Program
IMS Mini-Meeting on Functional Data Analysis

Department of Statistics
University of Florida

January 9–11, 2003

Contents

Sponsors

The National Science Foundation; Info Tech, Inc.; College of Liberal Arts and Sciences, University of Florida; and the Graduate School, University of Florida.

Organizing Committee

George Casella, Sam Wu, Jim Booth, Jim Hobert, Bhramar Mukherjee, Brett Presnell, Clyde Schoolfield, and Alex Trindade.

Invited Speakers

R. L. Eubank, Jianqing Fan, Iain Johnstone, Robert E. Kass, Charles Kooperberg, Mary J. Lindstrom, J. S. Marron, Ian McKeague, J. O. Ramsay, Bernard Silverman, Brani Vidakovic, Naisyin Wang, Colin Wu

Other Participants

Alan Agresti, Naomi Altman, Michael Andrew, Jay Beder, Sam Behseta, Dean Billheimer, Brian Caffo, Marinela Capanu, Keith Carlson, Fuxia Chen, Eric Chicken, Joel Dubin, David Finlay, Herwig Friedl, Daniel Gervini, Malay Ghosh, Mark Greenwood, Nels Grevstad, Serge Guillas, Jaroslaw Harezlak, David Hitchcock, Erin Hodgess, Rima Izem, Wolfgang Jank, Andre Khuri, Bernhard Klingenberg, Jack C. Lee, Ramon Littell, Fei Long, Sara Lopez-Pintado, Elizabeth Malloy, Eric Matzner-Lober, James McClave, Julie McIntyre, Yongyi Min, Masahiro Mizuta, Jeffrey Morris, Yolanda Munoz Maldonado, Todd Ogden, Robert Paige, Trevor Park, Emanuel Parzen, Debashis Paul, Marianna Pensky, Mohsen Pourahmadi, Ron Randles, Andrew Rosalsky, Ananya Roy, Pavlina Rumcheva, Dan Spitzner, Peter Wludyka, Rongling Wu, Jun Yan, Mark Yang Linda Young, Hao (Helen) Zhang, Xin Zhao, Yichuan Zhao

Acknowledgements

The organizers thank the Department of Statistics staff, and especially Carol Rozear, Marilyn Saddler, and Robyn Crawford, for their tremendous efforts in helping to set up this meeting and making it run smoothly.

Thursday, January 9

7:00–10:00 p.m. Reception — Keene Faculty Center (Dauer Hall)

Friday, January 10

8:00–8:45 a.m.	Breakfast	272/288 Reitz Union

8:45–10:20 a.m.	SESSION 1: MODELLING FUNCTIONAL DATA Chair: George Casella	282 Reitz Union

George Casella	Welcoming Remarks
J. O. Ramsay	From Data to Differential Equations
Mary J. Lindstrom	Semiparametric, mixed-effects models for functional data

10:20–11:10 a.m.	Group Photo/Break (refreshments)	272/288 Reitz Union

11:10–12:30 p.m.	SESSION 2: BAYESIAN METHODS AND FUNCTIONAL DATA Chair: Brett Presnell	282 Reitz Union

Ian McKeague	A Statistical Model for Signatures
Robert E. Kass	Bayesian Functional Data Analysis and its Application to Neuron Firing Patterns

12:30–2:00 p.m.	Lunch	Gator Corner Dining Center

2:00–4:00 p.m.	SESSION 3: DISCRIMINATION AND LONGITUDINAL DATA Chair: Alex Trindade	282 Reitz Union

J. S. Marron	Distance Weighted Discrimination
Naisyin Wang	Marginal Non- and Semi-parametric Regression for Longitudinal Data
Jianqing Fan	New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis

4:00–5:00 p.m.	Poster Session 1 (refreshments)	272/288 Reitz Union

Saturday, January 11

8:00–9:00 a.m.	Breakfast	272/288 Reitz Union

9:00–10:20 a.m. SESSION 5: SPARSE AND IRREGULARLY MEASURED DATA — 282 Reitz Union
Chair: Jim Booth

Bernard Silverman	Aspects of the lifting scheme for interpolation and smoothing on irregular grids
Iain Johnstone	Sparse Principal Components Analysis

10:20–11:10 a.m.	Break/Poster Session 2 (refreshments)	272/288 Reitz Union

11:10–12:30 p.m. SESSION 6: FUNCTIONAL REGRESSION AND ANOVA — 282 Reitz Union
Chair: Bhramar Mukherjee

R. L. Eubank	Inference in Functional Regression Models
Brani Vidakovic	Optimal Testing in Functional Analysis of Variance Models

12:30–2:30 p.m.	Lunch	

2:30–3:50 p.m. SESSION 7: FUNCTIONAL REGRESSION AND GENETICS — 282 Reitz Union
Chair: Sam Wu

Colin Wu	A Statistical Method for Adjusting Covariates in Linkage Analysis with Sib Pairs
Charles Kooperberg	Logic Regression

4:30–8:30 p.m.	Barbecue	Chez Casella

Abstracts

Inference in Functional Regression Models

R. L. Eubank
Texas A & M University

Smoothing spline estimators are considered for inference in finite dimensional functional regression models. A Kalman filter type recursion is used to produce an efficient computational algorithm for the coefficient estimators and other related quantities. In particular, efficient computational methods are developed for the Bayesian "confidence intervals" associated with the spline estimators and for cross-validation type criteria for selecting the levels of smoothing. The methodology is applied to a problem of sales prediction for the Texas Lotto game.

