Book of Abstracts
2007 Nonparametric Statistics Conference
"Current and Future Trends in Nonparametrics"
University of South Carolina, Columbia
October 10–12, 2007

Prepared By:
Organizing Committee

Sponsored by:

Department of Statistics, University of South Carolina
Academic Enrichment and Conferences, University of South Carolina
Section on Nonparametric Statistics of the ASA
National Institutes of Statistical Science (NISS)

Partial funding provided by the
US National Science Foundation

Session Number	02 (Plenary Session I)
Session Title	Nonparametrics: Perspectives and the Task Ahead
Time, Day	8:45–9:45AM, October 11, Thursday
Place	Lecture Room 1 (Lumpkin Auditorium)

Nonparametrics: Perspectives and Prospectives

Pranab K. Sen

University of North Carolina, Chapel Hill, NC 27599, USA

E-Mail: pksen@bios.unc.edu

Abstract: Stemming from distribution-free statistics and 'quick and dirty methods,' Nonparametrics has been going through an evolution which is far from being over at this juncture of time. The perspectives over the past 70 years provide good incentives to gauge the nature of changes in prospectives, and our task ahead to cope with some of these challenging statistical issues. Interdisciplinary research, specially arising in health sciences, large biological systems, genomics as well as environmental toxicologic studies have a greater appeal for nonparametrics than parametrics or even the semiparametrics. Even Bayesian methods amend well in nonparametrics, albeit amidst computational complexities. Such computational impasses are also in high-dimensional low sample size problems where parametrics can be quite misleading. Knowledge discovery and data mining tools are gaining popularity from application point of view, and it would be our mission to provide more adequate statistical interpretation and theoretical justification for such computational tools. A notable case of stochastic ordering in a microarray data model is focused as an illustration and some of the related nonparametrics disseminated.

Session Number	03
Session Title	Recent Developments in Semiparametrics
Time, Day	10:00–12:00noon, October 11, Thursday
Place	Lecture Room 1 (Lumpkin Auditorium)

Efficient linear regression for time-to-event data under length-biased sampling

Michael R. Kosorok[1]*, Rajat Mukherjee[2], and Jason P. Fine[3]

[1]University of North Carolina-Chapel Hill

[2]Public Health Foundation of India

[3]University of Wisconsin-Madison

E-Mail: kosorok@unc.edu

Abstract: In this talk, we develop both efficient and inefficient methods of inference for accelerated failure time regression under length-biased sampling and right censoring. This kind of data arises in both prevalent cohort and current duration study designs. An HIV sero-prevalence study can have a prevalent cohort design if the status but not the time of HIV infection is known at a fixed cross-sectional sampling time, after which patients are followed until they acquire AIDS or the study is censored. Certain time-to-pregnancy studies have a current duration design if time of initiation of unprotected intercourse is known but time to successful pregnancy is censored at a fixed cross-sectional sampling time. In both cases, there exists a bias in the sampling of the event time (the time between the initiating and terminal events) because the probability of observing the patient is proportional to the magnitude of the event time. This biased sampling structure induces a dependence between the covariates and the event time. We obtain the surprising result that this dependency can actually be ignored without decreasing the efficiency of the regression parameter estimator. We apply the proposed procedures to data from a French Fecundity study and present several simulation results for illustration.

Generalized self-consistency algorithms for mixture models

Alex Tsodikov
University of Michigan, Department of Biostatistics
E-Mail: tsodikov@umich.edu

Abstract: A generalization of the concept of mixture, self-consistency, expectation and imputation and associated Quasi-EM algorithms is presented and applied to multinomial logistic model, a family of univariate survival models, and multivariate survival models motivated by frailties. A subclass of Archimedian copula models is identified that is characterized by monotonically convergent Quasi-EM algorithms. A connection to recently proposed MM algorithms that extend the EM concept without using missing data arguments is established.

A New Class of Semiparametric Copula Mixture Cure Models: Analysis of SEER Prostate Cancer Data

Yi Li
School of Public Health, Harvard University
E-Mail: yili@hsph.harvard.edu

Abstract: This talk discusses cure detection among the prostate cancer patients in the NIH Surveillance Epidemiology and End Results (SEER) program, wherein the main endpoint (e.g. deaths from prostate cancer) and the censoring causes (e.g. deaths from heart diseases) may be dependent. While a number of authors have studied the mixture survival model to analyze survival data with nonnegligible cure fractions, none has studied the mixture cure model in the presence of dependent censoring. To account for such dependence, we propose a more general class of cure models that allow for dependent censoring. We derive the cure models from the perspective of competing risks and model the dependence between the censoring time and the survival time using a class of Archimedean copula models. Within this framework, we consider the parameter estimation, the cure detection, and the two- sample comparison of latency distributions in the presence of dependent censoring when a proportion of patients is deemed cured. Large sample results using the martingale theory are obtained. We examine the finite sample performance of the proposed methods via simulation and apply them to analyze the SEER prostate cancer data.

Asymptotic Theory for the Semiparametric Accelerated Failure Time Model with Missing Data

Bin Nan[1*], John D. Kalbfleisch[1], and Menggang Yu[2]
[1]University of Michigan
[2]Indiana University
E-Mail: bnan@umich.edu

Abstract: We consider a class of doubly weighted rank based estimating methods for the accelerated failure time model with missing data, for example case-cohort studies, where weights may not be predictable in a stochastic process formulation. We treat the general problem as a semiparametric estimating equation problem and provide proofs of asymptotic properties for the weighted estimators, with either true weights or estimated weights, by using the empirical process theory where the martingale theory may fail. Simulations show that the outcome-dependent weighted method works well for finite samples in case-cohort studies and improves efficiency comparing to the method using predictable weights, and that the method is even more efficient when estimated weights are used, which is commonly observed in the missing data literature.

Session Number	04
Session Title	Nonparametrics & Bioinformatics
Time, Day	10:00–12:00noon, October 11, Thursday
Place	Lecture Room 2 (Room 855)

Bayesian Clustering with Regression

Peter Mueller*, **Fernando Quintana, and Gary Rosner**
M.D. Anderson Cancer Center, Pontifica Univ Catolica de Chile
E-Mail: pm@wotan.mdacc.tmc.edu

Abstract: We propose a model for covariate-dependent clustering, i.e., we develop a probability model for random partitions that is indexed by covariates. The motivating application is inference for a clinical trial. As part of the desired inference we wish to define clusters of patients. Defining a prior probability model for cluster memberships should include a regression on patient baseline covariates. We build on product partition models (PPM). We define an extension of the PPM to include the desired regression. This is achieved by including in the cohesion function a new factor that increases the probability of experimental units with similar covariates to be included in the same cluster.

We discuss implementations suitable for continuous, categorical, count and ordinal covariates.

Applications of Wavelet-Based Functional Mixed Models to Proteomics and Genomics Data

Jeffrey S. Morris
Department of Biostatistics, The University of Texas MD Anderson Cancer Center
E-Mail: jefmorris@manderson.org

Abstract: Various genomic and proteomic assays yield high dimensional, irregular functional data. For example, MALDI-MS yields proteomics data consisting of one-dimensional spectra with many peaks, 2D gel electrophoresis and LC-MS yield two-dimensional images with spots that correspond to peptides present in the sample, and array CGH or SNP chip arrays yield one-dimensional functions of copy number information along the genome. In this talk, I will discuss how to identify candidate biomarkers for various types of proteomic and genomic data using Bayesian wavelet-based functional mixed models. This approach models the functions in their entirety, so avoid reliance on peak or spot detection methods. The flexibility of this framework in modeling nonparametric fixed and random effect functions enables it to model the effects of multiple factors simultaneously, allowing one to perform inference on multiple factors of interest using the same model fit, while adjusting for clinical for experimental covariates that may affect both the intensities

and locations of the peaks and spots in the data. I will demonstrate how to identify regions of the functions that are differentially expressed across experimental conditions, in a way that takes both statistical and clinical significance into account and controls the Bayesian false discovery rate to a pre-specified level. Time allowing, I will also demonstrate how to use this framework as the basis for classifying future samples based on their proteomic smf genomic profiles in a way that can also combine information across multiple sources of data, including proteomic, genomic, and clinical. These methods will be applied to a series of proteomic and genomic data sets from cancer-related studies.

Variable selection in penalized model-based clustering via regularization on grouped parameters

Benhuai Xie[1], Wei Pan*[1], Xiaotong Shen[2]
[1]Division of Biostatistics, School of Public Health, University of Minnesota
[2]School of Statistics, University of Minnesota
E-Mail: weip@biostat.umn.edu

Abstract: Penalized model-based clustering has been proposed for high-dimensional but small sample-sized data, such as arising from genomic studies; in particular, it can be used for variable selection. A new regularization scheme is proposed to group together multiple parameters of the same variable across clusters, which is shown both analytically and numerically to be more effective than the conventional L_1 penalty for variable selection. In addition, we develop a strategy to combine this grouping scheme with grouping structured variables. Simulation studies and applications to microarray gene expression data for cancer subtype discovery demonstrate the advantage of the new proposal over several existing approaches.

Bayesian Kernel Models: Theory and Applications

Sayan Mukherjee
Department of Statistical Science
Institue for Genome Sciences & Policy
Department of Computer Science
Duke University
E-Mail: sayan@stat.duke.edu

Abstract: Kernel methods have been very popular in the machine learning literature in the last ten years, often in the context of Tikhonov regularization algorithms. I will introduce a coherent Bayesian kernel model based on an integral operator whose domain is a space of signed measures. Priors on the signed measures induce prior distributions on their image functions under the integral operator. I will identify general classes of measures whose images are dense in the reproducing kernel Hilbert space (RKHS) induced by the kernel. This gives a function-theoretic foundation for some nonparametric prior specifications commonly-used in Bayesian modeling, including Gaussian processes and Dirichlet processes, and suggests generalizations. A general framework for the construction of priors on signed measures using Levy processes is described.

An application of this model to high-dimensional gene expression data will illustrate how this Bayesian kernel model can be used to illustrate the “when, why, and how” the incorporation of unlabelled data, semi-supervised learning, helps in predictive regression models. This talk is based upon the following papers:

http://ftp.stat.duke.edu/WorkingPapers/06-18.html
http://www.imstat.org/sts/future_papers.html

Session Number	05
Session Title	Nonparametrics: General & Pedagogical
Time, Day	10:00-12:00noon, October 11, Thursday
Place	Lecture Room 3 (Room 856)

A Century of Nonparametrics

A. K. Md. Ehsanes Saleh

School of Mathematics & Statistics, Carleton University, Canada

E-Mail: Saleh3422@rogers.com; esaleh@math.carleton.ca

Abstract: In this talk the author discusses the developments of nonparametric statistics from its infancy to modern methods through eyes of the text books that came about through the century. Some key mathematical techniques will be discussed enriching the *adhoc* methods. This discussion will help the new comers to see that simple ideas may lead to profound developments of modern times. Books are of historic value to the development of a subject.

Bootstrap Methods: Recent Advances and New Applications

Michael R. Chernick

United BioSource Corporation, Biotechnology Solutions

E-Mail: michael.chernick@unitedbiosource.com

Abstract: Since Efron's groundbreaking paper in the Annals of Statistics 1979 there has been a flood of research on the bootstrap theory and applications. In the 1980s the theory developed and there was an exponential growth in the number of journal articles on the topic. In the 1990s many important real world applications followed. The results have been published in a number of textbooks. Most notably Hall (1992), Efron and Tibshirani (1993), Davison and Hinkley (1997), Shao and Tu (1995) and Chernick (1999). The developments since 2000 have not been as extensively covered with the text Lahiri (2003) the main reference, but it emphasizes the theoretical developments with dependent data. In this presentation I will survey the developments of the years from the mid 1990s to the present where the theoretical difficulties of inconsistency of bootstrap estimates for certain estimation problems have been overcome by modifications to the "naive" bootstrap with techniques such as the m-out-of-n bootstrap. I will also highlight some biostatistical applications including testing for individual bioequivalence which was successful enough to appear as a recommended method by the FDA in a 1997 guidance document.

References:

1. Chernick, M. R. (1999). *Bootstrap Methods: A Practitioner's Guide.* Wiley, New York.
2. Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications.* Cambridge University Press, Cambridge UK.
3. Efron, B. (1979). *Bootstrap Methods: Another Look at the Jackknife.* Annals of Statistics 7, 1-26.
4. Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman & Hall, New York.
5. Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, Inc., New York.

