**Estimation Methods on Marginal Models for Clustered Data Using EM-Like Algorithms**

Tawanda Benesi,Statistics, University of Georgia

The EM algorithm is a useful tool in maximum likelihood estimation. However, in cases where the likelihood in intractable or difficult to formulate as in marginal models for clustered data, alternative methods can be used. Several authors have proposed EM-like methods which can be applicable is such cases were the quasi-likelihood or general estimating functions can be used. Examples include the expectation-solution(ES) algorithm by Rosen (2000), the projection-solution (PS) algorithm by Morton (1986) and the expectation-solution (ES) algorithm by Ryan (2004). Marginal ZIP models for clustered data of Hall and Zhang (2004) and marginal models for clustered angular/directional data of Hall and Shen (2015) use the expectation-solution algorithm by Rosen. However, this method does not account for within-cluster associations at the E-step since independence is assumed at this stage. A model based E-step which utilizes the relationship between linear discriminant analysis and multinomial logit models assuming underlying Gaussian distributions for the mixture components is proposed for the ZIP case to account for with-in cluster associations. The modified projection-solution (PS) algorithm will be used for the marginal models for clustered angular data.

**Inference on Interval-valued Data Regression by Measurement Error Models**

Yaotong Cai and Lynne Billard,Statistics, University of Georgia

Among different types of symbolic data, interval-valued data is one of the most common to be studied. Compared with several approaches of statistical inference on interval-valued linear regression model, measurement error theory provides a new perspective about how to estimate coefficient estimates as well as their variances. By means of measurement error, the constraint that values within interval observations follow a uniform distribution can be relaxed when conducting inference on interval-valued data. In this paper, we first introduce concepts of measurement error, derive the maximum likelihood

function with respective to regression coefficients, with the measurement error term to be considered, and then propose a method of point estimation and confidence interval by measurement error for regression models on interval-valued data, with several different distribution assumptions on the within values. A simulation study is conducted to evaluate the performances of the proposed method.

**Regression Analysis for Symbolic Interval Data: An overview of the Center, Center and Range, and Symbolic Covariance methods**

Natalia Costa Araujo and Lynne Billard,Statistics, University of Georgia

Symbolic data analysis was first introduced by Diday (1987) and presents an alternative approach to classical data when the data have a more complex formulation. It can be presented in a variety of types of data, but this project focus on the interval-valued data, that can happen naturally or due to the aggregation of large datasets.

In addition to a brief description of the symbolic interval-valued data, three different regression analysis approaches are presented, the Center, Center and Range and Symbolic Covariance methods. The center method (Billard and Diday, 2000) proposes that the interval-valued data should be centered and modeled as a classical linear model. The center and range method (Lima Neto and De Carvalho, 2008), on the other hand, suggests that the interval-valued data should be modeled by two independent regression equations, one for the center and another for the range of the intervals. Finally, the symbolic covariance method (Xu, 2010) introduces the use of symbolic covariance matrices to estimate the parameters, based on the least squares estimation.

**Random Dual Rotation: Generalized Permutation Test for High Dimension Low Sample Size Data**

Hee Cheol Chung and Jeongyoun Ahn, Statistics, University of Georgia

We propose a randomized hypothesis test for high dimension low sample size (HDLSS) data. Resampling methods such as the permutation and sign-change tests are popular in hypothesis testing with HDLSS data, especially

when the null distribution of a test statistic is not readily available. However, when the hypotheses are regarding the disposition of the data such as in clustering analysis, permutation-based tests are not applicable since they cannot change the data configuration. In this work, we propose a new randomization method called the random dual rotation (RDR) by considering a family of random matrices whose distribution is invariant under a specific group of transformation. The randomized test based on RDR generalizes permutation and sign-change tests, and it can change the configuration of the data while preserving the covariance structure. We provide a theoretical construction of the RDR test by studying the invariance measures on Stiefel and Grassmann manifolds. The proposed idea is applied to a outlier detection problem.

**Dimension Reduction in Time Series: A Simultaneous Estimation Approach**

Murilo Massaru Da Silva, Statistics, University of Georgia

This research on time series deals with the problem of finding which linear combinations of the past observations that influence the series present expected value. Additionally, we are also concerned with a heteroscedastic structure of the residuals that we assume to depend on its past squared values. The main difference between this approach and the classical methods is that we do not assume linearity or any fixed structure to our model before estimating the desired parameters, i.e., the dependence structure is assumed to be unknown, but dependent on the past. We make use of nonparametric and numerical techniques to estimate the parameter matrices we are interested in.