New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis

Jianqing Fan
University of North Carolina

Semiparametric regression models are very useful for longitudinal data analysis. The complexity of semiparametric models and the structure of longitudinal data pose new challenges to parametric inferences and model selection that frequently arise from longitudinal data analysis. In this paper, two new approaches are proposed for estimating the regression coefficients in a semiparametric model. The asymptotic normality of the resulting estimators is established. An innovative class of variable selection procedures is proposed to select significant variables in the semiparametric models. The proposed procedures are distinguished from others in that they simultaneously select significant variables and estimate unknown parameters. Rates of convergence of the resulting estimators are established. With a proper choice of regularization parameters and penalty functions, the proposed variable selection procedures are shown to perform as well as an oracle estimator. A robust standard error formula is derived using a sandwich formula, and empirically tested. Local polynomial regression techniques are used to estimate the baseline function in the semiparametric model.

Joint work with Runze Li, Pennsylvania State University

Sparse Principal Components Analysis

Iain Johnstone
University of California, Berkeley

The transient nature of certain signals, such as electrocardiogram (ECG) traces, suggests a modification of PCA that exploits sparsity. We developed algorithms of the following type: a) represent the data in a basis, so far wavelets, in which the 'true' coefficients are likely sparse, (b) select a reduced number of co-ordinates in this basis with an overwhelming fraction of sample variance, (c) run standard PCA on the selected coordinates, (d) use thresholding to filter out noise in the estimated eigenvectors, and (e) re-express the reconstruction in the original basis. The method is contrasted with quadratically regularized (functional) PCA, and illustrated on some treadmill test ECG data. Some (in)consistency results provide theoretical justifcation for the method. This is joint work with Arthur Lu.

Bayesian Functional Data Analysis and its Application to Neuron Firing Patterns

Robert E. Kass
Carnegie Mellon University

One of the most important techniques in learning about the functioning of the brain has involved examining neuronal activity in laboratory animals under varying experimental conditions. Neural information is represented and communicated through series of action potentials, or spike trains, and the central scientific issue in many studies concerns the physiological significance that should be attached to a particular neuron firing pattern in a particular part of the brain. Electrophysiological investigations typically involve recordings from dozens or hundreds of neurons. My colleagues and I have formalized specific scientific questions in terms of point process intensity functions, and have used Bayesian methods to fit the point process models to neuronal data. In my talk I will describe the neurophysiological setting in order to motivate a Bayesian approach to functional data analysis via free-knot splines. I will also try to assess the utility of our Bayesian methodology, and to characterize the situations in which it has something more to offer than simpler and more standard FDA methods. This is joint work with Sam Behseta and Garrick Wallstrom.

Logic Regression

Charles Kooperberg
Fred Hutchinson Cancer Research Center

We are concerned with (generalized) regression problems in which all (most) of the predictors are binary, and in which our interest is to discover potential high order interactions between these predictors. This is a situation that, for example, arises in the analysis of SNP data.

Our approach will be to construct new predictors that are logic (Boolean) combinations of the binary predictors. I will discuss how we can construct these models, as well as how to deal with the model selection issue. I will provide an example using (simulated) SNP data of the genetic analysis workshop, as well as an example in which we try to identify some simple risk factors.

This is joint work with Ingo Ruczinski and Michael LeBlanc.

Semiparametric, Mixed-Effects Models for Functional Data

Mary J. Lindstrom
University of Wisconsin - Madison

Hierarchical, mixed-effects models are an attractive approach to modeling functional data. However, in some cases the required parametric model relating the predictor variable and the response variable does not exist. In these situations it is often possible to use a technique called self-modeling to pool shape information across curves to estimate a common, fixed, non-parametric “shape function”. Individual data curves are then fit by combining random, individual-specific transformations and the common shape function. In the model’s simplest form the individual transformations are linear. The model can be extended to allow for nested curves, shape functions which vary systematically between groups of curves (treatment effects), shape functions which vary randomly over curves, and time-warping (flexible, non-parametric, individual-specific transformations of the time scale).

Distance Weighted Discrimination

Steve Marron
University of North Carolina

High Dimension Low Sample Size statistical analysis is becoming increasingly important in a wide range of applied functional data contexts. In such situations, it is seen that the appealing discrimination method called the Support Vector Machine can be improved. The revealing concept is ”data piling” at the margin. This leads naturally to the development of ”Distance Weighted Discrimination,” which also is based on modern computationally intensive optimization methods, and seems to give improved ”generalizability.” This is joint work with Michael Todd (Cornell University).

A Statistical Model for Signatures

Ian McKeague
Florida State University

A Bayesian model for off-line signature analysis involving the representation of a signature through its curvature is developed. The prior model makes use of a spatial point process for specifying the knots in an approximation restricted to a buffer region close to a template curvature, along with an independent time warping mechanism. In this way, prior shape information about the signature can be built into the analysis. The observation model is based on additive white noise superimposed on the underlying curvature. The approach is implemented using MCMC and applied to a collection of documented instances of Shakespeare's signature.

