

Division of Quantitative Sciences

Archive of Seminar Abstracts

2008 Seminars

December 10, 2008

Jong Soo Lee, Ph.D.
Visiting Assistant Professor, Department of Statistics
Carnegie Mellon University, Pittsburgh, Pennsylvania

Pointwise Testing with Functional Data Using the Westfall-Young Randomization Method

I will present a consideration of hypothesis testing with smooth functional data by performing pointwise tests and applying a multiple comparisons procedure. Methods based on general inequalities (e.g., Bonferroni's method) do not perform well because of the high correlation between observations at nearby points. I will consider the multiple comparison procedure proposed by Westfall and Young (1993) and show that it approximates a multiple comparison correction for a continuum of comparisons as the grid for pointwise comparisons becomes finer. I will describe simulations and an application to real data that have verified the method's applicability to practical settings. I will also discuss possible future directions of this research.

December 3, 2008

Melanie M. Wall, Ph.D.
Associate Professor, Division of Biostatistics
University of Minnesota, Minneapolis-St Paul, Minnesota

Structural Equation Modeling of Latent Classes

Latent class analysis typically involves modeling a set of several observed measurements via a single underlying (latent) categorical (class) variable that is meant to capture the associations found among the observed variables. This latent class variable can then be modeled as either an outcome or predictor variable in order to address some research question of interest. Latent class models can be seen applied within the health sciences to multiple diagnostic tests without a gold standard, multiple source or informant data, and multiple symptom assessments. As applications of this type of modeling of a single latent class variable are becoming more common, it is natural to consider models involving multiple latent class variables. In particular, structural equation models (SEM) of latent class variables will be considered, differing from traditional SEM in that all the latent variables are categorical rather than continuous. In addition to basic main effects type models, models involving interaction effects between different latent class variables on outcomes will be demonstrated as well as structural model relationships between multiple latent class processes (over time). Examples of traditional applications of the single latent class variable models will be given and an application relating social, familial, environmental, and personal factors associated with adolescent obesity will be used to demonstrate the new SEM of latent classes.

November 19, 2008

Mikyong Jun, Ph.D.
Assistant Professor, Department of Statistics
Texas A&M University, College Station, Texas

Nonstationary Covariance Models for Global Data

The widespread availability of satellite-based instruments has allowed investigators to measure many geophysical processes on a global scale. Such assessments often show strong nonstationarity in the covariance structure. I present a flexible class of parametric covariance models that can capture the nonstationarity in global data, and in particular, the strong dependency of covariance structure on latitudes. I apply the discrete Fourier transform to data on regular grids, which enables me to calculate the exact likelihood for large data sets. I apply the proposed covariance model to global total column ozone level data on a given day, and discuss how the model compares with some existing models.

November 12, 2008

Ori Rosen, D.Sc.
Associate Professor, Department of Mathematical Sciences
The University of Texas at El Paso

A Bayesian Regression Model for Multivariate Functional Data

I will describe a method for analyzing multivariate functional data with unequally spaced observation times that may differ among subjects. Fitting multivariate observations simultaneously rather than fitting each variable separately may be advantageous if the error terms corresponding to each variable are correlated. The proposed method is formulated as a Bayesian mixed-effects model in which the fixed part corresponds to the mean functions, and the random part corresponds to individual deviations from these mean functions. Covariates can be incorporated into both the fixed and the random effects. I will present the results of simulation studies, and will apply the methodology to real data for illustration.

