield new tests for covariate effects, parametric sub-modeling of these effects, and goodness-of-fit testing. An analysis of familial aggregation of alcoholism illustrates the methodology's practical utility.

W2 Invited Paper Session
Inference using computer simulation code

Name: Tilmann Gneiting
Affiliation: University of Washington
Email: tilmann@stat.washington.edu
Title: *Calibrated probabilistic weather forecasting using ensemble model output statistics (EMOS) and minimum CRPS estimation*

Abstract: The past 15 years have seen a culture change in the practice of numerical weather prediction. Up to the early 1990s, numerical weather forecasting was an intrinsically deterministic endeavor. National and international weather centers used sophisticated computing resources to run carefully designed numerical weather prediction models. While this is still the case today, the introduction of ensemble prediction systems forms a radical change. Ensemble systems generate an ensemble of initial conditions and run the simulation code forward in time from each of them in turn, thereby generating a sample from the predictive distribution of future weather events. Operational ensemble systems are subject to forecast bias and underdispersion, and therefore uncalibrated. To address these issues and to obtain calibrated predictive distributions, we propose the use of ensemble model output statistics (EMOS), an easy to implement statistical post-processing technique. The EMOS technique is based on multiple linear regression and yields probabilistic forecasts in the form of Gaussian predictive probability density functions. The EMOS predictive mean is an optimal, bias-corrected weighted average of the ensemble member forecasts, with coefficients that are constrained to be nonnegative and associated with the individual member model skill. The EMOS predictive variance is a linear function of the ensemble spread. For estimating the EMOS coefficients, we use minimum CRPS estimation, that is, we find the coefficient values that minimize the continuous ranked probability score (CRPS) for the training data, which can be understood as robust M-estimation. We applied the EMOS method to 48-hour ahead forecasts of sea-level pressure and surface temperature over the North American Pacific Northwest in Spring 2000, using the University of Washington mesoscale ensemble, and with good results.
This is joint work with Anton Westveld, Adrian E. Raftery and Tom Goldman.

Name: Dorin Drignei
Affiliation: National Center for Atmospheric Research
Email: drignei@iastate.edu
Title: *Statistical Analysis of Multivariate Computer Experiments with Large Temporal Dimension*
Abstract: Computer experiments are increasingly used in scientific investigations as substitutes for physical experiments in cases where the later are difficult or impossible to perform. A computer experiment consists of several runs of a computer model for the purpose of better understanding the input-output relationship. The practical difficulty in some situations is that a single computer model run may use a prohibitive amount of computational resources. A recent approach proposes to use statistical models as less expensive surrogates for such computer models; these provide both point predictors and uncertainty characterization of the outputs. In this talk we propose a two-stage statistical method for computer experiments which produce multivariate output on a spatio-temporal grid with large time dimension. A two-gyre ocean model will be used to illustrate the method.

Name: Bruno Sanso
Affiliation: University of California Santa Cruz
Email: bruno@ams.ucsc.edu
Title: *Statistical Inference for a Charged Particle Simulator*
Abstract: A beam of protons is produced by a linear charged particle accelerator, then focused through the use of successive quadrupoles. The initial state of the beam is unknown, in terms of particle position and momentum. Wire scans are used to collect data on the current state of the beam as it passes through and beyond the focusing region, and the goal is to infer the initial state from the wire trace data. This setup is that of a classic inverse problem, in which a computer simulator is used to link an initial state configuration to observable values (wire traces), and then inference is performed for the distribution of the initial state. We model the initial distribution (position and momentum) as two bivariate Gaussian processes, one for each of the x and y directions. The process convolution approach improves the computational efficiency.
Co-authors: Herbert Lee and Weining Zhou, University of California, Santa Cruz, Dave Higdon, Los Alamos National Labs.

Name: Rick Routledge
Affiliation: Simon Fraser University
Email: routledg@stat.sfu.ca

Title: *Hydrodynamic modelling of a British Columbia fjord: linking local climate change and sockeye salmon declines*
Abstract: Pacific salmon have shown signs of severe instability. Many more southerly populations have been listed as endangered or threatened under the United States Endangered Species Act. Three populations have also recently been listed under Canada's new Species at Risk Act. Rivers Inlet sockeye salmon ^Ö as yet unlisted ^Ö have declined from annual returns of around a million adults to a low of 3500 in 1999. Limited indirect evidence suggests that the cause is likely in the early-marine adjustment phase as the juvenile fish migrate down the inlet. Inlet conditions also appear to be strongly influenced by local climate and river discharge levels. The presentation will focus on a hydrodynamic model that is being developed for the inlet in an effort to gain further understanding of this influence.

Tuesday, June 29 8:30-10:15

W3 Invited Paper Session
Genetics: Phylogenetic Inference

Name: Dennis Pearl
Affiliation: Ohio State University
Email: Pearl.1@osu.edu
Title: *Diagnosing and Reporting an MCMC Phylogenetic Analysis*
Abstract: Markov chain Monte Carlo techniques are now heavily used in making phylogenetic inference because of the availability of good software and the benefits of the natural interpretation in the results provided. However, the parameter space for phylogenetic inference is huge and sufficient attention is seldom paid to assessing chain convergence and to methods for summarizing the posterior distribution over such a large space. This talk will address these two core issues using data from 131 amino acid sequences of the ubiquitous protein phosphoglycerate kinase, the major enzyme in energy metabolism.

Name: Li-Jung Liang
Affiliation: University of California, Los Angeles
Email: liangl@ucla.edu
Title: *Constructing Hierarchical Priors with Application to Bayesian Phylogenetic Analyses*
Abstract: We develop a statistical model and computational algorithm for combining analyses of a number of similar datasets. Individual analyses may be fit independently using previously written standalone software, such as MrBayes, that fits a computationally intensive Bayesian phylogeny model to aligned nucleotide sequence data using Markov Chain Monte Carlo (MCMC) simulation. We place a hierarchical regression model across the individual analyses for estimating parameters of interest within and across analyses. We use a Mixture of Dirichlet processes (MDP) prior for the parameters of interest to relax parametric assumptions and to ensure the prior distribution for the parameters of interest is continuous. Constructing a large complex model involving all the original datasets is time consuming and we would need to rewrite the existing standalone software. Instead, our strategy is to use the existing MCMC samples of the individual posteriors. We use an importance re-weighting algorithm within Gibbs to sample values of the individual parameters. We demonstrate our approach on a set of phylogenetic analyses of HIV-1 nucleotide sequence data.
This is joint work with Dr. Robert Weiss.

