Conference Abstracts

(Presented According to Order in Conference Program)

Latent Variables Conference 2016 (LV2016)

Organized by the Department of Statistics and Continuing Education and Conferences

Partial Support from the US National Science Foundation (NSF), the National Institute of Statistical Sciences (NISS), the College of Arts and Sciences (University of South Carolina), and the Office of the Vice President for Research (University of South Carolina

October 12-14, 2016
USC Alumni Center
University of South Carolina
Columbia, South Carolina, USA

Session Number	**02**
Session Title	**Plenary I**
Time and Day	**8:45–9:45 am, Thursday**
Place	**USC Alumni Center Ballroom 3**
Session Organizer	Organizers
Session Chair	**John M. Grego**, University of South Carolina

Functional and Very High Dimension Reduction

Yanyuan Ma

Department of Statistics
Pennsylvania State University
323 Thomas Building
University Park, PA 16802

E-Mail: yanyuanma@gmail.com

Abstract: The first half of the talk is to study the relation between a univariate response and multiple functional covariates via a functional single index model that is semiparametric. The parametric part of the model integrates the linear regression modeling for functional data and the sucient dimension reduction structure. The nonparametric part of the model further allows the response-index dependence or the link function to be unspecied. We use B-splines to approximate the coecient func- tion in the functional linear regression model part and reduce the problem to a familiar dimension folding model. We develop a new method to handle the subsequent dimension folding model by using kernel regression in combination with semiparametric treatment. The new method does not impose any special requirement on the inner product between the covariate function and the B-spline bases, and allows ecient estimation of both the index vector and the B-spline coecients. The estimation method is general and applicable to both continuous and discrete response variables. We further derive asymp- totic properties of the class of methods for both the index vector and the coecient function. We establish the semiparametric optimality, which has not been done before in a semiparametric model where both kernel and B-spline estimation are involved. This work can be viewed as an extension of the classical dimension reduction model where covariates are vectors.

To extend dimension reduction in a different direction to very high dimensional setting, inspired by Fan et al. (2015), we study the relation between a response variable and high dimensional covariates through a combination of factor analysis and sucient dimension reduction. To take advantage of the exibility in traditional factor models where the latent factors are not required to be normal, we recommend using semipara- metric sucient dimension reduction methods in the joint estimation of the combined model. The resulting estimator is more exible and has improved performance. We also quantify the asymptotic performance of the parameter estimation and analyze a GTEx data set concerning gene-SNPs relation in lung tissues and discovered new signicant SNPs.

Session Number	**03 (Concurrent Invited)**
Session Title	**Diagnostic Screening/ROC Curves**
Time and Day	**10:00–11:30 am, Thursday**
Place	**USC Alumni Center Ballroom 1A**
Session Organizer	**Tim Hanson**, University of South Carolina
Session Chair	**Tim Hanson**, University of South Carolina

Discrimination Surfaces for Region-Specific Brain Asymmetry Analysis

Miguel de Carvalho

School of Mathematics
University of Edinburgh
James Clerk Maxwell Building
The King's Buildings
Peter Guthrie Tait Road
Edinburgh EH9 3FD

E-Mail: mdecarvalho@mat.puc.cl

Abstract: In this talk, discrimination surfaces are introduced as a natural model for localizing brain regions where discrimination between diseased and non-diseased subjects is higher. An applied goal of interest will be on conducting a brain asymmetry analysis so to localize brain regions where schizophrenia patients may differ further from healthy controls. The value at which a discrimination surface is maximized carries important information on the brain zone at which distance to asymmetry is higher. Parametric models for the discrimination surface will be discussed, along with an empirical estimator which can be regarded as a Mann–Whitney statistic for stochastic processes. Joint work with Gabriel Martos.

Nonparametric Bayesian regression analysis of the Youden index

Vanda Inacio de Carvalho

School of Mathematics
University of Edinburgh
James Clerk Maxwell Building
The King's Buildings
Peter Guthrie Tait Road
Edinburgh EH9 3FD

E-Mail: icalhau@mat.puc.cl

Abstract: A novel nonparametric regression model is developed for evaluating the covariate-specific accuracy of a continuous biological marker. Accurately screening diseased from nondiseased individuals and correctly diagnosing disease stage are critically important to health care on several fronts, including guiding recommendations about combinations of treatments and their intensities. The accuracy of a continuous medical test or biomarker varies by the cutoff threshold, c, used to infer disease status. Accuracy can be measured by the probability of testing positive for diseased individuals (the true positive probability or sensitivity, Se(c), of the test) and the true negative probability (specificity, Sp(c)) of the test. A commonly used summary measure of test accuracy is the Youden index, YI=maxSe(c)+Sp(c)-1: c in R, which is popular due in part to its ease of interpretation and relevance to population health research. In addition, clinical practitioners benefit from having an estimate of the optimal cutoff that maximizes sensitivity plus specificity available as a byproduct of estimating YI. We develop a highly flexible nonparametric model to estimate YI and its associated optimal cutoff that can respond to unanticipated skewness, multimodality and other complexities because data distributions are modeled using dependent Dirichlet process mixtures. Inferences are available for the covariate-specific Youden index and its corresponding optimal cutoff threshold. The value of our nonparametric regression model is illustrated using multiple simulation studies and data on the age-specific accuracy of glucose as a biomarker of diabetes. Joint work with Miguel de Carvalho and Adam Branscum.

Covariate adjusted measures of diagnostic accuracy based on pooled biomarkers

Chris McMahan

Department of Mathematical Sciences
Clemson University
O-110 Martin Hall
PO Box 340975
Clemson, SC 29634

E-Mail: mcmaha2@clemson.edu

Abstract: There exists an undeniable need for epidemiological and medical researchers to identify new biomarkers (biological markers) which are useful in determining exposure levels and/or for the purposes of disease detection. Often this process is stunted by the testing cost associated with evaluating new biomarkers. Traditionally, biomarker assessments are made by collecting and testing specimens from individuals within a target population. In the recent literature, pooling has been proposed to help alleviate the cost of data collection. Pooling strategies test pools formed by amalgamating several individual specimens, rather than test them one-by-one. Methods of estimating measures of discriminatory ability (e.g., the receiver operating characteristic curve, area under the curve, and Youden's index) based on pooled biomarker assessments have been developed. Regretfully, all of these procedures fail to acknowledge confounding factors. Consequently, in this work, we propose a regression methodology, based on pooled biomarker measurements, that allows for the assessment of the discriminatory ability of a biomarker of interest. In particular, we develop covariate adjusted estimators of the receiver operating characteristic curve, the area under the curve, Youden's index, and the optimal cut-point. In addition, we establish the asymptotic properties of these estimators and develop inferential techniques that allow one to assess whether a biomarker is a good discriminator between cases and controls, while controlling for confounders. The proposed methodology is used to analyze myocardial infarction data, with the goal of determining whether the pro-inflammatory cytokine Interleukin-6 is a good predictor of myocardial infarction after controlling for the subjects cholesterol level.

