交大-印大应用统计双边论坛

Symposium of SJTU-IU on Applied Statistics.

Time: 8:30-12:30, Dec. 16, 2016

Venue: Large Lecture Room, Math Building.

Organizers: Dong Han and Weidong Liu, Dept. of Statistics, Shanghai Jiao Tong Univ.

Ying Zhang, Dept. of Biostatistics, Indiana Univ.

Speakers and Titles of Talks

ID	Title	Speakers	Affilliation
1	A Joint Model of an Internal Time-Dependent Covariate and Bivariate Time-to-Event Data with Application to MD STARnet Data	Ying Zhang	IU
2	Model Selection and Structural Discovery in Multivariate Semi-parametric Regression	Wanzhu Tu	IU
3	Bivariate Survival Data with Semi-competing Risk	Sujuan Gao	IU
4	Dynamic, Interactive data analysis/visualization with JavaScript: An introduction to Highcharts, D3, and open CPU	Spencer Lourens	IU
5	ADJUSTING FOR INCOMPLETE DEATH ASCERTAINMENT IN JOINTMODELS: A MULTIPLE-IMPUTATION APPROACH	Constantin T. Yiannoutsos	IU
6	Optimal High-dimensional multiclass linear discriminant analysis	Shan Luo	SJTU
7	Variable selection for mixture and promotion time cure rate models	Zhangsheng Yu	SJTU
8	Semi-parametric Spatial Model for Interval-censored Data with Time-Varying Covariate Effect	Yue Zhang	SJTU
9	Determining the number of factors based on the singular values	Cheng Wang	SJTU
10	Structured sub-composition selection in regression and its application to microbiome data analysis	Tao Wang	SJTU

Program

Chairman: Dong Han	
8:30-8:50	Sujuan Gao
8:50-9:10	Spencer Lourens
9:10-9:30	Shan Luo
9:30-9:50	Wanzhu Tu
9:50-10:10	Cheng Wang
10:10-10:30	**Tea Break**
Chairman: Weidong Liu	
10:30-10:50	Tao Wang
10:50-11:10	Constantin T. Yiannoutsos
11:10-11:30	Zhangsheng Yu
11:30-11:50	Ying Zhang
11:50-12:10	Yue Zhang
Lunch	

Abstract

A Joint Model of an Internal Time-Dependent Covariate and Bivariate Time-to-Event Data with Application to MD STARnet Data

Ke Liu and Ying Zhang
Nielsen Company and IU Fairbank of School of Public Health, Department of Biostatistics

Yz73@iu.edu

Abstract: Motivated by a study of muscular dystrophy in MD STARnet, a joint model of bivariate survival times and longitudinal data is developed. We propose to analyze correlated bivariate survival responses associated with a longitudinal biomarker in the Frequentist paradigm. A Gamma frailty variable is used to account for the correlation between the two correlated survival outcomes in addition to the random variables that account for the correlation between the survival times and longitudinal maker. The EM algorithm is adopted to compute the maximum profile likelihood estimate. The bootstrap method is applied to estimate the standard error of estimated model parameters. The simulation study is conducted to demonstrate the validity of the proposed methodology. Finally the method is applied to the MD STARnet for illustration.

Model Selection and Structural Discovery in Multivariate Semiparametric Regression

Zhuokai Li[1] , Hai Liu[2], Wanzhu Tu[3], ...

[1] Presenter, Duke Clinical Research Institute, Durham, NC 27705,
USA zhuokai.li@duke.edu
[2] Gilead Sciences, Inc., Foster City, CA 94404, USA
[3] Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN 46202, USA

Abstract: Model selection in multivariate semiparametric regression remains a challenge, especially for longitudinal data. We propose a model selection procedure that simultaneously selects fixed and random effects using a maximum penalized likelihood method with the adaptive least absolute shrink-age and selection operator (LASSO) penalty. We determine the correlation structure among multiple outcomes through random effects selection. Additionally, interactions of independent variables mod-eled by bivariate tensor product spline functions are selected using group LASSO. To implement the selection method, we propose a two-stage expectation-maximization (EM) procedure. We assess the operating characteristics of the proposed method through a simulation study. The method is illustrated in a clinical study of blood pressure development in children.