Name: Steve Poe
Affiliation: University of New Mexico
E-mail:anolis@unm.edu
Title: *Testing for an explosive evolutionary radiation in birds*
Abstract: All recent studies of bird phylogeny have produced poorly resolved relationships among the Orders of Neoaves, the lineage that includes most modern birds. This "bush" result suggests the possibility of an explosive and potentially unresolvable evolutionary radiation. However, such radiations are thought to be rare or nonexistent in nature and difficult to corroborate empirically because lack of phylogenetic resolution can also

be caused by analytical artifacts (1) . Here we examine the predictions of the explosive radiation hypothesis for five independent genetic data sets for Neoaves. We propose a methodology for testing for polytomies of evolutionary lineages, perform likelihood ratio tests to compare trees with zero-length branches to more resolved trees, compare topologies between independent gene trees, and propose a Power test for the SOWH (Swofford et al. 1996) test. The evidence of 1) extremely short (in some cases zero-length) branches for interordinal relationships across independent gene trees and 2) topological incongruence amongbetween gene trees suggests and show that the bird tree includes essentially simultaneous radiation of multiple lineages. This result explains why a robust phylogeny of birds has not been produced despite [many attempts and much data] much effort on the part of avian systematists.
Key words: polytomy, rapid radiation, birds, phylogeny, Neoaves, power test, avian evolution

C2 Contributed Paper Session
Spatiotemporal Data Analysis

Name: Geert Wenes
Affiliation: The National Center for Genome Resources (NCGR)
Email: gcw@ncgr.org
Title: *Process Detection in Large, Noisy, and Incomplete Data: Stochastic Hybrid Dynamical Systems*
Abstract: We develop accurate and robust methods for the detection of dynamical processes in large, noisy, and incomplete data. We formulate the process detection problem as one of identifying the state trajectory generated by a "super system" S based on the signature left by a specific process in various sets of measured data. S is capable of executing a number of dynamical processes simultaneously. The main challenges of our approach are: a) the uncertainty regarding the candidate processes; b) the nature of the available data sets (incomplete and noisy); and c) the existence of many similar environmental processes being executed contemporaneously. Our approach represents S as a stochastic hybrid dynamical system (S-HDS) and then formulates the process detection problem as a state estimation problem; the latter solved via a Hidden Markov Model (HMM) based analysis or through a control-theoretic approach. We illustrate our approach with two counter proliferation examples: a) a scientific research assessment study to detect and assess research activities and trends at a suspected Biological Warfare (BW) facility; and b) a facility function study to identify purpose and capability of suspected BW facilities.

Name: Margaret Short
Affiliation: Los Alamos National Labs
Email: mbshort@lanl.gov
Title: *Process convolution approach to reconstruction of binary fields*
Abstract: We discuss a process convolution approach to estimating binary fields, implemented via Markov chain Monte Carlo. An archeological data set is used to locate regions of human activity. A geological data set from an Italian aquitard is used to determine regions of high and low permeability. This represents on-going work with David Higdon, Daniel Tartakovsky and Alberto Guadagnini.

Name: Vera Bulaevskaya
Affiliation: Carnegie Mellon University
Email: vera@stat.cmu.edu
Title: *A Maximum Likelihood Approach to Assessing the Neuron/Muscle Relationship*
Abstract: Due to various factors, even in the presence of an association between a cortical neuron and a muscle group, a neuron spike leads to an impact in the muscle's electrical activity only a small fraction of the time. The standard method neuroscientists use to assess the relationship between neuron and muscle activity, spike-triggered averaging (STA), does not take this into account, as it simply averages electromyography (EMG) signals obtained from the muscle of interest at time points equal to the neuron's spike times plus the travel time

of the neuron's signal. We propose a mixture model for muscle response to neuron activity and show through a simulation study that maximum likelihood estimators of the relevant model parameters are up to 18 times as efficient as the STA estimator.
Co-authors: Robert Kass, Carnegie Mellon University; Andrew Schwartz, University of Pittsburgh

Name: Rachel Altman
Affiliation: University of Washington
Email: rachel@stat.ubc.ca
Title: *Parameter-Driven Models for Time Series of Count Data*
Abstract: Modelling correlated count data is a challenging problem. Unlike the situation for continuous data, for which the multivariate normal distribution is available, for count data there is no convenient and flexible class of multivariate distributions that can capture the shape of the distribution and the autocorrelation. Furthermore, techniques which have been developed for assessing the fit of models for normally distributed data do not extend readily to the count data setting. In this talk, we propose a general class of parameter-driven (latent variable) models for count data. This class includes the generalized linear mixed model, hierarchical generalized linear model, and the hidden Markov model. We consider the interpretation of these models, and discuss a parameter estimation method which yields estimates of the regression coefficients that are both efficient and robust to misspecification of the latent process. We apply these ideas to the analysis of multiple sclerosis and polio incidence data.
Joint with Brian Leroux, University of Washington.

Name: Christiana Drake
Affilation: University of California
Email: cmdrake@ucdavis.edu.
Title: *Missing Socio-Economic Status in the California Cancer Registry: How to adjust mortality rates*
Abstract: The California Cancer Registry is one of the biggest and probably ethnically most diverse cancer registries in the world. The California Health and Safety Code mandates statewide, population-based cancer reporting. The registry was established to enable the (California) Department of Health Services to "conduct a Program of epidemiological assessments of the incidence of cancer" with a view towards identifying potential risk factors. It publishes annual reports on cancer incidence and mortality in California. The registry calculates age-, race-, gender- adjusted as well as specific rates. The registry also adjusts for socio-economic status based on aggregate data published by the census bureau. The socio-economic status is determined by street address. A case is geocodable if street address is known and not geocodable otherwise. About 6\% of cases are not geocodable. Researchers at the California Cancer Registry have studied non-goecodable cases and believe excluding them from analysis biases cancer rates. The missing data mechanism and the feasibility of imputing missing geocodes with the data collected by the registry are explored as well as sensitivity analysis of the effects of excluding non-geocodable cases from the analysis.