November 7, 2008

Yijian Eugene Huang, Ph.D.
Associate Professor of Biostatistics and Bioinformatics, Rollins School of Public Health
Emory University, Atlanta, Georgia

Quantile Regression with Censored Data

Quantile regression has been advocated in survival analysis to assess evolving covariate effects. However, challenges arise when the censoring time is neither always observed nor independent of the covariates. In spite of several recent advances attempting to resolve this problem, existing methods either involve complicated algorithms, which lead to difficulties of implementation and asymptotics; or impose a cumulative-probability grid that introduces undesirable grid-dependence of the estimation. To resolve these issues, I introduce fundamental and general quantile calculus on a cumulative probability scale. These results give rise to a novel estimation procedure for censored quantile regression, based on estimating integral equations. I will propose a numerically reliable and efficient algorithm for the computation. This procedure reduces to the Kaplan-Meier method in the k-sample problem, and to standard uncensored quantile regression in the absence of censoring. The proposed regression quantile estimator is uniformly consistent and converges weakly to a Gaussian process. I will describe

simulation studies, which have shown good numerical and statistical performance of the proposed method. I will illustrate the method through its application to data from a clinical study.

October 22, 2008

Xihong Lin, Ph.D.
Professor, Department of Statistics, School of Public Health
Harvard University, Boston, Massachusetts

Genomic-Feature-Based Analysis of Genome-Wide Association Studies

Conducting a genome-wide association study (GWAS) has become an increasingly popular way to identify genetic variants of a disease through the examination of hundreds of thousands of SNPs across a genome. Investigators can use a GWAS to significantly accelerate the discovery of genetic variants associated with a disease. A common approach to analyzing a GWAS dataset is to test for a single SNP at a time and adjust for multiple comparisons. This approach has been found to have several limitations, including a lack of power and high false positives, which have made it difficult to replicate the findings of top SNPs in validation studies. I will present a biological feature-based GWAS analysis using the kernel machine method through its connection with generalized linear mixed models, and will illustrate the method using the CGEMS breast cancer GWAS data.

October 15, 2008

Robert McCulloch, Ph.D.
Professor, Risk Analysis and Decision Making, McCombs School of Business
The University of Texas at Austin

BART: Bayesian Additive Regression Trees, with Application to Classification

In Bayesian Additive Regression Trees (BART), Chipman, George, and McCulloch developed a fully Bayesian approach to the model: $y = f(x) + e$, where the errors may be drawn from any symmetric distribution. In the spirit of “ensemble models” the unknown function f was modeled as the sum of many simple tree models. The contribution of each individual tree was kept small through the use of a strong regularization prior. The BART methodology was shown to be very competitive in terms of out-of-sample prediction. However, the BART model, prior, and MCMC algorithm are all geared toward the case in which the response is numeric. I will explore the use of the BART methodology in classification problems, and will discuss different approaches to extending BART to classification.

October 14, 2008

Song Zhang, Ph.D.
Assistant Professor of Clinical Sciences
The University of Texas Southwestern Medical Center, Dallas, Texas

A Bayesian Approach to Ranking and Rater Evaluation with Application to Grant Reviews

I will describe a Bayesian hierarchical model for the analysis of ordinal data from multirater ranking studies. The model for an item's score includes four latent factors: one is a latent trait determining the true ordering of the items, and the other three are the rater's performance characteristics, including bias, discrimination, and measurement error. The fitted model can be used to rank items based on their latent trait and to evaluate the performance of raters based on their characteristics. I will also describe a simulation-based decision-theoretic approach to determining the optimal number of raters. A loss function is specified accounting for the penalty of

incorrect ranking and the cost of raters. I will identify the optimal number of raters for which the loss function is minimized. I will present the results of a simulation study and an application of this method to a grant review dataset.

October 8, 2008

Marco Ferreira, Ph.D.
Assistant Professor, Department of Statistics
University of Missouri - Columbia

Dynamic Multiscale Modeling

This represents a joint effort with Scott Holan and Adelmo Bertolde.

I will describe a new class of multiscale spatio-temporal models for Gaussian data. The framework we use decomposes the spatio-temporal observations and underlying process into several scales of resolution. Under this decomposition, the model evolves the multiscale coefficients through time with structural state-space equations. The multiscale decomposition we consider, which includes wavelet decompositions as a particular case, is able to accommodate irregular grids and heteroscedastic errors. The multiscale spatio-temporal framework we developed has several salient attributes. First, the multiscale decomposition leads to an extremely efficient divide-and-conquer estimation algorithm. Second, the multiscale coefficients have an interpretation of their own; thus, the multiscale spatio-temporal framework may offer new insight into understudied multiscale aspects of spatio-temporal observations. Finally, deterministic relationships between different resolution levels are automatically respected for the observations, the latent process, and the estimated latent process. I use two examples to illustrate the use of our multiscale framework. First, I will describe our analysis of a simulated dataset of functional data with temporally evolving functions; then, our analysis of a spatio-temporal dataset on agriculture production in the state of Espirito Santo, Brazil.

