Bayesian Modeling, Computation, and Applications

A Conference in Honor of Professor Lynn Kuo

Agenda – May 12, 2018 – University of Connecticut

9:00am **Registration at Lawrence D. McHugh Hall Room 101**

9:30am **Open Remarks**

Dipak Dey & Ming-Hui Chen, University of Connecticut

9:55am **Keynote Session**

Chair: Zhen Chen, Ph.D., National Institutes of Health

A Conditional Density Estimation Partition Model Using Logistic Gaussian Processes

--Bani Mallick, Texas A&M University

10:40am **Coffee Break**

10:50am **Session 1: Advanced Statistical Methods for Multivariate and High-Dimensional Data**

Chair: Jinjian Mu, M.S., University of Connecticut

Bayesian Modeling of Sparse High-Dimensional Data using Divergence Measures

--Dipak Dey, University of Connecticut

A Correlated Bayesian Rank Likelihood Approach to Multiple ROC Curves for Endometriosis

--Zhen Chen, National Institutes of Health

A Mixture Model and an Empirical Bayes Approach to Detect Edges in Sparse Co-expression Graphs

--Haim Bar, University of Connecticut

Statistical Learning for Biomedical Data Integration and Analysis

--Yuping Zhang, University of Connecticut

12:10pm **Lunch at McMahon Dining Hall**

1:30pm **Session 2: Bayesian Modelling and Application**

Chair: Wei Shi, M.S., University of Connecticut

On Spatial Disease Mapping

--Sudipto Banerjee, University of California, Los Angeles

Finite Population Unequal Probability Bayesian Bootstraps and Multiple Imputation

--Mike Cohen, American Institutes for Research

A Bayesian Regularized Mediation Analysis with Multiple Exposures

--Yu-Bo Wang, National Institutes of Health

Bayesian Joint Modelling of Response Times with Dynamic Latent Ability in Educational Testing

--Xiaojing Wang, University of Connecticut

Testing Congruence in Phylogenomic Data using Bayes Factors

--Suman Neupane, University of Connecticut

3:10pm **Social Mixer**

4:00pm **Session 3: Statistical Application in Clinical Trial and Studies**

Chair: Qi Qi, M.S., University of Connecticut

An Innovative Design to Combine a PoC Study with Dose Range

--Naitee Ting, Boehringer Ingelheim Pharmaceuticals, Inc

Statistical Methods for Evaluating Diagnostic Devices

--Changhong Song, Food & Drug Administration

Use of the VG (Virtual Twins Combined with GUIDE) Method in Development of Precision Medicines

--Wangang Xie, AbbVie Inc.

A Multi-Stage Stochastic Model in the Analysis of Longitudinal Dementia Data

--Qi Qi, University of Connecticut

5:20pm **Closing Remarks**

Depart for Banquet Dinner

--Han Restaurant, 310 Prospect Ave, Hartford, CT 06106

Abstracts

A Conditional Density Estimation Partition Model Using Logistic Gaussian Processes

Bani Mallick, Texas A&M University

Conditional density estimation (density regression) estimates the distribution of a response variable y conditional on covariates x. Utilizing a partition model framework, a conditional density estimation method is proposed using logistic Gaussian processes. The partition is created using a Voronoi tessellation and is learned from the data using a reversible jump Markov chain Monte Carlo algorithm. The Markov chain Monte Carlo algorithm is made possible through a Laplace approximation on the latent variables of the logistic Gaussian process model. This approximation marginalizes the parameters in each partition element, allowing an efficient search of the posterior distribution of the tessellation. The method has desirable consistency properties. In simulation and applications, the model successfully estimates the partition structure and conditional distribution of y.

Bayesian Modeling of Sparse High-Dimensional Data using Divergence Measures

Dipak Dey*, University of Connecticut
Gyuhyeong Goh, Kansas State University

In sparse and high-dimensional data analysis, a challenging task is a valid approximation of Lo-norm which has played a key role in recent Big data analytics. However, there is not much study on the Lo-norm approximation in Bayesian literatures. In this presentation, we introduce a new prior, called Gaussian and Diffused-gamma (GD) prior, which leads to a new Lo-norm approximation in a Bayesian framework. To develop a general likelihood function, we utilize a general class of divergence measures called Bregman divergence. The generality of Bregman divergence enables us to handle various types of data such as count, binary, continuous, etc. In addition, our Bayesian approach provides many computational advantages. To demonstrate the validity and reliability, we conduct simulation studies and real data analysis.