Tuesday, June 29 10:30-12:15

W4 Invited Paper Session
Homeland Security

Name: David Banks
Affiliation: Duke University
Email: banks@stat.duke.edu
Title: *Game Theory and Risk Analysis in Counterterrorism*
Abstract: Game theory is used in adversarial situations, but the classic approach requires unreasonable certainty about payoffs and probabilities. To address this problem, we propose the use of statistical risk

analysis to develop joint distributions for payoffs and to explore the stability of a game theory solution. We also address issues that arise from portfolio analysis considerations. The methodology is illustrated in the context of smallpox threat management.

Name: Michael Stoto
Affiliation: RAND
Email: mstoto@rand.org
Title: *Syndromic Surveillance: Is It Worth the Effort?*
Abstract: Syndromic surveillance systems are intended to give early warnings of bioterrorist attacks. However, even with access to the requisite data and perfect organizational coordination and cooperation, the statistical challenges in detecting an incident are formidable, involving a tradeoff among sensitivity, the false positive rate, and timeliness. We analyzed four statistical detection algorithms using daily counts of patients with influenza-like illness from a hospital emergency department. Under these conditions outside of the flu season, the detection algorithms had a roughly 50 percent chance of detecting a fast outbreak on the second day. The slow outbreak was more difficult to detect; detection algorithms that integrate data from more than one day were roughly equivalent to each other but have only a 50 percent chance at day 9. The benefits of syndromic surveillance also depend on how well these systems are integrated into public health. Even in the best of circumstances syndromic surveillance sets off an alarm that must be investigated before any action can be taken, and deciding what to do can be difficult. Given these results, city and state health departments should be cautious in investing in costly syndromic surveillance systems.

Name: Lawrence Ticknor
Affiliation: Los Alamos National Labs
Email: lot@lanl.gov
Title: *Analysis of Amplified Fragment Length Polymorphisms Genomic Data for Identification of Clinical and Environmental Isolates*
Abstract: With the threat of biological agents being used by terrorists, quickly and accurately identifying biological agents becomes important. Microbiologists hoped that DNA fragment data from a relatively rapid genomic technique called Amplified fragment length polymorphisms (AFLP) might be used as fingerprints for different microorganisms. Before an AFLP database could be useful, several analysis questions needed to be answered. The most pressing problem was being able to computerize the parts of the analysis that were done by hand. A realistic solution to this problem was provided. Once the computerization problem was solved the data analysis portion was no longer expensive and better science that involves replicates and controls could be pursued. Some improvements to the traditional AFLP phylogenetic analyses will also be mentioned. Finally, and most importantly, real world examples of how this AFLP data analysis is being used to characterize pathogens will be discussed.

I3 Invited Paper Session
Mixture Analysis: Theory and Applications

Name: Ramani S. Pilla
Affiliation: Case Western Reserve University
Email: pilla@cwru.edu
Title: *Testing for the Order of Mixture Models via Perturbation Theory*
Abstract: Fitting mixture models and performing statistical inference on the results is an important but a very challenging problem. An age-old and fundamental question is: "how many mixture components"? The asymptotic null distribution of the likelihood ratio test statistic is highly complex and very difficult to simulate from in practice. To date, there is no general theory along with a computationally feasible method to provide a satisfactory answer for a general family of mixture distributions. Recently, Pilla & Loader (2003) proposed the perturbation theory to address this problem. Building on this theory, inferential methods are developed to address the problem of order selection. It is shown that the limiting distribution of our test statistic is equivalent

to the supremum of a Gaussian process over a high-dimensional manifold with boundaries and singularities. A procedure to approximate the quantiles of the test statistic via the boundary crossing probabilities is derived. The general theory developed is applicable to testing for an arbitrary number of components from smooth families of distributions, including multivariate mixtures. Application of the resulting theory will be illustrated through data sets from astronomy and genetics.
[This is joint work with Catherine Loader, Department of Statistics, Case Western Reserve University.]

Name: Jia Li
Affiliation: Penn State University
Email: jiali@stat.psu.edu
Title: *Two-way Poisson Mixture Models for Simultaneous Document Classification and Word Clustering*
Abstract: An approach to simultaneous document classification and word clustering is developed using a two-way mixture model of Poisson distributions. Each document is represented by a vector with each dimension specifying the number of occurrences of a particular word in the document in question. As a collection of documents across several classes usually makes use of a large number of words, the document vectors are of high dimension. On the other hand, the number of distinct words in any single document is usually substantially smaller than the size of the vocabulary, leading to sparse document vectors. A mixture of Poisson distributions is used to model the multivariate distribution of the word counts in the documents within each class. To address the issues of high dimensionality and sparsity, the parameters in the mixture model are regularized by imposing a clustering structure on the set of words. An EM-style algorithm for the two-way mixture model will be derived for parameter estimation with the clustering of words part of the estimation process. The connection of the two-way mixture model with dimension reduction will also be elucidated. Experiments on the newsgroup data have demonstrated promising results.

Name: Guenther Walther
Affiliation: Stanford University
Email: walther@stat.stanford.edu
Title: *Analyzing flow cytometry mixture data*
Abstract: Flow cytometry instruments allow to measure multiple (currently up to 13) independent properties of each of a large number (10^6) of cells. Typically, cells belong to one of a number of subpopulations, so the observations constitute a mixture of those. I will discuss some of the problems arising in the analysis of such data, and possible ways to solve these problems.

C3. Contributed Paper Session
Bayesian Applications

Name: Raquel Prado
Affiliation: AMS, UCSC
Email: raquel@ams.ucsc.edu
Title: *Detecting positive selection in DNA sequences: a hierarchical model-based approach*
Abstract: A model-based Bayesian approach for detecting molecular adaptation in DNA protein coding sequences is presented. The statistical modeling is based on a class of hierarchical generalized linear models. The use of this class of models is motivated by the study of several DNA sequences that are closely related in evolutionary time and for which little or no phylogenetic information is available. In particular, DNA sequences encoding malaria antigens taken in various geographical locations in Asia, Africa and South America are studied. The proposed models allow the incorporation of information that might be relevant in inferring the pattern of substitutions in the sequences, such as information about the geographical location of the sequences or pairwise evolutionary distances. The results obtained using this methodology are compared to those obtained using traditional methods for identifying sites under positive selection.