From Data to Differential Equations

J. O. Ramsay
Mcgill University

Differential equations (DIFE's) can represent the underlying processes giving rise to observed functional data, and as such can offer a number of potential advantages over parametric or basis expansion models. They explicitly model the behavior of derivatives, and derivative estimates based on DIFE's are usually superior to those derived from conventional data smoothers. Solutions to a linear DIFE of order m span an m-dimensional space, and consequently have the capacity to model curve-to-curve variation as well as to fit the data. We can build known structure features into DIFE models more easily than is usually the case for conventional functional models. And finally a DIFE offers a wider range of ways to introduce stochastic behavior into models.

Earlier DIFE identification methods such as principal differential analysis required the intermediate step of first smoothing the data. We will discuss a technique for going directly from the discrete and noisy data to a DIFE that is based on the work of Heckman and Ramsay (2000). Some illustrations of its performance for simulated data will be offered as well as an example from process control in chemical engineering.

Joint work with R. Nuzzo, Mcgill University

Heckman, N. and Ramsay, J. O. (2000) Penalized regression with model-based penalties. *The Canadian Journal of Statistics*, 28, 241-258.

Aspects of the Lifting Scheme for Interpolation and Smoothing on Irregular Grids

Bernard Silverman
University of Bristol

A key ingredient of functional data analysis is the construction of functional observations from discrete data. In one dimension it is reasonable to assume that functions are observed on very fine grids, but once we move to two or more dimensions there are many contexts in which this will not be so. The lifting scheme constructs a basis appropriate for interpolation of data such as soil survey data collected on an irregular grid of points. The scheme essentially constructs a basis of functions of increasing scale. Work in progress with Maarten Jansen and Guy Nason includes the construction of lifting approaches based on minimum spanning trees, monotone thresholding approaches, and refinements of lifting algorithms based on Voronoi polygons.

Optimal Testing in Functional Analysis of Variance Models

Brani Vidakovic
Georgia Institute of Technology

The testing problem in a general functional analysis of variance model is considered. The null hypotheses that the main effects and/or the interactions are zeros against the composite nonparametric alternative hypotheses that they are separated away from zero in L^2-norm are tested. We adapt the minimax functional hypothesis testing procedures for testing a zero signal in a Gaussian "signal plus noise" model to derive asymptotically (as the noise level goes to zero) minimax nonadaptive and adaptive functional hypothesis testing procedures for the main effects and/or the interactions based on the empirical wavelet coefficients of the data. Wavelet decompositions allow one to characterize different types of smoothness conditions assumed on the response function by behavior of its wavelet coefficients for a wide range of function classes. In order to shade some light on the theoretical results obtained, a small simulation study will be presented to illustrate the finite sample performance of the proposed functional hypothesis testing procedures. We also apply these tests to several real-life data examples arising from medicine and geosciences. Concluding remarks and hints for possible extensions of the discussed methodology are also given.

This is joint work with Felix Abramovich, Anestis Antoniadis, and Theofanis Sapatinas.

Marginal Non- and Semi-parametric Regression for Longitudinal Data

Naisyin Wang
Texas A&M University

There has been a substantial recent interest in investigating the performance of non- and semiparametric marginal estimation using kernel methods. Most approaches adopt the strategy of ignoring the within-cluster correlation structure either in nonparametric curve estimation or throughout. When the cluster size m remains fixed, a result supporting the use of this "working independence" strategy indicates that under the conventional estimation procedure, a correct specification of the correlation structure actually diminishes the asymptotic efficiency. In this presentation, I will discuss an alternative kernel estimating equation that accounts for the within subject correlation. For nonparametric curve estimation, the variance of the proposed method is uniformly smaller than that of the most efficient working independence approach. Under the framework of marginal generalized partially linear models, the new estimator is semiparametric efficient in the Gaussian case, and is more efficient than the working independence estimator in non–Gaussian cases. Application of the proposed methods to medical data will be discussed.

Joint work with Ray Carroll, Zonghui Hu and Xihong Lin.

A Statistical Method for Adjusting Covariates in Linkage Analysis with Sib Pairs

Colin Wu
National Heart Lung and Blood Institute

We propose a statistical method, including models and estimation procedures, for adjusting the effects of age and other covariates of interest in linkage analysis involving data from sib pairs. Our methodology consists of three main components: (a) modeling the covariate adjusted population means of quantitative traits through regression; (b) estimating the covariate adjusted quantitative traits; and (c) evaluating the linkage between the adjusted trait values and the marker alleles shared identical by descent. We apply our method to a subset of the Framingham Heart Study in detecting linkage between marker alleles and systolic blood pressure within sib pairs, and compare our results with the ones obtained from the original Haseman-Elston method. Our regression estimates under the adjusted Haseman-Elston model are less variable than the unadjusted model, suggesting that covariate adjustment may be generally preferable.

This is joint work with Gang Zheng, JingPing Lin, Eric Leifer and Dean Follmann, all at NHLBI.

