THE FIELDS INSTITUTE
FOR RESEARCH IN MATHEMATICAL SCIENCES

ABSTRACTS 1.2

MICHAEL G. AKRITAS
Penn State University

Fully nonparametric models for random and mixed effects designs

The classical random and fixed effects models rely on the assumptions of normality and homoscedasticity. In some designs, these assumptions (and some additional ones) are needed to guarantee the assumed independence of the random effects. The key features of the new nonparametric formulation of random and mixed effects models are that a) they apply to all discrete and continuous ordinal data, b) the distribution of the response variable can depend in an arbitrary fashion on the levels of the random and fixed effects, and c) the random effects are shown to be uncorrelated. A number of open methodological problems will be discussed. An extension of the nonparametric formulation to two-level clustering, and application to root survival data will be presented.

EPIMACO A. CABANLIT, JR.
Mindanao State University

On The Mixture of the Displaced Exponential Distributions

The Displaced Exponential and the Schuhls Composite Distribution are considered as good models for traffic flow. The Schuhls Composite Distribution is a mixture of two Displaced Exponential Distributions. In this paper, we extend the Schuhls Composite Distribution for more than two Displaced Exponential Distributions. Important Summaries are then presented. The parameters of the said mixture are also estimated. Lastly, various graphs of the distribution are exhibited.

Keywords and Phrases: Exponential Distribution, Displaced Exponential Distribution, Schuhls Composite Distribution, Moment Generating Function, Characteristic Function and Maximum Likelihood Estimates

Y.P. CHAUBEY, A. SEN AND P. K. SEN
Concordia University and University of North Carolina

A New Smooth Density Estimator for Non-negative Random Variables

Commonly used kernel density estimators may not provide admissible values of the density or its functionals at the boundaries for densities with restricted support. For smoothing the empirical distribution a generalization of the Hille's lemma, considered here, alleviates some of the problems of kernel density estimator near the boundaries. For nonnegative ran- dom variables which crop up in reliability and survival analysis, the proposed procedure is thoroughly explored; its consistency and asymptotic distributional results are established un- der appropriate regularity assumptions. Analysis of obtaining smoothing parameters through cross-validation is also provided.

C. M. CRAINICEANU
John Hopkins University

Cox models with nonlinear effect of covariates measured with error: A case study of chronic kidney disease incidence

We propose, develop and implement the simulation extrapolation (SIMEX) methodology for Cox regression models when the log hazard function is linear in the model parameters but nonlinear in the variables measured with error (LPNE). The class of LPNE functions contains but is not limited to strata indicators, splines, quadratic and interaction terms. The first order bias correction method proposed here has the advantage that it remains computationally feasible even when the number of observations is very large and multiple models need to be explored. Theoretical and simulation results show that the SIMEX method outperforms the naive method even with small amounts of measurement error. Our methodology was motivated by and applied to the study of time to chronic kidney disease (CKD) progression as a function of baseline kidney function and applied to the Atherosclerosis Risk in Communities (ARIC), a large epidemiological cohort study.

CHANG C.Y. DOREA
Universidade de Brasilia

Strong Consistency of Kernel Density Estimates for Markov Chains Failure Rates

For a homogeneous and uniformly ergodic Markov chain, with transition kernel $P(x, A) = \int_A f(y|x)dy, x \in E \subset R^d$, we analyse some reliability measures and failure rates associated with the transition probabilities. Sufficient conditions for strong consistency are obtained for estimates based on kernel density estimators. Partially supported by CNPq, FAPDF/PRONEX and FINATEC/UnB

D.A.S. FRASER AND A.K.MD.E. SALEH
University of Toronto and Carleton University

Combining p-values from independent sources

We consider the combining of inference information from independent sources. If the full model and data information is available from each source then simplistically we would combine the models and combine the data and then analyze. Our interest however centres on the situation where the individual instances have been analyzed and simplified and replaced by inference summaries such as p-value functions or the requisites for such. Fisher(....) proposed pragmatically that each p-value be replaced by -2log(p-value) and that the sum be treeated as a chi-square variable with degrees of freedom equal to twice the number of instances and with large values as significant; this seems to handle primarily

the null distribution. In wide generality the information needed for p-values is provided by the log-likelihood say $\ell(\theta)$ and by a canonical type reparameterization say $\varphi(\theta)$; recent likelihood theory then provides routinely the p-value function for any scalar parameter of inference. For combining the information we would of course add the log-likelihood functions but this makes available only first order inference. We show that third order inference is then available by taking an appropriate weighted sum of the individual ϕ functions and that the weights are provided by a simple function of the parameter evaluted at the overall maximum likelihood value. Examples are provided.