Name: Paramjit Gill
Affiliation: Okanagan University College

E-mail: pgill@ouc.bc.ca
Title: *Bayesian Analysis for Dyadic Data*
Abstract: I will discuss Bayesian modelling of dyadic data with particular emphasis on applications in social psychology. Dyadic data are collected in studies on personal perception and other types of social relations amongst subjects. Each subject plays the dual role of an "actor" and a "partner". An analysis of dyadic behavior data examines three fundamental sources of variation: actor variance, pertner variance and relationship variance. Dyadic data have traditonally been analyzed using ANOVA and maximum likelihood (ML) methods. Bayesian analysis has certain advantages over the traditional methods. Missing, incomplete and unbalanced data problems are easliy handled and the inference is valid even for studies with few subjects having infrequent interactions. The vehicle for carrying out Bayesian analysis is using Markov chain Monte Caro (MCMC) with the help of freely available software package WinBUGS.

Name: Lurdes Inoue
Affiliation: University of Washington
Email: linoue@u.washington.edu
Title: *Combining Longitudinal Studies of PSA*
Abstract: Prostate-Specific Antigen (PSA) is a biomarker commonly used to screen for prostate cancer. Several studies have examined PSA growth rates prior to prostate cancer diagnosis. However, the resulting estimates are highly variable. In this talk we present a non-linear Bayesian hierarchical model to combine longitudinal data on PSA growth from three different studies. Our model enables novel investigations into patterns of PSA growth that were previously impossible due to sample size limitations. The goals of our analysis are two-fold. First, to characterize growth rates of PSA accounting for differences when combining data from different studies. Second, to investigate the impact of clinical covariates such as advanced disease and unfavorable histology on PSA growth rates.
This is a joint work with Ruth Etzioni, Elizabeth Slate, Christopher Morrell and David Penson.

Name: Tim Hanson
Affiliation: University of New Mexico
E-mail: hanson@math.unm.edu
Title: *Bayesian Semiparametric Survival Analysis with Time Dependent Covariates*
Abstract: Mixtures of Polya tree (MPT) models have been largely ignored relative to Dirichlet process mixture (DPM) models. MPT models can be fit where DPM models are impractical, for example in the presence of arbitrarily censored and/or truncated survival data, and full semiparametric or nonparametric inference is easily obtained. We present two generalizations of the accelerated failure time model and the standard generalization of the proportional hazards model for time dependent covariates, fit using a MPT prior on baseline survival. Full semiparametric inference is obtained using standard MCMC techniques. The models are applied to a data set on the time to cerebral edema in hospitalized children.

Tuesday, June 29 1:45-3:30

I4 Invited Paper Session
Markov Chain Monte Carlo Methodology

Name: Xiao-Li Meng
Affiliation: Harvard University
Email: meng@stat.harvard.edu
Title: *A Mutant Gibbs Sampler: Incompatibility, Instability, and Non-identifiability redeemed*
Abstract: Incompatible Gibbs Samplers, unstable Markov chains andnon-identifiable parameters are "devils" that we typically strive to avoid in statistical computation and in modeling in general. Thus, it was a pleasant

surprise to discover that, when used appropriately, these devils can dramatically improve existing two-step Gibbs sampler for some common statistical models, such as t models, probit regressions, and mixed-effect models, as demonstrated and reviewed in van Dyk and Meng (2001). Rigorous theoretical guarantees, however, are crucial in constructing and deriving algorithms of this type because common intuition and heuristics can be deceptive in the presence of these devils. For example, all of the fast samplers in van Dyk and Meng (2001) are obtained as the limit of a sequence of standard data-augmentation (DA) algorithms (i.e., two-step Gibbs samplers) indexed by a carefully chosen non-identifiable working hyperparameter. Each standard sampler in the sequence shares the same stationary distribution for the joint posterior distribution of the model parameter and the standard missing data, yet the limiting sampler shares only the same marginal stationary distribution for the model parameter. Although the resulting limiting samplers are Gibbs-like in that they alternately sample two conditional distributions, these conditionals are actually not compatible, even though the conditional distribution corresponding to all of the limiting sampler's "ancestors" (i.e., the standard samplers in the sequence) are compatible. In other words, a mutation occurs in the limit, due to the instability of the joint chain when the prior distribution on the non-identifiability working parameter becomes improper. In this article we develop strategies and theory to deal with such mutant Gibbs sampler without restricting to the DA case, namely, two-step Gibbs samplers. We illustrate the benefit of going through such methodological and theoretical exercises by providing a mutant Gibbs sampler that mixes substantially faster than the standard Gibbs sampler for posterior sampling from a common logistic mixed model, but with essentially the same computational complexity. This is a joint work with David van Dyk.

Name: Jim Hobert
Affiliation: University of Florida
Email: jhobert@stat.ufl.edu
Title: *A Mixture Representation of the Stationary Distribution*
Abstract: When a Markov chain satisfies a minorization condition, its stationary distribution can be represented as an infinite mixture. The distributions in the mixture are associated with the hitting times on an accessible atom introduced via the splitting construction of Athreya and Ney (1978) and Nummelin (1984). This mixture representation is closely related to perfect sampling and has potential applications in Markov chain Monte Carlo. (This is joint work with Christian Robert, Universite Paris Dauphine.)

Name: Galin Jones
Affiliation: University of Minnesota
Email: galin@stat.umn.edu
Title: *Output Analysis for Markov Chain Monte Carlo*
Abstract: Markov chain Monte Carlo is a method for producing correlated draws from a complicated target distribution. Features of this distribution are then approximated via ergodic averages. Hence, calculating the Monte Carlo standard errors of these ergodic averages is a critical step in assessing the output of the simulation. Two common methods of calculating a Monte Carlo standard error are batch means and regenerative simulation. Regenerative simulation is theoretically superior to batch means but is harder to program and sometimes experiences extremely long tours. Thus regeneration may not be as appealing as batch means from a practical point of view. In this talk, I will give an overview of both procedures and compare them from both practical and theoretical perspectives. (This is joint work with Brian Caffo of Johns Hopkins and Murali Haran of NISS/Penn State.)