October 1, 2008

Sining Chen, Ph.D.
Assistant Professor, Department of Environmental Health Sciences & Biostatistics
John Hopkins School of Public Health, Baltimore, Maryland

Estimation, Prediction and Screening of Colorectal Cancer Risk in Lynch Syndrome

I will give an overview of the statistical issues surrounding Lynch syndrome, the most common hereditary colorectal cancer syndrome, which also involves several other cancer sites. We will look at (1) the estimation of genetic risks from large, heavily ascertained families with cancer, (2) building mutation carrier probability models including MMRpro, and (3) individualized colonoscopy schedules based on personal risks for colorectal cancer.

September 10, 2008

Bradley P. Carlin, Ph.D.
Professor, Division of Biostatistics, School of Public Health
University of Minnesota, Minneapolis-St. Paul, Minnesota

Analysis of Marked Point Patterns with Spatial and Nonspatial Covariate Information

Hierarchical modeling of spatial point process data has historically been plagued by computational difficulties. Likelihoods feature intractable integrals that are themselves nested within a Markov chain Monte Carlo (MCMC)

algorithm. I extend customary spatial point pattern analysis in the context of a log-Gaussian Cox process model to accommodate spatially referenced covariates, individual-level risk factors, and individual-level covariates of interest that mark the process. I also use multivariate process realizations to capture dependence among the intensity surfaces across the marks. I illustrate this method using data of breast cancer case locations collected throughout the mostly rural northern part of Minnesota, which are marked by the selection of mastectomy or breast conserving surgery (“lumpectomy”) for breast cancer treatment. The key substantive covariate (driving distance to the nearest radiation treatment facility) is spatially referenced, but other important covariates (notably age and stage) are not. This approach facilitates the mapping of marginal log-relative intensity surfaces for the two treatment options, and resolves the issue of whether women who face long driving distances are significantly more likely to opt for mastectomy while still accounting for all sources of spatial and nonspatial variability in the data. I also briefly discuss methods for statistical boundary analysis (“wombling”) in such settings.

July 28, 2008

David Rossell, Ph.D.
Unit Manager, Biostatistics and Bioinformatics
IRB Barcelona

GaGa: Microarray Differential Expression, Gene Clustering and Class Prediction

A typical microarray study analysis requires multiple complementary analyses, such as gene differential expression, gene clustering, class prediction, gene ontology, network/pathway analyses. I propose a simple and computationally efficient model that unifies several of these tasks in a single framework. The model generalizes the hierarchical gamma/gamma model, first introduced by Newton and Kendziorski, to address several issues that limit the quality of the fit and therefore the reliability of the inference. The main advantage of a unified framework is that information can be shared between different analyses. When building a classifier, the model weights the contribution of each gene by the posterior probability that the gene is differentially expressed. When clustering genes, the clusters are defined according to biologically interpretable expression patterns. I illustrate the approach by walking through several real datasets.

May 29, 2008

María Eglée Pérez, Ph.D.
Associate Professor, Department of Mathematics
University of Puerto Rico - Rio Piedras Campus
San Juan, Puerto Rico

Intrinsic Priors for Testing the Hardy-Weinberg Equilibrium

Testing Hardy-Weinberg equilibrium is a relevant concern, for example, in studies relating genetical configurations with health conditions. The selection of prior distributions for testing Hardy-Weinberg equilibrium is a challenging issue as we are dealing with a low dimensional null hypothesis for a discrete model. In this work, intrinsic priors for testing Hardy-Weinberg equilibrium are calculated using hypothetical training samples from uniform and Haldane priors. Properties of both priors are discussed, and their performances are compared on hypothetical data sets, and on real data from a case-control study of risk factors for gastric cancer in Western Venezuela. Analysis of sensitivity to different training samples sizes is shown, and possible criteria for the selection of the training sample size are discussed.