MALAY GHOSH
University of Florida

Nonparametric Classification For High Dimension Low Sample Size Problems

The paper introduces a new method for classification in multivariate problems where the dimension of the vector is much larger than the sample size. The proposed method does not make any distributional assumptions, and requires only finiteness of certain moments. An oracle property of the proposed method is proved. The method has numerous applications in microarray analysis, and indeed for bioinformatics in general.

M. HALLIN, HANNA OJA AND D. PAINDAVEINE
Universite Libre de Bruxelles

Semiparametrically Efficient R-Estimation of Shape

A class of R-estimators, based on the concepts of multivariate signed ranks and the optimal rank-based tests developed in Hallin and Paindaveine (2006), is proposed for the estimation of the shape matrix of an elliptical distribution. These R-estimators are root-n consistent under any radial density g, without any moment assumptions, and semiparametrically efficient at some prespecified density f. When based on normal scores, they are uniformly more efficient than the traditional normal-theory estimator, based on empirical covariance matrices (the asymptotic normality of which moreover requires finite moments of order four), irrespective of the actual underlying elliptical density. They rely on an original rank-based version of Le Cam's one-step methodology, which avoids the unpleasant nonparametric estimation of cross-information quantities that is generally required in the context of R-estimation. Although they are not strictly equivariant, they are shown to be equivariant in a weak asymptotic sense. Simulations confirm their feasability and excellent finite-sample performances.

THE FIELDS INSTITUTE
FOR RESEARCH IN MATHEMATICAL SCIENCES

ABSTRACTS 1.2

GILES HOOKER
McGill University

Forcing Functions, Goodness of Fit and Latent Variables in Deterministic Diagnostics

Systems governed by deterministic differential equations have traditionally received little attention in statistics. Current methods for developing systems of differential equations rely either on a priori knowledge or on entirely nonparametric methods.

I present a suite of diagnostic tools for analyzing lack of fit in already-identified models. These tools are based on the estimation of unobserved forces for the proposed differential equation and I show how these may be used to identify the presence of unobserved dynamical and forcing components, and the mis-specification of model terms. The estimation of such components also implies a formal goodness of fit test and I analyze its power against a number of alternatives.

JANA JURECKOVA
Charles University in Prague

Regression rank scores in nonlinear models

Consider the nonlinear regression model

$$Yi = g(x_i, \theta) + e_i, i = 1, ..., n$$

with $x_i \in \mathbb{R}^k, \theta = (\theta_0, \theta_1, ..., \theta_p) \in \Theta$ (compact in $\mathbb{R}^p + 1$), where $g(x, \theta) = \theta_0 + g(x, \theta_1, ..., \theta_p)$ is continuous, twice differentiable in θ and monotone in components of θ. Following-up the results of Gutenbrunner and Jureckova (1992) and Jureckova and Prochazka (1994), we construct the regression rank scores for model (1), and consider their asymptotic behavior under some regularity conditions. The main application of this concept is in models with a nuisance nonlinear regression, because we can avoid estimation of the nuisance parameters. keywords: Nonlinear regression, regression quantile, regression rank scores.

H. M. KIM AND A.K.MD.EHSANES SALEH
University of Alberta and Carleton University

Improved Estimation of Regression Parameters in Measurement Error Models

The problem of simultaneous estimation of the regression parameters in a multiple regression model with measurement errors is considered. Specially, we address a situation when it is suspected, with some degree of uncertainty, that the regression parameter may be the null-vector. We propose four estimators: (i) the unrestricted estimator, (ii) the preliminary test estimator, (iii) the Stein-type estimator and (iv) the positive-rule Stein-type estimator. The properties of these estimators are studied based on asymptotic distributional risks under a sequence of local alternatives.

