Program

Eleventh Annual Winter Workshop
Semiparametric Methods

Department of Statistics
University of Florida
January 8-10, 2009

Contents

Sponsors

National Science Foundation, InfoTech, the College of Liberal Arts & Sciences, Graduate School, and the Department of Statistics of the University of Florida.

Organizing Committee

Michael Daniels
Ronald Randles
Linda Young

Invited Speakers

Raymond Carroll, Texas A&M University
Peter Hoff, University of Washington
Wes Johnson, University of California, Irvine
Susan Murphy, University of Michigan
Dave Ruppert, Cornell University
Butch Tsiatis, North Carolina State University
Matt Wand, University of Wollongong (Australia)
Naisyin Wang, Texas A&M University
L.J. Wei, Harvard University
Jon Wellner, University of Washington

Acknowledgements

The organizers thank the Department of Statistics staff, Robyn Crawford, Tina Greenly, Allison Pipkin, Summer Rozear, and Marilyn Saddler for their tremendous efforts in helping to set up this meeting and make it run smoothly.

Conference Schedule

Thursday, January 8, 2009

6:00 -7:30 pm Reception at Keene Faculty Center, Dauer Hall

All workshop sessions meet in the 209 Teaching Classroom, Emerson Alumni Hall

Friday, January 9, 2009

8:45-9:15am Continental Breakfast

9:15am Welcome by:
Bernard Mair, Associate Dean, College of Liberal Arts & Sciences
Mike Daniels, Conference Chair

9:30-11:30am **Session 1:** ***Semiparametric Regression: Overview and New Developments***
Chair: Mike Daniels, University of Florida

Speakers: Dave Ruppert, Cornell University
Semiparametric Regression, Penalized Splines, And Mixed Models

Matt Wand, University of Wollongong, Australia
Variational Approximations In Semiparametric Regression

11:30am **Conference photo on the Emerson Alumni Hall stairs to the second floor**

12:00-1:30pm **Lunch** (Gator Comer Dining Center)

1:30-3:30pm **Session 2:** ***Semiparametric Modeling And Theory With Applications To Copulas And Missing Data***
Chair: George Casella, University of Florida

Speakers: Jon Wellner, University of Washington
On And Off Semiparametric Models

Peter Hoff, University of Washington
Marginal Likelihoods Based On Sets, With Applications To Semiparametric Copula Estimation

3:30-5:30pm **Poster Session:** Conference rooms 207 and 208, Emerson Alumni Hall

All workshop sessions meet in the 209 Teaching Classroom, Emerson Alumni Hall

Saturday, January 10, 2009

8:00-8:30am Continental breakfast

8:30-10:30am **Session 3:** ***New Developments In Semiparametric Regression***
Chair: Linda Young, University of Florida

Speakers: Raymond Carroll, Texas A&M University
Hierarchical Functional Data: Semiparametric And Nonparametric Methods For Modeling Functional Dependence, With Application To Colonic Crypt Signaling

Naisyin Wang, Texas A&M University
Semiparametric Latent Feature Regression Models For Data With Longitudinal Covariate Process

10:30-ll:00am **Break**

11:00-1:00pm **Session 4:** ***Prediction In Semiparametric Settings***
Chair: Ron Randles, University of Florida

Speakers: Susan Murphy, University of Michigan
A Prediction Interval For The Misclassification Rate

L. J. Wei, Harvard University
How To Evaluate Added Value Of New Bio- And Genetic Markers Over The Conventional Risk Factors/Markers For Better Prediction Of Patient's Clinical Phenotypes?

1:00-2:30pm **Lunch** - free time

2:30-4:30pm **Session 5:** ***Semiparametric Methods In Biostatistics***
Chair: Hani Doss, University of Florida

Speakers: Butch Tsiatis, North Carolina State University
Improving Efficiency Of Inferences In Randomized Clinical Trials Using Auxiliary Covariates

Wes Johnson, University of California, Irvine
Bayesian Semi-parametric Methods in Biostatistics: A Selective Update

5:30-8:30pm. Barbecue at the home of Mary Christman and Bob Palmer
2219 NW 23rd Terrace, Gainesville, FL 32605

Invited Talks

Hierarchical Functional Data: Semiparametric And Nonparametric Methods For Modeling Functional Dependence, With Application To Colonic Crypt Signaling

Raymond J. Carroll, Texas A&M University

We consider hierarchical functional data experiments that arise naturally in our colon carcinogenesis experiments. The data are at three levels: individual rats have multiple colonic crypts, and within each colonic crypt, multiple measurements of biomarkers are made depending on the location of cells. In such experiments, it is typical to assume that conditional on the individual, the functions at the crypt level are independent. However, the biology suggests that the functions probably are not independent conditional on the individual, and it is of considerable interest to understand whether this is true and, if so, to quantify the degree of dependence.