May 28, 2008

William Rosenberger, Ph.D.
Professor and Chairman, Department of Statistics
George Mason University, Fairfax, Virginia

Handling Covariates in the Design of Clinical Trials

There has been a split in the statistics community about the need to take into account covariates when designing a clinical trial. There are many advocates of using stratification and covariate-adaptive randomization to promote balance on certain known covariates. However, balance does not always promote efficiency or ensure that more patients are assigned to the better treatment. I describe procedures, including model-based procedures, for incorporating covariates into the design of a clinical trial, and give examples where balance, efficiency, and ethical considerations may be in conflict. I advocate covariate-adjusted response-adaptive (CARA) randomization procedures, a new class of procedures that attempts to optimize both efficiency and ethical considerations while maintaining randomization. I review the philosophy and procedures, and present a few new simulation studies for illustration.

May 21, 2008

Alejandro A. Vallejos, Ph.D.
Postdoctoral Researcher, Department of Statistics
Pontificia Universidad Católica de Chile, Santiago, Chile

DPackage: An R Package for Bayesian Nonparametric Inference

Although Bayesian nonparametric methods are extremely powerful and have a wide range of applicability within several prominent domains of statistics, they are not as widely used as one might guess. At least part of the reason for this has been the gap between the type of software that many applied users would like to have for fitting models and the software that is currently available. I introduce an R package, DPpackage, that is designed to help bridge this gap. DPpackage allows the user to perform Bayesian inference via simulation from the posterior distributions for models considering Dirichlet processes (DP): Dirichlet process mixtures (DPM), Polya trees (PT), mixtures of triangular distributions, and random Bernstein polynomial priors. The package also includes generalized additive models considering penalized B-splines. I discuss the general syntax and design philosophy of the package, and demonstrate its usage and main features using several examples with an emphasis on semiparametric generalized linear mixed models.

May 19, 2008

Damien Chaussabel, Ph.D., Associate Investigator
Baylor Institute for Immunology Research, Dallas, Texas

A Modular Framework for Biomarker and Knowledge Discovery from Blood Transcriptional Profiling Studies

The analysis of patient blood transcriptional profiles offers a means to investigate immunological mechanisms relevant to human diseases on a genome-wide scale. Such studies also provide a basis for the discovery of clinically-relevant biomarker signatures. However, mining large-scale data for knowledge that is immunological and/or clinically relevant is a challenge. I present a strategy with the goal of reducing the dimension of microarray data. The strategy is based on the identification of transcriptional modules formed by genes coordinately expressed in multiple disease datasets.

May 14, 2008

Jing Cao, Ph.D.
Assistant Professor, Department of Statistical Science
Southern Methodist University, Dallas, Texas

Bayesian Chi-squared Goodness-of-fit Tests for Censored Data Models

Censored survival data can be viewed as a special case of missing data. In problems with missing data, it is common to first impute the unobserved data and then perform the model-checking procedure based on the complete data. For censored data, the complete data include both the uncensored data and the imputed censored data. When there is heavy censoring, it is possible that several partitioning cells of a goodness-of-fit test will contain a high proportion of counts that correspond to imputed data. Such cells can dramatically reduce the power of the resulting test. I present general methodology for testing the adequacy of parametric statistical models applied to data with censoring. The statistic is calculated from posterior samples of probability-transformed Bayesian residuals based on uncensored observations. Under the null hypothesis, the Bayesian residuals are independently and identically distributed according to a uniform distribution. These uniform deviates are then used to construct a statistic that has an asymptotic distribution (Johnson, 2004; 2007). I show that under heavy censoring, the test based on uncensored observations is more powerful than the test based on complete data. Under moderate or light censoring, the two tests are comparable in power. Another advantage of the proposed test is that the diagnostics apply for both simple and composite null hypotheses, and to i.i.d. and general regression settings.