Poster Abstracts

Inference for Time-Invariant Covariates in Self-Modeling Regression with Mixed Effects

Naomi S. Altman
Penn State University

In many longitudinal studies, the response curves are similar among subjects, while differences between subjects manifest themselves as differences in scaling of either the time or response axes or both. When the shape of the response curve is known a priori, the scaling factors may be incorporated into a nonlinear mixed model. This gives an efficient parametrization which allows for the estimation and testing of mixed model terms to summarize covariate and subject effects. Self-modeling regression model is a semi-parametric model in which the shape function is specified nonparametrically, while preserving the use of mixed model terms to summarize covariate and subject effects. This poster suggests an automatic method for fitting a special case of self-modeling regression, shape invariant regression, and discusses partial likelihood ratio tests for the nonlinear mixed model parameters.

Muscle-Like Activity of M1 Neurons During Multi-Joint Movements

Sam Behseta
Carnegie Mellon University

In an experiment concerning the function of primary motor cortex (M1), data from 347 neurons were collected in Dr. Peter Strick's laboratory at the Center for the Neural Basis of Cognition (CNBC) in Pittsburgh.

Monkeys performed sequential pointing movements guided by vision or part of a highly practiced repeating sequence. We analyzed the activity of 347 M1 neurons, and EMG recordings from multiple axial, shoulder, arm, forearm, wrist and finger muscles under those two conditions.

We fit the neuronal firing rate as a function of time using cubic splines by assuming the firing rate is a Poisson process intensity function and applying Poisson regression. We used Bayesian Adaptive Regression Splines (BARS) as described in DiMatteo, Genovese, Kass (2001). This approach produced 347 functions of time as representations of the neuronal firing rates.

A statistical clustering procedure based on the first few principal components applied to these functions, the fitted curves, identified a group of M1 neurons (n=16) with a distinct multi-phasic pattern of activity during the two types of pointing movements.

Two of the recorded muscles displayed a comparable pattern of activity during task performance. We computed the correlation between the EMG activity of these two muscles and a smoothed firing rate function for all 347 recorded neurons, after shifting the firing rate function to maximize the correlation. These results suggest that, even during multi-joint pointing movements, some M1 neurons display muscle-like properties.

A Functional Data Approach to MALDI-TOF MS Protein Analysis

Dean Billheimer
Vanderbilt University

Matrix-assisted laser desorption-ionization, time-of-flight (MALDI-TOF) mass spectrometry (MS) is emerging as a leading technology in the proteomics revolution. MALDI-TOF MS allows direct measurement of the "protein signature" of tissue, blood, or other biological samples, and holds tremendous potential for disease diagnosis and treatment. Despite recent technical advances in MALDI-TOF signal generation, key challenges remain in signal normalization and quantitation. Methods of functional data analysis (FDA) promise a principled approach for evaluating these complicated biological signals. I introduce an FDA approach to investigate protein mass spectra from brain tumors (gliomas) and normal white matter tissue samples. The goal is to characterize protein expression differences in tumor and normal brain tissue.

Wavelet Estimation of the Probability Density Function at the Boundary

Keith Carlson and Marianna Pensky
University of Central Florida

The wavelet estimator of the probability density function at the boundary is constructed by modifying the regular wavelet kernel. Performance of the estimator is studied analytically and numerically.

Error Density and Distribution Function Estimation in Nonparametric Regression

Fuxia Cheng
Illinois State University

This talk will discuss some asymptotics of some error density and distribution function estimators in nonparametric regression models. In particular, the histogram type density estimators based on nonparametric regression residuals obtained from the full sample are shown to be uniformly almost surely consistent. Uniform weak convergence with a rate is obtained for the empirical d.f. of these residuals. Furthermore, if one uses a part of the sample to estimate the regression function and the other part to estimate the error density, then the asymptotic distribution of the maximum of a suitably normalized deviation of the density estimator from the true error density funtion is the same as in the case of the one sample setup. Simiarly, a suitably standardized nonparametric residual empirical process based on the second part of the sample is shown to weakly converge to a time transformed Brownian bridge.

Nonparametric Regression on Random Samples via Wavelets

Eric Chicken
Florida State University

Wavelets have been shown to be very successful in nonparametric function estimation. When the positions of the sampling points are equispaced, wavelets excel in the areas of spatial adaptivity, optimality, and low computational cost. Alternate wavelet methods have been devised for fixed, nonequispaced sample points and samples with random design. Unfortunately, these methods are much more complex from a computational standpoint than their equispaced counterparts and can even lead to suboptimal estimators. To maintain computational efficiency, it is desirable to use equispaced methods.

Using term-by-term thresholding, it has been shown that equispaced wavelet methods can be directly applied to sample points distributed as independent uniform random variables without a loss in the rate of convergence, i.e., to within a logarithmic factor of the optimal minimax convergence rates. This method maintains the computational efficiency and simplicity of the equispaced algorithm.

In this session, these results for nonparametric regression with randomly placed sampling points are improved upon. First, through the use of block thresholding, the logarithmic penalty in the convergence rate has been removed. Second, the scope of the function spaces has been enlarged to include not only the Holder spaces studied previously, but many Besov spaces as well. Third, moment conditions are given that define a class of random distributions for which these results hold. Fourth, this fast convergence rate is also attained when the sample points come from a random process. Conditions are given defining the processes for which this method will result in optimal convergence rates. Finally, spatial adaptivity is maintained, and the computational cost remains low since the equispaced algorithm is used. Several numerical examples are examined, including sample points uniformly distributed over an interval, and sample points following the Poisson process.