A Correlated Bayesian Rank Likelihood Approach to Multiple ROC Curves for Endometriosis

Zhen Chen, National Institutes of Health

In analysis of diagnostic data with multiple tests, it is often the case that these tests are correlated. Modeling the correlation explicitly not only produces valid inference results, it also enables borrowing of information. Motivated by the Physician Reliability Study (PRS) that investigated diagnostic performance of physicians in diagnosing endometriosis, we construct a correlated modeling framework to estimate ROC curves and the associated area under the curves. The small sample sizes in later sessions of the PRS makes this correlated approach very appealing, as it enables information borrowing between physician groups and sessions. Given that the test score data appear to be non-normal even after logarithm transformation, we use the ranks of the data to conduct likelihood estimation and inference. We use the deviance information criterion to select competing models and conduct simulation studies to assess model performances. In application to the PRS dataset, we found that the physicians are not significantly different in their diagnostic performance between groups; however, they are different between the sessions. This suggests that clinical information may play more important role in physicians' diagnostic performance than whether the physicians are international experts or residents. Our empirical evidence also demonstrates that when using both woman- and physician-specific random effects, the model parameter estimates are much smoother.

A Mixture Model and an Empirical Bayes Approach to Detect Edges in Sparse Co-expression Graphs

Haim Bar, University of Connecticut

In the early days of microarray data, the medical and statistical communities focused on gene-level data, and particularly on finding differentially expressed genes. This usually involved making a simplifying assumption that genes are independent, which made likelihood derivations feasible and allowed for relatively simple implementations. However, this is not a realistic assumption, and in recent years the scope has expanded, and has come to include pathway and ‘gene set’ analysis in an attempt to understand the relationships between genes. In this talk we develop a method to recover a gene network’s structure from co-expression data, which we measure in terms of normalized Pearson’s correlation coefficients between gene pairs. We treat these co-expression measurements as weights in the complete graph in which nodes correspond to genes. We assume that the network is sparse and that only a small fraction of the putative edges are included (‘non-null’ edges). To decide which edges exist in the gene network, we fit three-component mixture model such that the observed weights of ‘null edges’ follow a normal distribution with mean 0, and the non-null edges follow a mixture of two log-normal distributions, one for positively- and one for negatively-correlated pairs. We show that this so-called L2N mixture model outperforms other methods in terms of power to detect edges. We also show that using the L2N model allows for the control of the false discovery rate. Importantly, the method makes no assumptions about the true network structure.

Statistical Learning for Biomedical Data Integration and Analysis

Yuping Zhang, University of Connecticut

Nowadays, unprecedented amounts and types of data are growing at a super-exponential rate in many fields such as biomedicine. The size, complexity and rate of availability of these modern massive data sets bring challenges and opportunities for the development of statistical methods. In this talk, I will present our recent research efforts on statistical learning method development for data integration and analysis motivated from biomedical research.

On Spatial Disease Mapping

Sudipto Banerjee, University of California, Los Angeles

In the fields of medicine and public health, a common application of a really-referenced (or regionally aggregated) spatial models is the study of geographical patterns of disease. Disease maps are used to highlight geographic areas with high and low prevalence, incidence, or mortality rates of a specific disease, and the variability of such rates over a spatial domain. They can also be used to detect "hot-spots" or spatial clusters which may arise due to common environmental, demographic, or cultural effects shared by neighboring regions. When we have several measurements recorded at each spatial location (for example, information on two or more diseases from the same population groups or regions), we need to consider multivariate disease mapping in order to handle the dependence among the different diseases as well as the spatial dependence between sites. This presents considerable challenges regarding the construction of valid probability models on the maps while maintaining flexibility and richness in the models. This talk will focus upon different strategies for constructing rich classes of models for disease mapping, including dynamic or a real-temporal models. These general classes subsume several special cases or sub-models that may be representing different hypothesis. Applications from different disease mapping contexts will be presented.