Session Number	**04 (Concurrent Invited)**
Session Title	**Model Misspecification Diagnosis**
Time and Day	**10:00–11:30 am, Thursday**
Place	**USC Alumni Center Ballroom 1B**
Session Organizer	**Xianzheng Huang**, University of South Carolina
Session Chair	**Xianzheng Huang** , University of South Carolina

A nonparametric goodness-of-fit test for random effects models via cross-validation Bayes factors

Jeffrey Hart

Department of Statistics
3143 TAMU
College Station, TX 77843

E-Mail: hart@stat.tamu.edu

Abstract: Consider the simple random effects model having just two random components, one of which is a main effect and the other an error effect. We consider testing the goodness of fit of a parametric model for the distributions of the two random components, which are assumed independent of each

other. By far the most common such model is one in which both components are normally distributed. A nonparametric Bayesian procedure is proposed for conducting the goodness-of-fit test. Alternatives to the hypothesized parametric model are based on bivariate kernel density estimates. Data splitting makes it possible to use kernel estimates for this purpose in a Bayesian setting. A kernel estimate indexed by bandwidth is computed from one part of the data, a training set, and then used as a model for the rest of the data, a validation set. A Bayes factor is calculated from the validation set by comparing the marginal for the kernel model with the marginal for the parametric model. Simulation results showing good behavior of the Bayes factor under both null and alternative hypotheses are presented. A real-data example illustrating the methodology is also provided.

Latent variable augmented sparse regression

Jinchi Lv

Data Sciences and Operations Department
Marshall School of Business
University of Southern California
Los Angeles, CA 90089

E-Mail: jinchilv@marshall.usc.edu

Abstract: As a powerful tool for producing interpretable models, sparse modeling has gained popularity for analyzing large-scale data sets. Most of existing methods assume implicitly that all features in a model are observable. Yet some latent confounding factors may potentially exist in the hidden structure of the original model. On the other hand, the key assumption of sparsity that enables high-dimensional inference has been questioned in many applications. In this paper, we propose a new framework, latent variable augmented sparse regression (LAVAR), based on the conditional sparsity assumption that the coefficient vector is sparse after taking out common latent features. In particular, we consider one potential family of latent variables that are linearly dependent on a group of observable features, represented by the population principal components. The latent factors are estimated by the sample score vectors and asymptotic properties are established for a wide class of distributions. With the aid of these properties, we prove that the proposed framework can enjoy model selection consistency and oracle inequalities under various prediction and variable selection losses for both observable predictors and latent confounding factors. Our new method and results are evidenced by simulation and real data examples. This is joint work with Wei Lin and Zemin Zheng.

Prediction risk for global-local shrinkage regression

Anindya Bhadra

Department of Statistics
Purdue University
250 N University Street
West Lafayette, IN 47907-2066

E-Mail: bhadra@purdue.edu

Abstract: Predictive performance in shrinkage regression suffers from two major difficulties: (i) the amount of relative shrinkage is monotone in the singular values of the design matrix and (ii) the amount of shrinkage does not depend on the response variables. Both of these factors can translate to a poor prediction performance, the risk of which can be explicitly quantified using Stein's unbiased risk estimate. We show that using a component-specific local shrinkage term that can be learned from the data under a suitable heavy-tailed prior, in combination with a global term providing shrinkage towards zero, can alleviate both these difficulties and consequently, can result in an improved risk for prediction. Demonstration of improved prediction performance over competing approaches in a simulation study and in a pharmacogenomics data set confirms the theoretical findings. Joint work with Jyotishka Datta, Yunnan Li, Nick Polson and Brandon Willard.

Session Number	**05 (Concurrent Invited)**
Session Title	**Factor Models**
Time and Day	**10:00–11:30 am, Thursday**
Place	**USC Alumni Center Ballroom 3**
Session Organizer	**Yanyuan Ma**, Pennsylvania State University
Session Chair	**Yanyuan Ma**, Pennsylvania State University

Variable selection for longitudinal data analysis in the presence of missing observations and measurement error

Grace Yi

Department of Statistics and Actuarial Science
University of Waterloo
200 University Avenue West
Waterloo, ON, Canada N2L 3G1

E-Mail: yyi@uwaterloo.ca

Abstract: Longitudinal studies have proven to be useful in studying changes of response over time, and have been widely conducted in practice. It is common that longitudinal studies collect a large number of covariates, some of which are unimportant in explaining the response. Including such covariates in modelling and inferential procedures would greatly degrade the quality of the results. Moreover, longitudinal data analysis is challenged by the presence of measurement error and missing observations. In this talk, I will discuss the issues induced from these features, and describe simultaneous variable selection and estimation procedures that handle high dimensional longitudinal data with missingness and measurement error.

A Bayesian Latent Variable Approach to Aggregation of Top-ranked Partial Gene lists in Genomic Studies

Sherry Wang

Department of Statistical Science
Southern Methodist University
104 Heroy Science Hall
3225 Daniel Avenue
Dallas, TX 75275-0332

E-Mail: swang@mail.smu.edu

Abstract: Rank aggregation has a rich history in the field of information retrieval, with applications to text mining, webpage ranking, meta-search engine building, etc. However, methods developed in such contexts are often ill suited for genomic applications, in which gene lists generated from individual studies are inherently noisy, due to various sources of heterogeneity. Further, because of missing or zero-count data, a portion of genes are not analyzed in all component studies, leading to partially ranked lists; and for some lists, only top-ranked genes are reported. In this study, we develop Bayesian latent variable approached to rank aggregation that formally deals with top and partial preference lists.

Identification of endogenous retrovirus integration sites using a mixture model

Le Bao

Department of Statistics
Pennsylvania State University
323 Thomas Building
University Park, PA 16802

E-Mail: lebao@psu.edu

Abstract: The presence or absence of (endogenous retroviruses) ERVs can be determined by identifying the junction with the host genome using high-throughput sequence technology. The resulting data are a matrix giving the number of sequence reads assigned to each ERV-host junction sequence for each sampled individual. We present a novel two-component mixture of negative binomial distributions to model these counts and to assign a probability that a given ERV is present in a given individual. We explain ways in which our approach is superior to existing alternatives, including another form of two-component mixture model and the much more common approach of selecting a threshold count for declaring the presence of an ERV. We apply our method to a data set of ERVs in mule deer from Oregon, Montana and Wyoming and blacktail deer from Oregon.

Session Number	**06**
Session Title	**Plenary II**
Time and Day	**1:00–2:00 pm, Thursday**
Place	**USC Alumni Center Ballroom 3**
Session Organizer	Organizers
Session Chair	**Don Edwards**, University of South Carolina

A review of Bayesian nonparametric regression through mixture models

Sonia Petrone

Department of Decision Sciences
Universitá Boccone
Via Roentgen, 1 (3rd floor)
20136 Milano Italy

E-Mail: sonia.petrone@unibocconi.it

Abstract: The talk will offer an overview of some modeling approaches for Bayesian nonparametric regression. The growth of Bayesian nonparametrics in the last decades has seen an explosion of models and methods, and the talk will restrict the attention on density regression, or estimation of a conditional density, through Dirichlet mixture models. Even so, the literature is vast and fragmented. We try to give a unifying viewpoint, and address comparison among different proposals. Our underlying point is that, even beyond frequentist asymptotic properties, fine details of the nonparametric prior may have a relevant impact on the finite sample properties and the predictive performance. A challenge that we also address is to provide a computational strategy that can be fairly easily adapted to the different models under examination. This is joint work with Sara Wade and Michele Peruzzi.

Session Number	**07 (Concurrent Invited)**
Session Title	**Measurement Error**
Time and Day	**2:15-3:45 pm, Thursday**
Place	**USC Alumni Center Ballroom 1A**
Session Organizer	**Josh Tebbs**, University of South Carolina
Session Chair	**Josh Tebbs**, University of South Carolina

Functional Semiparametric Bayesian Time Varying Coefficient ME Models in Matched Case-Crossover Studies

Inyoung Kim

Department of Statistics
Virginia Tech
Hutcheson Hall, Rm 406-A
250 Drillfield Drive
Blacksburg, VA 24061

E-Mail: inyoungk@vt.edu

Abstract: Matched case-crossover studies are typically studied using conditional logistic regression models. Using these models, any stratum effect is removed by the conditioning on the fixed number of sets of the case and controls in the stratum. Hence these models prevent the detection of any effects associated with the matching covariates by stratum. However, some matching covariates such as time, often play an important role in effect modification. The failure to include it can potentially yield incorrect statistical estimation. In addition, other factors such as spatial location can create heterogeneous subpopulations among strata. Often covariates in such studies are measured with error. Not accounting for this error can lead to incorrect inference for all covariates in the model. The methods for assessing and characterizing error-in-covariates in matched case-control studies are quite limited. Hence in this paper, we propose a nonparametric Bayesian approach constructed with Dirichlet process priors, which clusters subpopulations and assesses heterogeneity. Our functional semiparametric Bayesian approach is developed under semiparametric time varying coefficient models for matched case-crossover studies. This approach allows us to detect parametric relationships between the predictor and binary outcomes, nonparametric relationships between the predictor and time, as well as functional clusters of time varying coefficients among strata. We demonstrate the accuracy of our approach using a simulation study, as well as an example of a 1-4 bi-directional case-crossover study of childhood aseptic meningitis with drinking water turbidity.

Association Study of Children's Methylation and Growth Trajectory using Functional Mixed Models

Arnab Maity

NCSU Statistics Department
2311 Stinson Drive
Campus Box 8203
Raleigh, NC 27695-8203

E-Mail: arnab_maity@ncsu.edu

Abstract: Motivated from the Newborn Epigenetic STudy (NEST) data, we consider the problem of association study between children growth trajectory and children gene methylation profile while accounting for other confounders. We develop a functional semiparametric regression modeling framework where the response is functional variable (children growth trajectory measured over time), and scalar and vector valued covariates (gene methylation profile and other confounders). We model the joint effect of the gene methylation profile nonparametrically using the Gaussian process framework, and model the

remaining confounders parametrically. We develop estimation and hypothesis testing procedures for the effect of the gene methylation profiles using functional mixed effects models and variance components. We evaluate the performance of our method using a simulation study and via application to the NEST data. Joint work with Colleen McKendry and Jung-Ying Tzeng

Analysis of proportional odds models with censoring and errors-in-covariates

Samiran Sinha

Department of Statistics
3143 TAMU
College Station, TX 77843

E-Mail: simha@stat.tamu.edu

Abstract: In this talk I will describe a consistent method for analyzing time-to-event data in the presence of errors-in-covariates and right censoring. We shall assume that the time-to-event follows the proportional odds model. The proposed method does not rely on the distributional assumption of the true covariate which is not observed in the data. In addition, the proposed estimator does not require the measurement error to be normally distributed or to have any other specific distribution, and we do not attempt to assess the error distribution. Instead, we construct martingale based estimators through inversion, using only the moment properties of the error distribution, estimable from multiple erroneous measurements of the true covariate. The theoretical properties of the estimators are established and the finite sample performance is demonstrated via simulations. For illustration purpose, I will describe an application of the proposed method to a real dataset from a clinical study on AIDS.

Session Number	**08 (Concurrent Invited)**
Session Title	**Item Response Theory**
Time and Day	**2:15-3:45 pm, Thursday**
Place	**USC Alumni Center Ballroom 1B**
Session Organizer	**Brian Habing**, University of South Carolina
Session Chair	**Brian Habing**, University of South Carolina

IRT modelling of ordinal forced-choice data

Alberto Maydeu-Olivares

Department of Psychology
University of South Carolina
1512 Pendleton Street
Barnwell College, Suite #220
Columbia, SC 29208

E-Mail: amaydeu@sc.edu

Abstract: As an alternative to rating individual items in personality and similar questionnaires, researchers may use comparative response formats. The simplest of such formats is the popular forced choice, where items are presented in blocks and respondents are to provide their top choice and least preferred choice. The extent to which the items are preferred to each other can also be of interest; for example, respondents may indicate how strongly they prefer one item to another using several graded categories. We extend the Thurstonian IRT modeling approach (Brown & Maydeu-Olivares, 2011) beyond binary choice to graded preferences.

The Multidimensional Generalized Graded Unfolding Model: Issues and Applications

James Roberts

School of Psychology J.S. Coon Bldg
Georgia Institute of Technology
654 Cherry Street
Atlanta, Georgia 30332-0170

E-Mail: james.roberts@psych.gatech.edu

Abstract: This presentation will discuss a relatively new multidimensional IRT model for unfolding responses that follow from a proximity relation between persons and stimuli in a latent space. The model yields higher expected values to the extent that a given person is located close to a stimulus in the space. For two dimensions, the expected value surface is single peaked and symmetric about the stimulus in the directions of the two corresponding axes. These notions extend to higher dimensional spaces as well. The model will be illustrated with applications in the areas of facial affect and physical attraction. Issues surrounding dimensionality assessment with data that follow this model will also be briefly addressed.

A cautionary tale on equating

Louis Roussos

Measured Progress
100 Education Way
Dover, NH 03820

E-Mail: Roussos.Louis@measuredprogress.org

Abstract: One of the most common latent variable model uses is the application of item response theory (IRT) to model the psychometric properties of educational testing programs. Each such program typically involves at least a dozen, if not several dozen, exams that are given to thousands of students each year; and each exam attempts to maintain a consistent score scale across typically five or more years of testing, using different forms each year.

While establishing a psychometric scale for a single exam can be a complicated task, maintaining the same scale across multiple years using different forms requires important further assumptions that must be carefully understood and monitored.

In this talk I will present a case study of one such situation that will highlight the difficulty of this endeavor and how large errors can unwittingly occur if you are not careful. Joint work with Jian Jiang of Boston College and Unhee Ju of Michigan State University.

Session Number	**09 (Concurrent Invited)**
Session Title	**Latent Variables in Medical Testing**
Time and Day	**2:15-3:45 pm, Thursday**
Place	**USC Alumni Center Ballroom 3**
Session Organizer	**Don Edwards**, University of South Carolina
Session Chair	**Don Edwards**, University of South Carolina

Modeling rater diagnostic skills in binary classification processes

Xiaoyan Lin

Department of Statistics
University of South Carolina
Columbia, SC

E-Mail: lin@stat.sc.edu

Abstract: Many disease diagnoses involve subjective judgments by qualified raters. For example, through the inspection of a mammogram, MRI, or ultrasound image, the clinician himself becomes part of the measuring instrument. To reduce diagnostic errors and improve the quality of diagnoses, it is necessary to assess raters' diagnostic skills and to improve their skills over time. This paper focuses on a subjective binary classification process, proposing a hierarchical model linking data on rater opinions with patient true disease-development outcomes. The model allows for the quantification of the effects of rater diagnostic skills (bias and ability) and patient disease severity on the rating results. A Bayesian Markov chain Monte Carlo (MCMC) algorithm is developed to estimate these parameters. Linking to patient true disease outcomes, the rater-specific sensitivity and specificity can be estimated using MCMC samples. Cost theory is utilized to identify poor- and strong-performing raters and to guide adjustment of rater bias and diagnostic ability to improve the rating performance. An extensive simulation study is conducted to evaluate the proposed methods, and the methods are illustrated with a mammography example.

Modeling agreement between multiple raters' ordinal classifications

Kerrie Nelson

Department of Biostatistics
Boston University School of Public Health
Crosstown Building
801 Massachusetts Avenue 3rd Floor
Boston MA 02118

E-Mail: kerrie@bu.edu

Abstract: A screening test often involves the subjective interpretation of a patient's test result by a medical expert using an ordered categorical scale. Due to the subjectivity involved, wide discrepancies between experts' classifications have been reported, especially in diagnostic procedures such as mammography. These concerns have motivated large-scale studies to examine levels of agreement between experts in common screening settings and to investigate if factors such as rater experience affect consistency of ratings. However, limited statistical methods exist to assess agreement in large-scale studies such as these. In this talk, we introduce a chance-corrected approach based upon the ordinal class of generalized linear mixed models to assess agreement between multiple raters, assuming the true underlying disease status of a patient is an unobserved continuous latent variable. Our modeling approach also allows us to examine the impact of subject- and rater-related characteristics on agreement. We apply our approach to a recent large-scale mammography study. Joint work with Don Edwards.

A latent class modeling approach for predicting disease status using functional data in the absence of a gold standard

Amita Manatunga

Department of Biostatistics and Bioinformatics
Rollins School of Public Health
Emory University
1518 Clifton Road NE
Atlanta GA 30322

E-Mail: amanatu@sph.emory.edu

Abstract: We consider a latent class modeling approach for predicting disease status of a subject based on observed functional data in the absence of a gold standard. In our study, two consecutive curves over time are observed per subject and the second curve for some subjects are not observed. While there is no gold standard for determining disease status, there are ratings for disease status from multiple experts. This work is motivated by a renal study focused on evaluation of kidney obstruction based on diruresis renograhpy (curves). Three experts determined the obstruction status as obstructed, equivocal and non-obstructed. Conditioning on the latent status, we develop a random effects model for renogram curves and a probit model for expert ratings, which allows unstructured correlation for ratings among experts. A Bayesian procedure is developed for obtaining parameter estimates and subsequently, prediction of disease status. We describe the modeling procedure and provide simulation studies to evaluate the finite sample properties of our method and several prediction schemes. Finally, we apply our method to the motivating study and show that the pattern of estimated genogram curves for obstructed and non-obstructed kidneys are reasonably consistent with clinical interpretations. Research with Lijia Wang, Qi Long and Andrew Taylor.

Session Number	**10**
Session Title	**Plenary III**
Time and Day	**4:00–5:00 pm, Thursday**
Place	**Russell House Ballroom 3**
Session Organizer	Organizers
Session Chair	**John Grego**, University of South Carolina

Latent story of the stick breaking repesentation for the Dirichlet process

Jayaram Sethuraman

Department of Statistics
Florida State University
117 N. Woodward Avenue
P.O. Box 3064330
Tallahassee, FL 32306-4330

E-Mail: sethu@stat.fsu.edu

Abstract: When was the stick breaking representation for the Dirichlet process discovered? It is latent in two published papers of 1973! The final self-contained general version was discovered in 1978 while teaching a seminar course on Dirichlet processes with David Blackwell, Deb Basu and Jim Lynch as auditors and participants. It was published in 1994. This talk will provide details.

Session Number	**12 (Concurrent Invited)**
Session Title	**Frailty Models**
Time and Day	**8:00-9:30 am, Friday**
Place	**USC Alumni Center Ballroom 1A**
Session Organizer	**Edsel Peña**, University of South Carolina
Session Chair	**Edsel Peña**, University of South Carolina

Haiqun Lin
Yale School of Public Health
PO Box 208034
60 College Street
New Haven, CT 06520-8034
E-Mail: haiqun.lin@yale.edu

Abstract:

Frailty Models in the Social Sciences

Amanda Fairchild
Department of Psychology
University of South Carolina
1512 Pendleton Street
Barnwell College, Suite 220
Columbia, SC 29208
E-Mail: amanda.fairchild@sc.edu

Abstract: This talk considers the use of frailty models (and related approaches) in the social sciences. Applied examples of discrete-time survival mixture analysis, growth mixture modeling and a variety of other mixed models will be discussed. Emphasis will be placed on describing the contexts in which these models are being used, to facilitate an understanding of how statisticians can contribute to helping these scientists extend use of these tools to investigate their research questions. Outstanding needs and current concerns regarding these models in social science will also be discussed.

Elizabeth Slate
Department of Statistics
Florida State University
214 Rogers Building (OSB), 117 N. Woodward Avenue
PO Box 3064330
Tallahassee, FL 32306-4330
E-Mail: slate@stat.fsu.edu

Abstract:

Session Number	**13 (Concurrent Invited)**
Session Title	**Objective Bayes Analysis in Latent Variable Models**
Time and Day	**8:00–9:40 am, Friday**
Place	**USC Alumni Center Ballroom 1B**
Session Organizer	**Xiaoyan Lin**, University of South Carolina
Session Chair	**Xiaoyan Lin**, University of South Carolina

Latent-Variable Approaches for Accurate Computation in Bayesian Scale-Usage Models

Chris Hans

Department of Statistics
Ohio State University
1958 Neil Avenue
Columbus, OH 43210

E-Mail: hans@stat.osu.edu

Abstract: A common modeling approach for data collected on discrete scales is to introduce continuous latent variables, treated as missing data, that are linked to the observations via a censoring mechanism. Bayesian approaches typically use data augmentation within a Markov chain Monte Carlo (MCMC) algorithm that simplifies analysis by avoiding direct evaluation of the likelihood. However, if not carefully implemented, the cost of data augmentation is that Markov chain mixing can be severely degraded, rendering the method useless for inference. Motivated by the need to consider complex, high-dimensional models where the latent variables are not independently distributed, we introduce a covariance decomposition for the analysis of discrete data models with multivariate normal latent variables. The decomposition is shown to facilitate joint MCMC updates of the latent variables and censoring points. Our decomposition results in a sampling algorithm that only requires univariate normal integrals, which can be evaluated with high accuracy. We provide theoretical and practical guidance for choosing a good decomposition and present an illustration that demonstrates its effectiveness over a commonly used approach that employs numerical approximations of multivariate normal integrals over rectangular regions, which, when repeated many times in an MCMC algorithm, can have a substantial impact on the chains limiting distribution. This is joint work with Greg Allenby, Peter Craigmile, Ju Hee Lee (UC Santa Cruz), Steven MacEachern and Xinyi Xu.

Fast Bayesian Factor Analysis via Automatic Rotations to Sparsity

Veronika Ročková

Booth School of Business
University of Chicago
5807 Southlawn Avenue
Chicago, IL 60637

E-Mail: vrockova@wharton.upenn.edu

Abstract: Rotational post-hoc transformations have traditionally played a key role in enhancing the interpretability of factor analysis. Regularization methods also serve to achieve this goal by prioritizing sparse loading matrices. In this work, we bridge these two paradigms with a unifying Bayesian framework. Our approach deploys intermediate factor rotations throughout the learning process, greatly enhancing the effectiveness of sparsity inducing priors. These automatic rotations to sparsity are embedded within a PXL-EM algorithm, a Bayesian variant of parameter-expanded EM for posterior mode detection. By iterating between soft-thresholding of small factor loadings and transformations of the factor basis, we obtain (a) dramatic accelerations, (b) robustness against poor initializations and (c) better oriented sparse solutions. To avoid the pre-specification of the factor cardinality, we extend the loading matrix to have infinitely many columns with the Indian Buffet Process (IBP) prior. The factor dimensionality is learned from the posterior, which is shown to concentrate on sparse matrices. Our deployment of PXL-EM performs a dynamic posterior exploration, outputting a solution path indexed by a sequence of spike-and-slab priors. For accurate recovery of the factor loadings, we deploy the Spike-and-Slab LASSO prior, a two-component refinement of the Laplace prior (Ročková, 2015). A companion criterion, motivated as an integral lower bound, is provided to effectively select the best recovery. The potential of the proposed procedure is demonstrated on both simulated and real high-dimensional data, which would render posterior simulation impractical.

Permuted and Augmented Stick-Breaking Multinomial Regression

Mingyuan Zhou

Department of Information, Risk and Operations Management
McCombs School of Business
University of Texas at Austin
CBA 5.202
Austin, TX 78712

E-Mail: mingyuan.zhou@mccombs.utexas.edu

Abstract: To model categorical response variables given their covariates, we propose a permuted and augmented stick-breaking construction that one-to-one maps the S observed categories to S randomly permuted latent sticks. This new construction transforms multinomial regression into regression analysis of S stick-specific binary random variables, which are mutually independent given their covariate-dependent stick success probability parameters. We model these binary random variables using Bayesian support vector machines (SVM) to construct Bayesian multinomial SVM (MSVM). Furthermore, we parameterize the negative logarithms of the stick failure probabilities with a family of covariate-dependent softplus functions to construct nonparametric Bayesian multinomial softplus regression (MSR). Both Bayesian MSVM and MSR are not only capable of producing nonlinear classification decision boundaries, but also amenable to posterior simulation. Example results are used to demonstrate the attractive properties and appealing performance of both Bayesian MSVM and MSR.

An Objective Prior for Hyperparameters in Normal Hierarchical Models

Dongchu Sun

Department of Statistics
University of Missouri
146 Middlebush Hall
Columbia, MO 65211

E-Mail: sund@missouri.edu

Abstract: Normal latent models (or Normal hierarchical models) are the workhorse of much of Bayesian analysis, yet there is uncertainty as to which objective priors to use for latent variables. Formal approaches to objective Bayesian analysis, such as the Jeffreys-rule approach or reference prior approach, are only implementable in simple settings without latent variables. Thus it is common to use less formal approaches, such as utilizing formal priors without latent variables. This can be fraught with danger, however. For instance, non-hierarchical Jeffreys-rule priors for variances or covariance matrices result in improper posterior distributions if they are used at higher levels of a hierarchical model. Thus such less formal approaches must be carefully evaluated, and not just from the perspective of posterior propriety.

Berger, Strawderman and Tang (2005) approached the question of choice of latent priors in normal hierarchical models by looking at the frequentist notion of admissibility of resulting estimators. The motivation was that latent priors that are too diffuse result in inadmissible estimators, while hyperpriors that are concentrated enough result in admissible estimators. The priors for latent variables that are 'on the boundary of admissibility' are sensible choices for objective priors, being as diffuse as possible without resulting in inadmissible procedures. The admissibility (and propriety) properties of a number of priors were considered in the paper, but no overall conclusion was reached as to a specific prior to recommend, in part because they were not able to prove admissibility for the leading candidate prior.

In this talk, we complete the story and propose a particular objective prior for use in normal latent models, based on considerations of admissibility, ease of implementation (including computational considerations), and performance.

Joint work with James O. Berger, Duke University and Chengyong Song, East China Normal University

Session Number	**14**
Session Title	**Plenary IV**
Time and Day	**9:50–10:50 am, Friday**
Place	**USC Alumni Center Ballroom 3**
Session Organizer	Organizers
Session Chair	**David Hitchcock**, University of South Carolina

Scaling and Generalizing Variational Inference, with examples from Bayesian nonparametrics

David Blei

Department of Statistics
Columbia University
Room 1005 SSW, MC 4690
1255 Amsterdam Avenue
New York, NY 10027

E-Mail: david.blei@columbia.edu

Abstract: Latent variable models have become a key tool for the modern statistician, letting us express complex assumptions about the hidden structures that underlie our data. Latent variable models have been successfully applied in numerous fields.

The central computational problem in latent variable modeling is posterior inference, the problem of approximating the conditional distribution of the latent variables given the observations. Posterior inference is central to both exploratory tasks and predictive tasks. Approximate posterior inference algorithms have revolutionized Bayesian statistics, revealing its potential as a usable and general-purpose language for data analysis.

Bayesian statistics, however, has not yet reached this potential. First, statisticians and scientists regularly encounter massive data sets, but existing approximate inference algorithms do not scale well. Second, most approximate inference algorithms are not generic; each must be adapted to the specific model at hand.

In this talk I will discuss our recent research on addressing these two limitations. I will describe stochastic variational inference, an approximate inference algorithm for handling massive data sets. I will demonstrate its application to probabilistic topic models of text conditioned on millions of articles and to Bayesian nonparametric mixtures applied to massive data sets.

Then I will discuss black box variational inference. Black box inference is a generic algorithm for approximating the posterior. We can easily apply it to many models with little model-specific derivation and few restrictions on their properties. I will demonstrate its use on longitudinal models of healthcare data, deep exponential families, nonconjugate Bayesian nonparametric models, and discuss a new black-box variational inference algorithm in the Stan programming language.

Session Number	**15 (Concurrent Invited)**
Session Title	**Model-based Clustering**
Time and Day	**11:00 am–12:30 pm, Friday**
Place	**USC Alumni Center Ballroom 1A**
Session Organizer	**David Hitchcock**, University of South Carolina
Session Chair	**David Hitchcock**, University of South Carolina

Bayesian community detection with unknown number of communities

Debdeep Pati

Department of Statistics
Florida State University
PO Box 3064330
Tallahassee, FL 32306-4330

E-Mail: debdeep@stat.fsu.edu

Abstract: A fundamental problem in network analysis is clustering the nodes into groups which share a similar connectivity pattern. Existing algorithms for community detection assume the knowledge of the number of clusters or estimate it a priori using various selection criteria and subsequently estimate the community structure. Ignoring the uncertainty in the first stage may lead to erroneous clustering, particularly when the community structure is vague. We instead propose a coherent probabilistic framework (MFM-SBM) for simultaneous estimation of the number of communities and the community structure, adapting recently developed Bayesian nonparametric techniques to network models. An efficient Markov chain Monte Carlo (MCMC) algorithm is proposed which obviates the need to perform reversible jump MCMC on the number of clusters. The methodology is shown to outperform recently developed community detection algorithms in a variety of synthetic data examples and in benchmark real-datasets. We derive non-asymptotic bounds on the marginal posterior probability of the true configuration, and subsequently use it to prove a clustering consistency result which is novel in the Bayesian context to best of our knowledge.