6. Lahiri, S. N. (2003). *Resampling Methods for Dependent Data.* Springer-Verlag, Inc. New York.
7. Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap.* Springer-Verlag, Inc., New York.

Topics for a Nonparametric Statistics Course

Paul Kvam

Industrial and Systems Engineering, Georgia Tech

E-Mail: pkvam@isye.gatech.edu

Abstract: In the last thirty years, the topics of Nonparametric Statistics have evolved dramatically. There is a traditional set of topics, such as rank tests, that seem to serve as a basis for an introductory course in nonparametric statistics. But what topics really deserve to be under the umbrella of nonparametrics when teaching a graduate level course? We will review current controversies along with some survey results from our peers in the teaching and research fields. Discussion from the audience is encouraged.

Gaining Market Share for Nonparametric Statistics Ideas

Michael J. Schell

Moffitt Cancer Center, University of South Florida

E-Mail: Michael.Schell@moffitt.org

Abstract: Many nonparametric and semiparametric tests are not commanding the market share in applied usage that they should have. Examples of this are the Wilcoxon rank sum test, the normal scores test, the Kruskal-Wallis test, isotonic regression, and quantile regression. On the other hand, the Cox proportional hazards model and Kaplan-Meier curves dominate the survival analysis arena. In this talk, I will present estimates of the market share for these and other tests, speculate as to why this is the case, and suggest how and why statistical "tool builders" should fight for market share.

Session Number	06
Session Title	Nonparametric Constrained Inference
Time, Day	10:00–12:00noon, October 11, Thursday
Place	Lecture Room 4 (Room K-J)

Estimation of a convex support function using regression splines

Mary C. Meyer* and Johan Seger

Department of Statistics, Colorado State University

Université Catholique de Louvain, Belgium

E-Mail: marycmeyer@gmail.com

Abstract: Estimation of a convex set given noisy measurements of the support function is accomplished using regression splines with shape constraints. The problem in formulated in terms of a projection onto a convex cone and rates of convergence are obtained. A test statistic is derived for the null hypothesis that the convex set is a circle, and it is shown that its distribution under the null is that of a mixture of beta random variables.

Area based testing under likelihood ratio ordering

Christopher Carolan* **and Joshua M. Tebbs**
Department of Mathematics, East Carolina University
Department of Statistics, University of South Carolina
E-Mail: carolanc@mail.ecu.edu; tebbs@stat.sc.edu

Abstract: The ordinal dominance curve formed by the one sample and two sample MLE estimators of two cumulative distribution functions F and G under likelihood ratio ordering possesses a very appealing geometric interpretation in terms of the ordinal dominance curve formed by the traditional empirical estimates of F and G. This geometric interpretation, a least concave majorant, suggests some natural testing procedures. The Mann-Whitney test statistic is the area under the ordinal dominance curve formed by the traditional empirical estimates of F and G. Under a likelihood ratio ordering constraint, a natural test statistic is the area under the restricted or concave ordinal dominance curve. Tests are suggested for testing equality of F and G versus likelihood ratio ordering and likelihood ratio ordering versus all alternatives.

Peakedness and peakedness ordering in symmetric distributions

Hammou El Barmi* **and Hari Mukerjee**
Zicklin School of Business, Baruch College
Department of Mathematics and Statistics, Wichita State University
E-Mail: hammou_elbarmi@baruch.cuny.edu

Abstract: There are many ways to measure the dispersion of a random variable. One such method uses the concept of peakedness. If the random variable X is symmetric about a point μ, then Birnbaum (1948) defined the function $\mathcal{P}_\mu(x) = P(|X-\mu| \leq x),\ x \geq 0$, as the peakedness of X. If two random variables, X and Y, are symmetric about the points μ and ν, respectively, then X is said to be less peaked than Y, denoted by $X \leq_{pkd(\mu,\nu)} Y$, if $P(|X-\mu| \leq x) \leq P(|Y-\nu| \leq x)$ for all $x \geq 0$, i.e., $|X-\mu|$ is stochastically larger than $|Y-\nu|$. For normal distributions this is equivalent to variance ordering. Peakedness ordering can be generalized to the case where μ and ν are arbitrary points. However, in this paper we study the comparison of dispersions in two symmetric and continuous random variables using the peakedness concept where normality, and even moment assumptions are not necessary. We provide estimators of the distribution functions under the restriction of symmetry and peakedness ordering, show that they are consistent, derive the weak convergence of the estimators, compare them with the empirical estimators, and provide formulas for statistical inferences. An example is given to illustrate the theoretical results.

Statistical inference under order restrictions with applications to toxicology

Gregg E. Dinse* **and Shyamal Peddada**
National Institute of Environmental Health Sciences
E-Mail: dinse@niehs.nih.gov; peddada@niehs.nih.gov

Abstract: The US National Toxicology Program (NTP) routinely conducts carcinogenicity studies, which expose rodents to different doses of a chemical under investigation. The NTP researchers are interested in detecting trends in tumor rates across dose groups. Using tumor incidence data obtained from the current bioassay, along with control data from past experiments (known as historical control data), the NTP researchers want to determine if there is a dose-related trend in tumor incidence. An important feature of the NTP cancer bioassay data is that in addition to the risk of developing cancer, animals are subject to the risk of early death (not necessarily due to cancer). Although a formal survival-adjusted test is available for analyzing the current bioassay data, until now there did not exist a satisfactory test that formally incorporates the historical control data while adjusting for survival differences. Using order restricted inference techniques, we propose a survival-adjusted test for detecting dose-related trends in tumor incidence rates that incorporates data on historical control rates and formally accounts for variation in these rates among studies. An extensive simulation, based on a wide range of realistic situations, demonstrates that the proposed test performs well in comparison to the current NTP test, which does not incorporate historical control data. In particular, our test can aid in interpreting the occurrence of a few tumors in treated animals that are rarely seen in controls.

Session Number	07 (Plenary Talk II)
Session Title	Semiparametric Methods for Longitudinal Data
Time, Day	1:30–2:30PM, October 11, Thursday
Place	Lecture Room 1 (Lumpkin Auditorium)

New Developments of Semiparametric Modeling of Longitudinal Data

Jianqing Fan

Department of Operation Research and Financial Engineering, Princeton University, Princeton, NJ 08544

E-Mail: jqfan@Princeton.EDU

Abstract: Improving efficiency for regression coefficients and predicting trajectories of individuals are two important aspects in analysis of longitudinal data. Both involve estimation of the covariance function. Yet, challenges arise in estimating the covariance function of longitudinal data collected at irregular time points. A class of semiparametric models for the covariance function is proposed by imposing a parametric correlation structure while allowing a nonparametric variance function. A kernel estimator is developed for the estimation of the nonparametric variance function. Two methods, a quasi-likelihood approach and a minimum generalized variance method, are proposed for estimating parameters in the correlation structure. We introduce a semiparametric varying coefficient partially linear model for longitudinal data and propose an estimation procedure for model coefficients by using a profile weighted least squares approach. Sampling properties of the proposed estimation procedures are studied and asymptotic normality of the resulting estimators is established. Finite sample performance of the proposed procedures is assessed by Monte Carlo simulation studies. The proposed methodology is illustrated by an analysis of a real data example.

Session Number	08
Session Title	A Session in Honor of Pranab Kumar Sen
Time, Day	2:45–4:45PM, October 11, Thursday
Place	Lecture Room 1 (Lumpkin Auditorium)

Estimating the error distribution in multivariate heteroscedastic time series models

Mervyn J. Silvapulle

Department of Econometrics and Business Statistics, Monash University
Melbourne, Australia 3145

E-Mail: mervyn.silvapulle@buseco.monash.edu.au

Abstract: Copulas have attracted considerable interest for modelling multivariate observations and for stress testing in quantitative finance. In this paper, a semiparametric method is studied for estimating the copula parameter and the joint distribution of the error term in a class of multivariate time series models when the marginal distributions of the errors are unknown. The proposed method first obtains $\sqrt{n}$-consistent estimates of the parameters of each univariate marginal time-series, and computes the corresponding residuals. These are then used to estimate the joint distribution of the multivariate error terms, which is specified using a copula. The proposed estimator of the copula parameter of the multivariate error term is asymptotically normal, and a consistent estimator of its large sample variance is also given so that confidence intervals may be constructed. A simulation study was carried out to compare the estimators particularly when the error distributions are unknown. In this simulation study, our proposed semiparametric method performed better than the well-known parametric methods. An example on exchange rates is used to illustrate the method.

Estimating Medical Costs from a Transition Model

Joseph C. Gardiner*, Lin Liu, Zhehui Luo

Michigan State University, East Lansing, Michigan

E-Mail: jgardiner@epi.msu.edu

Abstract: Analyses of medical cost data have been proposed in a variety of settings. We describe a general longitudinal framework in which costs emanate from two streams, during sojourn in health states and in transition from one health state to another. We consider estimation of net present value for expenditures incurred over a finite time horizon from medical cost data that might be incompletely ascertained in some patients. Because patient specific demographic and clinical characteristics would influence total cost, we use a regression model to incorporate covariates. We discuss similarities and differences between our net present value estimator and other widely used estimators of total medical costs. Our model can accommodate heteroscedasticity, skewness and censoring in cost data and provides a flexible approach to analyses of health care cost.

A Bayesian Test for Excess Zeros in a Zero-inflated Power Series Distribution

A. Bhattacharya[1], Bertrand S. Clarke[2], Gauri S. Datta[1]*

[1]Department of Statistics, University of Georgia, Athens, GA 30602-1952, USA

[2]Department of Statistics, 333-6356 Agricultural Road, University of British Columbia, Vancouver, BC, V6T 1Z2, Canada

E-Mail: gaurisdatta@gmail.com

Abstract: Power series distributions form a useful subclass of a one-parameter discrete exponential families suitable for modeling count data. A zero-inflated power series distribution is a mixture of a power series distribution and a degenerate distribution at zero, with a mixing probability p for the degenerate distribution. This distribution is useful for modeling count data that may have extra zeros. One question is whether the mixture model can be reduced to the power series portion, corresponding to $p = 0$, or whether there are so many zeros in the data that zero inflation relative to the pure power series distribution must be included in the model i.e., $p \geq 0$. The problem is difficult partially because $p = 0$ is a boundary point.

Here, we present a Bayesian test for this problem based on recognizing that the parameter space can be expanded to allow p to be negative. Negative values of p are inconsistent with the interpretation of p as a mixing probability, however, they index distributions that are physically and probabilistically meaningful. We compare our Bayesian solution to two standard frequentist testing procedures and find that using a posterior probability as a test statistic has slightly higher power, on the most important ranges of the sample size n and parameter values, than the score test and likelihood ratio test in simulations. Our method also performs well on three real data sets.

Smooth estimation of mean residual life under random censoring

Yogendra P. Chaubey* **and Arusharka Sen**

Concordia University, Montreal, Quebec, Canada

E-Mail: chaubey@alcor.concordia.ca

Abstract: We propose here a smooth estimator of the mean residual life function based on randomly censored data. This is derived by smoothing the Product-Limit estimator using the Chaubey-Sen technique (Chaubey and Sen (1998)). The resulting estimator does not suffer from the boundary bias as is the case with standard kernel smoothing. The asymptotic properties of the estimator are investigated. We establish strong uniform consistency and asymptotic normality. This complements the work of Chaubey and Sen (1999) which considered a similar estimation procedure in the case of complete data. It is seen that the properties are similar, though technically more difficult to prove, to those in the complete data case with appropriate modifications due to censoring.

Session Number	09
Session Title	Nonparametric Regression and Density Estimation
Time, Day	2:45–4:45PM, October 11, Thursday
Place	Lecture Room 2 (Room 855)

Nonparametric Estimation of Distributions in Large-p, Small-n Settings

Jeffrey D. Hart

Texas A&M University

E-Mail: hart@stat.tamu.edu

Abstract: A distinguishing feature of a number of modern problems, including microarray analyses, is that the observed data consist of a large number, p, of quite small data sets. Nonparametric estimation of an error distribution common to all p data sets is very challenging in such settings, being akin to deconvolution. A canonical example of the models of interest is a random effects model in which means of data sets vary randomly, but otherwise all data are identically distributed. There are two distributions to estimate in this model: the common within-data-sets distribution, and the distribution of means. Both old and recent results on nonparametric estimators in this canonical model will be reviewed. We then consider a location-scale model in which, conditional on a data set, the data have a distribution of the form $F((x-\mu)/\sigma)$, where μ and

σ are the mean and standard deviation for the data set, and the joint distribution G of (μ, σ) across all data sets is arbitrary. We propose the use of Wolfowitz minimum distance methodology to obtain nonparametric estimates of F and G. A random search algorithm is used to approximate such estimates in the location-scale model. Simulation studies and a microarray example help to illustrate the efficacy of our approach.

