

A Conference on Nonparametric Inference and Probability with Applications to Science

Honoring Michael Woodroffe

Ann Arbor, Michigan

September 24–25, 2005

Sponsored by

National Science Foundation

National Security Agency

University of Michigan

Department of Statistics

College of Literature, Science, and the Arts

Office of the Vice President for Research

Rackham Graduate School

Co-Sponsored:

Institute of Mathematical Statistics

American Statistical Association

COMMITTEES

The organizers for the conference are Robert Keener and Jiayang Sun, assisted by the following committees:

Program Committee: Robert Keener (IMS), Jiayang Sun (IMS), Charles Hagwood, Mary Meyer, and Martha Aliaga (ASA)

Advisory Committee: Hermann Chernoff and David Siegmund

Publications Committee: Anirban Das Gupta, Vince Melfi, Connie Page, and Jiayang Sun

Contents

COMMITTEES	1
CONFERENCE SCHEDULE WITH TITLES	3
POSTER TITLES	5
ABSTRACTS	5
WHILE IN ANN ARBOR	19
Computing	19
Local Attractions	19
SHUTTLE SERVICE	20

Schedule at a Glance

All Talks in 1324 East Hall

Friday	7:30 pm	Registration, 450 West Hall
	8:00 pm	Reception in 450 West Hall (ends at 10:00 pm) Hosted by the Statistics Department
<hr/>		
	8:00 am	Registration and Continental Breakfast, 1324 East Hall
	8:45 am	Introductory Remarks
	9:00 am	Statistics in Astronomy and Physics
	10:30 am	Break
	10:45 am	Nonparametric Inference
Saturday	12:15 pm	Box Lunch and Posters, 1349 East Hall
	2:00 pm	Probability
	4:00 pm	Break
	4:30 pm	Contributed Talks
	6:10 pm	Break
	6:45 pm	Banquet, Great Lakes Room, fourth floor Palmer Commons (ends at 10:30 pm)
<hr/>		
	8:00 am	Continental Breakfast, 1324 East Hall
	8:30 am	Statistics in Biology
	10:00 am	Break
	10:15 am	Biased Sampling and Missing Data
Sunday	11:45 am	Box Lunch and Posters, 1349 East Hall
	12:45 pm	Shape Restricted Regression
	1:45 pm	Break
	2:00 pm	Modern Sequential Analysis and Clinical Trials
	3:30 pm	Conference ends

CONFERENCE SCHEDULE WITH TITLES

All Talks in 1324 East Hall

Friday, September 23

7:30 pm Registration, 450 West Hall

8:00–10:00 pm Reception, 450 West Hall

Hosted by the Statistics Department

Saturday, September 24

8:00 am Registration and Continental Breakfast, 1324 East Hall

8:45 am Introductory Remarks

Vijay Nair, *University of Michigan*

9:00 am Statistics in Astronomy and Physics

Chair: Byron Roe, *University of Michigan*

Dark matter in dwarf galaxies: non-parametric analyses.

♦Mario Mateo, Matthew Walker, Michael Woodroffe, and Xiao Wang, *University of Michigan*

On the behavior of Bayesian credible intervals for some restricted parameter space problems.

♦William E. Strawderman, *Rutgers University*

Heat, Burgers' and Navier-Stokes equations and their stochastic counterparts.

♦Anna Amirdjanova, *University of Michigan*

10:30 am Break

10:45 am Nonparametric Inference

Chair: Ron Butler, *Colorado State University*

Confidence sets for split points in decision trees.

♦Moulinath Banerjee, *University of Michigan*

Consistency of Bayes estimators in non-parametric regression.

♦Steve Lalley, *University of Chicago*

Confidence intervals in group sequential trials with random group sizes and applications to survival analysis.

♦Tze Leung Lai, *Stanford University*

12:15 pm Box Lunch and Posters, 1349 East Hall

2:00 pm Probability

Chair: Vince Melfi, *Michigan State University*

Small Value Phenomenons in Probability and Statistics.

♦Wenbo V. Li, *University of Delaware*

Empirical processes of stationary sequences.

♦Wei Biao Wu, *University of Chicago*

A weak law of large numbers for cooperative gamblers.

♦Gordon Simons, *University of North Carolina at Chapel Hill*

The cutoff conjecture for Markov chains.

♦Persi Diaconis, *Stanford University*

4:00 pm Break

4:30 pm Contributed Talks (selected)

Chair: Tom Sellke, *Purdue University*

Degradation modeling and nonparametric inference based on some nonhomogeneous Levy processes.

Vijay Nair and ♦Xiao Wang, *University of Michigan*

PfCluster: a new cluster analysis procedure for gene expression profiles.

♦Yaomin Xu, and Jiayang Sun, *The Cleveland Clinic Foundation and Case Western Reserve University*

Density estimation and clustering in astronomical sky surveys.

♦ Woncheol Jang, *Duke University*

Variance estimation in nonparametric regression: a possible approach.

♦ Michael Levine, *Purdue University*

A non-linear renewal theorem with stationary and slowly changing perturbations.

♦ Dong-Yun Kim and Michael Woodroofe, *Illinois State University and University of Michigan*

6:10 pm Break

6:45–10:30 pm Banquet, Great Lakes Room, fourth floor Palmer Commons

Remarks by Herman Chernoff, Ed Rothman, and others.

Sunday, September 25

8:00 am Continental Breakfast, 1324 East Hall

8:30 am Statistics in Biology

Chair: Martha Aliaga, *American Statistical Association*

First entrance of DNA into a nanopore.

♦ Charles Hagwood, *National Institute of Standards and Technology*

On the false discovery rates of a frequentist.

♦ Anirban DasGupta, *Purdue University*

Empirical Bayes and Conditional False Discovery Rate.