Key Words: adaptive LASSO, EM algorithm, mixed effects, multivariate data, L_1 penalty

Bivariate Survival Data with Semi-competing Risk

Sujuan Gao[1*], Ran Liao[2]

1*: Presenter, Department of Biostatistics, Indiana University School of Medicine, Indianapolis, Indiana, USA, 46202

2. Department of Biostatistics, Indiana University Richard M. Fairbank School of Public Health, Indianapolis, Indiana, USA, 46202

Abstract: Bivariate survival data often arise in medical research in the forms of events observed from studies on siblings or multiple diseases experienced by the same individual. Studies in aging research also encounter semi-competing risk when individuals under observation for a particular disease die from other causes. Statistical analysis ignoring the correlations from the bivariate survival times or the attrition from the semi-competing risk may lead to biased results. In this work, we propose a frailty-based model framework to simultaneously model both the bivariate survival times and the semi-competing risk. We propose a penalized pseudo-partial likelihood approach for parameter estimation and inference and compare the proposed approach to several alternative estimation approaches in simulation studies. Simulation results demonstrate adequate performances for the proposed method. We will illustrate the proposed model framework and estimation approach using data collected from a longitudinal aging cohort.

Key Words: bivariate survival, semi-competing, frailty model

Dynamic,Interactive data analysis/visualization with JavaScript: An introduction to Highcharts, D3, and openCPU

Spencer Lourens[1]

[1] Presenter, Indiana University Department of Biostatistics, 46202, United States slourens@iu.edu

Abstract: Due to the increased demand for statistical work and availability of data in the world to-day, end-users of statistical analyses come from a wider range of backgrounds than ever before. This places extra burden on statisticians and greatly increases the need for visualizations which are aes-thetically pleasing, dynamic, and easy for the user to interact with. JavaScript offers a wide range of highly customizable options, such as D3 and Highcharts, which can be used for data visualiza-tion. These libraries provide flexible interfaces and a higher level of designer control at the expense of requiring more development knowledge. We will highlight simple interactive visualization exam-ples from JavaScript, and subsequently show the integration of R libraries (even user defined ones) with JavaScript through the openCPU framework. The openCPU framework is very flexible, and can be extended beyond R by use of AJAX (asynchronous JavaScript and XML). As opposed to Shiny, openCPU is completely flexible and does not involve using predefined widgets, but puts extra burden on the user by requiring knowledge of JavaScript, html, and CSS.

Key Words: javaScript, Visualization, D3, Highcharts, openCPU, interactive

ADJUSTING FOR INCOMPLETE DEATH ASCERTAINMENT IN JOINT MODELS: A MULTIPLE-IMPUTATION APPROACH

Constantin T. Yiannoutsos[1] , Giorgos Bakoyannis[1], Dimitris Rizopoulos[2]

[1] Presenter, Department of Biostatistics, Indiana University, IN 46202, U.S.A cyiannou@iupui.edu

[2] Department of Biostatistics, Erasmus University, Rotterdam 2040, the Netherlands

Abstract: Monitoring and evaluation of the effectiveness of HIV treatment programs involves esti-mating precisely the clinical outcome of their clients. These efforts are complicated by lack of vital status information on a large proportion of patients who are lost to clinic (LTC). This is a particularly vexing issue afflicting programs in low and middle-income settings at the epicentre of the worldwide HIV epidemic. We and others have proposed methods to use vital status information available on a random sub-sample of LTC patients (double-sampled dropouts) to adjust mortality estimates as LTC is invariably informative. We use the same information to adjust mortality estimates within jointly evaluated longitudinal and failure-time (joint) models. Joint models are useful when longitudinal data on each patient (e.g., CD4 counts) are incorporated into the failure-time model. We use random ef-fects to capture the association between the longitudinal and the failure-time process (Rizopoulos J Stat Soft, 2010). The longitudinal sub-model involves linear mixed models of the CD4 trajectory (at the sq. root scale), while the failure-time sub-model is a Cox model with a piece-wise linear baseline hazard. Age and gender at ART initiation were incorporated in both models. We use the additional vital status information obtained on double-sampled LTC patients to impute the unknown vital status in non-double-sampled LTC patients. Monte Carlo estimates of the expected survival are averaged over the imputed data sets producing a single adjusted estimate. Unadjusted estimates, considering all LTC patients as administratively censored were also generated. A linear mixed model with a linear and quadratic effect for square-root CD4 count and random intercept plus linear and quadratic slopes was fit to the longitudinal CD4 count data. The model distinguishes among patients based on events after entry (here after initiation of ART) and can naturally be adapted to accommodate additional information on patients lost to clinic when loss is informative but when data can be assumed to be missing at random conditional on vital status information obtained after dropout. The same model is also amenable to competing-risk situations when additional information (e.g., re-engagement in or disengagement from care is available. (Data are available in our database, but analyses are not shown here due to space considerations).