WI2 Invited Paper Session
Social Network Analysis

Name: Steven Thompson
Affiliation: Pennsylvania State University
Email: skt@stat.psu.edu
Title: *Active set adaptive sampling in networks*

Abstract: An adaptive sampling design is one in which the procedure for selecting the sample depends on values of variables of interest observed during the sampling. In network settings, variables of interest include those associated with nodes and those associated with pairs of nodes. Link-tracing designs in networks may depend adaptively on the presence or absence of links, on relationship strengths, and on node values.
In this talk I'll describe adaptive sampling designs in which, at any point in the sampling, the next unit or set of units is with high probability selected from a distribution that depends on values of the variables of interest in an "active set" of units already selected. With some lower probability, the next selection is made from a distribution not dependent on those values. The active set may consist of only the most recently selected unit, or the entire current sample, or a wide range of other possibilities.
Unbiased estimation with such designs is based on a combination of initial and conditional selection probabilities, and these preliminary estimators are improved by taking conditional expectations over sample paths through the graph. Markov chain resampling estimators are used for larger sample sizes. The sampling strategies will be illustrated with examples in directed and undirected graphs as well as graphs arising from spatial settings.

Name: Mark Handcock
Affiliation: University of Washington
Email: handcock@stat.washington.edu
Title: *Assessing Degeneracy in Statistical Models of Social Networks*
Abstract: Statistical exponential family models for social networks recognize the complex dependencies within relational data structures. A major barrier to the application of random graph models to social networks has been the lack of a sound statistical theory to evaluate how closely the models capture structure in the observed graphs. This problem has at least two aspects: the specification of realistic models and assessing the degree to which the graph structure produced by the models matches that of the data. We show how the geometry of the exponential families provide insights into these factors and the related issues of inferential and model degeneracy for commonly used models and algorithms.

Name: James Jones
Affiliation: Stanford University
Email: jhj1@stanford.edu
Title: *Likelihood-Based Model Selection for Sexual Partnership Distributions*
Abstract: Social structure plays a fundamental role in the dynamics of infectious disease models. While this statement applies to any type of infectious disease, the effects of social structure are particularly strong in models for sexually transmitted infections (STIs). For example, heterogeneity in contact rates in a population can raise the effective reproduction number of the pathogen, thereby lowering the epidemic threshold and making control and eradication more difficult. One extreme case of this occurs when partnership heterogeneity is described by a power law with infinite variance. This model can arise when partnerships are formed by preferential attachment. In this case, there is no epidemic threshold and traditional control measures will inevitably fail to eradicate an STI from the population. We specify a series of competing stochastic models for partnership formation and using data on the sexual partnership distributions of three populations (Uganda, USA, Sweden), fit the data and choose the best model using likelihood-based model selection procedures (e.g., AIC). Our results do not, in general support the infinite-variance power law model and suggest that mechanisms other than preferential attachment should be considered for the formation of sexual networks. We suggest a heterogeneous stopping rule as one possibility. The most promising future directions for work in STI epidemic models moves beyond marginal distributions for partnerships and specifically models the network structure at the level of the actor and the partnership.

C4. Contributed Paper Session
Longitudinal and Regression Models

Name: Jonathan Schildcrout
Affiliation: University of Washington

Email: jschildc@u.washington.edu
Title: *Outcome Dependent Sampling with Longitudinal, Binary Response Data*
Abstract: Assume we have longitudinal series of binary response data, but the time-dependent exposure is expensive to ascertain. Such a situation could arise when conducting a panel study where subjects complete daily diary cards to report symptoms, and exposure is measured using blood samples that are collected and processed at the end of the study. In this situation, it may be desirable to retrospectively collect exposure on the subset of clusters that contribute the most information towards estimating the parameter of interest. For example, we might decide to measure exposure only on clusters that exhibit response variation (e.g., exclude subjects who did not experience symptoms during the study period). If subsampling of clusters is done in this manner, we must acknowledge the sampling design to make valid inference. In this talk, we focus on inference from marginally specified regression models. We discuss situations under which outcome dependent sampling may be a reasonable study design, circumstances in which estimates will be highly efficient, and the implications of dependence model misspecification.
Authors: Jonathan S. Schildcrout and Patrick J. Heagerty

Name: Rema Raman
Affiliation: University of California, San Diego
Email: rema@ucsd.edu
Title: *A Mixed-Effects Regression Model for Multi-level Ordinal Data that allows Heterogeneous Variances*
Abstract: Three-level data occur frequently in behavior and medical sciences. For example, in a multi-center trial, subjects within a given site are randomly assigned to treatments and then studied over time. Mixed-effects models have been developed to analyze such three-level data only when the response is binary, not ordinal. These models for binary data also assume that the variances at the second and/or the third level of data are the same. Unfortunately, this assumption does not hold in several situations. A mixed-effects model is described for the analysis of three-level ordinal response data. This model allows for either homogeneous or heterogeneous variances between groups at either higher level of data. A maximum marginal likelihood (MML) solution is described and Gauss-Hermite numerical quadrature is used to integrate over the distribution of random effects. Simulation studies will show that the fit of the heterogeneous model increases as the magnitude of the difference in variation between the groups increase. The features of this model will be illustrated using a real-life data set.
Authors: Rema Raman, PhD and Don Hedeker, PhD

Name: D. Keith Williams
Affiliation: University of Arkansas for Medical Sciences
Email: williamsdavidk@uams.edu
Title: *A Performance Comparison of Nonparametric Versus Generalized Linear Models in Longitudinal Studies*
Abstract: Recently there has been a unification of theory for non-parametric models that can be applied to longitudinal studies. These methods have a few trivial assumptions to be met for their application. Generalized linear models (GLM) are also in common use to model longitudinal data. Associated with these GLM models is a set of assumptions about the data. These assumptions include properties about the form, distribution, and covariance structure of the observations. One question that naturally arises in consideration of non-parametric methods is: In what situations are non-parametric methods more or less powerful than GLM methods, especially when the assumptions of the GLM are violated but GLM models applied regardless? This research focuses on comparing performance characteristics such as power for new non-parametric methods versus appropriately and inappropriately applied GLM models and offers performance information guidelines in the use of non-parametric methods. In simulation studies, it was found that for even moderate sample sizes, the performance of these non-parametric methods is comparable to GLM methods and in some instances outperforms them.