Finite Population Unequal Probability Bayesian Bootstraps and Multiple Imputation

Mike Cohen, American Institutes for Research

Efron’s bootstrap, Rubin’s Bayesian bootstrap, the finite population bootstrap of Gross, the finite population Bayesian bootstrap of Lo, and extensions to unequal probability sample designs are discussed. The connections between the Bayesian versions of the bootstrap and multiple imputation are described.

A Bayesian Regularized Mediation Analysis with Multiple Exposures

Yu-Bo Wang, National Institutes of Health

Mediation analysis assesses the effect of study exposures on an outcome both through and around specific mediators. While mediation analysis involving multiple mediators has been addressed in the recent literature, the case of multiple exposures has received little attention. With the presence of multiple exposures, we consider regularizations that allow simultaneous variable selection and effect estimation, while stabilizing model fit and accounting for model uncertainty. In the framework of linear structural-equation model, we analytically show that a two-stage approach that regularizes on regression coefficients does not guarantee a unimodal posterior distribution, and that a product-of-coefficient approach which regularizes on direct and indirect effects tends to penalize excessively. We propose a regularized difference-of-coefficient approach that avoids past limitations. Using connections between regularizations and Bayesian hierarchical models with Laplace priors, we develop an efficient Markov chain Monte Carlo (MCMC) algorithm for posterior estimation and inference. Through simulations, we show that the proposed approach has better empirical performances compared to some alternatives. The methodology is illustrated using data from two reproductive epidemiological studies.

Bayesian Joint Modelling of Response Times with Dynamic Latent Ability in Educational Testing

Xiaojing Wang, University of Connecticut

In educational testing, inferences of ability have been mainly based on item responses while the time taken to complete an item is often ignored. To better infer the ability, a new class of state space models, which conjointly model response time with time series of dichotomous responses, is developed. Simulations for the proposed models demonstrate the biases of ability estimation are reduced as well as the precisions of ability estimation are improved. An empirical study is conducted using EdSphere datasets, where the two competing relationships (i.e., monotone and inverted U-shape) are investigated for modeling response times with the distance between ability and difficulty. The results of model comparison support that the inverted U-shape relationship better captures the behaviors and psychology of examinees in exams.

Testing Congruence in Phylogenomic Data using Bayes Factors

Suman Neupane*, University of Connecticut
Karolina Fučíková, Assumption College
Louise A. Lewis, University of Connecticut
Lynn Kuo, University of Connecticut
Ming-Hui Chen, University of Connecticut
Paul Lewis, University of Connecticut

With the rapid reduction in sequencing costs of high-throughput genomic data, it has become commonplace to use hundreds of genes/sites to infer phylogeny of any study system. While sampling large number of genes has given us a tremendous opportunity to uncover previously unknown relationships and improve phylogenetic resolution, it also presents us with new challenges when the phylogenetic signal is confused by differences in the evolutionary histories of sampled genes. Given the addition of accurate marginal likelihood estimation methods into popular Bayesian software programs, it is natural to consider using the Bayes Factor (BF) to compare different partition models in which genes within any given partition subset share both tree topology and edge lengths. We explore using marginal likelihood to assess data subset combinability when data subsets have varying levels of phylogenetic discordance due to deep coalescence events among genes (simulated within a species tree), and compare the results with our recently-described phylogenetic informational dissonance index (D) estimated for each data set. BF effectively detects phylogenetic incongruence, and provides a way to assess the statistical significance of D values. We discuss methods for calibrating BFs, and use calibrated BFs to assess data combinability using an empirical data set comprising 56 plastid genes from green algae order Volvocales.

An Innovative Design to Combine a PoC Study with Dose Range

Naitee Ting, Boehringer Ingelheim Pharmaceuticals, Inc.