Key Words: Joint models, multiple imputation, unknown death ascertainment

Determining the number of factors based on the singular values

Cheng Wang

Department of Statistics, Shanghai Jiao Tong University, Shanghai, China

Abstract:In this work, we study the large approximate factor model where the cross-section dimension and the time dimension are both large. We revisit the existing estimation for the number of factors from the point of view the singular values of the data. The consistence of the estimation is proved with no restriction on the dependence structures between the factors and the idiosyncratic errors. Further, we study the model with time and/or individual effects and the related assumptions and consistence of the estimation are discussed. Finally, simulation experiments are conducted to demonstrate the results.

Optimal High-dimensional multiclass linear discriminant analysis

Shan Luo

Department of Statistics, Shanghai Jiao Tong University

Abstract: We reconsider the Bayes rule for multiclass linear discriminant analysis under high-dimensional situation. Theoretical results on the misclassification rate for LDA when $p<n$ but p is allowed to depend on n will be provided. We also propose a new method to estimate the population mean vectors and the common covariance matrix when p is much larger than n. The effectiveness of our method will be illustrated through extensive numerical results.

Variable selection for mixture and promotion time cure rate models

Abdullah Masud, Wanzhu Tu and Zhangsheng Yu[1]

[1]Department of Statistics, Shanghai Jiao Tong University

Abstract: Failure-time data with cured patients are common in clinical studies. Data from these studies are typically analyzed with cure rate models. Variable selection methods have not been well developed for cure rate models. In this research, we propose two least absolute shrinkage and selection operators based methods, for variable selection in mixture and promotion time cure models with parametric or nonparametric baseline hazards. We conduct an extensive simulation study to assess the operating characteristics of the proposed methods. We illustrate the use of the methods using data from a study of childhood wheezing.

Structured subcomposition selection in regression and its application to microbiome data analysis

Tao Wang

Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University

Abstract: Compositional data arise naturally in many practical problems and the analysis of such data presents many statistical challenges, especially in high dimensions. In this talk, we consider the problem of subcomposition selection in regression with compositional covariates, where the relationships among the covariates can be represented by a tree with leaf nodes corresponding to covariates. Assuming that the tree structure is available as prior knowledge, we adopt a symmetric version of the linear log contrast model, and propose a tree-guided regularization method for this structured subcomposition selection. Our method is based on a novel penalty function that incorporates the tree structure information node-by-node, encouraging the selection of subcompositions at subtree levels. We show that this optimization problem can be formulated as a generalized lasso problem, the solution of which can be computed efficiently using existing algorithms. An application to a human gut microbiome study and simulations are presented to compare the performance of the proposed method with an l_1 regularization method where the tree structure information is not utilized.

Semiparametric Spatial Model for Interval-censored Data with Time-Varying Covariate Effect

Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University

Yue Zhang, Xia Wang, Bin Zhang

Abstract: Cox regression model is one of the most commonly used methods in the analysis of interval-censored failure time data. In many practical studies, the covariate effects on the failure time may not be constant over time. In recent studies, time-varying coefficients are of great interest because of their flexibility in capturing the temporal covariate effects. In this paper, we propose a Bayesian approach to dynamic Cox regression model allowing for spatial correlation with interval-censored time-to-event data. With Bayesian approach, the coefficient curve is piecewise constant and the number of jump points are estimated from data. A conditional autoregressive distribution is employed to model the spatial dependency. The posterior summaries are obtained via an efficient reversible jump Markov chain Monte Carlo algorithm. The properties of our method are illustrated by simulation studies as well as an application to the smoking cessation data in southeastern Minnesota.