May 12, 2008

Su-Chun Cheng, Ph.D.
Associate Professor, Epidemiology and Biostatistics
University of California, San Francisco

Combination of Multiple Diagnostic Tests for Classifying Censored Event Times

When there are multiple sources of information available, it is often of interest to construct a composite score that can provide classification accuracy better than any individual measurement. In this collaboration, I present robust procedures for optimally combining tests when test results are measured prior to disease onset and disease status evolves over time. To account for censoring of the time of disease onset, the most commonly used approach to combine tests to detect subsequent disease status is to fit a proportional hazards model (Cox, 1972) and use the estimated risk score. However, simulation studies suggest that such a risk score may have poor accuracy when the proportional hazards assumption fails. I present a proposal using a nonparametric transformation model (Han, 1987) as a working model to derive an optimal composite score with theoretical justification. I demonstrate that the proposed score is the optimal score when the model holds and is optimal "on average" among linear scores even if the model fails. Time-dependent sensitivity, specificity, and receiver operating characteristic curve functions are used to quantify the accuracy of the resulting composite score. The model provides consistent and asymptotically Gaussian estimators of these accuracy measures. I present a simple model-free resampling procedure to obtain all consistent variance estimators, and illustrate the new proposals with simulation studies and an analysis of a breast cancer gene expression data set.

May 7, 2008

Lorenzo Trippa
Graduate Student
Department of Biostatistics
MD Anderson Cancer Center

A Truly Nonparametric Proportional Hazards Model

I present collaborative work of a novel Bayesian model for event time data. Many recent papers discuss Bayesian inference for the proportional hazards model and other semiparametric models. The semiparametric approach, by definition, still includes some rigid parametric assumptions. This study was motivated by practical limitations of such assumptions. A fully non-parametric prior is defined by extending the Polya tree model to a family of random survival functions indexed by covariates. An important feature of the proposed approach is that the random survival functions are a priori centered on a proportional hazards model. This allows us to report inference on covariate effects. I analyze a clinical study in which the semiparametric approach appears inadequate.

April 30, 2008

David B. Dahl, Ph.D.
Assistant Professor, Department of Statistics
Texas A&M University, College Station, Texas

Using Prior Information in Bayesian Nonparametric Models

Integration of data from several sources and technologies is a burgeoning field in bioinformatics. Some data naturally lead to formal statistical models, yet others may merely convey proximity among observations. In the context of clustering, methods are often either model-based or distanced-based. In many cases, however, both types of information are available. I propose a hybrid approach that is simultaneously model-based and distance-based. Specifically, I show how the usual Dirichlet process mixture model framework can be adapted to incorporate pairwise distances between observations. One application area is incorporating gene annotation information in statistical models for gene expression. Another application is protein structure prediction, wherein one can estimate protein torsion angle distributions using both (ϕ , ψ) angle pairs and RMSD distances from peptides.

April 9, 2008

Wesley O. Johnson, Ph.D.
Professor, Department of Statistics
University of California-Irvine

Non-Proportional Hazards Regression: Survival Curves Can (Be) Cross

This work represents a collaborative effort with Maria De Iorio, Peter Mueller, and Gary Rosner. I present a dependent Dirichlet process model for survival analysis data. The model extends the ANOVA DDP that was presented by De Iorio et al. in 2004 in JASA to handle continuous covariates and censored data. A major feature of the work is that there is no necessity for the resulting survival curve estimates to satisfy the ubiquitous proportional hazards assumption. I provide an illustration based on a cancer clinical trial in which the survival probabilities for times early in the study are estimated to be lower for those on the high-dose treatment regimen

than for those on the low-dose treatment regimen; and the reverse is true for later times. This is possibly due to the greater toxicity of the high dose in patients who are not as healthy at the beginning of the study.