♦ Cun-Hui Zhang, *Rutgers University*

10:00 am Break

10:15 am Biased Sampling and Missing Data

Chair: Bin Wang, *University of South Alabama*

Biased sampling problems: recent developments and challenges.

♦ Jiayang Sun, *Case Western Reserve University*

Regression analysis of truncated data.

♦ Zhiliang Ying, *Columbia University*

Goodness-of-fit testing in interval censoring case 1.

♦ Hira L. Koul, *Michigan State University*

11:45 am Box Lunch and Posters, 1349 East Hall

12:45 pm Shape Restricted Regression

Chair: Connie Page, *Michigan State University*

Bayes methods in shape-restricted regression.

♦ Mary C. Meyer, *University of Georgia*

A Kiefer-Wolfowitz comparison theorem for Wickell's problem.

♦ Michael Woodroofe, *University of Michigan*

1:45 pm Break

2:00 pm Modern Sequential Analysis and Clinical Trials

Chair: Dong-Yun Kim, *Illinois State University*

Corrected confidence intervals for secondary parameters following sequential tests.

♦ Steve Coad, *Queen Mary, University of London*

Large deviation asymptotics for randomized play the winner design.

♦ Anand N. Vidyashenkar, *Cornell University*

Improving Brownian approximations for boundary crossing problems.

♦ Robert Keener, *University of Michigan*

3:30 pm Conference Ends

POSTER TITLES

Efficient three-stage clinical trial designs. ♦Jay Bartroff, *Stanford University*

On wavelet estimation in censored regression. ♦Linyuan Li, *University of New Hampshire*

Bayesian approach to semi-parametric regression models. Liuxia Wang and ♦Yulin Li, *Carnegie Mellon University and University of Toledo, Toledo*

Locally efficient estimators for semiparametric models with measurement error. ♦Yanyuan Ma and Raymond J. Carroll, *Texas A&M University*

A concordance method for analysing biological categorical time series. An application for the search of hidden periodicities. M. Hassnaoui, R. Pupier, and ♦M. Rehailia, *University of Saint-Etienne*

A Markov model for defining sleep onset rapid eye movement periods (SOREMPs). An approach in sleeping sickness to help with stage diagnosis. Benot Berge, ♦Mohamed Rehailia, Alain Blanc, and Alain Buguet, *University of Saint-Etienne*

A random walk on a finite circular lattice. ♦Jyotirmoy Sarkar, *Indiana University Purdue University Indianapolis*

Wavelet-based monitoring for modern biosurveillance. ♦Galit Shmueli, *University of Maryland, Col-*

lege Park

Using invariant theory to obtain nonparametric motion estimates for rigid objects of unknown shape. ♦Mark A. Stuff, *General Dynamics Advanced Information Systems*

Estimation of rare event under selection bias. ♦Bin Wang and Jiayang Sun, *University of South Alabama and Case Western Reserve University*

Testing for nonlinearity in censored median regression model when the alternative is smooth. ♦Lan Wang, *School of Statistics, University of Minnesota*

Group invariant inferred distributions via non-commutative probability. Barbara Heller and ♦Mei Wang, *Illinois Institute of Technology and the University of Chicago*

Spatial-temporal data mining: LASR - a new procedure. ♦Xiaofeng Wang and Jiayang Sun, *Case Western Reserve University*, and Kath Bogie, *Cleveland VA FES Center*

Properties of local nondeterminism of self-similar random fields and applications. ♦Yimin Xiao, *Michigan State University*

Conditional Properties of a Parametric Bootstrap. ♦Russell Zaretzki, *University of Tennessee*

Existence of signal in the signal plus background model. ♦Tonglin Zhang, *Purdue University*

ABSTRACTS

Statistics in Astronomy and Physics

9:00 am, Saturday

Dark matter in dwarf galaxies: non-parametric analyses.

Mario Mateo, Matthew Walker, Michael Woodroffe, and Xiao Wang, *University of Michigan*

It has been known for decades that the kinematics of galaxies cannot be accounted by the visible stellar or gaseous content of these systems. Dwarf galaxies show this ‘Dark Matter’ problem especially strongly because they contain so little ‘normal’, or baryonic, matter yet their internal kinematics suggest large total masses. We have been obtaining high-precision velocities of individual stars in nearby galaxies with the aim of producing samples sufficiently large to be amenable to more detailed analyses than in the past. One

of the new approaches is to apply non-parametric analyses to these data to infer the mass distributions with a minimal number of restrictive geometric or astrophysical assumptions. In this talk, I will outline the observational project and the traditional and new techniques used to analyze these data.

On the behavior of Bayesian credible intervals for some restricted parameter space problems.

William E. Strawderman, *Rutgers University*

Following recent particle physics experiments aiming to estimate the mass of neutrinos, there has been a renewed interest in the problem of setting confidence bounds for constrained parameter models (see for instance Feldman and Cousins, 1988; Mandelkern, 2002; and Efron, 2004). For estimating a positive normal mean, Zhang and Woodroffe (2003) as well as Roe and Woodroffe (2000) investigate $100(1 - \alpha)\%$ HPD credible sets associated with priors obtained as the truncation of noninformative priors onto the restricted parameter space. Namely, they establish the attractive lower bound of $\frac{1-\alpha}{1+\alpha}$ for the frequentist coverage probability of these procedures. In this work, we establish that the lower bound of $\frac{1-\alpha}{1+\alpha}$ is applicable for a substantially more general setting with underlying distributional symmetry, and present various illustrations and related properties. Investigations of non-symmetric models are carried out and similar results are obtained.

Acknowledgement: This is joint work with Éric Marchand (Université de Sherbrooke).