SiZer Analysis for the Comparison of Regression Curves

Cheolwoo Park* and Kee-Hoon Kang

Department of Statistics, University of Georgia, Athens, GA 30602, USA

Department of Statistics, Hankuk University of Foreign Studies, Yongin 449-791, Korea

E-Mail: cpark@stat.uga.edu; khkang@hufs.ac.kr

Abstract: In this talk we introduce a graphical method for the test of the equality of two regression curves. Our method is based on SiZer (SIgnificant ZERo crossing of the differences) analysis, which is a scale-space visualization tool for statistical inferences. The proposed method does not require any specification of smoothing parameters, it offers a device to compare in a wide range of resolutions, instead. This enables us to find the differences between two curves that are really there at each resolution level. The extension of the proposed method to the comparison of more than two regression curves is also done using residual analysis. A broad simulation study is conducted to demonstrate the sample performance of the proposed tool. Applications with two real examples are also included.

Testing the Equality of Two Single-Index Models

K. B. Kulasekera

Department of Mathematical Sciences, Clemson University

E-Mail: kk@ces.clemson.edu

Abstract: Comparison of two nonparametric regression models with one dimensional covariates has been extensively discussed in the literature. Comparison problem largely remains open for completely nonparametric models for multi-dimensional covariates. We address this issue under the assumption that both models are single-index models (SIMs). We propose a test for checking the equality of the mean functions. The asymptotic normality of the test statistics is established under heteroscadastic errors and an empirical study is conducted to assess the finite-sample performance of the proposed procedure.

Modality Analysis with Higher-Order Density Estimation

Michael C. Minnotte

University of North Dakota

E-Mail: michael.minnotte@und.edu

Abstract: A number of approaches to mode detection and testing involve the intermediate use of kernel density estimation. These include Silverman's (1981) critical bandwidth test, Minnotte and Scott's (1993) mode tree, and Chaudhuri and Marron's (1999) SiZer plots. We investigate the use of higher-order (reduced bias) variants of kernel density estimates in mode testing and visualization. Fourth-order variable bandwidth density estimates appear to be especially favorable for bump hunting, while the use of fourth-order kernels

for the purpose should be strongly discouraged.

Session Number	10
Session Title	Recent Developments in Nonparametrics
Time, Day	2:45–4:45PM, October 11, Thursday
Place	Lecture Room 3 (Room 856)

Confidence Intervals for a Discrete Population Median

Denis Larocque and Ronald H. Randles*

HEC, Montreal and University of Florida

E-Mail: denis.larocque@hec.ca; rrandles@stat.ufl.edu

Abstract: In this paper, we consider the problem of constructing confidence intervals for a population median when the underlying population is discrete. We describe seven methods of assigning confidence levels to order statistic based confidence intervals, all of which are easy to implement. A simulation study shows that, with discrete populations, it is possible to obtain consistently more accurate confidence levels and shorter intervals compared to the ones reported by the classical method which is implemented in commercial software. More precisely, the best results are obtained by inverting a two-tailed sign test that properly takes into account tied observations. Some real data examples illustrate the use of these confidence intervals.

Ranked Set Sampling with Unequal Set Sizes

Douglas A. Wolfe

The Ohio State University

E-Mail: daw@stat.osu.edu

Abstract: Ranked set sampling (RSS) is a data collection technique that makes use of expert knowledge or concomitant information to rank sample units before selecting which ones to actually measure. Previously developed RSS methodology has emphasized balanced or unbalanced RSS with common set size. There will, however, be situations where the data collection opportunities present themselves naturally in the form of varying set sizes, which leads to a number of interesting questions.

1. What does it mean to have a balanced RSS in this setting and how would we go about it?
2. How might we optimize the collection of unbalanced RSS observations in such settings?
3. Even if we agree on the answers to 1 and 2, what is the optimal way to combine RSS observations collected from differing set sizes?

Even more challenging is the thought of optimizing over all three options when the data collection opportunities present themselves in a random, sequential order with differing set sizes. In this work we present the results of some preliminary work on these problems.

This is joint work with Nader Gemayel and Elizabeth Stasny.

A Test for Symmetry Using Ranked Set Samples

Reza Modarres

George Washington University

E-Mail: reza@gwu.edu

Abstract: Based on ranked set samples, we propose a sign test for the hypothesis of symmetry when the center of the distribution is assumed known (B-Test). We show that the test statistic follows a binomial distribution under the null hypothesis. We study the effect of estimating the center of symmetry by the sample mean on this test. Under the null hypothesis, the modified sign test will lose its distribution-free property. We obtain the efficacy of B-Test and compare it with the efficacy of the sign test of symmetry based on a simple random sample (S-test) and with the ranked set sample analog of the sign test (T-test). We show that for skewed distributions, B-Test is more efficient than the S-test and T-test. For shifted symmetric distributions, the B-Test is more efficient than the S-test and the T-test when the parent distributions do not have heavy tails. We assess the ability of the three tests to detect asymmetry in a Monte Carlo study.

On the Weak Convergence of Moment-Empirical Cumulative Distribution Function

Robert Mnatsakanov

Department of Statistics, P.O. Box 6330, West Viginia University, Morgantown, WV 26506, USA

E-Mail: rmnatsak@stat.wvu.edu

Abstract: In many indirect models, like convolutions, mixtures, multiplicative censoring, and biased sampling models, the moments of unobserved distributions of actual interest can be easily estimated from the transformed moments of the observed distributions. In all such models one can use the procedure which recovers a function via its moments. Some asymptotic properties of proposed moment-empirical cumulative distribution functions are derived for direct and some indirect models.

This is a joint work with Frits Ruymgaart from Department of Mathematics and Statistics, Texas Tech University.

Session Number	11
Session Title	Nonparametric Statistics on Manifolds and Applications
Time, Day	2:45–4:45PM, October 11, Thursday
Place	Lecture Room 4 (Room K-J)

Nonparametric Inference for Spaces of Shapes

Rabi Bhattacharya

The University of Arizona, Department of Mathematics

E-Mail: rabi@math.arizona.edu

Abstract: Given a finite number k of landmarks or positions on an object, the so-called k-ads, one may record the shape or image of an object whether the object is viewed in a two-dimensional section or projection, or in $3D$. The equivalence class of a k-ad invariant under translation and rigid body rotation is its size and shape, and when size is also scaled out, one obtains the shape of the k-ad. The space of k-ads is then a (Riemannian) manifold on which one can measure appropriate distances, etc. We consider nonparametric statistical inference for discriminating among different shape distributions for purposes of medical diagnostics, morphometrics, etc. Ideas on multivariate analysis and time series analysis on such non-Euclidean spaces are also presented. For $3D$ similarity (or, reflection-) shapes of k-ads, a recent methodology is presented.

A second topic for statistical inference we consider involves another notion of shape for k-ads, namely, projective shape, which is particularly appropriate in machine vision.

This talk is based on joint work with Vic Patrangenaru and Abhishek Bhattacharya.

Samples of functional data and image analysis

Frits Ruymgaart
Texas Tech University
E-Mail: h.ruymgaart@ttu.edu

Abstract: Given a sample of functional data, both a parametric and a nonparametric test is considered for testing the neighborhood hypothesis that the location of the sampled Hilbert space distribution is approximately equal to a given curve. These procedures are applied to the problem of identifying the projective shape of a planar contour.

Quantifying variation in manifolds using data depth

Rima Izem
Department of Statistics, Harvard University
E-Mail: izem@stat.harvard.edu

Abstract: Several statistical methods such as principal component analysis and analysis of variance are effective in analyzing variation in high dimensional data when the space of variation is linear. However, describing variability is much more difficult when the data varies along nonlinear modes. Simple examples of nonlinear variation in functional data are horizontal shift of curves of common shape, frequency change of acoustic signals of common shape, or lighting change in images of the same object. This presentation shows novel data depth functions that would extend data depth concepts to describe variation of multivariate data when the space of variation is a manifold or the result of nonlinear variation in the data. We propose new ways of defining depth in manifolds which respect the geometry of the support of the distribution. We illustrate these new depth measures for Riemannian manifolds of non-negative curvature.

A Moebius–Poincare Deconvolution Problem

Peter T. Kim
University of Guelph
E-Mail: pkim@uoguelph.ca

Abstract: Let $\mathbf{H}$ be the Poincare plane on which $\mathbf{SL}(2,\mathbf{R})$ acts on it by a Moebius transformation. Suppose that we have a random quantity X on $\mathbf{H}$, of which we only observe a version Y corrupted by a random Moebius transformation ε of known density f_ε on $\mathbf{SL}(2,\mathbf{R})$,

$$Y = \varepsilon X\,.$$

It is the objective of this work to propose a nonparametric deconvolution estimator for the density $f_X : \mathbf{H} \to \mathbf{R}$ of X based on the density $f_Y : \mathbf{H} \to \mathbf{R}$ of Y. The main technique will be through the use of the Helgason-Fourier transform.

Session Number	12
Session Title	Poster Session I
Time, Day	4:45–6:00PM, October 11, Thursday
Place	DMC Lobby (Set-Up Period: Morning to Noon)

Rank-based estimation for GARCH models

Beth Andrews
Department of Statistics, Northwestern University
E-Mail: bandrews@northwestern.edu

Abstract: Observed time series processes frequently appear uncorrelated, yet exhibit volatility clustering. Volatility clustering is the tendency of observations relatively small in absolute value to be followed by other small observations, and the tendency of observations relatively large in absolute value to be followed by other large observations. Hence, these series appear uncorrelated, but dependent. Nonlinear generalized autoregressive conditionally heteroskedastic (GARCH) models, which have time-dependent conditional variances, are often used to describe time series with these features. Processes exhibiting GARCH-type behavior have appeared, for example, in financial time series such as inflation rates and stock prices and, outside of finance, in speech signals, daily and monthly mean temperatures, wind speeds, and atmospheric carbon dioxide concentrations. In this paper, we use a rank-based technique to estimate GARCH model parameters. The rank (R) estimators we consider minimize the sum of mean-corrected model residuals weighted by a function of residual rank; they are similar to the R-estimators proposed by L.A. Jaeckel [Estimating regression coefficients by minimizing the dispersion of the residuals, Ann. Math. Statist. 43 (1972) 1449-1458] for estimating linear regression parameters. R-estimators are, in general, robust and relatively efficient. We show this is true in the case of GARCH parameter estimation. The estimation technique is robust because the R-estimators of GARCH model parameters are $n^{1/2}$-consistent (n represents sample size) and asymptotically normal under very mild conditions. Since the weight function can be chosen so that R-estimation has the same asymptotic efficiency as maximum likelihood estimation, the R-estimators are also relatively efficient. In addition, R-estimation dominates classical Gaussian quasi-maximum likelihood estimation with respect to both robustness and asymptotic efficiency. Simulation results for R-estimation show that the asymptotic theory is indicative of finite, large sample behavior.

Geometrical kernel selection for classification

Jeongyoun Ahn
Department of Statistics, University of Georgia
E-Mail: jyahn@uga.edu

Abstract: Kernel based classifications such as support vector machine are powerful. In practice, the choice of kernel function is an important issue. The common solution, cross-validation, is computationally expensive and suffers large sampling variability. The proposed method is derived from the geometry of the feature space embedded by the kernel function. It is empirically shown that it is robust to the sampling variability and fast in computation.