In Phase II clinical development of a new drug, the two most important deliverables are proof of concept (PoC), and dose ranging. Traditionally a PoC study is designed as the first Phase II clinical trial. In this PoC, there are two treatment groups – a high dose of the study medication, against the placebo control. After the concept is proven, the next Phase II study is a dose ranging design with many test doses. This manuscript proposes a two-stage design with the first stage attempting to generate an early signal of efficacy. If successful, the second stage will adopt a “Go Fast” plan to expand the current study and add lower study doses of the test drug to explore the efficacy dose range. Otherwise, a “Go Slow” strategy is triggered, and the study will stop at a reduced sample size with high dose and placebo only.

Statistical Methods for Evaluating Diagnostic Devices

Changhong Song, Food and Drug Administration

Diagnostic devices are devices that provide results that are used alone or with other information to help assess a subject’s health condition of interest. Study designs and statistical analysis that evaluate diagnostic devices can be quite different compared to other type of medical products. This presentation discusses study designs and statistical analysis methods issues for evaluating diagnostic devices.

Use of the VG (Virtual Twins Combined with GUIDE) Method in the Development of Precision Medicines

Jia Jia, AbbVie Inc.
Qi Tang, Sanofi US, Inc.
Wangang Xie*, AbbVie Inc.
Richard Rode, AbbVie Inc.

A lack of understanding of human biology creates a hurdle for the development of precision medicines. To overcome this hurdle we need to better understand the potential synergy between a given treatment (vs. placebo or active control) and various demographic or genetic factors, disease history and severity, etc., with the goal of identifying those patients at increased chance of exhibiting meaningful treatment benefit. For this reason we proposed the VG method, which combines the idea of individual treatment effect (ITE) from the Virtual Twins method (Foster et al 2013, Stat Med) and the unbiased variable selection and cutoff value determination algorithm from the GUIDE method (Loh 2015, Stat Med). Simulation results showed the VG method to have less variable selection bias than Virtual Twins and higher statistical power than GUIDE in the presence of prognostic variables with strong treatment effects. The type I error rate and predictive performance of Virtual Twins, GUIDE and VG were also compared through the use of simulation studies and a randomized clinical trial for Alzheimer’s disease.

Acknowledgement: This presentation was sponsored by AbbVie. AbbVie contributed to the design, research, and interpretation of data, writing, reviewing, and approving the publication.

A Multi-Stage Stochastic Model in the Analysis of Longitudinal Dementia Data

*Qi Qi, University of Connecticut
Lynn Kuo, University of Connecticut

Multi-stage transition model is playing an important role in dementia studies. Since death is a significant source of missing data in longitudinal epidemiological studies on elderly individuals, we consider four stages: normality, memory-impaired intermediate, dementia and death without dementia. To analyze longitudinal data, we develop the likelihood function based on a first order Markov chain model consisting of transitional probabilities between stages. Different from the typical illness-death model, we construct a reversible transition model between normality and memory-impaired intermediate. We use Kolmogorov’s backward equations to derive the probability of transition and ordinal logistic regression to investigate what covariates have significant influence on the transition.