Correlation for Multivariate Longitudinal Data

Joel Dubin
Yale University

In this poster we will describe methods to handle multivariate longitudinal data or multivariate responses that are followed repeatedly over time. These data are viewed as realizations of a random process. Dependency between the various components of the response is characterized by a non-parametric correlation technique which we refer to as dynamical correlation. The methods utilize smoothed curves constructed from the original data, a functional data analytic approach. The assessment of the dynamics of the underlying processes also includes the consideration of derivatives and of time lags. Our methods are illustrated with data on five acute phase blood proteins measured longitudinally for a sample of hemodialysis patients. Simulation results will also be presented.

Curve Registration via Self-Modeling Warping Functions

Daniel Gervini
University of Zurich

In this poster we introduce a new method of curve registration. The method consists in modeling the warping functions as linear combinations of a (small) number of basis functions or components, which are themselves estimated from the data. This provides a model that is both flexible and parsimonious. The components are, in most cases, easy to interpret, so that the corresponding scores are meaningful parameters. Subsequent statistical analysis can be carried out on these scores, using traditional statistical methods (such as repeated measures analysis of variance). It is important to mention that, unlike other registration methods commonly used, the method we propose requires neither pre-smoothing of the data nor identification of landmarks.

Modeling Shapes of Valley Profiles

Mark C. Greenwood
University of Wyoming

Shapes of valleys are a result of surficial processes ranging from the effects of glaciers to those of rivers. An important question in Geomorphology concerns inference about the dominant process of valley formation from observed cross-valley profile data. Valleys that have undergone extensive glaciation tend to be U-shaped while those formed by other processes, such as rivers, tend toward a V-shape. Specific questions of interest include curvature of observed valley profiles in order to describe the dominant surficial process and determination of transition from glaciated to non-glaciated regions within a valley. Additional information on direction of flow of the valley and bedrock type can be incorporated as covariates.

In the Geology literature, the above questions have been addressed using nave curve-fitting (eg. Harbor and Wheeler 1992; Pattyn and Van Heule 1998). Greenwood and Humphrey (2002) consider improvements using statistical methods of model selection in the context of nonlinear regression. However nonlinear regression may be too restrictive for the profiles that are observed. Improvements to nonlinear regression methods are possible by treating the observed valley profiles as functions of space and applying functional data analysis (FDA) methods as in Ramsay and Silverman (1997). The curvature of U-shaped valleys is distinct from that of V-shaped valleys. As a start, through FDA methods, we describe the curvature of observed profiles using the first two derivatives of the functions. These derivatives are compared to the theoretical derivatives of U and V shapes. Further research includes the use of model selection techniques and the use of covariate information to better describe the shape of profiles in determining the dominant surficial process.

Image Alignment Using Thin-Plate Splines

Nels Grevstad
Purdue University

To compare medical images of the same subject taken at slightly different orientations or of different subjects altogether, the first step is to realign (in the case of a single subject) or 'normalize' (in the case of different subjects) the images so that corresponding pixels (or voxels, in three dimensions) in the two images represent homologous biological points. In this article, an approach to such image realignment and normalization using thin-plate splines is presented. The pixels (voxels) are assumed to be noisy samples of underlying smooth functions on continuous domains, and the aim is to transform the coordinates of one image onto those of the other; no pixel imputation is performed in the estimation of the transformation as is done in other commonly used algorithms that work on discrete domains; and it is not necessary to identify fiducial markers, as is done prior to using the algorithms appearing in recent literature on the use of thin-plate splines to model biological shape deformations.

Among issues addressed are the isotropic invariance of the approach, the computation, and the smoothing parameter selection by generalized cross-validation. Simulated and real brain image data examples are used to illustrate the technique.

Functional Samples and Bootstrap for the Prediction of SO_2 Levels

S. Guillas
University of Chicago

In this paper, enhancements of several functional techniques are given in order to forecast SO_2 levels around a power plant. The data are considered as time series of curves. Assuming a lag one dependence, the predictions are computed using the functional kernel (see Besse and Cardot, 2000) but with local bandwith, and the linear autoregressive Hilbertian model (see Bosq, 2000). We carry out the estimation with a so-called historical matrix, that is a subsample of the original sample stressing the diversity of situations, but in terms of shapes. This idea was introduced in the real time series context in terms of levels by Garca Jurado et al. (1995). Finally, a bootstrap method, following Cao (1999) is introduced, making use of Fraiman and Muniz's order (see Fraiman and Muniz, 2001) for functional data.

Clustering Smoothed Functional Data

David Hitchcock
University of Florida

Cluster analysis is a common exploratory technique for multivariate data which may also be used for functional data. When clustering functional data, it is philosophically desirable to smooth the observed measurements and cluster the smooth curves rather than the raw data. The smoothing process involves a tradeoff between bias and variance which affects the cluster analysis. Simulation results are given to indicate the conditions under which pre-smoothing aids the eventual clustering of the data.