Heat, Burgers' and Navier-Stokes equations and their stochastic counterparts.

Anna Amirdjanova, *University of Michigan*

The talk is devoted to recent advances in the study of three fundamental equations of mathematical physics (heat, Burgers' and Navier-Stokes) when the classical models are perturbed by random forces. After reviewing the physical meaning of these equations, questions of existence, uniqueness, regularity and ergodicity of solutions will be discussed and several open problems presented.

Acknowledgement: This research was supported in part by the National Security Agency.

Nonparametric Inference

10:45 am, Saturday

Confidence sets for split points in decision trees.

Moulinath Banerjee, *University of Michigan*

In this talk, I will present a number of different inference strategies for estimating the optimal split point of a stump approximation to a regression function. The underlying regression function is assumed to be smooth; the stump approximation is used to identify "structure" and is especially relevant if the function has a monotonic trend and a region of sharp change, in which case the split point can be interpreted as a natural threshold. Optimal split points will be determined on the basis of two criteria: (a) least squares (which is used in CART) (b) likelihood function (when a likelihood function can be written down). The natural estimate of the optimal split point in either case will be shown to converge to the truth at rate $n^{1/3}$

with the limit being given (up to constants) by Chernoff's distribution. A residual sum of squares/likelihood ratio statistic will be shown to converge (up to constants) to the maximum of Brownian motion minus a quadratic drift. These results will be used to construct (asymptotic) confidence sets for the optimal split. Time permitting, I will also present alternative estimation strategies and illustrate the methods through simulation studies and applications to real life data.

Acknowledgement: The speaker's research was partially supported by NSF grant DMS-0306235. This is joint work with Ian McKeague (Columbia University).

Consistency of Bayes estimators in nonparametric regression.

Steve Lalley, *University of Chicago*

The consistency of Bayes procedures in nonparametric problems can depend on the degree of concentration of the prior: if the prior is not sufficiently concentrated on low-complexity models, then the posterior may concentrate on models that are "over-fit". For several natural classes of priors suited to use in regression problems, the degree of prior concentration necessary for consistency is connected to a large deviation problem associated with the model.

Techniques for analyzing the large deviations problems arising in connection with Bayes procedures based on hierarchical priors will be discussed. At least one useful class of priors, the "uniform mixture" priors introduced by Coram, will be discussed in some detail.

Confidence intervals in group sequential trials with random group sizes and applications to survival analysis.

Tze Leung Lai, *Stanford University*

A new ordering scheme for defining quantiles of the multivariate distribution of a stopping time and a stopped stochastic process is introduced herein. This ordering scheme is used in conjunction with resampling methods to construct confidence intervals for (i) a population mean following a group sequential test with random group sizes, and (ii) the regression parameter of a proportional hazards model following a time-sequential clinical trial with censored survival data. Asymptotic analysis and simulation studies show that the confidence intervals thus constructed have coverage probabilities close to the nominal values and provide marked improvements over those based on alternative ordering schemes and normal approximations.

Probability

2:00 pm, Saturday

Small Value Phenomenons in Probability and Statistics.

Wenbo V. Li, *University of Delaware*

Two fundamental problems in probability theory and statistical analysis are typical behaviors such as expectations, laws of large numbers, central limit theorems and approximated sampling distributions, and rare events such as large deviations, significant level and power. Small value phenomenons come from both typical behaviors and rare events of the type that positive random variables take smaller values. In the literature, small value probabilities of various types are studied and applied to many problems of interest under

different names such as small deviation, lower tail behaviors, boundary crossing probabilities, asymptotic evaluation of Laplace transform for large time, stopping rules in sequential analysis, etc. We will provide an overview on current and emerging opportunities in the area.

Empirical processes of stationary sequences.

Wei Biao Wu, *University of Chicago*

We will discuss sample path properties of empirical distribution functions of causal stationary processes. For short and long range dependent processes, their asymptotic behaviors are quite different. Weak convergence to Gaussian processes and multiple Wiener Ito integrals will be presented. Relations with stochastic prediction theory will also be explored.

A weak law of large numbers for cooperative gamblers.

Gordon Simons, *University of North Carolina at Chapel Hill*

In order to increase their chances of benefiting from a truly large outcome, n gamblers agree, before they play, to share their winnings among themselves. Under a simple sharing scheme, the winnings of each is distributed according to the weighted sum

$$T_n = p_{1,n}X_1 + \cdots + p_{n,n}X_n,$$

where X_1, \dots, X_n represent their independent and identically distributed individual winnings, and $p_n = (p_{1,n}, \dots, p_{n,n})$ is a probability distribution—with nonnegative components adding to unity—for $n \geq 1$.

The surprising and genuine benefit of this kind of sharing—for each of the gamblers—has already been demonstrated when the X 's have the St. Petersburg distribution ($P(X = 2^k) = 1/2^k$ for $k = 1, 2, \dots$). For this distribution, it can be shown that $T_n/H(p_n)$ converges in probability to 1, where $H(p_n)$ is the entropy of p_n , when $H(p_n) \rightarrow \infty$ as $n \rightarrow \infty$. The talk will describe similar kinds of weak laws of large numbers, valid for a large class of distributions.

The cutoff conjecture for Markov chains.

Persi Diaconis, *Stanford University*

Natural mixing processes sometimes show a sharp transition from start to stationarity. Yuval Peres observes that, in all cases, there is a sharp cutoff if, and only if, the product of the spectral gap and the relaxation time tends to infinity. Laurent Saloff-Coste and I have recently proved this for birth and death chains. I will review the history and examples, and explain the tools (duality and stochastic interpretation of eigenvalues).

Contributed Talks (selected)

4:30 pm, Saturday

Degradation modeling and nonparametric inference based on some nonhomogeneous Levy processes.