Nonparametric inference on shape spaces

Abhishek Bhattacharya
Department of Mathematics, University of Arizona
E-Mail: abhishek@math.arizona.edu

Abstract: The statistical analysis of shape distributions based on random samples is important in many areas such as morphometry, medical diagnostics, and machine vision. To measure the shape of an object, one may pick a suitable ordered set of k points or landmarks called k-ad on a two or three dimensional image of the object under consideration. The equivalence class of the k-ad identified modulo size and Euclidean motions of translation and rotation is called its similarity shape. This shape space can be given a metric tensor and hence a geodesic distance making it a Riemannian manifold. Thus, statistical analysis tools developed on general manifolds can be applied to estimate shape parameters and compare different shape distributions. Another notion of shape is the projective shape of a k-ad which is particularly appropriate in machine vision. This shape space consists of equivalence classes of k-ads invariant under all projective transformations. In this poster, I present certain recent methodologies and some new results for the statistical analysis of probability distributions on manifolds and apply them to the shape spaces.

A bootstrap resampling approach: Tests of significance on brain imaging data

Chung Chang
Department of Biostatistics, Columbia University
E-Mail: cc2240@columbia.edu

Abstract: In neuroimaging studies it is of interest to test for changes in imaging data among subjects in different groups. Testing hypotheses voxel by voxel results in a multiple comparisons problem for which solutions should take into account the spatial correlation structure inherent in the imaging. Statistical Parametric Mapping (SPM) and the permutation test have become popular in this setting but they rely on parametric and exchangeability assumptions, respectively, which are not always satisfied in practice. We propose a bootstrap approach (L1) that is free of the parametric assumptions made by SPM and also are more flexible than the permutation test. We compare the performance of the L1 method with that of SPM, the permutation test, and another bootstrap approach (L2). For the L2 method, we present sufficient conditions that ensure asymptotic control of the family-wise error rate.

Kernel-Density Estimation with Missing Data and Auxiliary Variables

Suzanne R. Dubnicka
Kansas State University
E-Mail: dubnicka@ksu.edu

Abstract: In most parametric statistical analyses, knowledge of the distribution of the response variable, or of the errors, is important. As this distribution is not typically known with certainty, one might initially construct a histogram or estimate the density of the variable of interest to gain insight regarding the distribution and its characteristics. However, when the response variable is incomplete, a histogram will only provide a representation of the distribution of the observed data. In the AIDS Clinical Trial Study protocol 175, interest lies in the difference in CD4 counts from baseline to final follow-up, but CD4 counts collected at final follow-up were incomplete. Therefore, we propose methods for estimating the density of an incomplete response variable when auxiliary data are available. Properties of the proposed density estimator will be presented. Simulation studies verify these properties and show that the proposed density estimator performs well.

Quantile Estimation for Discrete Data via Empirical Likelihood

Jien Chen* **and Nicole A. Lazar**
Department of Statistics, University of Georgia, Athens GA 30602
E-Mail: jienchen@uga.edu

Abstract: Quantile estimation for discrete distributions has not been well studied, although discrete data are common in practice. Under the assumption that data are drawn from a discrete distribution, we examine the consistency of the maximum empirical likelihood estimator (MELE) of the pth population quantile θ_p, with the assistance of a jittering method and results for continuous distributions. The MELE and the sample quantile estimator are closely related, and they may or may not be consistent for θ_p, depending on whether or not the underlying distribution has a plateau at the level of p. We propose an empirical likelihood-based categorization procedure which not only helps in determining the shape of the true distribution at level p, but also provides a way of formulating a new estimator that is consistent in any case. Analogous to confidence intervals in the continuous case, the probability of a correct estimate (PCE) accompanies the point estimator. Simulation results show that PCE can be estimated using the simple bootstrap method.

Robust semiparametric models for time-dependent longitudinal data

Daniel Gervini
Department of Mathematical Sciences, University of Wisconsin – Milwaukee
E-Mail: gervini@uwm.edu

Abstract: Longitudinal data can sometimes be seen as discrete observations of a continuous-time stochastic process. In those cases, it is of interest to estimate the mean and the covariance function (or the principal components) of the process. The most commonly used estimators are very sensitive to outliers, of which there are two types: isolated atypical observations whithin trajectories or whole outlying trajectories. Here we propose a semiparametric Student's t model for the mean and principal components of the process, and show that the resulting maximum likelihood estimators are resistant to both types of outliers. The estimators can be easily computed with a EM algorithm. As an example, we apply the proposed method to CD4-count trajectories of AIDS patients.

A semiparametric modeling approach for the development of metabonomic profile and biomarker discovery mechanism

Samiran Ghosh
Department of Mathematical Sciences, IUPUI
E-Mail: samiran@math.iupui.edu

Abstract: The discovery and validation of biomarkers is an important step towards the development of criteria for early diagnosis of disease status. Recently electrospray ionization (ESI) and matrix assisted laser desorption (MALDI) time-of-flight (TOF) mass spectrometry have been used to identify biomarkers both in proteomics and metabonomics studies. Data sets generated from such studies are generally very large in size and thus require the use of sophisticated statistical techniques to glean useful information. Most recent attempts to process these types of data model each compound's intensity either discretely by positional (mass to charge ratio) clustering or through each compounds' own intensity distribution. Traditionally data processing steps such as noise removal, background elimination and m/z alignment, are generally carried out separately resulting in unsatisfactory propagation of signals in the final model. It is more intuitive to develop models for patterns rather than discrete points following the basic principle of "borrowing strength" for such a scenario.

On control of the false discovery rate under no assumption of dependency

Wenge Guo[1*] **and M. Bhaskara Rao**[2]
[1]NIEHS
[2]Department of Environmental Health, University of Cincinnati
E-Mail: wenge.guo@gmail.edu

Abstract: Most false discovery rate (FDR) controlling procedures require certain assumptions on the joint distribution of p-values. Benjamini and Hochberg [J. Roy. Statist. Soc. Ser. B 57 (1995) 289-300] proposed a step-up procedure with critical constants $\alpha_i = \alpha(i/m)$, $1 \leq i \leq m$, for a given level $0 < \alpha < 1$ and showed that $FDR \leq \alpha(m_0/m)$ under the assumption of independence of p-values, where m is the total number of null hypotheses and m_0 the number of true null hypotheses. Benjamini and Yekutieli [Ann. Statist. 29 (2001) 1165-1188] showed that for the same procedure $FDR \leq \alpha(m_0/m)\sum_{j=1}^{m}(1/j)$, whatever may be the joint distribution of p-values. In one of the results in this paper, we show that this upper bound for FDR cannot be improved in the sense that there exists a joint distribution of p-values for which the upper bound is attained. A major thrust of this paper is to work in the realm of step-down procedures without imposing any condition on the joint distribution of the underlying p-values. As a starting point, we give an explicit expression for FDR specially tailored for step-down procedures. Using the same critical constants as those of the Benjamini-Hochberg procedure, we present a new step-down procedure for which the upper bound for FDR is much lower than what is given by Benjamini and Yekutieli. The explicit expression given for FDR and some optimization techniques stemming from the knapsack problem are instrumental in getting the main result. We also present some general results on stepwise procedures built on non-decreasing sequences of critical constants.

Nonparametric functional classification with sparsity and feature selection

Woncheol Jang
Department of Epidemiology and Biostatistics, University of Georgia
E-Mail: jang@uga.edu

Abstract: With the power of modern technology, it is increasingly common to collect functional data in scientific studies that are beyond the capabilities of traditional statistical methods. Here data are considered as functional if data are regularly measured on a fine grid. Dimension reduction or feature selection is a key issue to make statistical methods most effective in functional data such as profiles and curves and estimating profiles can be considered as making inferences for infinite dimensional objects. We develop a new classification method for functional data which we shall call CARDS, standing for Classification After Reduction of Dimension with Smoothed clipped absolute deviation penalty. This proposed method is novel because it

can be used for both prediction and feature selection by achieving sparsity at each step. More specifically, we want to keep as many nonzero coefficients as possible in the function estimation step yet with a spare representation and achieving sparsity in the classification step for the classification purpose. Examples in proteomics will be presented.

False discovery rate in microarray studies

Moonsu Kang* and Pranab K. Sen

Department of Biostatistics, University of North Carolina

Department of Biostatistics and Department of Statistics and Operations Research, University of North Carolina

E-Mail: mkang@email.unc.edu

Abstract: An important aim of DNA microarrays is the identification of differentially expressed genes affected by explanatory variables or covariates of interest: Multiple testing for each gene. The traditional approach to this multiplicity problem controls the familywise error rate(FWER). Control of the FWER is unduly conservative in analyzing large-screening microarray data. The false discovery rate(FDR) may be a promising alternative to the FWER which aims to control the expected proportion of Type I errors among the rejected hypotheses. Many FDR controlling procedures control the FDR under independence or restrictive dependence structures, resulting in an unreliable FDR estimation. For these problems, two-stage FDR controlling procedure under suitable dependence structures are proposed here. This proposed testing procedure is based on a Poisson distributional approximation. It eliminates the need to consider specific dependence structures of genes. We compare the performance of two-stage FDR controlling procedure with that of other FDR controlling procedures. The procedures are studied using Leukemia study of Golub et al. and simulated data. In these examples, proposed FDR procedure has greater power without much elevation of FDR.

Simple location models for microarray data

John Kloke and Patrick Kimes

Department of Mathematics, Pomona College

E-Mail: John.Kloke@pomona.edu

Abstract: Analysis of microarray experiments has become an important component of biological research. Even though much work has gone into normalization of microarray data, the expression values for many of the genes do not follow a normal distribution. For the genes with heavier tailed distributions, using Wilcoxon or sign scores results in more efficient use of the microarrays. Using simple data-based simulation techniques applied to a null microarray dataset we explore efficiency and power for the one and two sample location problems.

Semiparametric Least Squares ROC Analysis of Correlated Biomarker Data

Liansheng (Larry) Tang

Department of Statistics, George Mason University

E-Mail: ltang1@gmu.edu

Abstract: The receiver operating characteristic (ROC) curve is a popular tool to evaluate the accuracy of continuous-scale biomarkers. Least squares methods have been proposed in the literature by various authors. These methods either estimate a single parametric ROC curve of a single biomarker or several parametric ROC curves from independent biomarker data. In the paper we propose a semiparametric least squares method to estimate ROC curves from correlated biomarker data. Our new method has several advantages over existing ROC methods. First, unlike most existing methods our method does not require iterations and is simple to implement. Second, our method allows unknown baseline functions. Third, our method includes interaction terms between discrete covariates and false positive rates. Such interaction terms are important in the ROC modeling because without them the ROC curves of biomarkers are not allowed to intersect. In addition, based on our semiparametric method we propose a separation curve method to identify the range of false positive rates for which two ROC curves differ or one ROC curve is superior to the other. We compared the finite sample performance of the newly proposed semiparametric method with a parametric least squares method and a semiparametric method in large-scale simulation studies. Finally, our method is illustrated through two real life examples.

Session Number	13 (Plenary Talk III)
Session Title	Rank-Based Methods in Regression
Time, Day	8:35–9:45AM, October 12, Friday
Place	Lecture Room 1 (Lumpkin Auditorium)

Ranks and quantiles in regression and autoregressive models

Jana Jurečková

Charles University in Prague, Czech Republic

E-Mail: jurecko@karlin.mff.cuni.cz

Abstract: Ranks and quantiles are important tools in the nonparametric statistical inference. Ranks are very convenient for testing, while the quantiles/order statistics are appealing for estimation. Natural is the idea to extend these entities to more general models, preserving their advantages to the largest possible extent. The notion of sample quantiles was first extended to linear models by Koenker and Bassett under the name *regression quantiles* (RQ). They formulated the RQ as a solution of a special parametric linear programming problem, and used the dual linear program as a computational device. Gutenbrunner and Jurečková found that the optimal dual solutions share some properties with the vector of ranks, thus called them the *regression rank scores* (RRS). The main advantages of RQ and RRS are the equivariance of the former and the invariance of the latter, what is not true for the residuals from an estimator and their ranks. This naturally led to the L-estimation in the linear regression model based on the RQ's and to the extension of the rank tests in models with the nuisance linear regression based on the RRS. Both RQ and RRS were extended to the linear autoregressive model by Koul and Saleh.

The regression quantiles were extended to some nonlinear regression models by Jurečková and Procházka. The RRQ now form a basic tool in econometrics, where they became a part of a general consciousness and are independently developing in various directions, suitable for econometric problems. *Quantile regression* is an often subject of econometric conferences and workshops.