Attendees

Abidemi Adeniji, Abidemi.Adeniji@gmail.com, EMD Serono

Prince Allotey, prince.allotey@uconn.edu, University of Connecticut

Sudipto Banerjee, sudipto@ucla.edu, University of California, Los Angeles

Haim Bar, haim.bar@uconn.edu, University of Connecticut

Jorge Bazan, jlbazan@uconn.edu, University of Connecticut

Ming-Hui Chen, Ming-hui.chen@uconn.edu, University of Connecticut

Renjie Chen, renjie.chen@uconn.edu, University of Connecticut

Zhen Chen, zhen.chen@nih.gov, National Institutes of Health

Kun Chen, kun.chen@uconn.edu, University of Connecticut

Robert Chiang, robert.chiang@live.com, Birch Creek Capital LLC

Michael P. Cohen, mpcohen@juno.com, American Institutes for Research

Lijiang Geng, lijiang.geng@uconn.edu, University of Connecticut

Yuwen Gu, yuwen.gu@uconn.edu, University of Connecticut

Yeongjin Gwon, yeongjin.gwon@uconn.edu, University of Connecticut

Aritra Halder, aritrah.halder@uconn.edu, University of Connecticut

Charles Hall, charles.hall@einstein.yu.edu, Albert Einstein College of Medicine

Guanyu Hu, guanyu.hu@uconn.edu, University of Connecticut

Jun Hu, jun.hu@uconn.edu, University of Connecticut

Siddesh Kulkarni, siddhesh.kulkarni@uconn.edu, University of Connecticut

Lynn Kuo, Lynn.kuo@uconn.edu, University of Connecticut

Yan Li, yan.4.li@uconn.edu, University of Connecticut

Yujia Li, yujia.li@uconn.edu, University of Connecticut

Hongfei Li, hongfei.2.li@uconn.edu, University of Connecticut

Henry Linder, mhlinder@gmail.com, University of Connecticut

Xiaokang Liu, xiaokang.liu@uconn.edu, University of Connecticut

Yang Liu, yang.5.liu@uconn.edu, University of Connecticut

Yueqi Liu, yueqi.liu@uconn.edu, University of Connecticut

Zhihua Ma, zhihua.ma@uconn.edu, University of Connecticut

Bani Mallick, bmallick@stat.tamu.edu, Texas A&M University

Shariq Mohammed, shariq.mohammed@uconn.edu, University of Connecticut

Jinjian Mu, Jinjian.mu@uconn.edu, University of Connecticut

Suman Neupane, suman.neupane@uconn.edu, University of Connecticut

Zhengqing Ouyang, zhengqing.ouyang@jax.org, The Jackson Laboratory

Qi Qi, qi.2.qi@uconn.edu, University of Connecticut

Vishal Kumar Sarsani, vsarsani@umass.edu, University of Massachusetts

Md. Tuhin Sheikh, mdtuhin.sheikh@uconn.edu, University of Connecticut

Daoyuan Shi, daoyuan.shi@uconn.edu, Vertex Pharmaceuticals

Wei Shi, wei.shi@uconn.edu, University of Connecticut

Changhong Song, changhongsong@hotmail.com, Food & Drug Administration

Naitee Ting, naitee.ting@boehringer-ingelheim.com, Boehringer Ingelheim Pharmaceuticals, Inc.

Haiying Wang, Haiying.wang@uconn.edu, University of Connecticut

Xiaojing Wang, xiaojing.wang@uconn.edu, University of Connecticut

Wenjie Wang, wenjie.2.wang@uconn.edu, University of Connecticut

Zhe Wang, zhe.4.wang@uconn.edu, University of Connecticut

Hong Wang, hongmuw@hotmail.com, Boehringer Ingelheim Pharmaceuticals, Inc.

Yu-Bo Wang, yu-bo.wang@nih.gov, National Institutes of Health

Rui Wu, rui_2.wu@boehringer-ingelheim.com, Boehringer Ingelheim Pharmaceuticals, Inc.

Wangang Xie, wangang.xie@gmail.com, AbbVie Inc.

Yishu Xue, yishu.xue@uconn.edu, University of Connecticut

Jun Yan, jun.yan@uconn.edu, University of Connecticut

Tae Young Yang, tyang@mju.ac.kr, Myongji University

Jun Ying, yingj@uc.edu, University of Cincinnati

Fang Yu, fangyu@unmc.edu, University of Nebraska Medical Center

Wenlin Yuan, wenlin.yuan2305@gmail.com

Yuping Zhang, yuping.zhang@uconn.edu, University of Connecticut

Chen Zhang, chen.zhang@uconn.edu, University of Connecticut

Planning Committee

- Kun Chen (Co-chair), University of Connecticut
- Zhen Chen (Co-chair), National Institutes of Health
- Jinjian Mu, University of Connecticut
- Qi Qi, University of Connecticut
- Wei Shi, University of Connecticut
- Jun Ying, University of Cincinnati
- Fang Yu, University of Nebraska
- Yu-Bo Wang, National Institues of Health
- Rui Wu, Boehringer Ingelheim Pharmaceuticals, Inc.