Functional Data Analysis in Continuous Reaction Norms : Identifying Nonlinear Variations

Rima Izem
University of North Carolina, Chapel Hill

A continuous reaction norm in evolutionary biology is a trait or characteristic of an individual or genotype that varies as a continuous function of some aspect of the environment. The trait value as a function of the environment for an individual could be represented by a curve, and the set of curves for distinct clones, families or genotypes in the population shows the genetic variations in that population. Directions of variations of the curves are of importance to evolutionary biologists because they provide information on the genetic constraints to evolution. In this poster, I will first define three directions of variations of biological interest for continuous reaction norms: Faster-Slower direction or vertical shift, Hotter-Colder direction or horizontal shift, and Generalist-Specialist direction or change in width. Then, I will propose a method to identify and quantify those directions. The Faster-Slower direction allows for standard linear analysis (e.g. by PCA), but effective analysis of the other directions has motivated the development of a new non-linear methodology. Results of this methodology will be illustrated on curves of growth rate of caterpillars as a function of temperature.

Exploring Depth for Functional Data

Sara López-Pintado
Universidad Carlos III de Madrid

The main goal of this paper is to introduce a new definition of data depth for functional observations. Given a collection of curves, this idea allows to measure the centrality of a function and it provides a natural center-outward order for the sample functions; thus, statistics based on this order can be defined. The finite dimensional version of the proposed ideas of depth can be seen as a new definition of depth for multivariate data. Zuo and Serfling (2000) introduced in a general framework four key properties a depth should verify: affine invariance, maximality at center, monotonicity and vanishing at infinity. The depth introduced here verifies all these properties. Some other statistical properties such as consistency are also established. Simulated results show that the trimmed mean gives a better performance than the mean when we consider contaminated models. A real data set is analyzed to illustrate our method.

Similarity Analysis of Curves

Yolanda Munoz Maldonado
Texas A&M University

A statistical method for comparing samples of curves from two populations is presented. The approach easily generalizes to comparison of sample curves from more than two populations. The method is used to analyze data obtained by subjecting Gangliosides extracted from brain tissue to a process called thin layer chromatography (TLC). Gangliosides are complex glycolipids found in their highest concentration in the outer membrane of vertebrae neurons of the central nervous system. Density profiles are obtained by scanning densitometry applied to lanes of the TLC gel plate. Regression splines are used to generate a sample of ”curves” from the density profiles. Once in functional form, the sample curves are subjected to the process of registration. The concept of correlation is used to develop several test statistics, which measure similarity of curves within curves. The permutation distribution is used to obtain p-values for testing the null hypothesis of no differences between groups. Simulations are conducted to explore the power of the test.

Diagnostics for Functional Predictors in Logistic Regression

Elizabeth J. Malloy
University of New Mexico

We present a collection of tools for examining the appropriate scale of the functional covariate in a functional logistic regression model. A transformation of the functional covariate is considered. These are a power transformation, a fractional derivative, and a cubic spline transformation with known knots.

We derive a score test for testing the appropriateness of the transformation through the parameter of interest and find estimates for the parameter using an iterative algorithm. For the one dimensional cases profile plots are also examined to find an estimate of the scalar parameter.

These methods are applied to pilot evaluation grades during simulated flights. This data consists of 680 pilots who were graded (pass/fail) on the quality of their landing in a flight simulator. Functional data on various features of the flights were collected and are used to model the pilots' grades. Further illustrations of these techniques are also carried out on simulated data.

Clustering Functional Data

Eric Matzner-Lober
CREST-ENSAI

Data in many different fields come to practitioners through a process naturally described as functional. Although data are gathered as finite dimensional vectors and may contain measurement errors, the functional form has to be taken into account. We propose a clustering procedure of such data emphasizing the functional nature of the objects. The new clustering method consists of two stages: fitting the functional data by B-splines and partitioning the estimated model coefficients using a k-means algorithm. Strong consistency of the clustering method is proved and a real-world example from the food industry is given.

Density Estimation with Replicate Heteroscedastic Measurements

Julie P. McIntyre
North Carolina State University

We present an estimator for the density function of a random variable X given independent observations $\{W_{r,j}\}_{r=1,j=1}^{n,\ m_r}$, where $W_{r,j} = X_r + U_{r,j}$. We derive our estimator under the assumption that $U_{r,j}$ is a normally distributed measurement error having unknown variance σ_r^2. The estimator is a generalization of the deconvolution estimator of Stefanski and Carroll (1990), with the measurement error variances estimated from replicate observations. We derive an expression for the integrated mean squared error and bounds on the integrated variance. The rate of convergence of the estimator is examined for fixed numbers of replicate measurements. For the important case of two replicates, we show under certain assumptions on the measurement error variances, that the integrated mean squared error converges to zero at a rate of $\{\log(n)\}^{-2}$, the optimal rate for deconvolution estimators when the measurement error variance is known. A simulation study is presented illustrating the performance of the estimator.

Wavelet-Based Nonparametric Modeling of Hierarchical Functions in Colon Carcinogenesis

Jeffrey S. Morris
University of Texas

In this work, we develop new methods for analyzing the data from an experiment using rodent models to investigate the effect of type of dietary fat on MGMT, an important biomarker in early colon carcinogenesis. The data consist of observed profiles over a spatial variable contained within a two-stage hierarchy, a structure we dub "hierarchical functional data". We present a new method providing a unified framework for modeling these data, simultaneously yielding estimates and posterior samples for mean, individual and subsample-level profiles, as well as covariance parameters at the various hierarchical levels. Our method is nonparametric in that it does not require the prespecification of parametric forms for the functions, and involves modeling in the wavelet space, which is especially effective for spatially heterogeneous functions as encountered in the MGMT data. Our approach is Bayesian, with the only informative hyperparameters in our model being effectively smoothing parameters. Analysis of this data set yields interesting new insights into how MGMT operates in early colon carcinogenesis, and may depend on diet. Our method is general so can be applied to other settings where hierarchical functional data are encountered.