We shall describe some applications of the RRS in testing problems with nuisance regression (including nonlinear) and autoregression, without estimating the nuisance parameters. Surprisingly, the tests work well numerically even under distributions for which we do not yet have a rigorous proof. The studentizing scale statistics based on the RRS enjoy a desirable invariance with respect to the regression. The close relation of the ranks and quantiles extends to a multiple setup: The two-step regression quantile, estimating the slope components separately with the aid of a suitable rank-estimate and then ordering the residuals to get

the intercept component, is surprisingly close to the ordinary regression quantile, while it well illustrates its structure. It is even more appealing in the case of the *extreme regression quantile*, where the extreme two-step and the ordinary extreme regression quantiles coincide. For the Gumbel domain of attraction and some other distribution tails, the asymptotic distribution of the intercept component of the extreme regression quantile depends on the design matrix only through the diagonal elements of the hat matrix.

There are still many appealing open questions, and we believe that these concepts will still find many interesting applications.

Session Number	14
Session Title	Rank-Based Methods
Time, Day	10:00–12:00noon, October 12, Friday
Place	Lecture Room 1 (Lumpkin Auditorium)

Semi- and Non-parametric Mixture Models

T. P. Hettmansperger*, Tatiana Benaglia, Tracey Wrobel
Department of Statistics, Penn State University
E-Mail: tph@stat.psu.edu

Abstract: We will review the current status of fitting semi- or non-parametric mixture models to multivariate or univariate data. Past difficulties include identifiability and computability. We will discuss and illustrate an approach that seems to overcome these difficulties.

Rank and Sign Methods for Clustered and Multilevel Data

Denis Larocque*, Jaakko Nevalainen, Hannu Oja
Department of Management Sciences
HEC Montréal, 3000, chemin de la Côte-Sainte-Catherine,
Montréal (Québec), Canada

Tampere School of Public Health,
33014 University of Tampere, Finland

Tampere School of Public Health,
33014 University of Tampere, Finland

E-Mail: denis.larocque@hec.ca; jaakko.nevalainen@uta.fi; Hannu.Oja@uta.fi

Abstract: Situations involving cluster correlated data are often encountered in practice. A typical example is children clustered into schools. It is usually assumed that observations are possibly correlated within clusters and independent across clusters. Generalizations of classical rank and sign based methods for the case of clustered data have appeared in the literature recently. In this talk, I will provide a brief review on some of this work by focusing mainly on the one-sample estimation and testing problem for univariate and multivariate data. I will then present more details on some recent development dealing with multilevel (or hierarchical) data.

Comparisons of Rank-Based Estimators for The Fixed Effects of a Mixed Model

Joseph W. McKean*[1], **Asheber Abebe**[2], **John D. Kloke**[3], **M. Mushfiqur Rashid**[4]
[1]Western Michigan University, [2]Auburn University, [3]Pomona College, [4]FDA
E-Mail: joseph.mckean@wmich.edu; abebeas@auburn.edu; johnkloke@gmail.com;
mushfiqur.rashid@fda.hhs.gov

Abstract: Several recently proposed rank-based estimators of the fixed effects in a mixed model are discussed. These include the R estimator for the independent error case, where the associated standard errors are corrected for the dependencies in the model. Based on its residuals, robust, practical estimates of the variance components are obtained, which in turn determine in the usual way generalized estimators for the fixed effects. Discussion is based on general score functions and highly efficient estimators, but high breakdown estimators can easily be obtained. A third estimator is based on an iterated reweighted least squares estimator for the general linear and nonlinear models. The weights can be chosen to bound the influence in both the response and factor spaces. When the random effect is a grouping (block) factor, R estimators based on separate (group) rankings of residuals are also a possibility. Illustrative examples and small sample investigations of these estimators are presented.

Rank Tests for Trend in Genetic Association Studies with Quantitative Traits

William R. Schucany* **and Julia V. Kozlitina**
Statistical Science, Southern Methodist University
E-Mail: schucany@mail.smu.edu

Abstract: Testing against order-restricted alternatives in a k-sample location problem arises frequently in the sciences, biomedical applications, and more recently in genetics. Nonparametric tests for trend, such as the Jonckheere test, have higher power than normal theory tests when the underlying distribution is non-normal. These tests are well-known to statisticians however have not been widely used in genetic practice. We apply Jonckheere's test and its modification (first introduced by Tryon and Hettmansperger, 1973) to the problem of detecting association between genetic markers and quantitative phenotypes. Our simulations show that the test maintains its size and is asymptotically normal even in the presence of extreme disparities in sample sizes.

These rank tests are optimal for detecting linear trends, expected under the assumption of a co-dominant genetic model. For dominant or recessive models however, (i.e., when the alternative is on the boundary of the parameter space) other tests may be more powerful than Jonckheere's. Since the underlying genetic model is often unknown, we explore robust procedures, including a linear combination (MERT) and a maximum (MAX) of the optimum statistics for linear, convex and concave alternative shapes. While these more computationally involved procedures offer better protection against model misspecification, we show that because of the imbalance in sample sizes and the unequal probability of different models, the simpler Jonckheere or weighted Jonckheere tests may perform just as well in practice as their robust counterparts.

Session Number	15
Session Title	Nonparametric and Semiparametric Modeling and Inference
Time, Day	10:00–12:00noon, October 12, Friday
Place	Lecture Room 2 (Room 855)

Partial Linear Transformation Models for Censored Survival Data

Wenbin Lu* and Hao Helen Zhang
Department of Statistics, North Carolina State University
E-Mail: lu@stat.ncsu.edu

Abstract: We propose a class of partially linear transformation models, which extends linear transformation models by allowing flexible estimation of covariate effects semiparametrically. A local estimating equation approach, motivated by martingale representation, is proposed for estimating parametric and nonparametric covariates effects in a unified manner. In particular, a locally weighted polynomial regression is used to estimate the nonparametric component. We show that, with a proper choice of the kernel bandwidth parameter, the resulting estimators for the finite-dimensional regression parameters are root-n consistent and asymptotically normal. Furthermore, a new resampling method is developed for estimating the asymptotic variances of the estimators. Numerical studies are conducted to evaluate the finite-sample performance of the proposed estimators.

Semiparametric Additive Isotonic Regression

Guang Cheng
SAMSI & Duke University
E-Mail: chengg@stat.duke.edu

Abstract: Abstract. This paper is about the efficient estimation of semiparametric additive isotonic regression model, i.e. $Y = X'\beta + \sum_{j=1}^{J} h_j(W_j) + \epsilon$. Each additive component h_j is assumed to be a monotone function. It is shown that the least square estimator of the parametric component is asymptotically normal. Moreover, the isotonic estimator for each additive functional component is proved to have the oracle property, which means it can be estimated with the highest asymptotic accuracy, equivalently, as if the other components were known.

Nonlinear Nonparametric Regression Models

Chunlei Ke and Yuedong Wang*
University of California at Santa Barbara
E-Mail: yuedong@pstat.ucsb.edu

Abstract: Almost all of the current nonparametric regression methods such as smoothing splines, generalized additive models and varying coefficients models assume a linear relationship when nonparametric functions are regarded as parameters. In this talk we present a general class of smoothing spline nonlinear nonparametric models that allow nonparametric functions to act nonlinearly. They arise in many fields as either theoretical or empirical models. Our new estimation methods are based on an extension of the Gauss-Newton method to infinite dimensional spaces and the backfitting procedure. We extend the generalized cross validation and the generalized maximum likelihood methods to estimate smoothing parameters. We establish connections between some nonlinear nonparametric models and nonlinear mixed effects models. Approximate Bayesian confidence intervals are derived for inference. We illustrate the methods with an application to term structure of interest rates. Simulations are conducted to evaluate finite-sample performance of our methods.

On Modeling and Forecasting Time Series of Curves

Jianhua Huang
Texas A & M University
E-Mail: jianhua@stat.tamu.edu

Abstract: Our work is motivated by the need for interday and dynamic intraday forecasting of call arrival rates of a telephone call center. The forecasts are useful for staffing and agent scheduling purpose. Our approach combines a dimension reduction through smoothed functional principal comments analysis and time series modeling. Using simulation and real data, we shall show that (i) smoothing is beneficial and (ii) our dynamic intraday forecasting method is effective in producing good forecasts.

This is joint work with Haipeng Shen of UNC.

Session Number	16
Session Title	Nonparametrics in Reliability and Survival Analysis
Time, Day	10:00–12:00noon, October 12, Friday
Place	Lecture Room 3 (Room 856)

Goodness-of-fit tests for recurrent event data

Russell Stocker
Department of Mathematics and Statistics, Mississippi State University
E-Mail: rstocker@math.msstate.edu

Abstract: We consider a repairable system with n units. The ith unit is observed over a time period $[0, \tau_i]$, where τ_i is a right-censoring random variable. When a unit fails it is perfectly repaired. We propose a class of goodness-of-fit tests for the hypothesis that the hazard rate function belongs to a parametric family $\mathcal{C} = \{\lambda(\cdot; \theta | \theta \in \Theta\}$. The asymptotic properties of the test statistic are given. Finite sample properties are examined using computer simulation studies. A real data set is used to illustrate the proposed tests.

Nonparametric Density Estimation for Flowgraph Models With Applications to System Reliability

David H. Collins*, Aparna V. Huzurbazar
Department of Mathematics and Statistics, University of New Mexico
Albuquerque, New Mexico

Statistical Sciences Group, Los Alamos National Laboratory
Los Alamos, New Mexico

E-Mail: dcollins@unm.edu; aparna@lanl.gov

Abstract: Statistical flowgraphs (Huzurbazar 2005) provide a methodology for analysis of multistate stochastic models with arbitrary transition time distributions. Flowgraph models are solved using integral transforms (e.g., Laplace transforms) of the transition time densities between adjacent states, which are then combined algebraically to derive transforms for transitions between arbitrary pairs of states; e.g., to derive the transform for the time of passage between two states considered “initial” and “final”. This latter transform is then inverted to recover the density or distribution function.

As an example, the process of interest may be the operation of a complex system with repairable components. Its states represent various partial failures, recurrent transitions represent fail/repair cycles, and a final state represents complete system failure. The flowgraph methodology allows data on transitions between adjacent states to be used for inference about total system performance and reliability.

Most prior work on flowgraphs has been parametric: Distribution families (gamma, inverse Gaussian, etc.) for transitions are selected based on prior information and inspection of sample histograms, and parameter values are estimated by maximum likelihood; transforms can then be determined in closed form and inverted either analytically or numerically.

This talk presents nonparametric approaches to flowgraphs using empirical transforms. For example, where the Laplace transform of the density for a transition time T is defined as

$$L(s) = E[\exp(-sT)] = \int_0^\infty \exp(-st) dF(t)$$

(F being the distribution function of T), the empirical transform based on a sample $t_1, \ldots, t_n$ is

$$\tilde{L}(s) = \frac{1}{n} \sum_{i=1}^{n} \exp(-st_i).$$

Empirical transforms can be combined algebraically just as in the parametric case; inversion can only be done numerically, using saddlepoint approximation techniques starting from the (empirical) moment generating function, or by Fourier inversion of the (empirical) Laplace transform.

After reviewing the theory of empirical transforms, we present three examples of nonparametric flowgraphs in a reliability context:

- A three-state repairable redundant system, involving failure and repair of identical units in parallel.
- A model in which a component may fail either randomly with constant hazard rate, or due to wearout centered around a known point in time.
- The repairable redundant system model with incomplete (censored) data on transition times, which requires imputing transition times for the censored observations.

Approximations derived from these models using empirical transforms are compared with the exact densities. We discuss computational issues, factors affecting accuracy of the results relative to parametric methods, and conclude with a discussion of how nonparametric flowgraph methods may be used either alone or in conjunction with parametric methods as a form of cross-validation.

Reference: Huzurbazar, Aparna V. *Flowgraph Models for Multistate Time-to-Event Data.* Hoboken, N. J.: Wiley-Interscience, 2005.

A New Generalized Test Statistic for Multinomial Goodness-of-Fit

Sunil Mathur

Department of Mathematics, University of Mississippi, University, MS 38677

E-Mail: skmathur@olemiss.edu

Abstract: Several goodness-of-fit tests are proposed for testing the goodness-of-fit of discrete multivariate data. We propose a unified analysis of goodness-of-fit using a generalized test statistic(s). Depending on the choice of parameters, the generalized test statistics results into some of the known statistics such as Chi-square, and likelihood ratio generalized test, it has been observed that the proposed test statistic is highly robust against extreme values and does not assume the distribution of parent population. The asymptotic distribution of the proposed test statistic and the p-value function are discussed. The power comparisons have been made, and it is found that, the proposed statistic dominates over the Cressie and Read (1984) test statistic however the comparisons are made with respect to change in the location only. The application of proposed method has been attempted by using a real-life data, of Cranor and Christensen (2003).