Vijay Nair and Xiao Wang, *University of Michigan*

Recent advances in sensing and measurement technologies are making it feasible to collect extensive amounts of data on degradation and other performance measures associated with components, systems, and manufacturing processes. Degradation data are a very rich source of reliability information and offer many advantages over the analysis of time-to-failure data. In this talk, we propose some models for degradation data and describe their application for reliability inference. Here, time-to-failure is defined as the level crossing (first-passage time) of a specified degradation threshold. The models we consider are based on non-homogeneous Gaussian and Gamma processes. These models can accommodate a variety of degradation rates and shapes. They also lead naturally to a wide class of time-to-failure distributions. We will discuss inference based various types of data sources and in particular some interesting problems dealing with nonparametric inference.

PfCluster: a new cluster analysis procedure for gene expression profiles.

Yaomin Xu, and Jiayang Sun, *Case Western Reserve University and The Cleveland Clinic Foundation*

Properly clustering gene profiles from a longitudinal study or under different conditions is important in building gene association networks and for finding leading genes for certain diseases. In this paper, we propose a new clustering technique that is efficient and flexible in uncovering clusters determined by a class of biologically meaningful distances. We develop within- and between-group coherence indices for the resulting clusters in the spirit of p-values. An efficient filtering method is also provided for preprocessing the data to filter out the idled genes with flat profile patterns. Several fresh real data are analyzed by utilizing the new technique, and implications are discussed.

Acknowledgement: This research is supported in part by an NSF grant. We would also like to acknowledge Toshimori Kitami and Dr. Joe Nadeau in the department of genetics at Case Western Reserve University and Dr. Arun D.Singh in the department of Ophthalmic Oncology, Cole Eye Institute at the Cleveland Clinic Foundation for their helpful discussions on biological and clinical background of the experiments and access to their data to test the methods developed in this paper.

Density estimation and clustering in astronomical sky surveys.

Woncheol Jang, *Institute of Statistics and Decision Sciences, Duke University*

The Universe is homogeneous and isotropic on large scales, but on small scales, one can find significant deviations from homogeneity and isotropy. Clusters of galaxies play an important role in finding where the local structure fades away into a homogeneous and isotropic distribution. From statistical point of view, finding clusters of galaxies is equivalent to finding density contour clusters (Hartigan, 1975), connected components of a level set $S_c \equiv \{f > c\}$. We present a nonparametric method to find density contour clusters. To extract connected components of the estimated level set, we propose to use a union of balls to approximate the estimated level set which is a modified version of Cuevas, et al. (2000). The method is applied to studies of the Edinburgh-Durham Southern Galaxy Catalogue (EDSGC) and Mock 2df catalogue. Results are compared with the existing galaxy catalogs.

Acknowledgement: This is joint work with Larry Wasserman and Bob Nichols.

Variance estimation in nonparametric regression: a possible approach.

Michael Levine, *Purdue University*

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

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

for $i = 1, \dots, n$ 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

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

for a set of numbers $\{d_i\}$ such that $\sum_j d_j = 0$ and $\sum_j d_j^2 = 1$ and $i = 1, \dots, n - r$. 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 than 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 $r \rightarrow \infty$. 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.

Acknowledgement: This is joint work with Prof. Lawrence D. Brown.

A non-linear renewal theorem with stationary and slowly changing perturbations.

Dong-Yun Kim and Michael Woodroffe, *Illinois State University and University of Michigan*

Non-linear renewal theory is extended to include random walks perturbed by both a slowly changing sequence and a stationary one. Main results include, a version of the Key Renewal Theorem, a derivation of the limiting distribution of the excess over a boundary, and an expansion for the expected first passage time. The formulation is motivated by problems in sequential analysis with staggered entry, where subjects enter a study a random times. This is illustrated by an example.

Acknowledgement: This is joint work with Michael Woodroffe.

Statistics in Biology

3:30 am, Sunday

First entrance of DNA into a nanopore.

Charles Hagwood, *National Institute of Standards and Technology*

Single stranded DNA is modeled as a flexible polymer using Zimm's model for hydrodynamic interaction. The motion of DNA in a box with all sides reflecting except one absorbing side on which the pore resides is modeled using the Langevin equation. The transition probability is derived using the Fokker Planck equation. From this, the probability the DNA polymer is found at the entrance to a nanopore is computed.

On the false discovery rates of a frequentist.

Anirban DasGupta, *Purdue University*

Consider testing an abstract null hypothesis $H_0 : \theta \in \Theta_0$. A very common practice is to reject the null when the P-value computed from the observed sample using a suitable statistic falls below some threshold number α , typically .05. In well known foundational work, Berger, Bayarri, Casella, the present author, Hall, Sellke, and others have spiritedly discussed the often contentious issue of the evidentiary value of P-values. Concerns have been raised that the frequentist rejects the null too soon.

We ask the question *how many of the null hypotheses a frequentist rejects are actually true*. Precisely, with a proper prior π on the parameter θ , we look at the Bayesian False Discovery Rate $P_\pi(\theta \in \Theta_0 | \text{P-value} < \alpha)$. This is a function of the sample size n , the prior π , the threshold value α , and the choice of the test statistic used to compute the P-value. For one sided nulls, we derive a third order asymptotic expansion for this false discovery rate, with the natural sufficient statistic in the continuous exponential family, with the median in location parameter families, and with the MLE in nonregular families. The expansions differ qualitatively in the regular and nonregular cases.

The expansions are derived by putting together Edgeworth expansions for the CDFs of the test statistics with Cornish-Fisher expansions for their quantile functions. The terms are complex, but explicit. We use the three term expansions to analyze the intended question. We reach the conclusion that unless the prior π is too sharp, the frequentist does not falsely discover an effect in any alarming numbers. In fact, the expansions show that the frequentist needs to protect against overlooking an effect when it was actually there.