Clustering of Functional Data for Profiling Placebo Responders

R. Todd Ogden
Columbia University

Identification of placebo responders among subjects treated with active drug has significant clinical and research implications. This paper presents a framework for studying placebo response in diverse areas of medicine. In order to identify placebo responders among drug treated patients, a profile of the clinical status over time (outcome profile) is estimated for each subject. Semi-parametric clustering techniques are used to group subjects based on the amount of curvature in the profile as well as the overall trend in the profile. The resulting clusters produce representative profiles for subjects in the drug group which can subsequently be used to classify patients. To extend this analysis, the cluster results are then compared to data from the placebo treated subjects in the study. The identification of placebo responders in the drug treatment group is based on the common clusters of profiles in the two groups. The proposed method is applied to data from a clinical trial for treatment of depression involving placebo and the active drug phenelzine.

Joint work with Thaddeus Tarpey, Wright State University and Eva Petkova, Columbia University.

Rotation of Principal Components for Stabilization: A Penalized Likelihood Approach

Trevor Park
Cornell University, School of Operations Research and Industrial Engineering

Principal component analysis remains a standard tool for exploration of functional data. To facilitate interpretation of individual components, investigators in many applied sciences sometimes choose to perform *rotations* on selected groups of components — usually orthogonal transformations of the component direction vectors that preserve the subspace they span, yet bring them into closer alignment with an easily interpretable basis for the space, as defined by a rotation criterion like Varimax (Ramsay and Silverman 1997; Richman 1986; Basilevsky 1994). The choice of which components to rotate is not always clear, but especially suitable candidates for rotation are components that are ill-defined, in the sense of having nearly equal eigenvalues (Jolliffe 1989). Estimates of such components can be quite unstable, and this can greatly reduce interpretability of functional principal components (North et al. 1982).

Presented here for the first time is a general rotation method designed to focus on ill-defined components. The method is derived from a penalized likelihood approach to rotation, which suggests maximization of the expression (in the d-dimensional case, with n observations)

$$-\tfrac{n}{2}\sum_{i=1}^{d}\log(\mathbf{u}_i{}^{\mathsf{T}}\hat{\mathbf{\Sigma}}\mathbf{u}_i) \;+\; \rho\, V([\mathbf{u}_1\cdots\mathbf{u}_d]),$$

over orthogonal unit column vectors $\mathbf{u}_1,\ldots,\mathbf{u}_d$ (representing the component directions), where $\hat{\mathbf{\Sigma}}$ is the empirical covariance matrix, V is an orthogonal rotation criterion (like Varimax), and $\rho \geq 0$ is a parameter controlling the degree of rotation. (This method is superficially similar to that of Jolliffe and Uddin (2000), although that method does not specifically treat the problem of ill-defined components and does not optimize simultaneously over all components.) For $\rho = 0$, the solution is the principal components; as $\rho \to \infty$, the components rotate toward the cardinal axes, with ill-defined components having greater susceptibility to rotation. The optimization can be effectively performed with the assistance of an algorithm recently proposed by Jennrich (2001).

A Method for Nonparametric Function Estimation by Wavelets when the Function is Sampled on a Smooth Non-Uniform Grid

Debashis Paul
Stanford University

Wavelets have now become a very useful tool in nonparametric function estimation since it is quite efficient in representing functions which have certain singularities (e.g. discontinuities) at different spatial locations. The practical application of the method is somewhat constrained by the fact that in order to apply the fast wavelet transform algorithms to the discrete sample, the sample must be taken on a regularly spaced grid. While this may be alright in applications like signal and image processing, it does restrict the applicability of this method in usual statistical problems where the underlying function may be sampled at unequally spaced (and often random) points in space. In this presentation we consider the case of fixed (or non-random) design. In functional data analysis context the analysis of a possibly large number of sampled noisy curves necessitates some fast computational technique. A scenerio that is quite relevant there is the use of some smooth, non-uniform design to collect the data (this may be due to some apriori idea about the important regions in the spatial scale where some phenomea take place). Our method is desgined to cater to such needs. It also generalizes readily to higher dimensions, and does not require any inversion of functions (which is a must in rank-based "naive" method). In this method the final estimate of the unknown function is obtained by thresholding the wavelet coefficients of a suitably transformed function. It has been shown method produces estimate of f which has suboptimal rate of convergence when measured in squared error loss, but the suboptimality tends to vanish as the sampling design function H becomes smoother. In this poster we illustrate the use of this methodology in terms of various examples. We also study the effect when the smoothness assumptions on the sampling design are not satisfied. Our method is in parts motivated by the work of Cai and Brown (1998).