Sequential test for censored data with linear transformation models

Lin Huang and Zhezhen Jin*
Department of Biostatistics
Mailman School of Public Health
Columbia University
New York, NY 10032
E-Mail: zj7@columbia.edu

Abstract: In this talk, we will present a sequential test for censored data with linear transformation models which include Cox proportional hazards model and proportional odds model. The approach is based on a score process motivated by Chen, Jin and Ying's (2002) estimation procedure. We show that for given interim analyses time points, the score process is approximated by a mean 0 multivariate normal distribution and the covariance matrix can be consistently estimated. Then with the boundaries procedure proposed by Slud and Wei (1982), a repeated significant test can be conducted. Numerical studies will also be presented.

Session Number	17
Session Title	Empirical Likelihood
Time, Day	10:00–12:00noon, October 12, Friday
Place	Lecture Room 4 (Room K-J)

Empirical likelihood based inference for the calibration regression model with lifetime medical cost

Yichuan Zhao
Department of Mathematics and Statistics, Georgia State University
E-Mail: matyiz@langate.gsu.edu

Abstract: In recent years, medical cost has received increasing interest in Biostatistics and public health. Statistical analysis and inference of lifetime medical cost have been challenging by the fact that the survival times are censored on some study subjects and their subsequent cost are unknown. Recently, the calibration regression model has been proposed to study the medical cost associated with covariates. However, the accuracy of the inference procedure may be low when the sample size is small due to highly skewed nature of the medical cost and non-uniform rate of cost accumulation over time. In this paper, we develop an empirical likelihood ratio method for the calibration regression model. The adjusted empirical likelihood confidence regions are constructed for the regression parameters accordingly. Furthermore, we compare the proposed empirical likelihood method with normal approximation based method. Simulation results show that the proposed method outperforms the existing method in terms of coverage probability. In particular, the proposed adjusted empirical likelihood method overcomes the under coverage problem.

Convergence of the nonparametric EM Algorithm

John Fresen

Department of Statistics, University of Missouri at Columbia

E-Mail: fresenj@missouri.edu

Abstract: In this presentation we consider the nonparametric EM algorithm, the EM algorithm based on empirical likelihood rather than the parametric likelihood. We develop the theoretical underpinnings and assess the convergence rates of the nonparametric EM algorithm. We provide a comparison of the convergence rates with that of the parametric EM algorithm. We also explore the effect of initial estimates on the convergence and explore if the geometric progression provides robust starting values. Examples will be given.

Parametric/Nonparametric Hybrid Two Sample Problem by Empirical Likelihood with Censored Data

Mai Zhou

Department of Statistics, University of Kentucky

E-Mail: mai@ms.uky.edu

Abstract: We use the hazard formulation of the censored data empirical likelihood to study the two sample parametric/nonparametric hybrid model. We demonstrate that a proper Empirical Likelihood definition that takes into account of censoring will result an empirical likelihood ratio test with a proper chi-square limiting distribution under null hypothesis. We illustrate the use of the proposed test by way of testing the ROC curve with censored data, among others. Results are compared to Zhou & Liang (2005 Biometrika).

Joint work with Hua Liang, Department of Biostatistics, University of Rochester.

Empirical Likelihood in the framework of Parametric Likelihood

Mi-Ok Kim

Cincinnati Children's Hospital Medical Center

E-Mail: miok.kim@cchmc.org

Abstract: We use a method of parameterized sub-family of probability distributions to connect empirical likelihood (EL) with parametric likelihoods and discuss the EL inference in the framework of parametric likelihood inference. The EL inference benefits from theoretical developments in the parametric case. We illustrate the method with general estimating equations and consider M-type linear regression as an example of practical applications where the proposed method promotes conditional EL inference with parameter orthogonality in place of profile EL inference.

Session Number	18 (Plenary Talk IV)
Session Title	Non and Semiparametrics in Biomedical Settings
Time, Day	1:30–2:30PM, October 12, Friday
Place	Lecture Room 1 (Lumpkin Auditorium)

Semiparametric Analysis of Case-Control Studies, With Applications to Gene-Environment Interactions

Raymond J. Carroll

Department of Statistics
3143 TAMU, Texas A&M University, College Station TX 77843-3143
E-Mail: carroll@stat.tamu.edu

Abstract: Consider a standard retrospective case-control study involving interactions, with covariates (G, X) and logits of the form $\beta_0 + m(X, G, \beta_1)$ for an arbitrary known function $m(\cdot)$. If the function $m(\cdot)$ is known, and if the distribution of the covariates (X, G) is modeled nonparametrically, then only β_1 is identifiable and Prentice and Pyke (1979, Biometrika) showed that the semiparametric efficient estimator for β_1 is obtained via ordinary logistic regression, ignoring the case-control sampling scheme. The same result applies for partially linear models, with logits $m(G, \beta_1) + \theta(X)$, where $\theta(\cdot)$ is modeled nonparametrically. Indeed, if the distribution of (X, G) is nonparametric, inference that ignores the case-control sampling scheme and pretends that the observed data are from a prospective random sample is asymptotically efficient and asymptotically correct.

In genetic epidemiology, however, it is often reasonable to make assumptions about the distribution of the gene, G, given the environment X, while still treating the distribution of X nonparametrically. For example, suppose that G is binary (you have a mutation or you do not) while X is multivariate. There is a considerable methodological and applied literature that assumes that genetic status is independent of environment in the population. If you make this assumption, and no other, and then compute the semiparametric MLE for retrospective case-control data in the parametric model with logits $\beta_0 + m(X, G, \beta_1)$, the resulting efficient estimator is very different from ordinary logistic regression. In practical cases, we have seen decreases in standard errors for interactions that are a fraction of 2 or more. Put less precisely, the assumption that G and X are independent, without modeling either component, is effectively the same as having 4 times more data for certain parameters. Similar statements apply if the logistic model has nonparametric components, e.g., varying coefficient interaction models.

I will describe a very general theory when things are known about the distribution of G given X, and apply it to an example of testing whether oral contraceptive use among those carrying the BRCA1/2 gene mutation is protective against cervical cancer.

Session Number	19
Session Title	Bayesian Nonparametrics
Time, Day	2:45–4:45PM, October 12, Friday
Place	Lecture Room 1 (Lumpkin Auditorium)

Partition Based Nonparametric Priors

Jayaram Sethuraman

Florida State University
Tallahassee, FL 32306-4330
E-Mail: sethu@stat.fsu.edu

Abstract: Failure data, in Reliability and Survival analysis, may consist of exactly observed failure times or censored failure times or failure times restricted by previous failure times. The first case is usually referred to as complete observations. The second case occurs when there are censoring random variables. The third case occurs when repairs to a system restrict the range of the failure times to sets based on previous failures.

For repair models, when the data consists of X with a distribution P restricted to a set A depending on previous failure times, Sethuraman and Hollander introduced a class of priors called *Partition Based* **PB** priors for P and showed that the posterior distributions are also **PB** distributions. They also show how to simplify the calculations when using Dirichlet priors. We will summarize some of these results.

The second situation occurs in survival analysis and covers all kinds of censoring. The exact value of the failure time X is not known, but is only known to belong to a set A. In this talk, we will show that the same class of **PB** priors are the natural priors in this problem. This simplifies and clarifies previous work on Bayes methods in survival analysis.

Dimension Augmenting Vector Machine: A new General Classifier System for Large p Small n problem

Dipak K. Dey*, Samiran Ghosh, Yazhen Wang
Department of Statististics, Uuniversity of Connecticut, Storrs
Department of Mathematical Sciences, Indiana University-Purdue University, Indianapolis
Department of Statistics, University of Connecticut, Storrs
E-Mail: dipak.dey@uconn.edu

Abstract: Support vector machine (SVM) and other reproducing kernel Hilbert space (RKHS) based classifier systems are drawing much attention recently due to its robustness and generalization capability. All of these approaches construct classifier based on training sample in a high dimensional space by using all available dimensions. SVM achieves huge data compression by selecting only few observations lying in the boundary of the classifier function. However when the number of observations is not very large (small n) but the number of dimensions are very large (large p) then it is not necessary that all available dimensions are carrying equal information in the classification context. Selection of only useful fraction of available dimensions will result in huge data compression. In this paper we have come up with an algorithmic approach by means of which such an optimal set of dimensions could be selected. We have reversed and modified the solution proposed by Zhu and Hastie in the context of Import Vector Machine (IVM), to select an optimal sub model by using only few observations. For large p small n domain (e.g. Bioinformatics) our method compares different trans-dimensional model to come up with optimal set of dimensions to build the final classifier.

Semiparametric Bayesian Analysis of Studies of Gene-Environment Interaction

Bhramar Mukherjee
University of Michigan
E-Mail: bhramar@umich.edu

Abstract: We will discuss the problem of retrospective modeling of case-control data for studying gene-environment interactions in a semiparametric Bayesian framework. The special feature of gene-environment interaction studies is that in many situations it is scientifically plausible to assume that the genetic and environmental factors are independent in the underlying population. Under this additional constraint of gene-environment independence, one can derive more efficient estimation techniques than the traditional prospective logistic regression analysis (Piegorsch et al, 1994; Chatterjee and Carroll, 2005). However, the efficient estimates from the retrospective likelihood may be severely biased under the violation of the independence assumption. Stratification effects present in the population could potentially induce non-independence among genetic factors and environmental exposures. We will first provide a semiparametric Bayesian approach to model stratification effects under the assumption of gene-environment independence.

We will then propose an alternative to relax the constraint of gene-environment independence in a natural Bayesian framework to strike a compromise between efficiency and robustness. We will analyze data from a case-control study on ovarian cancer patients conducted in Israel to illustrate our methods. Collaborators on material related to the presented work are: Nilanjan Chatterjee (National Cancer Institute), Malay Ghosh (University of Florida), Li Zhang (Cleveland Clinic Foundation) and Samiran Sinha (Texas A & M).

Current Methods for Recurrent Events Data with Dependent Termination: A Bayesian Perspective

Debajyoti Sinha*[1], Joseph G. Ibrahim[2], Tapabrata Maiti[3], Bichun Yang[4]
[1]Florida State University
[2]University of North Carolina
[3]Medical University of South Carolina
[4]Iowa State University
E-Mail: sinhad@stat.fsu.edu

Abstract: There has been a recent surge of interest in modeling and methods for analyzing recurrent events data with risk of termination dependent on the history of the recurrent events. To aid the future users deciding about appropriate models for the application at hand, we review state of the art statistical methods and present novel theoretical properties, identifiability results and practical consequences of key modeling assumptions of several fully specified stochastic models. After introducing stochastic models with non-informative termination process, we focus on a class of models which allows both negative and positive association between the risk of termination and the rate of recurrent events via a frailty variable. We also discuss the relationship as well as the major differences between these models in terms of their motivations and physical interpretations. We discuss associated Bayesian methods based on Markov chain Monte Carlo tools, and novel model diagnostic tools to perform inference based on fully specified models. We demonstrate the usefulness of current methodology through an analysis of a data set from a clinical trial. In conclusion, we explore possible future extensions and limitations of the methodology.

Session Number	20
Session Title	Recent Advances in Non- and Semi-parametrics
Time, Day	2:45–4:45, October 12, Friday
Place	Lecture Room 2 (Room 855)

Evaluating the reproducibility of two studies of a large number of objects: Modified Kendall Rank-Order Association Test

Tian Zheng* and Shaw-Hwa Lo
Department of Statistics, Columbia University
E-Mail: tzheng@stat.columbia.edu

Abstract: Assessing the reproducibility of research studies can be difficult, especially when the number of objects involved is large. In such situations, there is only a small set of those objects that are truly relevant to the scientific questions. For example, in microarray analysis, despite data sets containing expression levels for tens of thousands of genes, it is expected that only a small fraction of these genes are regulated by the treatment in a single experiment. In such cases, it is acknowledged that reproducibility of two studies is high

only for objects with real signals. One way to assess reproducibility is to measure the associations between the two sets of data. The traditional association methods suffered from the lack of adequate power to detect the real signals, however. We present in this talk the use of a modified Kendall rank-order test of association, based on truncated ranks. Simulation results show that the proposed procedure increases the capacity to detect the real signals considerably. Applications to gene expression analysis and genetic epidemiology will be discussed.

Regression models with functional predictors

Todd Ogden
Mailman School of Public Health, Columbia University
E-Mail: ogden@neuron.cpmc.columbia.edu

Abstract: Regression of a scalar response on functional predictors (or signals), such as spectra or images, presents a major challenge when, as is typically the case, the dimension of the signals far exceeds the number of signals in the dataset. Fitting such a model meaningfully requires some form of dimension reduction. A proposed approach to this problem extends common multivariate methods (principal component regression (PCR) and partial least squares (PLS)) to handle functional data by also incorporating a roughness penalty. A number of alternative estimation strategies are available and these will be discussed briefly, as well as sufficient conditions for consistency. These methods are illustrated using data from near infrared (NIR) spectra from chemical samples and data from a brain imaging study.

SIMEX and Variance Estimation in Semiparametric Measurement Error Models

Tatiyana V. Apanasovich
Cornell University
E-Mail: tva2@cornell.edu

Abstract: SIMEX is a general-purpose technique for measurement error correction. There is a substantial literature on the application and theory of SIMEX for purely parametric problems, as well as for purely nonparametric regression problems, but there is neither application nor theory for semiparametric problems. Motivated by an example involving radiation dosimetry, we develop the basic theory for SIMEX in semiparametric problems using kernel-based estimation methods. This includes situations that the mismeasured variable is modeled purely parametrically, purely nonparametrically, or that the mismeasured variable has components that are modeled both parametrically and nonparametrically. Using our asymptotic expansions, easily computed standard error formulae are derived, as are the bias properties of the nonparametric estimator. The standard error method represents a new method in general for semiparametric problems, and we show in our example that it improves dramatically on the first order methods.

Levene's Family of Tests for Equality of Variances: historical perspective, impact and modifications

Yulia Gel
University of Waterloo, Canada
E-Mail: ygl@math.uwaterloo.ca

Abstract: In many applications, the underlying scientific question concerns whether the variances of k samples are equal. The problem of assessing homogeneity of variances has a long history and there exists a substantial number of related tests. Many such tests rely on the assumption of normality and are not robust to its violation. In 1960, Prof. Howard Levene proposed a new approach to the problem of obtaining a valid test for equality of variances which is essentially the F-test computed on the absolute deviations of observations from the group mean. Levene's approach is shown to be a powerful and robust to non-normality test and quickly became a very popular tool for assessing homegeneity of variances. The practical relevance and importance of Levene's approach is demonstrated by the fact that Levene's (1960) article has been cited over 750 times in the scientific literature. The test has been used in a wide variety of applications, e.g. clinical trials, astronomy, marine pollution, business, auditing and law cases.

This talk reviews the original statistic proposed by Levene and various modifications which replace the group means by robust measures of location. We also propose a new Levene-based test for assessing monotonic trends in variances. The new test can be useful when one concerns with an increasing or decreasing variability, for example, increasing volatility of stocks in financial time series, "open or closed gramophones" in regression residual analysis and patterns of anisotropy in spatial statistics. This is a joint work with Professors Joseph L. Gastwirth and Weiwen Miao.

Session Number	21
Session Title	Nonparametrics in Biomedical Settings
Time, Day	2:45–4:45PM, October 12, Friday
Place	Lecture Room 3 (Room 856)

Longitudinal Data with Nonignorable Dropout

Mengling Liu
Division of Biostatistics, School of Medicine, New York University
E-Mail: mengling.liu@med.nyu.edu

Abstract: We investigate marginal inference for longitudinal data subject to nonignorable dropout.The proposed marginalised model directly specifies marginal associations between longitudinal responses and covariates while leaving the within-subject dependence unspecified. It incorporates nonignorable dropout events, which marginally follow a semiparametric transformation model, through a flexible conditional mean model. We develop an estimation procedure based on a series of asymptotically unbiased estimating equations. The resulting estimators for the marginal regression parameters are consistent and asymptotically normal, with a sandwich-type variance-covariance matrix that can be consistently estimated by the usual plug-in rule. The proposed approach is evaluated by simulations and illustrated by a real data application.

Nonparametric tests for two group comparisons of dependent observations obtained at varying time points with application to RNA viral load decline

Susanne May
University of California, San Diego
E-Mail: smay@ucsd.edu

Abstract: We propose new tests for two-group comparisons of repeated measures that might be obtained at arbitrary time points and differ over individuals. The tests do not make any assumptions regarding the distribution of the repeated measures except that one of them assumes that the repeated measures can be grouped into distinct periods of observations such that the covariance between scores only depends on the periods the observations belong to and that the covariance matrices are the same in the two groups. The tests remain valid even if the probability that a response is observed depends on the level of response if missing data mechanisms are the same in both groups. Inference can be based on resampling. We use the tests to assess differences in viral load decline for drug resistant and drug sensitive human immunodeficiency virus (HIV)-1 infected patients.

A semiparametric Bayesian model for examiner agreement in periodontal research

Elizabeth H. Slate* and Elizabeth G. Hill

Medical University of South Carolina

E-Mail: slate@musc.edu; hille@musc.edu

Abstract: An important measure of the severity of periodontal disease is the probing pocket depth (PPD), which is measured on up to 6 sites for each tooth in the mouth. Establishing and monitoring agreement among multiple examiners is critical to high quality periodontal research. We develop a Bayesian hierarchical model that links the true, observed and recorded values of PPD, permitting correlation among the measures within patient. Tooth-site-specific examiner effects are modeled as arising from a Dirichlet process mixture, facilitating discovery of subgroups among the periodontal sites according to degree of agreement with a reference examiner. We analyze data from a PPD calibration study and illustrate the effects of correlation on assessments of examiner agreement.

Nonparametric Estimation of a Distribution Function Under Biased Sampling and Multiplicative Censoring

Micha Mandel

The Hebrew University of Jerusalem, Israel

E-Mail: msmic@mscc.huji.ac.il

Abstract: Suppose that instead of observing realizations of $X \sim F$, one observes realizations of X^w from a weighted distribution $F^w(dx) \propto w(x)F(dx)$. Furthermore, observations may be multiplicative censored, that is, part of the data are realizations of UX^w, where U is an independent uniform (0,1) random variable. Giving these data, one aims at estimating the law F without any further parametric assumptions. The model with $w(x) = x$ was studied by Vardi (1989, Biometrika) and it naturally arises when sampling stationary renewal processes or when collecting cross-sectional data. However, in many cases data are more complex. Examples are when entrances to the population are governed by an inhomogeneous Poisson process or when lifetimes are composed of several phases. In such cases, data are subject to different biases and censoring mechanisms and require new models and estimation methods. Motivating by data on promotions in the Hebrew University, I present in this talk new multiplicative censoring models. I offer algorithms to estimate non-parametrically the distribution function of lifetimes, F, and discuss their properties.

Session Number	22
Session Title	Nonparametrics from the 'Younger' Generation
Time, Day	2:45–4:45PM, October 12, Friday
Place	Lecture Room 4 (Room K-J)

Nonparametric variable selection via sufficient dimension reduction

Lexin Li* and Howard D. Bondell

Department of Statistics, North Carolina State University

Department of Statistics, North Carolina State University

E-Mail: li@stat.ncsu.edu

Abstract: Sufficient dimension reduction (SDR) has proven effective to transform high dimensional problems to low dimensional projections, while losing no regression information and pre-specifying no parametric model during the phase of dimension reduction. However, existing SDR methods suffer from the fact that each dimension reduction component is a linear combination of all the original predictors, and thus can not perform variable selection. In this talk, we propose a regularized SDR estimation strategy, which is capable of simultaneous dimension reduction and variable selection. We demonstrate that the new estimator achieves consistency in variable selection without requiring any traditional model, meanwhile retaining $n^{1/2}$ estimation consistency of the dimension reduction basis. Both simulation studies and real data analyses are reported.

Wilcoxon-type generalized Bayesian information criterion

Lan Wang

School of Statistics, University of Minnesota

E-Mail: lan@stat.umn.edu

Abstract: We extend the basic idea of Schwarz (1978) and develop a generalized Bayesian information criterion for regression model selection. The new criterion relaxes the usually strong distributional assumption associated with Schwarz's BIC by adopting a Wilcoxon-type dispersion function and appropriately adjusting the penalty term. We establish that the Wilcoxon-type generalized BIC preserves the consistency property of Schwarz's BIC without its need to assume a parametric likelihood. We also show that it outperforms Schwarz's BIC with heavier-tailed data in the sense that asymptotically it can yield substantially smaller L_2 risk. On the other hand, when the data are normally distributed, both criteria have similar L_2 risk. The new criterion function is convex and can be conveniently computed via existing statistical software.

Smoothing Dissimilarities for Cluster Analysis: Binary Data and Functional Data

David Hitchcock

University of South Carolina

E-Mail: hitchcock@stat.sc.edu

Abstract: Cluster analysis attempts to group data objects into homogeneous clusters, often on the basis of the pairwise dissimilarities among the objects. When the data contain noise, we might consider performing a smoothing operation, either on the data themselves or on the dissimilarities, before implementing the clustering algorithm. Possible benefits to such pre-smoothing are discussed in the context of binary data, and we also make connections to previous research on smoothing functional data for cluster analysis. We suggest a method for cluster analysis of binary data based on "smoothed" dissimilarities. The smoothing method presented borrows ideas from shrinkage estimation of cell probabilities. The method is illustrated with an example involving binary item response data.

Nonparametric density estimation from covariate information

Ryan Elmore
Department of Statistics, Colorado State University
E-Mail: elmore@stat.colostate.edu

Abstract: An increasing number of statistical problems arise in connection with functional calibration. In each case, inexpensive, indirect data in a particular context are combined with direct, expensive-to-acquire data from different but related settings, so as to estimate quantities in the former case. We observe data which give us access to the distribution of U given V, and, from these and data on U, we wish to estimate the density of V. The motivating real datasets are of age and covariate information in fish populations. We suggest two methodologies, each based on transforming the problem to one which involves inversion of a symmetric, linear operator. Our techniques have connections to methods for functional data analysis and for a variety of mixture and deconvolution problems, as well as to calibration techniques.

Session Number	23
Session Title	Poster Session II
Time, Day	4:45–6:00PM, October 12, Friday
Place	DMC Lobby (Set-Up Period: Morning to Noon)

Nonparametric Estimation of Conditional Expectation

Jiexiang Li* and Lanh Tat Tran
Department of Mathematics, College of Charleston
Department of Mathematics, Indiana University
E-Mail: LiJ@cofc.edu

Abstract: Denote the integer lattice points in the N dimensional Euclidean space by $\mathcal{Z}^N$ and assume that $(X_i, Y_i), i \in \mathcal{Z}^N$ is a mixing random field. Estimators of the conditional expectation $r(x) = E[Y_i|X_i = x]$ by nearest neighbor methods are established and investigated. The main analytical result of this study is that, under general mixing assumptions, the estimators considered are asymptotically normal. Many difficulties arise since points in higher dimensional space $N \geq 2$ cannot be linearly ordered. Our result applies to many situations where parametric methods cannot be adopted with confidence.

Bayesian inference in semiparametric mixed models for longitudinal data

Yisheng Li
Department of Biostatistics, The University of Texas, M.D. Anderson Cancer Center
E-Mail: ysli@mdanderson.org

Abstract: We consider Bayesian inference in semiparametric mixed models (SPMMs) for longitudinal data. SPMMs are a class of models that use a nonparametric function to model a covariate effect, e.g., time effect, a parametric function to model other covariate effects, and parametric or nonparametric random effects to account for the within-subject correlation. We model the nonparametric function using the Bayesian formulation of a cubic smoothing spline, and the random effect distribution using a normal distribution and alternatively with a nonparametric Dirichlet process (DP) prior. When the random effect distribution is assumed to be normal, we propose a uniform shrinkage prior (USP) for the variance components and the smoothing parameter. When the random effect distribution is modeled nonparametrically, we use a DP prior with a normal base measure and propose a USP for the hyperparameters of the DP base measure. We argue that the commonly assumed DP prior implies a non-zero mean of the random effect distribution, even when a base measure with mean zero is specified. This leads to biases in Bayesian inference for the regression coefficients and the spline. We propose a correction using a post-processing technique. We show that under mild conditions the posterior is proper under the proposed USPs, a flat prior for the fixed effects parameters, and an improper prior for the residual variance. We illustrate the proposed approach using a longitudinal hormone dataset, and carry out extensive simulation studies to compare its finite sample performance with that of the existing methods.