Acknowledgement: This work is in collaboration and consultation with Michael Woodroffe and Tonglin Zhang.

Empirical Bayes and conditional false discovery rate.

Cun-Hui Zhang, *Rutgers University*

The false discovery rate has been widely used in large scale multiple testing problems as it seems to attain a balanced compromise between the more liberal per comparison error rate and the more conservative familywise error rate. Examples include microarray studies with a large number of genes and imaging problems

involving selection from a large number of pixels or potential regions of interest. We formulate an optimization problem as the maximization of the total amount of statistical discovery subject to a preassigned level of false discovery rate conditionally on test statistics, and propose an empirical Bayes approach based on the Bayes solution of the optimization problem. Asymptotic optimality of certain empirical Bayes procedures is proved and the results of a simulation study are presented.

Biased Sampling and Missing Data

10:15 am, Sunday

Biased sampling problems: recent developments and challenges.

Jiayang Sun, *Case Western Reserve University*

In observational studies, data are often incomplete. If whether a subject is included in a study depends on its true value, the resulting sample for the study is subject to a sampling or selection bias. In this talk, a variant of selection bias models is surveyed, some recent developments in estimation and hypothesis testing about the bias are presented, challenges and open problems are discussed. A connection with pattern mixture models will also be shown, some comparative studies will be given. A few applications will be illustrated via read data analyses.

Acknowledgement: Part of the talk is based on the joint work over last 10 years with Michael Woodroofe and Bin Wang. The work is supported in part by NSF grants from DMS.

Regression analysis of truncated data.

Zhiliang Ying, *Columbia University*

I will give an overview about various methods for the linear regression model when the response variable is subject to truncation. Particular attention will be given to some recent developments when the truncation is complex. Some new proposals in regard to computation and inference will be described and their applications to analysis of astronomical data be discussed.

Goodness-of-fit testing in interval censoring case 1.

Hira L. Koul, *Michigan State University*

In the interval censoring case 1, instead of observing an event occurrence time X time T and $I(X \leq T)$. Here we shall discuss ADF tests of goodness-of-fit hypotheses pertaining to the d.f. F of X . These tests are shown to be also consistent against a large class of fixed alternatives and have nontrivial asymptotic power against a large class of local alternatives. A simulation study is included to exhibit the finite sample level and power behavior.

Shape Restricted Regression

12:45 pm, Sunday

Bayes methods in shape-restricted regression.

Mary C. Meyer, *University of Georgia*

Least-squares regression using shape restrictions for the regression function is a type of nonparametric regression that is useful in many applications. The maximum likelihood solution involves a projection onto a polyhedral convex cone rather than onto a linear subspace. Inference methods are difficult in part due to a lack of obvious definition of error degrees of freedom and difficulty in finding distributions of the test statistics. A Bayesian formulation provides smooth function estimates, and inference methods involve sampling from the posterior. methods are feasible using Bayes credible intervals and Bayes factors. Extensions to generalized linear models and hazard function estimation are discussed.

Acknowledgement: Part of this work was supported by NSF Grant DMS 0204572. This is joint work with Prakash Laud.

A Kiefer-Wolfowitz comparison theorem for Wickell's problem.

Michael Woodroffe, *University of Michigan*

A well known result of Kiefer and Wolfowitz states: given a sample from a decreasing density, the difference between the empirical distribution function and its least concave majorant is of order $n^{-2/3} \log(n)$ w.p.1. A result of this nature is obtained in the context of Wickell's Problem, but with a faster rate of convergence. Wickell's Problem is to infer the distribution of the radii of three-dimensional spheres from a sample of the radii of their two-dimensional cross-sections. Let F and G denote the distribution functions of squared radii, and let $U(y) = 2 \int_0^y [\sqrt{y} - \sqrt{y-z}] G\{dz\}$. Then it can be shown that $U'(y) = \pi^2 [1 - F(y)]$, so that U is an non-decreasing, concave function. There is a natural unbiased estimator of U which is non-decreasing, but not concave. Letting $U_n^\#$ and \hat{U}_n denote the unbiased estimator and its least concave majorant, and $\epsilon_n^2 = \log(n)/n$, it is shown that $\sup_y |U_n^\# - \hat{U}_n| = O_p(\epsilon_n)$. A local version of the result is obtained too, with O replaced by o , and both results are developed in a more general context.

Modern Sequential Analysis and Clinical Trials

2:00 pm, Sunday

Corrected confidence intervals for secondary parameters following sequential tests.

Steve Coad, *Queen Mary, University of London*

Corrected confidence intervals are developed for the mean of the second component of a bivariate normal process when the first component is monitored sequentially. This is accomplished by constructing a first approximation to a pivotal quantity, and then using very weak expansions to determine the correction terms. The asymptotic sampling distribution of the renormalised pivotal quantity is established in both the case where the covariance matrix is known and when it is unknown. The resulting approximations have a simple form and the results of a simulation study of two well-known sequential tests show that they are very accurate. The practical usefulness of the approach is illustrated by a real example of bivariate data.

Acknowledgement: This is joint work with Ruby Weng at the National Chengchi University in Taipei. Part of the work was carried out while she was visiting the University of Sussex during July and August 2003, and in receipt of Study Visit Grant 15600 from The Royal Society. She is also partially supported by the National Science Council of Taiwan. The authors are grateful to John Whitehead for suggesting the bivariate data used.

Large deviation asymptotics for randomized play the winner design.