This problem is an example of a very general class of semiparametric problems that have a nonparametric component that is evaluated multiple times at the individual level, as in for example longitudinal data, clustered data, and matched studies. We provide the general framework, exhibit the semiparametric efficient solution, and describe computationally convenient variants. We then describe the analysis of p27 marker data (p27 is a predictor of programmed cell death). In our first approach, the multiple functions for each individual are modeled through a low-order penalized spline regression with a separable correlation structure. We exhibit frequentisi and Bayesian analyses of these problems, analyses that suggest surprisingly strong functional correlations. We also provide an alternative, completely nonparametric approach.

Marginal Likelihoods Based On Sets, With Applications To Semiparametric Copula Estimation

Peter Hoff, University of Washington

Often the primary interest of an analysis of multivariate data is in the associations among the variables, and not the scale on which the individual variables are measured. In such situations it is appropriate to analyze the data using a copula model, in which the associations among the variables are parameterized separately from their univariate marginal distributions. For continuous data a likelihood based on the observed ranks can be used to estimate the copula parameters, and the univariate marginal distributions need not be specified or estimated. If the data include discrete ordinal variables, then the rank likelihood must be modified to accommodate the possibility of ties. The resulting "extended rank likelihood" is based upon a subset of the information of the observed data, but is not based on the sampling distribution of a statistic. In this talk we will present an application of the extended rank likelihood, show how parameter estimates can be obtained, and discuss some of its theoretical properties. A few other semiparametric problems will be presented, for which a similar type of likelihood can be used.

Bayesian Semi-Parametric Methods In Biostatistics: A Selective Update

Wesley Johnson, University of California, Irvine

We review some recent developments in the application of Bayesian nonparametric methodology to semi-parametric problems in the areas of receiver operating characteristic curve estimation, survival analysis with and without time dependent covariates, modeling longitudinal data and jointly modeling longitudinal and survival data. We begin with a brief review of Mixtures of Polya Trees and Dirichlet Process Mixtures, followed by illustrations based on real data. An emphasis is given to selecting among classes of semi-parametric models eg. in survival analysis with time dependent covariates, we may wish to choose among proportional hazards, proportional odds and Cox and Oaks accelerated failure time models.

*A Prediction Interval For The Misclassification Rate**

Susan Murphy*, *University of Michigan

When small data sets are used for classification it is crucial that we provide some measure of confidence for the estimated misclassification rate. However the misclassification rate is a non-smooth function of the classifier. Furthermore the estimated rate suffers from bias due to over-fitting and the is classification rate is a "minimized" quantity. For all of these reasons the construction of measures of confidence such as estimates of variance and confidence/prediction intervals are challenging. We discuss this problem and propose a method based on the use of a smooth upper bound combined with the bootstrap. This upper bound utilizes the surrogate loss that is used in the construction of the classifier.

*Co-author Eric Laber

Semiparametric Regression, Penalized Splines, And Mixed Models

Dave Ruppert, Cornell University

A semiparametric regression models combines parametric and nonparametric components. Penalized splines can model the nonparametric components using a pre-determined basis that is rich enough to avoid under-fitting. Over-fitting is prevented by a roughness penalty, and penalized splines include classical smoothing splines as a special case. A penalized splines can be viewed as a BLUP in a mixed model or as an empirical Bayes estimator. The mixed mixed viewpoint is especially convenient for applications because of its conceptual simplicity, because it allows the use of readily available software, and especially because it can incorporate random subject-specific effects as well. Additive models, single-index models, and nonlinear regression models can be fit relatively simply. Penalized spline methods are arguably the most effective methods known for nonparametric regression with covariate measurement errors.

The first part of this talk will be an overview of mixed-model splines for semiparametric regression circa 2003 when "Semiparametric Regression" was published by Ruppert, Wand, and Carroll. The second half will survey work that has been published since then.

One particularly interesting development in the past few years has been an asymptotic theory for penalized splines. There have been two parallel developments. In one the number of knots is a smoothing parameter and the asymptotics are similar to those of un-penalized least-squares splines. In the second, the number of knots increase sufficiently fast that it does not play the role of a smoothing parameter. In this case, the asymptotics are similar to those of smoothing splines and, somewhat surprisingly, the asymptotic distribution does not depend on the degree of the spline, only the order of the penalty.