March 20, 2008

Guosheng Yin, Ph.D.
Assistant Professor, Department of Biostatistics
MD Anderson Cancer Center

Bayesian Dose-Finding Trial Designs for Drug Combinations

Treating patients with a combination of agents is becoming commonplace in clinical trials, with biochemical synergism often the primary focus. In a typical drug combination trial, the toxicity profile of each individual drug has already been thoroughly studied in single-agent trials, which naturally offers rich prior information. We propose Bayesian adaptive designs for dose finding to account for the synergistic effect of two or more drugs in combination. To search for the maximum tolerated dose combination, we continuously update the posterior estimates for the toxicity probabilities of the combined doses. By reordering the dose toxicities in the two-dimensional probability space, we adaptively assign each new cohort of patients to the most appropriate dose. We conduct extensive simulation studies to examine the operating characteristics of the designs.

March 12, 2008

Valen E. Johnson, Ph.D.
Professor and Deputy Chair, Department of Biostatistics
MD Anderson Cancer Center

Better Bayes Factors, with Applications to Clinical Trial Design

Most Bayesian hypothesis tests result in exponential accumulation of evidence in favor of the alternative hypothesis when the alternative hypothesis is true, but only sub-linear accumulation of evidence in favor of the point null hypothesis when the null hypothesis is true. Thus, it is often impossible for an experiment to provide “strong evidence” in favor of the null hypothesis even when moderately large sample sizes have been obtained. Because Bayesian hypothesis tests yield probability statements regarding the truth of the null hypothesis (rather than a frequentist decision to simply “not reject”), this imbalance in the rates of accumulation of evidence is highly problematic. I review and contrast asymptotic convergence rates of Bayes factors for different classes of objective prior distributions and propose two new classes of prior densities that correct the imbalance inherited by standard objective priors. I illustrate the performance of hypothesis tests defined using these new prior distributions in context of phase II clinical trials.

March 5, 2008

Yu Ryan Yue
Doctoral Candidate, Department of Statistics
University of Missouri at Columbia
Columbia, Missouri

Nonstationary Gaussian Markov Random Fields for Regression and Spatial Modeling

The smoothing spline is one of the most popular curve-fitting methods. Its two-dimensional version, the thin-plate spline, is a well-known surface fitting model that has been used intensively in spatial smoothing areas. These two splines, however, both suffer from having only one global smoothing parameter that controls the smoothness of

the fit function. This becomes an issue when the function of interest is highly variable through the input space. To overcome this inadequacy, we have developed a class of priors for smoothing splines and thin-plate splines that are spatially adaptive. These priors extend Gaussian Markov random fields (GMRFs) by using a spatially adaptive variance component and taking a further GMRF prior for this variance function. Fully Bayesian inference can be carried out through efficient Markov chain Monte Carlo simulation. The performance is demonstrated with simulation examples and an application to a set of U. S. rainfall data.

February 27, 2008

Song Yang, Ph.D.
Senior Mathematical Statistician, Office of Biostatistics Research
National Heart, Lung and Blood Institute
Bethesda, Maryland

Semiparametric Estimation of the Hazard Ratio Function

The hazard ratio provides a valuable tool for assessing a treatment effect with survival data, with the proportional hazards special case of the Cox model as a widely used example. In general, the hazard ratio is a function of time, and provides a visual display of the temporal pattern of the treatment effect. The proportional hazards assumption is often too restrictive, at least for the initial exploration of a treatment effect, while a nonparametric estimate of the hazard ratio function requires a bandwidth selection, and may result in increased variance or bias. On the other hand, most semiparametric hazards models proposed so far imply certain restrictions on the hazard ratio that limit their utility. We investigate a model that allows monotone increasing or decreasing hazard ratio functions, including crossing hazards. This model provides a sufficient level of flexibility for many applications. The point estimates, point-wise confidence intervals and simultaneous confidence intervals, or confidence bands, of the hazard ratio, are proposed under this model. We demonstrate the inference procedures using data of coronary heart disease from the Women's Health Initiative clinical trial on estrogen plus progestin, in addition to other data examples. These examples, with a diverse range of time dependence of the hazard ratio from mild to severe, suggest that the hazard ratio under this class of models, its confidence intervals and confidence bands, provide very useful visual display tools for assessing the treatment effect with survival data.