Anand N. Vidyashankar, *Cornell University*

Randomized Play the Winner design has been used to allocate patients to one of the treatments in a clinical trial with several competing treatments. Law of large numbers and the Central limit theorem for number of patients allocated to each treatment have been developed under fairly general conditions. These results have been used to establish the asymptotic properties of the maximum likelihood estimators of the design parameters. In this paper, I will describe results concerning the conditional and unconditional large deviations for the maximum likelihood estimators of the design parameters. These results, in turn, yield results concerning the large deviations for proportions in generalized Polya's urn model. Finally, I will use these results to establish precise connections to various aspects of statistical efficiency. Related problems concerning randomly stopped averages will also be described.

Acknowledgement: Research Supported in part by a grant from NSF DMS 000-03-07057 and also by grants from the NDCHealth Corporation.

Improving Brownian approximations for boundary crossing problems.

Robert Keener, *University of Michigan*

Donsker's Theorem shows that random walks behave like Brownian motion in an asymptotic sense. This result can be used to approximate certain expectations associated with the time and location of a random walk when it first crosses a nonlinear boundary. We will derive correction terms to improve the accuracy of Brownian motion approximations.

Acknowledgement: Research supported in part by NSA grant F012499.

Posters

Lunch Time, Saturday and Sunday

Efficient three-stage clinical trial designs.

Jay Bartroff, *Stanford University*

Since there is often little information about a new treatment at the beginning of a clinical trial, methods of re-estimating the sample size midway through the trial are of much interest. However, these adaptive designs typically are inefficient, using non-sufficient test statistics in order to maintain desired error probabilities. On the other hand, group sequential designs are efficient but don't use information available midway through the trial to adjust the sample size and possibly stop early. We present a three-stage design that is both efficient, using a sufficient statistic and achieving given error probabilities, as well as adaptive, adjusting the sample size after the first stage based on the data observed so far.

On wavelet estimation in censored regression.

Linyuan Li, *University of New Hampshire*

The Cox proportional hazards model has become the model of choice to use in analyzing the effect of covariates on survival data. However, this assumption has significant restrictions on the behavior of the

conditional survival function. The accelerated failure time model, which models the survival time and covariates directly through regression, provides an alternative approach to interpret the relationship between survival times and covariates. We consider here the estimation of the nonparametric regression function under the right random censorship model and investigate the asymptotic rates of convergence of estimators based on thresholding of empirical wavelet coefficients. We show that the estimators achieve nearly optimal minimax convergence rates within logarithmic terms over a large range of Besov function classes, a feature not available for the linear estimators when $p < 2$. The performance of the estimators is tested via simulation and the method is applied to the Stanford Heart Transplant data and other interesting data sets.

Acknowledgement: This is a joint work with Brenda MacGibbon and Christopher Valenta.

Bayesian approach to semi-parametric regression models.

Liuxia Wang and Yulin Li, *Carnegie Mellon University and University of Toledo, Toledo*

Semi-parametric regression model is a generalization of the regression model. It can not only keep the flexibility of the nonparametric models for the baseline function, but also maintain the explanatory power of parametric models. Various approaches to evaluate the baseline function have been proposed. In this article, a full Bayesian approach is proposed. B-spline is used to estimate the baseline function. Reversible jump Markov chain Monte Carlo is adopted to automatically determine both the number of knots and their locations, together with the unknown basis coefficients. Simulation results suggest that this approach performs well. We also illustrate the method using the Munich rental rates data.

Locally efficient estimators for semiparametric models with measurement error.

Yanyuan Ma and Raymond J. Carroll, *Texas A&M University*

We derive constructive locally efficient estimators in semiparametric measurement error models. The setting is one where the likelihood function depends on variables measured with and without error, where one of the variables measured without error is modeled nonparametrically. The algorithm is based on backfitting. This problem includes the partially linear measurement error model as a special case. We show that if one assumes a parametric model for the latent variable measured with error and if this model is correct, then our estimators are semiparametric efficient. In contrast to standard problems, our methods enjoy the property that even if the latent variable model is misspecified, our methods still lead to consistent and asymptotically normal estimators. We illustrate the methods with a logistic regression problem where the latent variable measured with error is modeled by a quadratic regression, while a variable measured without error is modeled nonparametrically.

We demonstrate the method in a partially linear measurement error model, where the measurement error and the model error are both normal. In such setting, the resulting estimator has a closed form. When a normal model for the unobservable variable is posited, the estimator gains extra robustness against the normality assumption on the response variable and the measurement, resulting in an estimator that is consistent for the general partially linear measurement model without any normal assumptions.

A concordance method for analysing biological categorical time series. An application for the search of hidden periodicities.

M. Hassnaoui, R. Pupier, and M. Rehailia, *University of Saint-Etienne*

A simple method for searching for hidden periods in biological time series is proposed. Because the method is based on an auto-comparison of the observations within a series we call it the concordance method. It requires very few theoretical assumptions. In fact, even the often present stationarity condition is not used.

The performances of the method are compared with competing methods based on the chi-square periodogram. It is shown that the concordance method is more efficient for analysing multimodal and noisy data. Rhythms presenting simultaneously circadian and ultradian components can also be analysed with this method.

Several illustrations related to the study of sleep patterns of healthy and trypanosomiasis infected rats are given. A fairly comprehensive understanding of sleep rhythmicity alteration during trypanosomiasis is obtained. Investigations were conducted by studying overall sleep patterns as well as every single sleep/wake state separately.

A Markov model for defining sleep onset rapid eye movement periods (SOREMPs). An approach in sleeping sickness to help with stage diagnosis.