*Improving Efficiency Of Inferences In Randomized Clinical Trials Using Auxiliary Covariates**

Anastasios A. Tsiatis, North Carolina State University

The primary goal of a randomized clinical trial is to make comparisons among two or more treatments. For example, in a two-arm trial with continuous response, the focus may be on the difference in treatment means; with more than two treatments, the comparison may be based on pairwise differences. With binary outcome, pairwise odds-ratios or log-odds ratios may be used. In general, comparisons may be based on meaningful parameters in a relevant statistical model. Standard analyses for estimation and testing in this context typically are based on the data collected on response and treatment assignment only. In many trials, auxiliary baseline covariate information may also be available, and there has been considerable debate regarding whether and how these data should be used to improve the efficiency of inferences. Taking a semiparametric theory perspective, we propose a broadly-applicable approach to achieving more efficient estimators and tests in the analysis of randomized clinical trials, where “adjustment" for auxiliary covariates is carried out in such a way that concerns over the potential for bias and subjectivity often raised for other covariate adjustment methods may be obviated. Simulations and applications demonstrate the performance of the methods.

*This is joint work with Marie Davidian, Min Zhang, and Xiaomin Lu.

Variational Approximations In Semiparametric Regression

Matt Wand, University of Wollongong, Australia

Variational approximations are a body of analytic procedures for handling difficult probability calculus problems. They have been used extensively in Statistical Physics and Computer Science. Variational approximations offer an alternative to Markov chain Monte Carlo methods and have the advantage of being faster and not requiring convergence diagnoses, albeit with some loss in accuracy. Despite the growing literature on variational approximations, they currently have little presence in mainstream Statistics. We describe recent work on the transferral and adaptation of variational approximation methodology to contemporary Statistics settings such as generalised linear mixed models and semiparametric regression. This talk represents joint research with and Professor Peter Hall and Dr John T. Ormerod.

Semiparametric Latent Feature Regression Models For Data With Longitudinal Covariate Process

Naisyin Wang, Texas A&M University

We consider a joint model approach to study the association of nonparametric latent features of multiple longitudinal processes with a primary endpoint. We propose estimation procedures and corresponding supportive theory that allow one to perform investigation without making distributional assumption of the latent features. The uncertainty which is associated with accounting for these latent features is also properly taken into consideration. We investigate the practical implications behind certain theoretical assumptions, which aims at having a better understanding of where the estimation variation lies. Numerical performances of the proposed approach were illustrated through simulations and a hypertension study.

How To Evaluate Added Value Of New Bio- And Genetic Markers Over The Conventional Risk Factors/Markers For Better Prediction Of Patient's Clinical Phenotypes?

L.J. Wei, Harvard University

Recent technology advancements for obtaining bio- and genetic-markers have drastically enhanced the knowledge of certain disease processes and the potential for accurately predicting patient's clinical outcomes. Traditional statistical methods for the so-called individualized/ personalized medicine with such markers are derived under a rather strong assumption, that is, one can accurately identify the true model (at least for the large sample case), which relates the predictors to their corresponding clinical phenotype variable(s). In practice, however, it is difficult if not impossible, even to identify the class of models which contains the true one. Therefore, it is interesting and important to investigate whether the standard statistical methods for model estimation, evaluation and comparisons can be modified when the fitted model may not be correctly specified. In this talk, we discuss new procedures for predicting future observations and for evaluating and comparing prediction rules. One key feature of the proposals is that their validity does not require that assumption that the fitted models are correct. Moreover, the new proposal provides a reliability measure of the estimated prediction precision, an important component for model evaluation and checking. The new methods are illustrated with examples with continuous, binary and censored responses. We also use these examples to show how to estimate the added value from bio- and genetic markers over the routinely obtained clinical markers for predicting the clinical outcomes. Lastly even if, on an average sense, the markers are useful (or not useful), it is important to identify subgroup of patients who would benefit from the new markers. We will discuss new proposals how to locate such a subgroup.

On And Off Semiparametric Models

Jon A. Wellner, University of Washington

The theory of semiparametric models gives a wealth of results concerning efficient estimation when various semiparametric models hold. In this talk I will present several recent results concerning estimation on neighborhoods of semiparametric models. The main concern will be stability of the efficiency property on local and non-local neighborhoods. I will also make connections to some problems involving missing data and empirical likelihood estimation.