February 21, 2008

Guoqing Diao, Ph.D.
Assistant Professor, Department of Statistics
George Mason University
Fairfax, Virginia

Semiparametric Cure Rate Models with Random Effects

Joint work with Dr. Guosheng Yin, MD Anderson Cancer Center

We propose a novel class of cure rate models for multivariate failure time data with a survival fraction. The class is formulated through a transformation on the unknown population survival function. It incorporates random effects to account for the underlying correlation, and includes the mixture cure model structure and the proportional hazards cure model structure as two special cases. We propose a general form of the covariate structure that automatically satisfies an inherent parameter constraint. Moreover, it accommodates the corresponding binomial and exponential covariate structures in the two main formulations of cure models. The proposed class provides a natural link between the mixture and proportional hazards cure models, and it offers a wide variety of new modeling structures. We show that the nonparametric maximum likelihood estimators for the parameters of these models are consistent and asymptotically normal. The limiting variances achieve the

semiparametric efficiency bounds and can be consistently estimated. Simulation studies demonstrated that the proposed methods perform well in practical situations. We use real data to illustrate this class of models.

February 20, 2008

Sonia Petrone, Ph.D.
Associate Professor, Department of Decision Sciences
Bocconi University
Milan, Italy

Bayesian Nonparametric Methods for Complex Heterogeneous Data

Bayesian nonparametric methods have more and more application in treating heterogeneous data in an extremely wide range of fields. I present recent developments of nonparametric allocation rules with multivariate and functional data, where several kinds of heterogeneity have to be taken into account.

February 13, 2008

Wen Ye, Ph.D.
Research Assistant Professor, Department of Biostatistics
University of Michigan
Ann Arbor, Michigan

Semi-Parametric Joint Modeling of Longitudinal and Time-to-Event Data Using P-Spline: A Penalized Likelihood Approach

Longitudinal studies in medical research often generate repeated measurements of biomarkers, and possibly censored survival data. Several joint models recently developed deal with the challenges arising from this type of data. A linear mixed model is commonly used to model the longitudinal covariate in joint models. However, in some cases, the longitudinal covariate time trajectory is not linear. We propose a joint model using penalized cubic B-splines to accommodate the nonlinear trajectory of longitudinal covariate measurements. To ease computation, the estimation procedure maximizes a penalized joint likelihood generated by a Laplace approximation of the joint likelihood, which combines the likelihood of the longitudinal data and the partial likelihood of the time-to-event data. We investigated the properties of the parameter estimators in simulation studies.

February 5, 2008

Sudipto Banerjee, Ph.D.
Associate Professor, Division of Biostatistics
School of Public Health
University of Minnesota
Minneapolis, Minnesota

Gaussian Predictive Processes Models for Large Spatial Datasets

With accessibility to geocoded locations that involve the use of geographical information systems (GIS) to collect scientific data, investigators are increasingly turning to spatial process models to carry out statistical inference. Over the last decade, hierarchical models implemented through Markov chain Monte Carlo (MCMC) methods have become especially popular for spatial modeling due to their flexibility and power to estimate models (and, hence, to address scientific hypotheses) that would be infeasible using classical methods. However, fitting

hierarchical spatial models often involves expensive matrix decompositions, the computational complexity of which increases exponentially with the number of spatial locations. This renders them infeasible for large spatial data sets.

I propose using a predictive process derived from the original spatial process that projects process realizations to a lower-dimensional subspace, thereby reducing the computational burden. I discuss the attractive theoretical properties of this predictive process, as well as its greater modeling flexibility compared to that of existing methods. In particular, I show how the predictive process seamlessly adapts to settings with nonstationary processes, with richer and more complex space-varying regression models, and with multivariate spatial models. I present a computationally feasible template that encompasses these diverse settings.