Nonparametric transfer function models: A polynomial spline approach

Jun Liu and Jing Wang

Department of Finance and Quantitative Analysis, Georgia Southern University

Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago

E-Mail: jliu@georgiasouthern.edu

Abstract: In this paper, we propose a polynomial splines-based method to model nonlinear relationship between input and output time series. The functional form of the underlying relationship (the transfer function) is unknown but smooth. The noise is serially correlated and assumed to follow a parametric AR model. The transfer function is modeled using polynomial splines and estimated jointedly with the AR parameters. By modeling the transfer function nonparametrically, the model is flexible and can be applied on highly nonlinear relationship of unknown functional forms; by modeling the noise explicitly, the correlation in the data is removed so the transfer function can be estimated more efficiently. Additionally, the estimated AR parameters can be used to improve the forecasting performance. Compared with existing local polynomial-based approaches, the proposed polynomial splines-based estimator is much less computationally intensive, more importantly, it can be easily extended to model non-stationary noise. The estimation procedures are introduced and the asymptotic properties of the estimators are discussed. The finite-sample properties of the estimators are studied through simulations and one real example.

Spline-backfitted kernel smoothing of additive coefficient model

Rong Liu and Lijian Yang

Department of Statistics, Michigan State University

E-Mail: liurong@stt.msu.edu

Abstract: Additive coefficient model (Xue and Yang 2006a, b) is a flexible tool for multivariate regression and time series analysis that circumvents the “curse of dimensionality.” We propose spline-backfitted kernel (SBK) and spline-backfitted local linear (SBLL) estimators for the component functions in the additive coefficient model that is both (i) computationally expedient so it is usable for analyzing very high dimensional data, and (ii) theoretically reliable so inference can be made on the component functions with confidence. In addition, it is (iii) intuitively appealing so it does not intimidate practitioners. Simulation experiments have provided strong evidence that corroborates with the asymptotic theory. The SBLL procedure is applied to a varying coefficient extension of the Cobb-Douglas model for the US GDP that allows non neutral effects of the R&D on capital and labor as well as in the Total Factor Productivity (TFP).

SIMEX and variance estimation in semiparametric measurement error models

Arnab Maity

Department of Statistics, Texas A & M University

E-Mail: amaity@stat.tamu.edu

Abstract: SIMEX is a general-purpose technique for measurement error correction. There is a substantial literature on the application and theory of SIMEX for purely parametric problems, as well as for purely nonparametric regression problems, but there is neither application nor theory for semiparametric problems. Motivated by an example involving radiation dosimetry, we develop the basic theory for SIMEX in semiparametric problems using kernel-based estimation methods. This includes situations that the mismeasured variable is modeled purely parametrically, purely nonparametrically, or that the mismeasured variable has components that are modeled both parametrically and nonparametrically. Using our asymptotic expansions, easily computed standard error formulae are derived, as are the bias properties of the nonparametric estimator. The standard error method represents a new method in general for semiparametric problems, and we show in our example that it improves dramatically on first order methods. We find that for estimating the parametric part of the model, standard bandwidth choices of order $O(n^{-1/5})$ are sufficient to ensure asymptotic normality, and undersmoothing is not required. We also include results on uniform expansions of nonparametric function estimators. SIMEX has the property that it fits misspecified models, namely ones that ignore the measurement error. Our work thus also more generally describes the behavior of kernel-based methods in misspecified semiparametric problems.

An Asymptotically distribution-free Aligned Rank Test for Location in a Repeated Observation Model

Bernard Omolo

University of South Carolina Upstate

E-Mail: bomolo@uscupstate.edu

Abstract: Rank tests are known to be distribution-free for simple linear models, where the observations are i.i.d. For general linear models with nuisance parameters, however, the alignment principle can be useful in deriving asymptotically distribution-free rank tests. This is especially so when the centered design matrices have full rank. However, when the centered design matrices are not of full rank, the classical Chernoff - Savage approach, as opposed to Hájek's approach, can be applied to yield asymptotically distribution-free rank tests. The asymptotic distribution of an aligned rank test for location in a repeated observations model where the centered orthonormal design matrix is not of full rank is derived. The distribution turns out to be chi-square under the null hypothesis as well as local alternatives. Simulation studies regarding the Type I error rate and power in testing for linearity in a nonparametric regression model with standard Cauchy random errors corroborate these theoretical results.

Time series central subspace

Jin-Hong Park
Department of Mathematics, College of Charleston
Charleston, South Carolina
E-Mail: ParkJ@cofc.edu

Abstract: We develop a sufficient dimension reduction theory for time series, which does not require specification of a model but seeks to find a $p \times d$ matrix Φ_d with smallest possible number $d \leq p$ such that the conditional distribution of $x_t|X_{t-1}$ is the same as that of $x_t|\Phi_d^T X_{t-1}$, where $X_{t-1} = (x_{t-1}, ..., x_{t-p})^T$, resulting in no loss of information about the conditional distribution of the series given its past p values. We define the subspace spanned by the columns of Φ_d as the time series central subspace and estimate it using Kullback-Leibler distance. We show that the estimator is consistent when p and d are known. In addition, we propose a consistent estimate of d and a graphical method to determine the lag p. Finally, we present examples and real data analysis to illustrate the proposed theory, which may open new research avenues in time series data analysis.

Bayesian analysis of physiologically based pharmacokinetics modelling of perchloroethylene in humans

Junshan Qiu
Department of Statistics, University of Georgia
E-Mail: qiuj@mail.rx.uga.edu

Abstract: This study is to estimate population distributions of PBPK model parameters and to make a dose reconstruction with clinical data from uncontrolled studies. Perchloroethylene (PCE) is a widely distributed pollutant in the environment. The cancer risks of PCE at low exposures are uncertain. PCE occurs widely in the dry cleaning establishments and also can be found in indoor air. However, the concentrations of PCE are mostly below 1ppm. Therefore, it is very important to assess cancer risks at these low concentrations. A human physiologically based pharmacokinetics (PBPK) model was used to quantify tissue doses of PCE and its key metabolite, Trichloroacetic Acid (TCA) after inhalation exposures. This PBPK model was integrated with a statistical hierarchical model to acknowledge variations due to intraindividual variation, interindividual variation, measurement error and difference between study methods. A Bayesian approach, Markov chain Monte Carlo analysis, was employed to analyze clinical data obtained from controlled studies. The data are on alveolar or exhaled breath concentrations of PCE, blood concentrations of PCE and TCA, urinary excretion of TCA. The posterior distributions of PBPK model parameters were obtained. Predictive ability of posteriors was satisfactory. Posterior predictions are much better than prior fit.

The Mann-Whitney-Wilcoxon random field, with applications to brain mapping

Farzan Rohani*[1], **Masoud Asgharian**[1], **Keith J. Worsley**[1,2]
[1]Department of Mathematics and Statistics, McGill University, Montreal, Canada
[2]Montreal Neurological Institute, McGill University, Montreal, Canada
E-Mail: rohani@math.mcgill.ca

Abstract: Suppose we have two groups of observations at each point in a subset S of Euclidean space, so that, instead of being scalars, the observations are smooth random fields. We define the Mann-Whitney-Wilcoxon

(MWW) random field as the MWW test statistic evaluated at each point in S. We are interested in the null distribution of the maximum MWW over all points in S.

The motivation comes from brain mapping, where the observations may be functional magnetic resonance images (fMRI) or anatomical images such as density of grey matter or multiple sclerosis lesions. The two groups may be fMRI measurements made under a control and a task condition, or in the case of anatomical data, a control group and a disease group. The difference between the two groups is thought to be confined to a small number of localised regions in the brain, so we calculate a test statistic $T(s)$ at each point s in the brain S, then determine a threshold so that the p-value $\mathrm{P}(\max_S T(s) \geq t) = \alpha$, where α is a desirable level of significance. The brain regions where $T(s) \geq t$ are the places where there is a difference between the two groups.

It is commonly assumed that such observations are Gaussian random fields, and there is a well-developed theory for approximations to

$$\mathrm{P}(\max_S T(s) \geq t) = \alpha$$

where $T(s)$ is for example a two-sample T statistic [5]. However there is considerable evidence that even in large samples these approximations are not accurate when the data is not Gaussian. The reason is that the threshold t is so far out in the tails of the null distribution of $T(s)$ that the usual Gaussian approximation (by appealing to the Central Limit Theorem) is not accurate. One solution is to base inference on the permutation distribution of $\max_S T(s)$, found by permuting the group labels of the data [4], but this requires time-consuming simulations. Instead we propose replacing $T(s)$ by the MWW statistic at each point.

In the Gaussian case, the random field theory approximation to the p-value is the expected Euler characteristic (EC) of the excursion set $A_t = \{s : T(s) \geq t\}$ [1,2]. We will find a closed form expression for the expected EC of 1D and 2D MWW random fields. Although the resulting test is not fully non-parametric, it depends on a particular smoothness parameter of the data which is estimable from within the groups.

We apply our method to a data set on the difference in multiple sclerosis (MS) lesion density between a group of 212 patients with high disability (EDSS>2) and 213 patients with low disability (EDSS<=2)[3].

Improved estimation in multiple linear regression model with measurement error and general constraint

Weixing Song and Hua Liang
Department of Statistics, Kansas State University
Department of Biostatistics, University of Rochester
E-Mail: weixing@ksu.edu

Abstract: When uncertain prior information is available for regression parameters in the multiple linear regression model, two restricted estimators are constructed in the multiple linear regression models with measurement errors in the predictors. Based on these restricted estimators, two sets of estimators which include the preliminary test estimator, the stein-type estimator and the positive rule stein type estimator, are constructed for both slopes and intercept. Their asymptotic properties, including the asymptotic distributional quadratic biases, the asymptotic distributional quadratic risks, are discussed. The performances based on the asymptotic distributional quadratic risks are compared among these estimators. Finally, a simulation study illustrates the finite sample performance of the proposed estimators.

Estimating the distance distribution of subpopulations for a large-scale complex survey

Jianqiang Wang
Iowa State University
E-Mail: jqwang@iastate.edu

Abstract: Many finite populations targeted by sample surveys consist of homogenous subpopulations with respect to the variables being collected. We propose a sample-based fully nonparametric estimator for the subpopulation distribution functions of the distances between elements and the subpopulation centers, and explore a general definition of distance metric and different ways to define the subpopulation centers. The usual mean vector and multivariate L_1 median are discussed as measures of center. We describe the theoretical properties of the estimator. The asymptotic variance of the estimator involves a gradient vector for which we propose a nonparametric estimator and this plug-in estimator is examined in simulation study. And a jackknife variance estimator is proposed as an alternative.

Efficient and Robust Estimation in Semiparametric Models

Jingjing Wu[1*], Anton Schick[2] and R.J. Karunamuni[3]
[1]University of Alberta
[2]Binghamton University
[3]University of Calgary
E-Mail: jinwu@math.ucalgary.ca

Abstract: The successful application of Hellinger distance approach to fully parametric models is well known. The corresponding optimal estimators, known as minimum Hellinger distance estimators, are robust and efficient (Beran, 1977). In this paper, we extend this approach to general semiparametric models. We obtain minimum Hellinger distance and minimum "profile" Hellinger distance estimators in this context. The asymptotic properties such as consistency, asymptotic normality, efficiency and adaptivity of the proposed estimators are investigated. The robustness and small sample properties of the proposed estimators are also studied using a Monte Carlo study.

Efficient linear programming algorithm for functional component pursuit

Yonggang Yao* and Yoonkyung Lee
Department of Statistics, Ohio State University
E-Mail: yao@stat.osu.edu

Abstract: Based on reproducing-kernel Hilbert space theory, a large class of feature spaces can be characterized through their kernel functions, which have been widely applied in statistics and engineering. By further parameterizing the kernel structure, many current nonparametric regularization methods can be extended to achieve both model-fitting and feature selection objectives. In this paper, we focus on the nonparametric regularization problems associated with a kernel collection, and introduce a fast adaptive kernel selection algorithm rooted in parametric linear programming.

Index