**Computer Model Calibration for Design, with an Application to Wind Turbine Blades**

Carl Ehrett, Mathematical Sciences, Clemson University

Computer simulations have become a common means of studying phenomena for which it is difficult to acquire data through direct physical experimentation. Often these computer models contain unknown inputs, called

calibration inputs, the values of which must be estimated for successful simulation. The value of a calibration input may be estimated, for example, by combining observations of the simulator output with real-world experimental data. Previous explorations of computer model calibration have approached calibration as a matter of bringing a computer model into agreement with physical reality. In the present work, we consider computer model calibration as a method for design. Under this framework, we calibrate a computer model not using physical experimental data, but rather using desired data which describes the performance one hopes to achieve in the simulated system. We illustrate this technique using a finite element simulation of wind turbine blade performance. We create a Gaussian process emulator of the finite element output and use Markov chain Monte Carlo sampling to calibrate the parameters of the emulator. Whereas in traditional model calibration, the result of calibration would be to discover the settings that allow a simulation to approximate reality, here the result of calibration is to discover settings that allow the simulation to approximate the desired performance outcome.

**From Mixed-Effects Modeling to Spike-and-Slab Variable Selection: A Bayesian Regression Model for Group Testing Data**

Chase Joyner, Mathematical Sciences, Clemson University

Due to dramatic reductions in both time and cost, group (pooled) testing is becoming a popular alternative to individual level testing. These reductions are gained by testing pooled bio-specimen (e.g., blood, urine, saliva, etc.) for the presence of an infectious agent. Though this process may reduce the cost of classification, it comes at the expense of data complexity, thus making the complimentary task of conducting disease surveillance more tenuous. More to the point, the statistical analysis of data arising from a group testing procedures is a nontrivial task due to the fact that an individual's disease status is obscured by the effect of imperfect testing, and in some instances the group testing protocol. Further, unlike individual level testing, a given participant could be involved in multiple testing outcomes, but may never be tested individually. To circumvent these hurdles and to incorporate all of the available data, we propose a Bayesian generalized linear mixed model which can accommodate data arising from any group testing

procedure, estimate unknown assay accuracies, and account for the potential heterogeneity in the covariate effects across clusters (e.g., clinic sites), with the latter feature being of key interest to practitioners tasked with conducting diseases surveillance. To achieve model selection, the proposed approach makes use of spike and slab priors for both the fixed and random effects. The proposed methodology is illustrated through extensive numerical studies and is applied to a motivating chlamydia data set collected by the State Hygienic Laboratory in Iowa City.

**Acknowledging the Dilution Effect in Group Testing Regression: A New Approach**

Stefani Mokalled, Mathematical Sciences, Clemson University

From screening for infectious diseases to detecting bioterrorism, group (pooled) testing of bio-specimen is a cost efficient alternative to individual level testing. Group testing has been utilized for both classification (identifying positive individuals) and estimation (fitting regression models using covariate measurements). A concern with the estimation process is the possible dilution of one individual's positive signal past an assay's threshold of detection. To account and correct for this dilution effect, we develop a new group testing regression model which explicitly acknowledges the effect. Unlike previous work in this area, this is accomplished by considering the continuous outcome that the assay measures, the individuals' latent biological marker (biomarker) levels, and the distributions of the biomarker levels of the cases and controls without requiring a priori knowledge of these distributions. We develop a novel mixture model and an expectation-maximization algorithm to complete model fitting. The performance of the methodology is evaluated through numerical studies and is illustrated using Hepatitis B data on Irish prisoners.

**Statistical Inference for Porous Materials using Persistent Homology**

Chul Moon, Statistics, University of Georgia

We propose a porous materials analysis pipeline using persistent homology. We first compute persistent homology of binarized 3D images of sampled material subvolumes. For each image we compute sets of homology intervals, which are represented as summary graphics called persistence diagrams. We convert persistence diagrams into image vectors in order to analyze the similarity of the homology of the material images using the mature tools for image analysis. Each image is treated as a vector and we compute its principal components to extract features. We fit a statistical model using the loadings of principal components to estimate material porosity, permeability, anisotropy, and tortuosity. We also propose an adaptive version of the structural similarity index (SSIM), a similarity metric for images, as a measure to determine the statistical representative elementary volumes (sREV) for persistence homology. Thus we provide a capability for making a statistical inference of the fluid flow and transport properties of porous materials based on their geometry and connectivity.