Benot Berge, Mohamed Rehalia, Alain Blanc, and Alain Buguet, *University of Saint-Etienne*

The analysis of duration and transition between the various sleep-wake states over the nycthemeron help characterize sleep architecture deregulation and is at the basis of a sleep onset REM periods (SOREMP) definition in the human African trypanosomiasis (HAT) context. Twenty-three African volunteers were included in the study: 6 healthy subjects and 17 meningoencephalitic patients, infected by *Trypanosoma brucei gambiense*. Sleep-wake states were scored in 20-sec epochs following standard criteria. The 24-hour polysomnographic recordings were regarded as a dynamical process and a continuous time Markov chain (CTMC) model was adopted. In both healthy subjects and meningoencephalitic patients, wakefulness duration distribution was best described by a mixture of exponentials, suggesting that several underlying processes generate wakefulness. Transitions between sleep-wake states are altered during HAT. Wakefulness generation and related sleep architecture disruptions are markers of the meningoencephalitic stage. REM latencies are defined by distinguishing between short and long (≥ 2 minutes) wakefulness episodes. The comparison of REM sleep latency distributions between healthy subjects and meningoencephalitic patients led to define SOREMPs as corresponding to REM latencies shorter than 12 minutes.

A random walk on a finite circular lattice.

Jyotirmoy Sarkar, *Indiana University Purdue University Indianapolis*

A particle moves among a set of $m + 1$ nodes, labeled $0, 1, \dots, m$, that are arranged around a circle. The particle starts at 0 and henceforth, at each step the particle moves one position in the clockwise direction with probability p , or one position in the counterclockwise direction with probability $1 - p$, where $0 < p < 1$ is known. The directions of successive movements are independent of one another.

We answer the following questions: (1) What is the probability that the last state visited is i ($i = 1, 2, \dots, m$)? (2) What is the expected number of moves needed to visit all states? (3) What is the expected number of moves needed to return to 0 after visiting all states? Suppose now that the process continues *ad infinitum*. (4) What proportion of transitions enter into each node? (5) What is the expected number of moves between successive visits to each node?

Acknowledgement: This is joint work with Taiwo Salau, St. Mary's College of Maryland. The authors thank Benzion Boukai for his special interest, guidance and many discussions.

Wavelet-based monitoring for modern biosurveillance.

Galit Shmueli, *University of Maryland, College Park*

Current surveillance methods rely mostly on traditional statistical monitoring methods such as statistical process control and autoregressive time series models. However, these methods are not always suitable for monitoring syndromic and especially non-traditional biosurveillance data. Assumptions such as normality, independence, and stationarity are the backbone of such methods, whereas the types of data that are monitored in bio-surveillance almost always violate these assumptions. Furthermore, outbreak signatures in such data are of unknown patterns, and therefore methods that are tuned to particular anomaly patterns (such as classic control charts) become much less powerful. In light of this we propose the use of the non-parametric wavelet-based methods. These methods make much less assumptions and are suitable for detecting abnormalities of unknown form. They are also computationally cheap and becoming more available in commercial software. Although wavelets have been widely used for data denoising and compression, there is very little work done on using them for monitoring, and especially for prospective monitoring. We discuss various issues that are relevant to wavelet-based monitoring and illustrate the methods using real data on military outpatient clinic visits in Charleston.

Using invariant theory to obtain nonparametric motion estimates for rigid objects of unknown shape.

Mark A. Stuff, *General Dynamics Advanced Information Systems*

When a sensor collects data from a moving extended object, estimation of the direction vectors from the object to the sensor is often essential to the extraction of useful information from the sensor data. If the object or the sensor moves as result of uncontrolled or unknown forces, simple parametric models for the angular motions often rapidly loose fidelity. So, even if the object can be modeled parametrically, nonparametric motion estimates are desirable. In one example of such a problem, a direct approach to estimating all the unknowns leads to difficult nonlinear optimization problems. But a characterization of the shape of the object, using the right choice of geometric invariants, can decouple the problem, temporarily isolating the object shape estimation from the motion estimation. This facilitates the extraction of nonparametric motion estimates both by subdividing the parameter space, and by enabling parts of the problem to be solved using linear methods.

Acknowledgement: This is joint work with thesis advisors, Michael Woodroffe and Robert Keener.

Estimation of rare event under selection bias.

Bin Wang and Jiayang Sun, *University of South Alabama and Case Western Reserve University*

This paper is motivated by a study of the cancer risks of Vietnamese Americans in Mobile Alabama. In the study, the researchers encountered two difficulties: (1) biasing problem: the detection rate of cancer may vary for individuals at different risk levels. (2) the incidence rate of cancer is pretty small in the surveyed sample. This may cause under-estimate in estimation by a logistic regression model. In this paper, method to correct the possible selection bias will be considered. Also, we will propose a new model to justify the estimation bias in the logistic regression model. The new method will be compared with King and Zeng's (2000) and YSSA DeWoody etc's (1999) methods. Simulation will also be performed to illustrate the performance of the new estimators.

Testing for nonlinearity in censored median regression model when the alternative is smooth.

Lan Wang, *School of Statistics, University of Minnesota*

We propose a nonparametric test for diagnosing the adequacy of linearity assumption in censored median regression model. The main advantage of the new test over existing methods is that it allows the alternative to be any smooth function. In addition, the error distribution is not required to comply with any parametric assumption. The test statistic has asymptotic normal distribution under the null hypothesis. A Monte Carlo study demonstrates its favorable performance compared to the classical Wald test.

Group invariant inferred distributions via noncommutative probability.

Barbara Heller and Mei Wang, *Illinois Institute of Technology and the University of Chicago*

In answering the question what is the probability distribution of the parameter given observed data when there is little or no prior knowledge on the parameter values, one may consider three types of statistical inference: Bayesian, frequentist, and group invariance-based. The focus here is on the latter method. We use three one-parameter probability families (the Poisson, normal and binomial distributions) to illustrate a group-invariance method to obtain inferred distributions on the parameter spaces conditional upon observed results. The families are constructed according to group theoretic methods involving so-called coherent states. These particular inferred distributions coincide with Bayesian posteriors. In that sense, this context provides a method for obtaining noninformative prior measures.