Name: Loki Natarajan
Affiliation: University of California at San Diego
Email: loki@math.ucsd.edu
Title: *Measurement Error Models: an application to dietary data*
Abstract: Self-report dietary assessment methods are known to be biased and subject to measurement error. We present a measurement error model for dietary intake measured by two instruments, repeat 24-hour recalls and a food frequency questionnaire (FFQ). A plasma biomarker is also included in the analysis thereby allowing for correlated error terms between the self-report instruments and on repeats of the same instrument, and hence quantification of both random and systematic error. Graphical diagnostics are developed to assess the goodness of fit. The model is applied to carotenoid data from the Women's Healthy Eating and Living (WHEL) Study. The results indicate that both dietary assessment methods performed inadequately with over 75% of variance attributable to error. In addition, the systematic error component was non-negligible, particularly for the FFQ, where for many carotenoids systematic error accounted for over 40% of measurement error variance. These findings underscore the need for more accurate modes of dietary assessment and imply that regression dilution due to measurement error in self-report instruments may be larger than previously estimated.

Name: Osana Shckerbak
Affiliation: Salford Systems
Email: lisas@salford-systems.com
Title: *An Alternative Methodology to Linear Regression and Neural Networks*
Abstract: One of the most effective predictive tools ever invented is linear regression. Linear Regression, however, has a number of shortcomings, including the inability to accommodate highly non-linear relationships, intolerance for missing values, and sensitivity to outliers. In this presentation, Senior Statistician Mikhail Golovnya will discuss a non-linear, fully automated regression methodology called MARS (Multivariate Adaptive Regression Splines). Mars was initially designed to address the most challenging of predictive modeling problems. Stanford University's Jerome Friedman developed the methodology. He took regression from the statisticians, borrowed splines from the mathematicians and took Binary Recursive Partitioning techniques from his own work on CART decision tree.

Wednesday, June 30 8:30-10:15

W5 Invited Paper Session
Statistical Learning and Data Mining

Name: Paul Gustafson
Affiliation: University of British Columbia
Email: gustaf@stat.ubc.ca
Title: *Bayesian model selection for semi-supervised and unsupervised learning*
Abstract: It is well known that Bayesian methods are good at preventing overfitting. Both the use of shrinkage priors to downweight some variables, and the use of Bayesian model averaging to eliminate some variables, have been well-studied in the literature. Most of this work, however, has been in the context of regression modelling to do supervised learning. In this talk we discuss modelling and computational issues that arise when applying these techniques to semi-supervised or unsupervised learning scenarios. Ideas are illustrated in the context of an object-recognition problem from machine learning. This is joint work with Nando de Freitas and Natalie Thompson.

Name: Dustin Lang
Affiliation: Univeristy of British Columbia
Email: dalang@cs.ubc.ca
Title: *Fast inference in probabilistic graphical models*
Abstract: We present probabilistic graphical models for beat tracking in recorded music. We are interested in determining the smoothed posterior distribution of tempo and phase at each point in time given all the audio

observations. Inference in these models is intractable so we develop two approximation strategies. In the first strategy, we discretize the continuous state consisting of tempo and phase. This enables us to apply standard discrete belief propagation to compute the smoothing distribution of the states given the observations. In the second strategy, we develop a particle smoother. Both of these strategies can be computationally expensive. The cost of the belief propagation scheme is quadratic in the number of discrete levels, while the cost of particle smoothing is quadratic in the number of particles. This quadratic cost arises when computing the messages in belief propagation and the smoothed importance weights in particle smoothing. That is, it arises when computing the interaction between each object (particle or discrete level) at one point in time and each object at the next point in time. Using the fast Gauss transform and efficient data structures, we are able to solve both of these problems in linear time. Our experiments demonstrate a remarkable improvement in the results for the same computational cost. Joint work with Nando de Freitas.

Name: Firas Hamze and Nando de Freitas
Affiliation: University of British Columbia
Email: fhamze@cs.ubc.ca
Title: *From Fields to Trees*
Abstract: We present new MCMC algorithms for computing the posterior distributions and expectations of the unknown variables in undirected graphical models with regular structure. For demonstration purposes, we focus on Markov Random Fields (MRFs). By partitioning the MRFs into non-overlapping trees, it is possible to compute the posterior distribution of a particular tree exactly by conditioning on the remaining tree. These exact solutions allow us to construct efficient blocked and Rao-Blackwellised MCMC algorithms. We show empirically that tree sampling is considerably more efficient than other partitioned sampling schemes and the naive Gibbs sampler, even in cases where loopy belief propagation fails to converge. We prove that tree sampling exhibits lower variance than the naive Gibbs sampler and other naive partitioning schemes using the theoretical measure of maximal correlation. We also construct new information theory tools for comparing different MCMC schemes and show that, under these, tree sampling is more efficient.

C5. Contributed Paper Session
Exploratory Data Analysis/Data Mining

Name: Shenghan Lai
Affiliation: Johns Hopkins Bloomsberg School of Public Health
Email: lisas@salford-systems.com
Title: *Examples in Epidemiology Using Advanced Data Mining Techniques: CART, MARS and TreeNet/MART*
Abstract: Advanced data mining tools can be exceptionally powerful techniques in analyzing massive epidemiologic data. In this presentation, several examples from our epidemiologic research at Johns Hopkins University's Bloomsburg School of Public Health are used to illustrate the usefulness of MARS, CART and TreeNet/MART data mining techniques. The first example uses MARS to explore the association between regional heart function and coronary calcification. This example demonstrates that without MARS analysis, the conventional approach fails to identify the association. The second example uses CART to classify the study participants. This example shows that CART is more powerful than the conventional logistic regression analysis. The third example uses MART to explore the association between vitamin E and the development of myocardial infarction. Again, this example suggests that without MART, the relationship may never be able to be identified.
Co-presenter: Mikhail Golovnya, Salford Systems

Name: Mikhail Golovnya
Affiliation: Salford Systems
Email: lisas@salford-systems.com
Title: *Advanced Data Mining Techniques and how to build and interpret TreeNet/MART and Random Forests models: The Evolution of Data Mining from CART to Ensembles of Trees*

Abstract: Learn how to use Data Mining software recently developed at Stanford University and Berkeley by world-renowned statisticians Leo Breiman and Jerome Friedman. You will learn how to use TreeNet/MART and Random Forests. Both TreeNet/MART and Random Forests attempt to leverage predictive power of traditional CART (Classification and Regression Trees) models by combining a large number of trees together using either bootstrap aggregation or boosting approaches.