Poster Abstracts

Resampling Methods For Gmm Models

Mihai Giurcanu, University of Louisiana at Lafayette

In this research project, I develop consistent resampling methods for generalized method of moments models (Hansen, 1982). I prove that both the biased-bootstrap and the uniform bootstrap estimators of the null distributions of the t-test and Wald test statistics are consistent. Moreover, I prove that, the classical bootstrap estimator of the null distribution of the test of over identifying restrictions is inconsistent. This result shows that the asymptotic level of the *naive* bootstrap test is not equal to the nominal level, result which is in agreement with the empirical findings in the econometrics literature of over-inflated critical values of the uniform bootstrap test. Moreover, I prove that if the moment conditions are recentered, then the bootstrap estimator of the distribution of the null distribution is consistent even though the data may fail to satisfy the null hypothesis, and that the resulting modified bootstrap test is consistent as well. Because of its parametric nature, importance sampling can be successfully used when the biased-bootstrap is iterated, and the resulting procedure yields an efficient and computationally feasible bootstrap recycling algorithm. In a small empirical study, the bootstrap confidence intervals based on the biased-bootstrap recycling perform well compared with other resampling procedures in terms of empirical coverage errors.

*A Semiparametric Approach To Incorporating Systematic Uncertainties Into Bayesian X-Ray Spectral Fitting**

Hyunsook Lee, Harvard-Smithsonian Center for Astrophysics

We develop a unique methodology to incorporate systematic uncertainties into X-ray spectral fitting analysis. These uncertainties have been ignored in calibrating noisy astronomical data and as a consequence, error bars of interesting parameters are generally underestimated. Our strategy combines parametric Bayesian spectral fitting and nonparametric approximation of the detector characteristics, the source of systematic uncertainties. We describe our implementation of this method here, in the context of recently codified *Chandra* effective area uncertainties. We estimate the posterior probability densities of absorbed power-law model parameters that include the effects of systematic uncertainties. We apply our method to both simulated as well as actual *Chandra* ACIS-S data. Because of the modular structure of the Bayesian spectral fitting technique, incorporating such uncertainties can be executed simultaneously within the Markov chain Monte Carlo method. Therefore, our strategy itself does not significantly affect the overall computing time but offers adequate parameter estimates and error bars.

*This is a joint work with Hyunsook Lee, Vinay L. Kashyap, Jeremy J. Drake, Alanna Connors,Rima Izem, Taeyoung Park, Pete Ratzlaff, Aneta Siemiginowska, David A. van Dyk, Andreas Zezas

*Bayesian Inference in Semiparametric Mixed Models for Longitudinal Data**

Yisheng Li, University of Texas M. D. Anderson Cancer Center

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., a 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 a Bayesian formulation of a cubic smoothing spline, and the random effect distribution using a normal distribution and alternatively 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 implies weak identifiability for the fixed effects, and can therefore lead to biased estimators and poor inference for the regression coefficients and the spline estimator of the nonparametric function. We propose an adjustment using a post-processing technique. We show that under mild conditions the posterior is proper under the proposed USP, a flat prior for the fixed effect 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.

*This is a joint work with Xihong Lin and Peter Mueller.

*A Statistical Approach to Understand Natural Language**

Dongyu Lin, University of Pennsylvania

Human languages are so complex that understanding natural language is highly challenging. Currently many of the best systems for language are statistical in nature: the best parsing models use hidden state Markov chains, and the best topic models are based heavily on statistics. We build a model that allows combining both of these models in one linear model. Ours is a state-space model where we view each document as a sequence of observations (words) living in a high dimensional space. To allow for many different *topics* to be represented, we need the hidden state to be of very high dimension (at least several thousand). So in spite of the fact that the model could be estimated simply by classical methods, for example, Kalman filters, the massive nature of the model requires new techniques. To allow tractability, these new techniques have to be simple and fast. In particular we are using, exponential smooths, and CCA's (canonical correlation analysis) in estimating our system.

We test our system by analyzing articles from the wikipedia. The problem we set ourselves is to identify the correct page to link an internal link to. We focus on words that have ambiguous meanings. Hence our system must resolve the "meaning" of the word before it can disambiguate it and determining the correct linking page. We do this by determining a high dimensional hidden state the goes along with each word and then matching it to the average hidden state in the possible pages to link to. To help understand how our model works, we applied it to two other datasets, one from education and the other from global warming.

*This is a joint work with Dean P. Foster, Sham M. Kakade and Lyle H. Ungar.

Classification And Gene Selection Of Cancer Microarrays By Penalized Conditional Logistic Regression

June Luo, Clemson University

Classification of patient samples is an important aspect of cancer diagnosis and treatment. The support vector machine (SVM) has been successfully applied to microarray cancer diagnosis problems. However, one weakness of the SVM is that given a tumor sample, it only predicts a cancer class but does not provide any estimate of the underlying probability. The penalized logistic regression (PLR) has the advantage of additionally providing an estimate of the underlying probability with which the sample will be assigned to a class, but it does not offer an estimate of the probability of the class conditional upon an individual gene expression level. We propose the penalized conditional logistic regression (PCLR) model, which is an alternative method for the microarray cancer classification, and for estimating the underlying probability of the outcome given an individual gene expression level. Since the gene selection acts as a primary goal in microarray cancer diagnosis problem, we propose a new method called modified univariate ranking (MUR) for dimension reduction besides the application of univariate ranking (UR), the ration of between-group to within-group sum of squares (BSS/WSS) and recursive feature elimination (RFE). Empirical results on leukemia and breast cancer prognosis data indicate the PCLR combined with one gene selection method (MUR, BSS/WSS or RFE) tends to perform superior on both CV-error and test error rate than SVM and PLR.

*Spline-Backfitted Kernel Smoothing Of Partially Linear Additive Model**

Shujie Ma, Michigan State University

Under weak conditions of smoothness and mixing, we propose spline-backfitted kernel (SBK) and spline-backfitted local linear (SBLL) estimators of the component functions for a partially linear additive model that is both computationally expedient for analyzing high dimensional large time series, and theoretically reliable as the estimator is oracally efficient and comes with asymptotically confidence intervals. Simulation experiments have provided strong evidence that supports the asymptotic theory.

*This is a joint work with Lijian Yang, Michigan State University

Improvements To Kaplan Meier With Proportional Hazards Applications

Kagba Suaray, California State University

The problem of estimating survival probabilities has been of utmost importance to statisticians, engineers, and epidemiologists alike. The occurrence of censoring renders classical statistical inference infeasible. We discuss improvements to the classical Kaplan Meier estimator achieved via an innovative application of kernel smoothing. Applications to the Cox proportional hazards model are investigated.

Semiparametric Estimation Of ARCH(co) Model

Lily Wang, University of Georgia

We analyze a class of semiparametric ARCH models that nests the simple GARCH(1,1) model but has flexible news impact function. A simple estimation method is proposed based on profiled polynomial spline smoothing. Under regular conditions, the proposed estimator of the dynamic coeffcient is shown to be root-n consistent and asymptotically normal. A fast and efficient algorithm based on fast fourier transform (FFT) has been developed to analyze volatility functions with infinitely many lagged variables within seconds. We compare the performance of our method with the commonly used GARCH(1, 1) model, the GJR model and the method in Linton and Mammen (2005) through simulated data and various stock return series. For the S&P 500 index returns, we find further statistical evidence of the nonlinear and asymmetric news impact functions.

Linear Mixed Regression Models For A Functional Response And A Functional Predictor

Wei Wang, University of Pennsylvania

The functional regression model with both the predictor and the response being functions has been studied under different settings with different emphases. Dependence of the response on the predictor is through a bivariate functional regression coefficient which is of particular interest. A standard approach studying the model is to approximate the predictor with basis function expansions. In our approach, we use the eigenfunction expansion with random effects. We assume the random effects have an unstructured covariance matrix as empirically the estimated eigenfunctions, particularly smoothed, may not be perpendicular to each other.

Our modeling is in a linear mixed models framework and thus model identifiability checking is a necessary step. The model identifiability verification shows that under very mild conditions, the modeling error in both the predictor and the response can have covariance structures other than the conventionally homogeneous structure.

Previous research work of the model has been focused on estimation of the regression coefficient. In our study, we also find the standard errors of the estimate which is a bivariate function as well. An interesting observation is that the estimated regression coefficient is smooth but the standard errors are generally not. Penalty terms need to be added to ensure smoothness of the standard

errors. Based on the estimate and its standard errors, we proposed a test statistic to test nullity of the regression coefficient.

Semiparametric Inference On A Class Of Wiener Processes

Xiao Wang, University of Maryland, Baltimore

This article studies the estimation of nonhomogeneous Wiener process model for degradation data. A pseudo-likelihood method is proposed to estimate the unknown parameters. An attractive algorithm is established to compute the estimator under this pseudo-likelihood formulation. We establish the asymptotic properties of the estimator, including consistency, convergence rate and asymptotic distribution. Random effects can be incorporated into the model to represent the heterogeneity of degradation paths by letting the mean function be random. The Wiener process model is extended naturally to a normal inverse Gaussian process model and similar pseudo-likelihood inference is developed. A score test is used to test the presence of the random effects. Simulation studies are conducted to validate the method and we apply our method to a real dataset in the area of health structure monitoring.