Generalized linear mixed models with gaussian mixture random effects: an application to nationwide kidney transplant center evaluation

Yehua Li

Department of Statistics
3214 Snedecor Hall
Iowa State University
Ames, IA 50011

E-Mail: yehuali@iastate.edu

Abstract: Five year post-transplant survival rate is an important indicator on quality of care delivered by kidney transplant centers in the United States. To provide a fair assessment of each transplant center, an effect that represents the center-specific care quality, along with patient level risk factors, is often included in the risk adjustment model. In the past, the center effects have been modeled as either fixed effects or Gaussian random effects, with various pros and cons. Our numerical analyses reveal that the distributional assumptions do impact the prediction of center effects especially when the effect is extreme. To bridge the gap between these two approaches, we propose to model the transplant center effect as a latent random variable with a finite Gaussian mixture distribution. Such latent Gaussian mixture models provide a convenient framework to study the heterogeneity among the transplant centers. To overcome the weak identifiability issues, we propose to estimate the latent Gaussian mixture model using a penalized likelihood approach, and develop sequential locally restricted likelihood ratio tests to determine the number of components in the Gaussian mixture distribution. The fitted mixture model provides a convenient means of controlling the false discovery rate when screening for underperforming or outperforming transplant centers. The performance of the methods is verified by simulations and by the analysis of the motivating data example.

Edge-exchangeable graphs, sparsity, and power laws

Tamara Broderick

Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Room 38-401
77 Massachusetts Avenue
Cambridge, MA 02139

E-Mail: tbroderick@csail.mit.edu

Abstract: Many popular network models rely on the assumption of (vertex) exchangeability, in which the distribution of the graph is invariant to relabelings of the vertices. However, the Aldous-Hoover theorem guarantees that these graphs are dense or empty with probability one, whereas many real-world graphs are sparse. We present an alternative notion of exchangeability for random graphs, which we call edge exchangeability, in which the distribution of a graph sequence is invariant to the order of the edges. By expanding on results from clustering, we characterize the class of edge-exchangeable models with a paintbox construction and probability functions. And we demonstrate that edge-exchangeable models, unlike models that are traditionally vertex exchangeable, can exhibit sparsity and power laws. To do so, we outline a general framework for graph generative models; by contrast to previous work, models within our framework are stationary across steps of the graph sequence. In particular, our model grows the graph by instantiating more latent atoms of a single random measure as the dataset size increases, rather than adding new atoms to the measure.

Session Number	**16 (Concurrent Invited)**
Session Title	**Group Testing and Biomarker Pooling**
Time and Day	**11:00 am-12:30 pm, Friday**
Place	**USC Alumni Center Ballroom 1B**
Session Organizer	**Dewei Wang**, University of South Carolina
Session Chair	**Dewei Wang**, University of South Carolina

A general framework for the regression analysis of pooled biomarker assessments

Colin Gallagher

Department of Mathematical Sciences
Clemson University
O-110 Martin Hall
PO Box 340975
Clemson, SC 29634

E-Mail: cgallag@clemson.edu

Abstract: As a cost-efficient data collection mechanism, the process of assaying pooled biospecimens is becoming increasingly common in epidemiological research; e.g. pooling has been proposed for the purpose of evaluating the diagnostic efficacy of biological markers (biomarkers). To this end, several authors have proposed techniques that allow for the analysis of continuous pooled biomarker assessments. Regretfully, most of these techniques proceed under restrictive assumptions, are unable to account for the effects of measurement error, and fail to control for confounding variables. These limitations are understandably attributable to the complex structure that is inherent to measurements taken on pooled specimens. Consequently, in order to provide practitioners with the tools necessary to accurately and

efficiently analyze pooled biomarker assessments, herein a general Monte Carlo maximum likelihood-based procedure is presented. The proposed approach allows for the regression analysis of pooled data under practically all parametric models and can be used to directly account for the effects of measurement error. Through simulation, it is shown that the proposed approach can accurately and efficiently estimate all unknown parameters and is more computationally efficient than existing techniques. This new methodology is further illustrated using monocyte chemotactic protein-1 data collected by the Collaborative Perinatal Project in an effort to assess the relationship between this chemokine and the risk of miscarriage.

Revisiting Nested Group Testing Procedures: New Results, Comparisons, Robustness

Yaakov Malinovsky

Department of Mathematics and Statistics
University of Maryland, Baltimore County
1000 Hilltop Circle
Baltimore, MD 21250

E-Mail: yaakovm@umbc.edu

Abstract: Group testing has its origin in the identification of syphilis in the U.S. army during World War II. Much of the theoretical framework of group testing was developed starting in the late 1950's with continued work into the 1990's. Recently, with the advent of new laboratory and genetic technologies, there has been increasing attention to group testing applications for cost savings. In this paper we compare different nested designs including Dorfman, Sterrett, and an optimal nested procedure. To elucidate these comparisons we develop closed form expressions for the optimal Sterrett procedure as well as provide a concise review of the prior literature for other commonly used procedures. We also investigate the robustness of these procedures.

Adjusting for Processing or Measurement Error in Regression Analyses with Biomarker Exposure Levels Assessed on Pooled Samples

Robert Lyles

Rollins School of Public Health
Emory University
1518 Clifton Road NE
Atlanta GA 30322

E-Mail: rlyles@sph.emory.edu

Abstract: When laboratory assay costs are high, potential benefits associated with the pooling of biological specimens motivate statistical considerations to facilitate regression analysis involving group-level exposure measurements. However, the pooling of samples can introduce errors in measurement due to processing, possibly in addition to error that may be present when the assay is applied to individual samples. We look into methods that might be applied to address this type of measurement error problem in common regression settings. As suggested by prior research addressing overall mean and variance estimation, hybrid designs consisting of individual as well as pooled samples facilitate the estimation of processing (or pooling) error, while further variation in pool sizes may be called for to identify a potential underlying measurement error variance. For continuous outcomes, one can consider maximum likelihood (ML) or approaches based on regression calibration in conjunction with ordinary or weighted least squares under hybrid designs. For binary outcomes, we assess the potential applicability of discriminant function analysis in addition to alternatives such as ML based on logistic regression. As time permits, we will consider an example involving actual outcome and laboratory assay data and discuss the potential benefits of incorporating replicate measurements.

Session Number	**17**
Session Title	**Plenary V**
Time and Day	**2:15-3:15 pm, Friday**
Place	**USC Alumni Center Ballroom 3**
Session Organizer	Organizers
Session Chair	**John Grego**, University of South Carolina

From Latent Variable and Mixture Models to Inference in Subgroup Analysis

Xuming He

Department of Statistics
University of Michigan
311 West Hall
1085 South University
Ann Arbor, MI 48109-1107

E-Mail: xmhe@umich.edu

Abstract: We propose a statistical model for the purpose of identifying a subgroup that has an enhanced treatment effect as well as the variables that are predictive of the subgroup membership. The need for such subgroup identification arises in clinical trials and in market segmentation analysis. By using a latent variable for group membership in a normal-logistic mixture model, our proposed framework enables us to perform a confirmatory statistical test for the existence of subgroups, and at the same time, to construct predictive scores for the subgroup membership. In this talk we will discuss subgroup analysis in clinical studies, present the asymptotic theory for the proposed test and provide some empirical results to demonstrate how the test can be used to reduce false positive reporting in subgroup identification. The talk is based on joint work with Dr. Juan Shen.

Session Number	**18 (Concurrent Invited)**
Session Title	**Recent Advances in Regression Analysis of** Interval-censored Data
Time and Day	**3:30-5:00 pm, Friday**
Place	**USC Alumni Center Ballroom 1A**
Session Organizer	Organizers
Session Chair	**Lianming Wang**, University of South Carolina

Regression analysis of informatively interval-censored failure time data

Tony Sun

Department of Statistics
University of Missouri
146 Middlebush Hall
Columbia, MO 65211

E-Mail: sunj@missouri.edu

Abstract: Interval-censored failure time data occur in many fields such as demography, economics, medical research and reliability, and many inference procedures on them have been developed (Chen et al., 2012; Sun, 2006). However, most of the existing approaches assume that the mechanism that yields interval censoring is independent of the failure time of interest and it is clear that this may not be true in practice. In this talk, we will discuss this latter situation and present some inference procedures for the problem.

Accounting for measurement uncertainty in environmental preterm studies

Shuangge Ma

Laboratory of Epidemiology and Public Health
Yale University
60 College Street, Suite 201
New Haven, CT 06510

E-Mail: shuangge.ma@yale.edu

Abstract: Environmental exposures, especially including air pollution, have been identified as risk factors for preterm birth. Although a large amount of data has been collected, there has been insufficient attention to statistical methodological development. Specifically, the existing methods ignore the uncertainty in the measurements of pregnancy days (which directly defines preterm) as well as air pollution (e.g., PM2.5). In this study, we adopt a semiparametric single index model to describe pregnancy days. An imputation approach is developed to accommodate measurement uncertainty in both exposure and response variable. Inference is conducted using a weighted bootstrap approach. This new approach leads to additional insights into the effect of air pollution on preterm birth.

Bayesian semiparametric models for spatially correlated arbitrarily censored data

Haiming Zhou

Division of Statistics
Northern Illinois University
1425 West Lincoln Highway
DeKalb, IL 60115

E-Mail: zhouh@niu.edu

Abstract: A comprehensive, unified approach to modeling arbitrarily censored spatial survival data is presented for the three most commonly-used semiparametric models: proportional hazards, proportional odds, and accelerated failure time. Unlike many other approaches, all manner of censored survival times are simultaneously accommodated including uncensored, interval censored, current-status, left and right censored, and mixtures of these. Both georeferenced (location observed exactly) and areally observed (location known up to a geographic unit such as a county) spatial locations are handled. Georeferenced data are modeled with a discrete process convolution whereas areal data are modeled with a Markov random field. Variable selection is also incorporated. All models are fit via new functions which call efficient compiled C++ in the R package `spBayesSurv`. The methodology is broadly illustrated with simulations and real data applications.

Session Number	**19 (Concurrent Invited)**
Session Title	**Large Sample Testing and Model Selection**
Time and Day	**3:30-5:00 pm, Friday**
Place	**USC Alumni Center Ballroom 1B**
Session Organizer	Organizers
Session Chair	**John Grego**, University of South Carolina

A modified mixed model approach to the large scale multiple testing problem

Paramita Chakraborty

Department of Statistics
University of South Carolina
Columbia, SC

E-Mail: chakrabp@stat.sc.edu

Abstract: In modern big data situations, such as in microarray analysis, one source of lack of reproducibility is the voluminous number of false positives/discoveries that occur. Recently a number of studies have been done to propose new methods of handling large scale multiple tests to counter that problem. An effective way of dealing with the situation is: use a mixture contamination model for the entire data to identify real significant (contaminants) cases. This is done on the basis of the false discovery rate (FDR) associated with each case. This method has been presented as an empirical Bayes approach; in the light of the fact that the FDR is actually a posterior probability derived from the fitted mixture model.

We propose a further modified approach using the mixture model with cross-validation style data partitioning. Which not only gives us a way to identify significant cases but also helps us to balance out other sources of variation in the data. Moreover, this method gives us an insight towards the hierarchical inter-relation between the significant cases.

Multiple testing approaches for removing background noise from images

Subhashis Ghoshal

Department of Statistics
NC State University
5109 SAS Hall
2311 Stinson Drive
Raleigh, NC 27695-8203

E-Mail: ghoshal@stat.ncsu.edu

Abstract: Images arising from low-intensity settings such as in X-ray astronomy and computed tomography scan often show a relatively weak but constant background noise across the frame. The background noise can result from various uncontrollable sources. In such a situation, it has been observed that the performance of a denoising algorithm can be improved considerably if an additional thresholding procedure is performed on the processed image to set low intensity values to zero. The threshold is typically chosen by an ad-hoc method, such as 5% of the maximum intensity. In this talk, we formalize the choice of thresholding through a multiple testing approach. At each pixel, the null hypothesis that the underlying intensity parameter equals the intensity of the background noise is tested, with due consideration of the mulltiplicity factor. Pixels where the null hypothesis is not rejected, the estimated intensity will be set to zero, thus creating a sharper contrast with the background. The main difference of the present context with the usual multiple testing applications is that in our setup, the null value in the hypotheses is not known, and must be estimated from the data. We employ a Gaussian mixture model to estimate the unknown common null value of the background intensity level. We discuss three approaches to solving the problem and we compare them through simulation studies. The methods are

applied on noisy X-ray images of a supernova remnant.

A double empirical Bayes approach for high-dimensional problems

Ryan Martin

Department of Statistics
NC State University
5109 SAS Hall
2311 Stinson Drive
Raleigh, NC 27695-8203

E-Mail: rgmarti3@ncsu.edu

Abstract: In high-dimensional problems, selecting a good prior, i.e., one that leads to optimal posterior concentration properties and efficient computation, can be a challenge. In this talk, I will present a new kind of empirical Bayes that uses data in the prior in two ways: first, the prior is suitably centered on the data and, second, a regularization step is taken to prevent the greedy centering from driving the posterior behavior. In the context of a sparse high-dimensional linear model, a variety of posterior concentration results will be presented, along with simulations that demonstrate the method's quality model selection performance. Extensions to other high-dimensional problems will also be discussed.

Poster Session Presenters

October 14, 2016 (Friday)
USC Alumni Center Ballroom Two
Period: 12:00–6:00
(Set-Up Time: Friday Morning)

Latent promotion time cure rate model using dependent tailfree mixtures

Li Li

Department of Mathematics & Statistics
University of New Mexico
Science and Math Learning Center 230
311 Terrace NE MSC01 1115 Albuquerque, NM 87131

E-Mail: llis@umn.edu

Abstract: The paper presented in this poster extends the latent promotion time cure rate marker model (Kim et al., 2009) for right-censored survival data. Instead of modeling the cure rate parameter as a deterministic function of risk factors, Kim et al. (2009) assumed the cure rate parameter of a targeted population be distributed over a number of ordinal levels according to the probabilities governed by the risk factors. In this work, we propose to use a mixture of linear dependent tail-free processes as the prior for the distribution of the cure rate parameter, resulting in a latent promotion time cure rate model (LPTCR). This approach provides an immediate answer to perhaps one of the most pressing questions: "what is the probability that a targeted population has high proportions (e.g. ¿70%) of being cured?" The proposed approach can accommodate a rich class of distributions for the cure rate parameter, while centered at Gamma densities. The algorithms developed in this work allow the fitting of LPTCR with several survival models for metastatic tumor cells.

Identifying a PTSD resilient group based on a latent class analysis of childhood trauma and adult PTSD symptoms

Megan Warnock

2301 Briarcliff Rd NE Apt C
Atlanta, GA 30329

E-Mail: megan.warnock@emory.edu

Abstract: The goal of this study was to investigate and better understand the structure of heterogeneity in post-traumatic stress disorder (PTSD) based on reported childhood trauma and current PTSD symptoms using latent class analysis (LCA). Participants were recruited from an urban hospital in Atlanta, GA. LCA were conducted on data collected from an early cohort of participants using 25-item childhood trauma questionnaire (CTQ) and 17-item modified PTSD symptom scale (PSS) to identify subclasses. Robustness of the latent subclasses was evaluated using a recently collected, independent group of participants. To better understand the heterogeneity between subclasses, resilience and affect differences between subclasses were analyzed.

LCA on the early cohort of 3940 subjects suggested four subclasses provided the best fit with meaningful, distinct subclasses: 1. Low childhood trauma, high PTSD symptoms; 2. High childhood trauma, high PTSD symptoms; 3. High childhood trauma, low PTSD symptoms; 4. Low childhood trauma, low PTSD symptoms. LCA on the recently collected dataset of 1299 subjects supported these results as robust. Subclass 3 experienced childhood trauma and yet had low PTSD symptoms, highlighting this group as psychologically resilient to PTSD. In comparison to subclasses 1 and 2, who had high PTSD symptoms, subclass 3 was less negative and depressed and had higher resilience measures compared to subclass 2. In future studies wed expect to see a psychologically resilient group, which may help in identifying persons at lower risk for PTSD. Joint work with Amita Manatunga, Ying Guo, Limin Peng, Tanja Jovanovic.

Group testing regression models with dilution submodels

Md Warasi

Department of Mathematics and Statistics
Walker Hall 209
Radford University
Radford, VA 24142

E-Mail: msarker@radford.edu

Abstract: Group testing, where specimens are tested initially in pools, is widely used to screen individuals for sexually transmitted diseases. However, a common problem found in practice is that group testing can increase the number of false negative test results. This occurs primarily when positive individual specimens within a pool are diluted by negative ones, resulting in positive pools testing negatively. If the goal is to estimate a population regression model relating individual disease status to covariates, regression parameters and individual-level disease probabilities can be severely biased if an adjustment for this dilution effect is not made. Recognizing this as a critical issue for estimation, recent binary regression approaches in group testing have utilized additional continuous biomarker information to incorporate the dilution effect. In this paper, we have the same overall modeling goals but we take a different approach. We augment existing group testing regression models (that assume no dilution) with a parametric dilution submodel for pooled-level sensitivity and estimate all parameters using maximum likelihood. The primary advantage of our approach is that it does not rely on external biomarker test data, which may not be available in seroprevalence studies. Furthermore, unlike previous approaches, our flexible framework allows us to formally test whether dilution is present based on observed group testing data. We use simulation to illustrate the finite-sample performance of our regression methods, and we apply our estimation and inference techniques to two sexually transmitted disease data sets. The poster is based on joint work with Dr. Chris McMahan and Dr. Josh Tebbs.

Size Investing Strategy on Multiple Confidence Intervals under FWER

Taeho Kim

Department of Statistics
1523 Greene Street
Columbia, SC 29208

E-Mail: taeho@email.sc.edu

Abstract: For confidence interval problems, researchers generally attempt to minimize the length of the interval, maintaining the coverage probability. By extending this approach to multiple confidence intervals, the optimal Size Investing Strategy is investigated given the global size ? (equivalently global coverage probability 1-?) under FWER. Simulation results show that 98% and 92% of the total lengths decrease on the 1,000 location parameters of normal and t random variables, respectively. However, the resulting investing strategy shows less dynamic structure than that of the multiple testings (Peña et al., 2011). This is because the classical way of individual confidence interval does not equip the minimization procedure in itself. Joint work with Edsel Peña

Adaptive posterior contraction rates in non-linear latent variable models

Shuang Zhou

Department of Statistics // 117 N Woodward Ave // Tallahassee Fl 32303-4330

E-Mail: shuang.zhou@stat.fsu.edu

Abstract: Non-linear latent variable models have become increasingly popular in a variety of machine learning applications. In this poster, we are interested in studying the theoretical properties of

the posterior distribution arising from a fully Bayesian specification of the non-linear latent variable model. Specifically, we study the rates of posterior contraction in univariate density estimation where the unobserved uniformly distributed latent variables are related by the response variables via a random non-linear regression with an additive error. We characterize the space of densities as kernel convolution with a class of continuous mixing measures, and obtain the density of response variable conditional on the unknown regression function by marginalizing out the latent variables. Our key idea is to approximate the true density using a normal kernel convolution and then allow the prior on the regression function to appropriately concentrate around the true quantile function. Using a Gaussian process prior on the regression function with a random bandwidth parameter, we obtain optimal posterior contraction rate adaptively over the space of β-Hölder densities with compact support up to a logarithmic factor. The proof techniques rely on the construction of an approximator to the true density which lies in the prior support. With appropriate smoothness conditions on the true function, the approximator is obtained by defining a new density using convolution with a twicing kernel. To achieve the optimal convergence rate, we obtain sharp upper bounds for the Hellinger distance between any two densities lying in the prior support and sharp lower-bound to the prior concentration probability. In addition, the proof requires careful calibration of the unit reproducing kernel Hilbert space balls of the Gaussian process prior.

Non-compensatory MIRT with Rotation

Xinchu Zhao

Department of Statistics
1523 Greene Street
Columbia, SC 29208

E-Mail: xinchu@email.sc.edu

Abstract: Compensatory multidimensional item response theory (MIRT) models, allow for rotation of axes as in factor analysis. In those models, the correlation of the underlying abilities is arbitrary, and parameters from one correlation structure have a one-to-one link to others, as such it is not necessary to specify or estimate the correlation to recover the model. This is not the case for the non-compensatory MIRT models (NCM) due to its multiplicative structure.

The purpose of this study was to develop a non-compensatory multidimensional item response model that allows for transformation between different correlation structures. This non-compensatory model with additional discrimination parameters was created to allow for rotation as the compensatory model and estimation without specifying the correlation. The performance of the model was evaluated by simulation. In the simulation, the data was simulated for the standard two dimensional non-compensatory model with correlated abilities. It was estimated using the new model as if the abilities were uncorrelated. Parameters were estimated via Metropolis-Hasting algorithm within Gibbs sampler for both the previous NCM and the created model. The parameter recovery was evaluated after transforming the estimates back onto scales where the abilities the correlated, and it performed well.