### B-Scaling: A Novel Nonparametric Data Fusion Method

Yiwen Liu, Statistics, University of Georgia

With the rapid development in science and technology, massive data has been collected from different sources, which leads to a large amount of data with different types and formats, such as the image data and omics data. Each type of the data only captures part of the contained information, and the data has to be integrated or fused to provide a complete understanding of the whole picture. Thus, there is an urgent call of powerful data fusion method. In this presentation, I introduce a B-scaling method to integrate multisource data. The asymptotic property of the B-scaling method is discussed to provide theoretical underpinning of the method. The application of the method on epigenetic and biomedical research will be highlighted in the presentation.

### Model-based Clustering with Application of Copulas for Symbolic Data

Wenhao Pan and Lynne Billard, Statistics, University of Georgia

Contemporary data sets can be too large or complex for traditional statistical methods to handle. One approach is to use symbolic data first introduced by Diday (1987). Our interest is the study of model-based clustering for symbolic data, especially for distributions (i.e., observations are not single numerical point values). We will describe symbolic data and consider differences between symbolic data and classical data. For multivariate data, with p ¿ 1, we only have the marginal distributions; so we do not know the dependence relationship between random variables. One approach to measure these dependencies is that of Vrac et al. (2012) in which a copula function is used to describe the cumulative joint distribution function of random variables in a mixture model. We further develop the algorithm from various perspective and make it applicable.

**The Persistent Homology of Rossby Waves**

Richard Ross, Statistics, University of Georgia

Recent changes in both climate and weather patterns have caused researchers to more closely examine the Jet Stream and its patterns of perturbation known as Rossby waves. The frequency of these waves depends on season, but some researchers believe that the overall frequency of Rossby waves has increased over time, likely due to a decrease in the temperature gradient between the Equator and North Pole. Unfortunately, there is no publicly available dataset reporting the number of waves present on a given date despite an abundance of raw data which can be analyzed for this purpose. The noisy nature of this data presents some difficulty in reporting these waves. We propose the use of persistent homology (a branch of Topological Data Analysis) on isobaric surface data from NOAA to fill this need. The resulting output allows us to report the quantity of Rossby waves on a given date, with the potential for researchers to find waves of specified magnitudes or levels of persistence. This analysis focuses on using sequential Morse filtrations to quickly capture the presence and persistence of formations with a given number of waves over a suitable range of latitudes and filtration heights. We also present longitudinal summaries of this data, illustrating the change in frequency and persistence of 5-wave structures (of interest in winter months) over the past 60 years. The results of this work show a gradual increase in the number of waves overall, especially in winter months. These

results will help climate researchers to more consistently study these waves and how they change in frequency and amplitude.

### BrainPack: A Suite of Advanced Statistical Techniques for Multi-Subject and Multi-Group FMRI Data Analysis

Arunava Samaddar, Nicole A. Lazar, Jennifer E. McDowell, Cheolwoo Park, University of Georgia

We aim to evaluate brain activation using functional Magnetic Resonance Imaging (fMRI) data and activation changes across time associated with practice related cognitive control during tasks. FMR images are acquired from participants engaged in 1 block design and 5 probability event related designs at two scan sessions: 1) pre-practice before any exposure to the task, and 2) post-practice, after 4 days of daily practice on either general antisaccade (generating a glance away from the cue) tasks or specific probability related event runs, which are a mixture of antisaccade and prosaccade (generating a glance towards the cue) task. The clustering technique is composed of several steps: detrending, data aggregation, wavelet transform and thresholding, the adaptive pivotal thresholding test, principal component analysis and $K$-medoids clustering. We use the structural similarity index to compare similarity between pre- and post- scan session images on the probability event related runs. Also, we apply the semiparametric model under shape invariance to test the differences between the two sessions and the two practice groups in regions of interest for the block run.

### A Bayesian Binomial GLMM for Modeling Spatially Varying Trends in Large Datasets with a Computationally Efficient Model Fitting Procedure

Stella Watson Self, Mathematical Sciences, Clemson University

A methodology for modeling large binary datasets with spatial and temporal correlation is presented. A Bayesian GLMM for a binomial response is constructed. Gaussian predictive processes are used to efficiently model spatially varying covariate effects. A conditional autoregressive (CAR) structure

is used to model spatio-temporal correlation. The model is fit using MCMC techniques. Latent Polya Gamma random variables are introduced to improve computational efficiency. The model is applied to a dataset consisting of 16,000,000+ Lyme disease test results from domestic dogs. The data is collected across the contiguous United States and aggregated by county and by month. The model is used to assess where Lyme disease prevalence is increasing and decreasing. This model provides a computationally efficient way to model binary data with spatio-temporal structure while allowing covariate effects to change with space.

**Optimal Penalized Function-on-Function Regression under a Reproducing Kernel Hilbert Space Framework**

Xiaoxiao Sun, Statistics, University of Georgia

Many studies collect data with response and predictor variables both being functions of some covariate. Their common goal is to understand the relationship between these functional variables. Motivated from two real-life examples, we propose a new function-on-function regression model that can be used to analyze such kind of functional data. Our estimator of the 2D coefficient function is the optimizer of a form of penalized least squares where the penalty enforces certain level of smoothness on the estimator. Our first result is the Representer Theorem which states that the exact optimizer of the penalized least squares actually resides in a data-adaptive finite dimensional subspace although the optimization problem is defined on a function space of infinite dimensions. This theorem then allows us an easy incorporation of the Gaussian quadrature into the optimization of the penalized least squares, which can be carried out through standard numerical procedures. We also show that our estimator achieves the minimax convergence rate in mean prediction under the framework of function-on-function regression. Extensive simulation studies demonstrate the numerical advantages of our method over the existing ones. The method is then applied to a histone regulation study.

**Identifying and Extracting a Seasonal Streamflow Signal from Remotely Sensed Snow Cover in the Columbia River Basin**

Benjamin Washington, Lynne Seymour, Thomas Mote, David Robinson,
and Thomas Estilow, University of Georgia

In the western United States, meltwater from mountain snowpacks serves as the dominant water supply for many communities. Efficient distribution and use of this renewable, yet temporally and spatially variable resource relies critically on accurate forecasting of future water availability. Here we report on initial efforts to use Interactive Multisensor Snow and Ice Mapping System (IMS) data on snow coverage to forecast flow in six selected watersheds within the Columbia River Basin. Little research has been done on identifying the relationship between seasonal discharge volume and these satellite-derived snow cover data. In the Yakima watershed within the Columbia River Basin, we could explain 52% of the spring discharge (April July total streamflow volume) variance by selecting specific 24-km grid cells that exhibit both strong correlation with historical flows as well as high inter-annual variation. This approach yielded reasonable success in other watersheds. Of the six Columbia River subbasins examined in this paper, five of them give statistically significant predictors of April July streamflow volume at the $= 0.05$ level. When comparing this optimized specific-cell technique to the overall average across the entire watershed of interest, we observe improvements in each of our six subbasins, although in some regions, improvements were minimal. Clearly, this optimization technique is inherently limited by the role of snow cover variation in determining streamflow discharges in different subbasins. For both mountainous regions with extensive and stable snow cover as well as low-elevation regions with consistently minimal snow, the snow cover variation only accounts for a small inter-annual streamflow discharge variance. Our methodology shows that the IMS provides remotely-sensed data that are ready to plug and play into existing streamflow forecast models such as the Natural Resources Conservation Services (NRCS) Visual Interactive Prediction and Estimation Routines (VIPER).

## A Gamma-Frailty Proportional Hazards Model for Bivariate Interval-Censored Data

Prabhashi Wickramasingha, Mathematical Sciences, Clemson University

The Gamma-frailty proportional hazards (PH) model is commonly used to analyze correlated survival data. Under this model, the regression param-

eters have marginal interpretations and the statistical association between the failure times can be explicitly quantified via Kendalls tau. In this work, a Gamma-frailty PH model for bivariate interval-censored data is presented and an expectation-maximization (EM) algorithm for model fitting is developed. The proposed model adopts a monotone spline representation for the purposes of approximating the unknown conditional cumulative baseline hazard functions, significantly reducing the number of unknown parameters while retaining modeling flexibility. The EM algorithm was derived from a novel data augmentation procedure involving latent Poisson random variables. The algorithm is easy to implement, robust to initialization, and enjoys quick convergence. Simulation results suggest that the proposed method provides reliable estimation and valid inference, and is robust to the misspecification of the frailty distribution. To further illustrate its use, the proposed method is used to analyze data from an epidemiological study of sexually transmitted infections.

**Online Sequential Leveraging Sampling Method for Streaming Time Series Data**

Rui Xie, Statistics, University of Georgia

Advances in data acquisition technology pose challenges in analyzing large volumes of streaming data. Sampling is a natural yet powerful tool for analyzing such data sets due to their competent estimation accuracy and low computational cost. Unfortunately, sampling methods and their statistical properties for streaming data, especially streaming time series data, are not well studied in the literature. In this article, we propose an online leverage-based sequential sampling algorithm for streaming time series data, which is assumed to come from an autoregressive model of order $p \geq 1$ (AR($p$)). The proposed *sequential leveraging sampling* method samples only one consecutively recorded block from the data stream for inference. While the starting point of the sequential sampling scheme is chosen using a random mechanism based on leverage scores of the data, the subsample size is decided by a sequential sampling threshold. We show that an appropriately normalized sequential least squares estimator of the AR parameter vector is uniformly asymptotically normally distributed for non-explosive AR($p$) model. Simu-

lation studies and real data examples are presented to evaluate the empirical performance of the proposed sequential leveraging sampling method.

**Decentralized algorithm for estimating dimension reduction space**

Jingyi Zhang, Wenxuan Zhong and Ping Ma, Statistics, University of Georgia

Due to the data transmission cost and data privacy, traditional statistical tools cannot be directly applied to the scattered datasets. Decentralized algorithms tackle this problem by keeping the data in local nodes, and exchanging only the estimator in each optimization step, thus received significant attention recently. In this paper, we consider the problem that how to search an effective dimension reduction space when the data are scattered in different nodes. We propose a decentralized algorithm which extending the minimum average variance estimation (MAVE) method. Theoretical results show the proposed method has the same efficiency as the full sample MAVE approach, even when there exists batch effect on different nodes. Simulation study and real data examples indicate the proposed method dominates the existing distributed algorithms.

**MetaGen: Reference-Free Learning with Multiple Metagenomic Samples**

Xin Xing, Statistics, University of Georgia

Metagenomics refers to the study of a collection of genomes, typically microbial genomes, presenting in environmental samples. By sequencing bulk DNA that is directly extracted from environmental samples, one can bypass the difficulties arising in cell cultivation. Moreover, one can identify novel microbial species and study their distributional variations in different samples. However, these advantages cannot directly benefit biological researchers without a high-resolution and reference-free Metagenomics tool. We propose a statistically-based algorithm, MetaGen, to simultaneously identify microbial species and estimate their abundances in multiple metagenomics samples

without using any reference genome. The novelty in this project is that we propose to use a new marker, the relative abundance pattern cross samples, to distinguish different species in multiple metagenomic samples. By using this new marker, MetaGen can handle data with fairly low sequencing coverage, which can be extremely challenging with the currently available methods for metagenomic analysis. Also, MetaGen shows higher power in distinguishing genetically similar species comparing with existing methods. In addition, MetaGen can estimate the relative abundance of the microbial species in each sample without utilizing any reference information, which provides a way to study the composition of environmental microbial communities quantitatively. We have demonstrated the performance of MetaGen on simulated data sets and successfully implemented our algorithm on three large-scale biological data sets related to inflammatory bowel disease, type 2 diabetes, and obesity.

**Symbolic Data Analysis and Image Recognition**

Jiankun Zhu and Lynne Billard, Statistics, University of Georgia

There are many basic symbolic data types, such as modal categorical data, interval data, histogram data and distributions. One symbolic observation can be treated as a combination of ordinary single observations. Thus, when the sample size of a dataset is too large to analyze efficiently, we can first transform the ordinary data into symbolic data and then conduct symbolic data analysis on this transformed dataset. After the transformation, we can obtain a dataset with much smaller sample size and analyzing this dataset will be much easier and more efficient. However, when transforming the original dataset into a symbolic dataset, we may lose some information for each individual graph. Thus, we make some experiment designs to keep as much information of the original images as possible. We use histograms to catch the distribution of different pixels values of an image and the distribution of the locations of those dark pixels in a image. The final results show our method of transformation can keep most information of the images.

**Bayesian Spline Smoothing with Ambiguous Penalties**

Xinlian Zhang, Statistics, University of Georgia

A popular approach for flexible function estimation in nonparametric models is through spline smoothing using the general penalized likelihood method. In applying this method, one needs to specify a penalty functional which puts a soft constraint on the function to be estimated. A good choice of penalty functional is of key importance. In practice, specifying the penalty functional is mostly based on expert knowledge of the system. However, for many dynamic systems there naturally exist more than one sets of well-studied theory that explains the dynamics systems, i.e., there exist more than one sensible choices of penalties. To tackle this problem, we propose an approach that takes into consideration of all candidate penalties as well as the ambiguity in choosing among them. We take a fully Bayesian perspective, made use of the connection between penalized least squares and Bayesian estimation, and model the uncertainty of choosing penalty through introducing a mixture distribution as prior for parameters to be estimated. We also propose efficient sampling algorithm for making inference based on taking samples from posterior distribution.