Spatial-temporal data mining: LASR - a new procedure.

Xiaofeng Wang and Jiayang Sun, *Case Western Reserve University*, and Kath Bogie, *Cleveland VA FES Center*

This paper is concerned with the spatial-temporal data mining motivated by analyzing data from our “Neuromuscular Electrical Stimulation” (NMES) experiment. We develop an efficient procedure for mining spatial-temporal data – *Longitudinal Analysis with Self-Registration* (LASR, pronounced “laser”). This new procedure is a statistical ensemble built on following modern or newly developed components: (1) data segmentation for separating heterogeneous data and for distinguishing outliers, (2) automatic approaches for spatial and temporal data registration, (3) statistical smoothing mapping for identifying “activated” regions based on generalized false discovery rate (FDR) controlled p -maps/movies from “large- p -small- n ” data sets. Our new procedure should be applicable to other types of spatial-temporal data sets beyond those from the NMES experiment. It has the potential to be used in the analysis of time-series images and functional images such as those from fMRI.

Properties of local nondeterminism of self-similar random fields and applications.

Yimin Xiao, *Michigan State University*

In this talk we discuss properties of strong local nondeterminism and sectorial local nondeterminism of Gaussian and stable random fields. We give conditions for such local nondeterministic properties to hold and show their applications in studying sample path properties of N -parameter Gaussian and/or stable random

fields with values in \mathbf{R}^d . The class of random fields that we consider includes fractional Brownian motion, the Brownian sheet, fractional Brownian sheets and self-similar stable random fields with stationary increments and so on.

Conditional Properties of a Parametric Bootstrap.

Russell Zaretzki, *University of Tennessee*

DiCiccio, Martin and Stern (2001) introduced the parametric bootstrap of the signed root statistic as a useful computational alternative to analytic approximations when highly accurate statistical inference is desired. This performance is equivalent to asymptotic techniques such as the r^* formula; see Barndorff-Nielsen(1986). In addition, simulation examples contained in DiCiccio (2001) suggest that this bootstrap technique can produce extremely accurate conditional inferences. The present work further investigates these conditional properties. In particular, we prove that, in the presence of nuisance parameters, inferences based on a parametric bootstrap of the signed root are conditionally accurate to $O_p(n^{-1})$.

Acknowledgment: This is joint work with Thomas J. DiCiccio of Cornell University and G.A. Young of Imperial College, London.

Existence of signal in the signal plus background model.

Tonglin Zhang, *Purdue University*

Suppose that an observed count X is of the form $X = B + S$, where the background B and the signal S are independent Poisson random variables with parameters b and θ respectively, b is known but θ is not. The model arises in astronomy and high energy physics, and some recent articles have suggested conditioning on the observed bound for B ; that is, if $X = n$ is observed, the suggestion is to base the inference on the conditional distribution of X given $B \leq n$. This suggestion is used here to derive an estimator of the probability of the existence of the signal. The estimator is examined from the view of decision theory and is shown to be admissible.

WHILE IN ANN ARBOR

Computing

Computing is available in B760 East Hall. Several computers will be set up in the computing lab for Guest account use.

Wireless computing is available at near by coffee shops. Espresso Royale Caffe: 1101 South University (adjacent to the conference location) or 324 South State Street.

Local Attractions

Kelsey Museum of Archaeology 434 S. State Street. Representative pieces from U-M archaeological digs in the Middle East, Greece, and Rome. Monday: Closed; Tuesday–Friday: 9:00 am to 4:00 pm; Saturday and Sunday: 1:00 pm to 4:00 pm There is no admission charge, although donations are accepted.

Exhibit Museum of Natural History 1109 Geddes Avenue. Permanent exhibits on dinosaurs and other prehistoric life, Michigan wildlife, Native American culture, anthropology, geology, and a planetarium. We also offer changing temporary exhibits on various themes reflecting current research of our sister museums and other relevant topics. Monday–Saturday 9:00–5:00; Sunday 12:00–5:00; While admission to the museum is free, the suggested donation is \$5 for adults and \$3 for children.

Museum of Art 525 S. State Street. The University of Michigan Museum of Art houses one of the finest university art collections in the country and the second largest art collection in the state of Michigan. A community museum in a university setting, the Museum of Art offers visitors a rich and diverse permanent collection, supplemented by a lively, provocative series of special exhibitions and a full complement of interpretive programs. Monday closed; Tuesday–Saturday 10:00–5:00; Thursday 10:00–9:00; Sunday 12:00–5:00; admission is free, although a donation of \$5 is suggested.

SHUTTLE SERVICE

Shuttle service (mini vans) will be provided during the conference. The pick up and drop off locations in the schedule below refer to:

East and West Hall. (east entrance—Psychology—Church Street)

Palmer Commons. (first level Palmer Commons Parking Structure)

Hotels. Kensington Court and Holiday Inn Express

Friday	Shuttle to East Hall from Hotel 7:15 pm, 7:30 pm, 7:45 pm	Shuttle to Hotels from East Hall 9:30 pm, 9:45 pm, 10:00 pm, 10:15 pm
Saturday	Shuttle to East Hall from Hotel 7:45 am, 8:00 am, 8:15 am	Shuttle to Hotels from Palmer Commons 9:30 pm, 9:45 pm, 10:00 pm, 10:15 pm
Sunday	Shuttle to East Hall from Hotel 7:45 am, 8:00 am, 8:15 am	Shuttle to Hotel from East Hall 3:30 pm, 3:45 pm, 4:00 pm, 4:15 pm