Frequentist Assessment of Bayesian Wavelet Shrinkage Rules

Marianna Pensky
University of Central Florida

The present paper investigates theoretical performance of Bayesian wavelet shrinkage in nonparametric regression model with iid errors which are not necessarily normally distributed. The main purpose is comparison of various Bayesian models in terms of frequentist asymptotic optimality. We establish relationships between hyperparameters, verify that majority of Bayesian models studied so far achieve theoretical optimality, state which Bayesian models cannot achieve optimal convergence rate and explain why it happens. We conclude that even if the errors are normal, it is often advantageous in Bayesian wavelet regreesion to use other distributions for errors, for example, the ones resulting from mixing normal distributions over the scale parameter.

Functional Regression Models and Temporal Processes

Jun Yan
University of Wisconsin–Madison

We consider response and covariates which are temporal processes observed in overlapping intervals and can be viewed as functional data. A functional generalized linear model is proposed for the mean of the response over time. The process means are modeled marginally, without a Markov assumption. The framework is rich, extending standard parametric and semiparametric models in multistate survival analysis. Based on the abundance of cross-sectional data provided by the continuously observed temporal processes, we propose simple estimators of the time-varying coefficients using easy-to-implement moment methods. The functional estimators are shown to be uniformly consistent and to converge weakly to Gaussian processes. The estimation procedure does not require smoothing, unlike most approaches to varying-coefficient models.

The nonparametric estimators are the basis for new tests of the covariate effects. They also enable estimation of submodels in which greater structure is imposed on the parameters, resulting in partly parametric models. The methodology is useful in goodness-of-fit testing for the partly parametric models, and permits predictions involving estimated components from both the functional model and the submodels.

The usefulness of the modeling strategy is illustrated in recurrent event simulations, where the proposed methodology is comparable in efficiency with existing methods when the true model coefficient is constant, and performs well estimating the time-varying covariate effects when it is not. An analysis of the prevalence of the chronic graft versus host disease in bone marrow transplant patients is carried out, and the effects of treatment and gender are found to be time-varying.

Variable Selection and Model Building via Likelihood Basis Pursuit

Hao (Helen) Zhang
North Carolina State University

Variable selection is fundamental to multivariate statistical model building, validation and selection. Traditional approaches include the best subset selection, the forward, backward and stepwise selection, often validated with selection criteria like Mallow's C_p, AIC and BIC. Shrinkage methods such as ridge regression, the non-negative garrotte, the least absolute shrinkage and selection operator (LASSO) and the smoothly clipped absolute deviation (SCAD) penalized approach are proposed recently. By solving a certain constrained linear regression problem, these methods tend to mitigate the instability and high variability of subset selection. Much progress has been made on variable selection in linear models. In this work we propose a nonparametric approach for variable selection and model building called "likelihood basis pursuit" (LBP), which selects important variables and provides a flexible regression fit to the data simultaneously.

Likelihood basis pursuit (LBP) is a penalized likelihood approach constructed in the framework of the smoothing spline analysis of variance (SS-ANOVA). The SS-ANOVA has been intensively used for nonparametric multivariate function estimation. LBP decomposes the regression function into the sum of different functional components, such as the constant, the main effects, the two-factor interactions and all higher-order interactions. In a tensor product reproducing kernel Hilbert space, each functional component in the decomposition falls into a certain subspace and can be represented in terms of some basis functions. The basis functions are chosen to be compatible with variable selection and model building in the context of SS-ANOVA. Basis pursuit is a penalized model selection for wavelet regression. We apply the basis pursuit to obtain the optimal decomposition which has the smallest l_1 norm on the coefficients. To decide which variables or which ANOVA components are included in the model, an importance measure for the functional components is proposed, with the threshold value being determined by a sequential Monte Carlo bootstrap testing algorithm. The generalized approximate cross validation (GACV) is developed as a criterion for the adaptive choice of the regulation parameters in the model. As a nonparametric generalization of the LASSO, the LBP produces shrinkage estimates for the coefficients which greatly facilitates the variable selection process. At the same time, it provides a flexible and interpretable nonparametric model fitting for the data. A special algorithm slice modeling is applied for efficient optimization.

Variable selection plays an important role in analyzing clinical and epidemiological studies, where a large number of medical, demographic and other covariates are encountered. In this work LBP is applied to two large on-going epidemiological studies: Wisconsin Epidemiological Study of Diabetic Retinopathy (WESDR) and Beaver Dam Eye Study (BDES). Very recently variable selection has become the focus of intensive research in several areas, for which massive datasets with tens or hundreds of thousands of variables are available, such as text processing and genomics, particularly gene expression array data. LBP has great potential in these fields. A freeware for public use is in plan.

Analysis of Periodicity in a cDNA Microarray Time-Course Experiment

Xin Zhao
Cornell University

Microarray time-course genome-wide data are typically HDLSS (High Dimension Low Sample Size). Gene expression profiles over time could be seen as functional data. The functional approach could provide powerful new insights for this type of data. Successful data reduction from a functional viewpoint is used in an analysis of periodicities for a microarray gene expression data set. For the purpose of analyzing periodicity, an appropriate Fourier Transformation followed by PCA (Principal Component Analysis) reduces the dimension of data from 18 to 2. The 2-dimensional Fourier subspace spanned by the sine and cosine functions with 2 periods captures the main feature of periodicity in the data. The distance to the origin in the subspace could be used to measure the degree of periodicity for genes.