Name: Chuan Zhou
Affiliation: University of Washington
Email: czhou@u.washington.edu
Title: *A Bayesian Hierarchical Mixture Model for Curve Clustering*
Abstract: Curve data are commonly encountered in appllications. In recent years there has been an increasing interest in clustering such data, especially in the fields of gene expression analysis and biomedical studies. In this paper, we propose a general Bayesian hierarchical mixture model for clustering curve data. Under this model, in stead of clustering based on the high dimensional observed curve data, we construct the hierarchy in such a way that lower dimensional random effects, which characterize the curves, form the basis for clustering. This model provides a flexible framework that can be tuned to the specific context, and allows information regarding curve forms, measurement errors and other prior knowledge to be incorporated. Under this model, the order of observations within curve is explicitly taken into account, and the number of clusters can be treated as unknown and inferred from the data. Computation is carried out via an implementation of birth-death MCMC algorithm. For fixed number of clusters, this model can be viewed as a model-based partitioning method. A simulation example and a real example on gene expression data are used to illustrate the method.
This is joint work with Jon Wakefield, Department of Biostatistics, University of Washington

C6. Contributed Paper Session
Topics in Mathematical Statistics and Statistical Computing

Name: Sam Efromovich
Affiliation: University of New Mexico
Email: efrom@math.unm.edu
Title: *On Blockwise Wavelet Estimation*
Abstract: A blockwise shrinkage, used as an adaptive procedure for nonparametricwavelet estimation, is a popular method of a data-driven estimation.Exact lower and upper bounds for Mean Integrated Squared Error and Mean Squared Error at a point are obtained via the analysis of specific corner functions. The results bridge known computational and theoretical approaches and shed a new light on familar conjectures and phenomena including the Lepski penalty for adaptive pointwise estimation. Numerical simulations are also presented.

Name: Hari Mukerjee
Affiliation: Wichita State University
Email: mukerjee@math.twsu.edu
Title: *Statistical Inferences for a decreasing mean residual life distribution*
Abstract: In survival analysis and in the analysis of life tables, an important biometric function of interest is the life expectancy at age x, defined by $M(x)=E[X-x|X>x]$, where X is a life distribution. M is called the mean residual life function. Yang (1978) produced an empirical estimator, showed that it is strongly uniformly consistent, and that it converges weakly to a mean-zero Gaussian process after normalization. In many applications it is reasonable to assume that M is decreasing. Kochar, Mukerjee and Samaniego (2000)developed a projection type estimator of M under this order restriction, showed that it is strongly uniformly consistent, and that the weak convergence is to the same Gaussian limit as in the unrestricted case, but under some heavy analytic assumptions on M. In this talk we generalize the weak convergence results with none of these analytic assumptions. We also provide a new test for a decreasing mean residual life distribution against exponentiality and compare it with a test due to Hollander and Proschan (1975).This is joint work with Edgardo Lorenzo.

N**ame:** Claudia Schmegner
Affiliation: DePaul University
Email: cschmegn@condor.depaul.edu
Title: *Principles of Optimal Sequential Planning*
Abstract: Even though sequential analysis was introduced to save on the expected cost plus loss, sampling one observation at a time is most of the times, expensive and impractical. A concept of sequential planning is presented as extension and generalization of ``pure'' sequential procedures. According to it, observations are collected in groups of variable sizes. The article discusses optimality of sequential plans in terms of a suitable risk function that balances an observation cost and a group cost. It is shown that only non-randomized sequential plans based on a sufficient statistic need to be considered in order to achieve optimality. Performance of several classes of plans is evaluated.
Key words and Phrases: cost, risk, sequential plan, sequentially planned probability ratio test

Name: Jeffrey Pontius
Affiliation: Kansas State University
Email: jpont@ksu.edu
Title: *Conditioning Plots Based on Experimental Designs*
Abstract: Conditioning plots (coplots) are statistical graphics that display subsets of data from some variables in panels, where the panels are arranged conditionally on data (or estimates) from other variables. Many experimental designs have conditioning embedded. For example, in a split-unit design, the response values from subunits can be thought of as conditional on the main unit responses. Displaying data or estimates from experimental designs based on the nesting (conditioning) of the design and the data characteristics can lead to effective interpretations of the response variables of interest.

Name: Ahmad Reza Soltani
Affiliation: University of Kuwait
Email: soltani@kuc01.kuniv.edu.kw
Title: *On Threshold of Moving Averages With Prescribed On Target Significant Levels*
Abstract: A sequence of random variables represents successive inputs to a moving average process of finite order which will be off target by the nth input if it exceeds a threshold. By introducing two states Markov chain, we define "on target significant level" and establish a technique for evaluating and estimating the threshold corresponding to a prescribed on target significant level. It is proved that in such circumstances for exponential and normal inputs, the threshold is a linear function in the mean of the input variable, where slops and intercepts are also specified. These linear relationships can be easily applied for estimating the thresholds.

Wednesday June 30 10:30AM-12:15PM

I5 Invited Paper Session
Model Selection

Name: Peter Bickel
Affiliation: University of California, Berkeley
Email: bickel@stat.berkeley.edu
Title: *Cross validation for constructing adaptive minmax procedures*
Abstract: In work parallel to Dudoit and van der Laan(2003) we show how crossvalidation can be used to construct adaptive minmax procedures in a variety of contexts.In particular this approach leads to results at least as strong as those of Abramovich,Benyamini,Donoho, and Johnstone(2000) in the Gaussian white noise model and optimal results for L2 boosting.

Name: Jonathan Taylor
Affiliation: Stanford University

Email: jonathan.taylor@stanford.edu
Title: *Connections between stagewise algorithms and the LASSO*
Abstract: The Lasso (Tibshirani 1996) is a method for regularizing least squares regression via L1 constraints. The LAR (Least angle regression) algorithm of (Efron et al 2003) provides an efficient method for computing the entire sequence of Lasso solutions. In the process, the LAR algorithm also provides a conceptual link between the Lasso and Forward Stagewise regression. The latter strategy is an important component in adaptive regression procedures like boosting, and hence this link helps us understand how boosting works. In this talk we a sequential criterion that is optimized by Forward Stagewise regression: it features a sequential minimimum L1 arc-length penalty. We also characterize problems for which the coefficient curves for Lasso are monotone as a function of the L1 norm; this is the situation where all three procedures (LAR, Lasso, and Forward Stagewise) coincide. This is joint work with Trevor Hastie, Rob Tibshirani, and GuentherWalther.