H. L. KOUL AND W. SONG
Michigan State University

Minimum Distance Regression Model Checking with Berkson Measurement Errors

This talk will discuss a class of minimum distance tests for fitting a parametric regression model to a regression function in the Berkson model. These tests are based on certain minimized distances between a nonparametric regression function estimator and the parametric model being fitted. The paper investigates the asymptotic normality of the null distribution of the proposed test statistics and of the corresponding minimum distance estimators under minimal conditions on the model being fitted. A simulation study shows very desirable finite sample behavior of the proposed inference procedures.

T. TONY CAI AND M. LEVINE AND M. LEVINE AND L. WANG
University of Pennsylvania and Purdue University and University of Pennsylvania

Effect of the Mean on the Variance Estimation in the Multivariate Nonparametric Regression

Variance estimation in multivariate nonparametric regression is considered and the minimax rate of convergence is derived. Our interest lies in finding the effect of the unknown mean function on the estimation of the variance function. We find that there exists the mean function smoothness cutoff such that the minimax rate of convergence for the variance estimator is completely determined by the variance function smoothness if the mean function has at least the pre-specified number of derivatives. This cutoff depends on the number of dimensions we consider. In this case it is not necessary to know the mean function to achieve the minimax rate of convergence for the variance function estimator. On the other hand, if the mean function does not have this pre-specified number of

derivatives, we have the situation where the minimax rate of convergence for the variance is completely driven by the roughness of the mean function. When the number of dimensions is high, even a relatively smooth mean function can easily become the dominant factor in determining the minimax rate of convergence.

Our results also indicate that, contrary to the common practice, it is not desirable to base the variance estimator on the residuals from an optimal estimator of the mean. Instead, it is better to use estimators of the mean with minimal bias. The higher the number of dimensions, the less mean-related bias reduction is possible.

JAERIN CHO AND BORIS LEVIT
Queen's University

On Application of Splines in Functional Estimation

As a numerical method of approximation, *splines* seem to occupy an unparalleled place in Applied Statistics. They are a favorite tool of choice in such diverse areas as *Time Series, Communication Theory, Numerical Mathematics.* There is a huge literature exclusively dealing with splines in Approximation Theory. Yet, in comparison with other nonparametric techniques, such as kernel-type estimators or wavelets, Theoretical Statistics offers little in the way of explaining why are splines doing so well.

We provide a theoretical framework allowing to establish asymptotic optimality of nonparametric regression estimates based on splines. There are several interesting questions specifically targeting splines. One is based on the folklore evidence that, in any practical situation, only *cubic splines* need to be considered. Could any theoretical justification of this principle be found? Some explanation as to why cubic splines are sufficient in most cases will be provided.

Two methods of spline approximation visibly stand out. One method is based on so-called *basic splines,* or B-*splines*; another method is using so-called *cardinal interpolating splines,* or C-*splines.* Both methods fare well, and the problem of determining their relative efficiency is not a trivial one. Our results indicate that C-*splines* are generally more efficient. There is sufficient evidence that, even for moderately large sample sizes, this method can be 50% more efficient.

F. LI
Indiana-Purdue University

Testing for the equality of two nonparametric regression functions with long memory errors

This paper discusses the problem of testing the equality of two nonparametric regression functions against two-sided alternatives for uniform design on [0, 1] with long memory moving average errors. The standard deviations and the long memory parameters are

possibly different for the two errors. The paper adapts the partial sum process idea used in the independent observations settings to construct the tests and derives their asymptotic null distributions. The paper also shows that these tests are consistent for general alternatives and obtains their limiting distributions under a sequence of local alternatives. Since the limiting null distributions of these tests are unknown, we first conducted a Monte Carlo simulation study to obtain a few selected critical values of the proposed tests. Then based on these critical values, another Monte Carlo simulation is conducted to study the finite sample level and power behavior of these tests at some alternatives. The paper also contains a simulation study that assesses the effect of estimating the nonparametric regression function on an estimate of the long memory parameter of the errors. It is observed that the estimate based on direct observations is generally preferable over the one based on the estimated nonparametric residuals.

R. CARROLL AND H. LIANG
Texas A & M and University of Rochester Medical Center

Statistical Inference in Partially Linear Models with Missing Response Variables and Error-prone Covariates

We investigated partially linear models when the response variable is sometimes missing with missing probability depending on the covariates, and the linear covariate is measured with error. We proposed a class of semi-parametric estimators for parameter of interest. The resulting estimators were shown to be consistent and asymptotically normal under general assumptions. To construct a confidence region of the parameter and to avoid estimating covariance matrix because of its complexity, we also proposed an empirical likelihood based statistic, which was shown to have an asymptotic chi-squared distribution. The proposed methods were applied to analyze an AIDS clinical trial dataset. A simulation study was also reported to illustrate our approach.

Joint work with Drs. Raymond Carroll and Suojin Wang

Z.Q. JOHN LU
National Institute of Standards and Technology

Statistical Methods for Holistic Mass Spectral Data Analysis

MALDI-TOF mass spectrometry offers a promising approach for measuring Molecular Mass Distributions (MMDs) of synthetic polymers for the first time. As being typical of many other high throughput experiments, the MMD data set from a given experiment has only a few replicates at each setting, while the dimensionality of the mass spectra ranges from 70s to hundreds. There are challenging and interesting metrology problems for statisticians such as assessment of data reproducibility and detection of effects due to instrumental setting changes. This paper reviews existing approaches in the literature and

proposes several new approaches for holistic mass spectral data analysis, including adaptation of Neymans smooth lack-of-fit tests, analysis of diversity measures using entropy and differential metrics, and component analysis using singular value decomposition.

DAVY PAINDAVEINE AND M. HALLIN
Universite Libre de Bruxelles

Optimal Rank-Based Tests for Homogeneity of Scatter

We propose a class of locally and asymptotically optimal tests, based on multivariate ranks and signs, for the homogeneity of scatter matrices in m elliptical populations. Contrary to the existing parametric procedures, these tests remain valid without any moment assumptions, and thus are perfectly robust against heavy- tailed distributions (validity robustness). Nevertheless, they reach semiparametric efficiency bounds at correctly specified densities (efficiency robustness). They are also affine-invariant. We compute local powers and asymptotic relative efficiencies of the proposed tests with respect to the Schott (2001) pseudo-Gaussian test, which actually is a robustified version of the traditional Gaussian likelihood ratio test. As we show, the normal-score version of our tests outperforms Schott's test in most cases.

M. PANDEY
Banaras Hindu University

Ganga Water Quality in Varanasi : An ARIMA Model

Varanasi , one of the oldest living cities of the world, situated on the north bank of river Ganga, is a religious and cultural capital of India as well as a seat of learning of world fame. From time immemorial Hindus regard the Ganga as the holiest of all the rivers with spiritual power of purifying not only body but redeeming soul of all sins if ever committed. However, the fast growing urbanization, industrialization, green revolution with chemical fertilizers & pesticides, and armament race and related factors have rendered the water of sacred Ganga unfit even for bathing though drinking it is a part of worship. Clean Ganga movement was launched in Varanasi in early 1980s whereas, the Govt. of India initiated Ganga Action Plan in 1985 to clean the holy river, which was scheduled to be completed by December 2001, with no apparent improvement, according to some experts.

Continuous monitoring of Ganga water quality parameters is undertaken at Govt. level by Central Water Commission(CWC) along with many other water bodies also. Central region Office of CWC at Varanasi maintains this data for Varanasi apart from some voluntary Organizations, such as Swatcha Ganga Research Lab. Reports are published with this data but without proper statistical analysis and interpretation. Secondary data on 21 parameters have been collected from CWC Varanasi office, for the last 20 years with the objective of an in depth analysis to find out an effective way of expressing water

quality level. A water quality index for river Ganga has been obtained, based on rating curves on five parameters. The indices of Ganga water quality obtained from 1995 to 1999 indicate a deterioration. Using CWC data, this paper also attempts to provide an ARIMA model of the Ganga river water quality and to predict its status on the basis of this model.

MADAN L. PURI
Indiana University

Conditional U-Statistics with Applications in Discriminant Analysis, ARMA Processes and Hidden Markov Models

Stute (Ann. Probab. (1991), Ann. Statist. (1994)) introduced a class of conditional U-statistics which generalize the Nadaraya-Watson estimate of a regression function. Under the usual iid set-up, Stute proved the asymptotic normality, weak and strong consistency and the universal consistency of the estimate in the rth mean. Here we extend Stutes results from the independent case to the dependent case. Applications to discriminant analysis, ARMA processes and hidden Markov models are provided. The work is in collaboration with Professor Michel Harel (C.N.R.S. Toulouse, France).

TIMOTHY RAMSAY
University of Ottawa

Concurvity and bias in semiparametric additive models

The generalized additive model (GAM) offers the ability to fit extremely flexible models with minimal prior assumptions concerning the nature of individual additive components, but the resulting models can be difficult to interpret. Semiparametric models are an increasingly popular alternative whereby a covariate of interest in a GAM is constrained to be linear, thus yielding an easily interpretable effect estimate while still flexibly controlling for confounding by nuisance covariates. These semiparametric models have the potentially dangerous, and largely unknown, property that the linear coefficient is asymptotically biased. This talk will discuss this bias and how it arises from concurvity (nonparametric collinearity) in the data. A prominent applied example from the environmental epidemiology literature will demonstrate that the bias can have serious practical consequences.

J. N. K. RAO
Carleton University

Some Recent Advances in Linear and Generalized Linear Mixed Models With Applications to Small Area Estimation

In this talk I will present some recent advances in the estimation of fixed and random effects under linear and generalized linear mixed models. In particular, I will consider EBLUP and EB estimators and estimation of associated variability using Taylor linearization, jackknife and bootstrap methods as well as confidence interval estimation. I will also present some applications of the results to small area estimation.

WATARU SAKAMOTO
Osaka University

Selecting basis and knots in MARS with an empirical Bayes method

The multivariate adaptive regression spline (MARS), proposed by Friedman (1991), estimates regression structure including interaction terms adaptively with truncated power spline basis functions. However, it adopts the generalized cross-validation criterion to add and prune basis functions, and hence it tends to choose such large numbers of basis functions that estimated regression structure may not be easily interpreted. On the other hand, some Bayesian approaches incorporated in MARS have been proposed, in which the reversible jump MCMC algorithm is adopted. However, they generate enormous combinations of basis functions, from which it would be difficult to obtain clear interpretation on regression structure. An empirical Bayes method to select basis functions and knots in MARS is proposed, with taking both advantages of the frequentist model selection approach and the Bayesian approach. The penalized likelihood is maximized to estimate regression coefficients for given basis functions, and a Laplace approximation of the marginal likelihood is maximized to select knots and variables involved in basis functions. Moreover, the Akaike Bayes information criterion (ABIC) is used to determine the number of basis functions. It is shown that the proposed method gives estimation of regression structure which is relatively parsimonious and easy to interpret for some example data sets. Keywords: Akaike Bayes information criterion, estimation of interaction terms, marginal likelihood, multivariate adaptive regression, penalized likelihood approach.

ARUSHARKA SEN AND FANG TAN
Concordia University

Cure-rate estimation under Case-1 interval censoring

Estimation of cure-rate, or probability of long-term survival, has received considerable attention lately. The issue seems to be of particular importance when data are subject to some kind of heavy censoring such as Case-1 interval censoring (current-status data). We show that the NPMLE of cure-rate under this model is non-unique when no cures are explicitly observed, but is unique and given by a max-min formula when some cases of cure are known. We also give an asymptotic confidence interval and present a simulation study.

PRANAB K. SEN
University of North Carolina at Chapel Hill

Robust Statistical Inference for High-Dimension Low Sample Size Models with Applications

In high-deminsion (K) low sample size (n) environments, often, nonlinear, inequality, order, or general shape constraints crop up in rather complex ways. As a result, likelihood principle based optimal statistiical inference procedures either may not be in a closed, manageable form or may not even exist. While some of these complex statistical inference problems can be treated in suitable asymptotic setups, the curse of dimensionality (that is, K ¿ ¿ n , with n possibly small) calls for a different asymptotics route (wherein K is large) having possibly different perspectives. Roy's (1953) union-intersection principle (having genesis in the likelihood principle) provides some alternative approaches which are generally more amenable for the K ¿ ¿ n environment. This scenario is appraised with a few important applications in neuronal spike-train models and genomics. In the later context, there may be a very large number of genes with plausible dependence and heterogeneity amidst a small sample size (sequences), thus creating impasses for standard robust statistical inference. These statistical perspectives are appraised in some nonstandard ways.

ROBERT SERFLING
University of Texas at Dallas

Nonparametric Multivarate Outlier Detection via Depth and Quantile Functions

In recent years the study of multivariate quantile functions, depth functions, rank functions, and outlyingness functions has followed somewhat differing lines of pursuit. These separate developments will be reviewed and it will be seen that they may be viewed in a unified framework as equivalent. Within this setting, a general approach for nonparametric multivariate outlier detection will be described. It competes, for example, with popular methods based on the Mahalanobis distance, which however imposes elliptical outlyingness contours. Robustness of the nonparametric outlier identifiers will be characterized in terms of masking and swamping breakdown points and influence functions. Outlier identification methods based on the halfspace, simplicial, spatial, and projection depth functions will be characterized and compared. Besides applications in multivariate data analysis, also discussed will be the extension of the ideas to outlier detection and handling in shape fitting problems in computational geometry.

B. K. SINHA
University of Maryland-Baltimore County

Generalized P=values with applications

Generalized P-values were introduced by Weerahandi (1989)to deal with what are known as nonstandard testing problems. Such testing problems include a) inference for a variance component, b) inference for the ratio of variance components, c) Behrens-Fisher problem, d)reliability function, e)difference of two exponential means, f) ANOVA problem in presence of error heterogeneity, and so on. In this talk this concept will be explained and illustrated with a few of the above examples.

WINFRIED STUTE, LIUGEN XUE AND LIXING ZHU
University of Giessen

Empirical Likelihood Inference In Nonlinear Errors-In-Covariables Models With Validation Data

In this talk we study inference in parametric-nonparametric errors-in- covariables regression models using an empirical likelihood approach based on validation data. It is shown that the asymptotic behavior of the pro- posed estimator depends on the ratio of the sizes of the primary and the validation sample, respectively. Unlike cases without measurement errors, the limit distribution of the estimator is no longer tractable and cannot be used for constructing condence regions. Monte Carlo approximations are employed to simulate the limit distribution. To increase the coverage accuracy of condence regions, two adjusted empirical likelihood estima- tors are recommended which in the limit have a standard chi-squared distribution. A simulation study is carried out to compare the proposed methods with other existing methods. The new methods outperform the least-squares method, and one of them works better than SIMEX even when the restrictive model assumptions needed for SIMEX are satised. An application to a real data set illustrates our new approach.

A. K. MD. EHSANES SALEH AND ZHENGMIN ZHANG
Carleton University

A Decomposition Method for Minimizing Neyman χ^2 Distance in Finite Mixture Models

n finite mixture models, we study a geometric decomposition method of minimizing Neyman χ^2 distance which is equivalent to algebric method.

Neyman proved that the minimization procedure of Neyman χ^2 distance leads to a BANE (Best Asymptotically Normal Estimate).

Saleh proposed that for finite mixtures Neyman χ^2 distance can be decomposed into two parts and one of them depends only on component parameters, to minimize Neyman χ^2 distance is essentially to minimize this part. Consequently our decomposition method makes Neyman theorem having a special form for finite mixture models.

THE FIELDS INSTITUTE
FOR RESEARCH IN MATHEMATICAL SCIENCES

ABSTRACTS 1.2

MUNI S. SRIVASTAVA
University of Toronto

A Review Of Multivariate Theory For High Dimensional Data With Fewer Observations