Name: Sunduz Keles
Affiliation: University of California, Berkeley
Email: sunduz@stat.berkeley.edu
Title: *Asymptotic optimality of cross-validation in density estimator selection and in model selection with right censored outcomes*
Abstract: In this talk, we consider the cross-validation method, which is one of the widely used model selection methods, in two different contexts. Firstly, we focus on the problem of selecting a density estimate among a collection of candidate density estimators. We establish a finite sample result for a general class of likelihood-based cross-validation procedures. This result implies that the cross-validation selector performs asymptotically as well (w.r.t. to the Kullback-Leibler distance to the true density) as a benchmark density selector that depends on the true density. Crucial conditions of our theorem are that the size of the validation sample converges to infinity with sample size, which excludes leave-one-out cross-validation, and that the candidate density estimates are bounded away from zero and infinity. Secondly, we develop a new cross-validation based model selection method to select among predictors of right censored outcomes such as survival times. The proposed method considers the risk of a given predictor based on the training sample as a parameter of the full data distribution in a right censored data model. Then, the doubly robust locally efficient estimation method or an inverse probability of censoring weighting method is used to estimate this conditional risk parameter using the validation sample. We prove that, under general conditions, the proposed cross-validated selector is asymptotically equivalent with an oracle benchmark selector based on the true data generating distribution. The presented method covers model selection with right censored data in prediction (univariate and multivariate) and density/hazard estimation problems. Some of the applications of these methods to genomic data analysis include the prediction of biological and clinical outcomes (possibly censored) using microarray gene expression measures, and the identification of regulatory motifs in DNA sequences. This talk will also present examples of such applications. Joint work with Mark J. van der Laan and Sandrine Dudoit.

C7. Contributed Paper Session
Ecological/Population Dynamics

Name: Christopher Williams
Affiliation: University of Idaho
Email: chrisw@uidaho.edu
Title: *Estimation of Fish Disease Prevalence from Imperfect Diagnostic Tests*
Abstract: A Bayesian model is used for estimation of disease prevalence with imperfect diagnostic tests. A latent variable approach leads to an easy-to-implement Gibbs sampling scheme for sampling from the joint posterior distribution of prevalence and the sensitivity and specificity of the tests. These models have been widely adopted for use with human medical data, and we apply them here for prevalence estimation for diseases in fish from the northwest United States. In this setting where less information is available about sensitivity and specificity of tests, some interesting differences occur in the behavior of the Gibbs sampler and in the conclusions that result from the analyses.

Name: Grace Chiu
Affiliation: University of Washington
Email: grace@stat.washington.edu
Title: *Why SHIPSL?*
Abstract: We developed the stream health index for the Puget Sound Lowland (SHIPSL) based on ideas of the index of biotic integrity (IBI), which is commonly used to gauge the biological conditions, or "health," of freshwater systems. Unlike most existing indices of water quality, SHIPSL is not a regionalized version of the IBI. For example, SHIPSL has a simple metric scoring mechanism that (1) requires minimal subjective and painstaking input for its calibration, and (2) can be made "portable" between any geographical regions. Moreover, SHIPSL's natural, unbounded continuous scale leads to highly desirable statistical and biological properties for the index. In this talk, we discuss the rationale behind SHIPSL and the need for SHIPSL-like indices used in developing environmental policies.

Name: Ling Xu
Affiliation: University of New Mexico
Email: lxu93@math.unm.edu
Title: *Detecting Multimodality in Ecological Data*
Abstract: We discuss Bayesian analysis of mixture models to detect multimodalities in body mass. A weakly informative prior and a slightly modified partially proper prior were used for normal models and both priors yield similar posteriors. Reversible jump methods were also used for Bayesian model determination when the number of components is not fixed. Comparisons with frequentist methods are also discussed.
Keywords: Bayesian methods, Normal mixture model, Gibbs sampling, Bayes factors, Posterior model probability, Reversible jump, Kernel density estimation.
Coauthors: Edward J. Bedrick is Professor, Department of Mathematics and Statistics, University of New Mexico. Timothy Hanson is Assistant Professor, Department of Mathematics and Statistics, University of New Mexico. Carla Restrepo is Assistant Professor, Department of Biology, University of Puerto Rico, San Juan, Puerto Rico.

Name: Wayne W. Chen
Affiliation: Minnesota State University
Email: chenw@mnstate.edu
Title: *Parasite Population Dynamics in Malaria*
Abstract: Cerebral malaria results in 1.5 to 3 million deaths each year. Clinical evaluation of P. falciparum parasite clearance, in response to therapy, requires consideration of parasites circulating in the peripheral blood and those sequestered in the lining of blood vessels.While peripheral blood samples can provide an estimate of the number of circulating young parasites, there is no direct measure of the sequestered parasite population. We evaluated a 2-compartment model of parasite population dynamics (Gravenor, et al., Proc. Natl. Acad. Sci., 1998) for prediction of parasite clearance in children treated for malaria in Bangkok, Thailand. Objectives were to determine the minimum number of blood samples needed and to estimate the population size of the sequestered malarial parasites during therapy. For 28 of 37 patients (76%), for whom the number of circulating parasites had decreased from baseline by 24 hours, excellent goodness-of-fit was found to models of parasite counts taken at four 12-hourly intervals. In contrast, poor fit was found to models of data from the remaining 9 patients (24%) for whom the number of circulating parasites remained stable or had increased from baseline. The estimated ratio of initial sequestered to circulating parasite density was similar in 13 patients treated with artesunate orally for uncomplicated malara (3.3) and in 10 patients with severe malaria treated with arteether intramuscularly (3.3) but decreased in 14 patients with cerebral malaria treated with artesunate intravenously (2.9). The average death rate of circulating parasites was similar in the patients with uncomplicated and severe malaria, but faster in patients with cerebral malaria. We conclude that this mathematical model can be used to estimate the baseline and subsequent number of sequestered parasites with only 4 measurements collected at 12-hourly intervals. The modeling technique is promising for evaluation for the effect of antimalarial drug therapy.

Co-authors: C. E. McLaren, Department of Medicine, Epidemiology Division, University of California, Irvine, CA; G. M. Brittenham, College of Physicians and Surgeons, Columbia University, New York, NY; S. Looareesuwan, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand