2017 Seminars

12/06/2017

Deep Learning and Its Application in Predicting Enhancer in Human Genome

Xiaohui Niu, PhD

Abstract: Enhancer sequences contain short DNA motifs that act as binding sites for sequence-specific transcription factors. The crucial roles of enhancers in generating cell-type and state-specific transcriptional programs, further understanding of the process of enhancer transcription and its contribution to the overall functionality of enhancers will offer crucial insights into gene regulation, cell identity control, development and disease. However, this is a challenging problem because the very long distance of enhancer with its target gene increases the searching difficulty. Moreover, unlike promoter located in the upstream of its target gene, enhancer can act its regulatory role bidirectionally, which makes the problem more challenging. To address this need, we propose a novel hybrid convolutional and Gated Recurrent Unit (GRU) recurrent neural network framework for predicting enhancer de novo from sequence. In the model, the convolution layer captures regulatory motifs, while the recurrent layer captures long-term dependencies between the motifs in order to learn a regulatory 'grammar' to improve predictions. This model improves considerably in several benchmark datasets.

Bio: Dr. Xiaohui Niu is a visiting scholar in Department of Family Medicine and Public Health at UC San Diego and an associate professor in Huazhong Agricultural University in China. His research interests are machine learning methods and their applications in Bioinformatics, especially proteomics, including protein (gene) function prediction, protein binding site prediction, methods to construct phylogenetic tree, protein structure prediction etc.

11/01/2017

Ethical Issues in Statistical Practice - Case Examples

Larry Shen, PhD

Abstract: In this presentation, I discuss statistical integrity and some ethical issues in our practice in the bio-pharmaceutical industry. Some common ethical issues include data integrity, validity of statistical testing and conclusions, presentation of data, and post-hoc analyses. I am going to highlight a few ethical guidelines from the American Statistical Association and use a few case examples to illustrate ethical dilemmas that we often face in our daily work.

Bio: Dr. Larry Shen has a highly accomplished career in leading clinical organizations to support drug development and clinical research programs. He has directly worked on over 20 investigational new drug projects and played leading roles in regulatory submissions that had led to 6 drug approvals in both the US and European Union. He has authored or co-authored many articles on statistical methodology and their applications to drug development. His work on dose titration received the Thomas W Teal award at the 2007 Drug Information Association annual meeting. Dr. Shen also served as past President of the San Diego Chapter of the American Statistical Association (ASA). In 2014, Dr. Shen was elected as fellow of the ASA for his leadership in applying statistics to drug development and for his contributions to the statistics profession. Prior to co-founding Pharmapace, Dr. Shen was Vice President at Amylin Pharmaceuticals in charge of their clinical development organizations including Statistics, Programming, Data Management, PK/PD modeling, and Medical Writing. He had worked at Amylin since 1997 and had implemented rigorous procedures for data processing, analyses, and reporting to ensure data integrity and statistical excellence. Under his leadership, his department had played a critical role in the development and approval of four innovative medicines. Dr.

Shen obtained his Ph.D. in Statistics from the University of California at Berkeley and both BS and MS degrees in mathematics/statistics from Beijing University in China.

10/04/2017

Multi-Block Models for Psychiatric and Brain Imaging Data

Wesley Thompson, PhD

Abstract: Modern large-scale observational psychiatric studies collect data in a plethora of modalities, including questionnaires, structured clinical interviews, life histories, and many biological variables, including, e.g., structural and functional brain imaging, genetics, inflammatory measures. An important goal of such studies is to obtain a biological foundation for psychiatric diagnoses that are predictive of outcomes and/or response to specific treatments. However, a major difficulty in analyzing data from these studies is reducing dimensionality via revealing latent structures that inform about relationships across modalities, while simultaneously accounting for "batch" effects and method variance within modalities of measurement. Here, we present a Bayesian multi-level model that uncovers both shared and idiosyncratic factors within blocks (data modalities). We demonstrate that this methodology is effective in uncovering latent structure and predicting clinical outcomes in the T-1000 data, a large-scale of psychiatric disorders collecting data in scores of domains, including structural and functional imaging.

Bio:Dr. Wesley Thompson earned his Ph.D. in Statistics from Rutgers University in 2003, with a focus on statistical methods for longitudinal data analysis. He was appointed Assistant Professor of Statistics and Psychiatry at the University of Pittsburgh in 2005, where he received a five year NIH K25 Career Development Award to develop novel methods for studying co-variation in brain function and depression. Dr. Thompson joined UCSD in 2008, and is currently an Associate Professor of Family Medicine and Public Health within the Division of Biostatistics and Bioinformatics. His current work involves Bayesian semi-parametric and mixture models with applications to (i) improving effect size estimation, replication, and prediction in genome-wide association studies, (ii) predicting onset of illness from multivariate biomarker trajectories, (iii) applications of to functional and structural MRI data.

09/06/2017

Are Tumors Predictable? Inherited Genetic Variation Constrains Tumor Evolution

Hannah Carter, PhD

Abstract: Recent studies have characterized the extensive somatic alterations that arise during cancer and various studies have probed rare inherited mutations that lead to early onset cancer syndromes. However, little is understood about the role of genetic background in ‘sporadic’ adulthood cancers. It is possible that the somatic evolution of a tumor may be significantly affected by inherited polymorphisms carried in the germline. To investigate this, we analyzed genomic data for thousands of tumors from The Cancer Genome Atlas to reveal and systematically validate hundreds of genetic interactions between germline polymorphisms and major somatic events, including tumor formation in specific tissues and alteration of specific cancer genes. Among germline–somatic interactions, we found germline variants in RBFOX1 that increased incidence of SF3B1 somatic mutation by 8-fold via functional alterations in RNA splicing. Similarly, 19p13.3 variants were associated with a 4-fold increased likelihood of somatic mutations in PTEN. In support of this association, we found that PTEN knockdown sensitizes the MTOR pathway to high expression of the 19p13.3 gene GNA11. Finally, we observed that stratifying patients by germline polymorphisms exposed distinct somatic mutation landscapes, implicating new cancer genes. Our findings suggest that individual genomic data can help to forecast the trajectory of tumor evolution, including where and how cancer develops, opening avenues for prevention research.

Bio:Dr. Hannah Carter is an Assistant Professor in the UCSD Department of Medicine. She received her M.Eng in Electrical Engineering at the University of Louisville and her PhD in Biomedical Engineering from Johns Hopkins University. The Carter Lab uses bioinformatics and computational biology to study the role of inherited and acquired genetic variation in cancer. The goals of her research are to advance precision cancer medicine by developing approaches to discriminate drivers from passengers, predict cancer cell-specific therapeutic vulnerabilities and identify germline variation that contributes to the emergence or progression of tumors. Dr. Carter is a Siebel Scholar and a recipient of a 2013 NIH Director's Early Independence Award.

04/12/2017

"Efron's Rules" for Inference after Imputation and Model Selection

(Joint work with Lin Liu and Loki Natarajan)

Karen Messer, PhD

Abstract: We address the practical problem of model selection in the presence of imputation for missing data. Our focus is on valid inference, in particular on confidence intervals that incorporate both the imputation mechanism and the model selection mechanism. We investigate commonly used resampling-based approaches - multiple imputation and the bootstrap - and incorporate Efron's 2014 computationally efficient variance estimate for bootstrap-smoothed estimates. We compare the resulting `Efron's rules' estimator to a 'Rubin's rules' estimator based on multiple imputation. These turn out to be versions of frequentist model averaged estimators, and are compared to an un-averaged selection estimator using the framework of Claeskens and Hjort. Simulation and real data examples are drawn from the related literature. Practical recommendations are given, including circumstances where the new Efron's rules estimator is seen to work well.

Bio: Dr. Karen Messer is Professor and Chief of Division of Biostatistics and Bioinformatics in Department of Family Medicine and Public Health at UCSD. She went to Clairemont high school here in San Diego, and she was an undergraduate math major at Harvard university. Dr. Messer got her PhD in mathematical statistics at UCSD under Dr. Murray Rosenblatt. Before joining UCSD in 2006, Dr. Messer was assistant professor of mathematics at UCLA, then associate professor and professor of mathematics at California State University Fullerton.

03/17/2017

Controlling Epidemics: Challenges and Opportunities for Quantitative Scientists

Victor De Gruttola, ScD

Abstract: Recent developments in biomedical science, such as those in molecular epidemiology and surveillance, vaccinology, and antimicrobial treatment, can greatly aid in devising effective responses to epidemic and endemic diseases. To take maximal advantage of such successes requires advances in quantitative science that combine across different disciplinary domains. For example in investigation and scale-up of HIV prevention interventions, challenges arise from the complex dependencies that characterize data from clinical studies and that reflecting the spread of HIV along sexual contact networks. Both randomized and observational studies often collect data on HIV incidence in different subpopulations, risk behavior, and viral genetic sequences. New methods are required to make maximal use of this very useful, but incomplete information to estimate quantities that will be useful in guiding scale-up of successful interventions. These include not only effects of randomized interventions—for trials randomized at both individual and cluster level--but also expected effects under implementation policies likely to be used practice. We propose methods that make use of baseline data to improve estimation of intervention effects and of their modification by factors measured at individual and network levels. We show their advantages in settings with complete or missing data—for design of both randomized and observational studies with or without missing data. Cluster randomized trials are also useful for controlling outbreaks. We propose and demonstrate properties a novel design for settings like the Ebola epidemic, where a proof-of-principle vaccine trials provided evidence of efficacy, but where questions remain about the effectiveness of different possible modes of implementation. Our goal for these studies is not only to generate information about intervention effects but also to provide public health benefit. To do so, we leverage information about contact networks – in particular the degree of connection across randomized units obtained at study baseline – and develop a novel class of connectivity-informed cluster trial designs. We investigate the performance of these designs in terms of epidemic control outcomes (time to end of epidemic and cumulative incidence) and power to detect intervention effect, by simulating vaccination trials during an SEIR-type epidemic outbreak using a network-structured agent-based model.

Bio: Dr. Victor De Gruttola Professor of Biostatistics Department of Biostatistics Harvard T.H. Chan School of Public Health Dr. Victor De Gruttola has spent the past 30 years working with junior colleagues and in collaborating with clinical and laboratory investigators to develop and apply methods for advancing the HIV prevention and treatment research agendas. He also has managed large projects devoted to improving the public health response to the AIDS epidemic, both within the US and internationally. The aspects of the HIV epidemic on which he has worked include transmission and natural history of infection with the Human Immunodeficiency Virus (HIV), as well as investigation of antiretroviral treatments, including the development and consequences of resistance to them. The broad goals of his research have included developing treatment strategies that provide durable virologic suppression while preserving treatment options after failure, and evaluating the community-level impact of packages of prevention interventions, including antiviral treatment itself. He served as the Director of the Statistics and Data Analysis Center of the Adult Project of the AIDS Clinical Trials Group during the period in which highly active antiretroviral treatment was developed, and was instrumental in designing and analyzing studies of the best means of providing such therapy. He has also served as the Co-PI (with Max Essex) for a cluster-randomized trial of an HIV combination prevention program in Botswana. His methods research activity is focused HIV prevention research, especially with regard to the development of methods for analyses of sexual contact networks, for viral genetic linkage analyses in the presence of missing data, and for improving validity and efficiency of analyses of HIV prevention trials.

03/15/2017

Modeling Time-Varying Trends in ERP Data with Applications to an Implicit Learning Paradigm in Autism

Kyle Hasenstab, PhD

Abstract: Event-related potential (ERP) studies are a set of experimental frameworks that use electroencephalography (EEG) to study the electrical potential outputted by a subject's brain when presented with an implicit task in the form of stimuli. Data consist of a temporally recorded functional ERP curve repeatedly observed over a sequence of stimuli and across a set of electrodes placed on the scalp, producing a complex data structure consisting of a functional (ERP curve), longitudinal (stimulus repetition), and spatial (electrode) dimension. In typical ERP studies, the dimension of data is reduced into a single measure for each subject by cross-sectionally averaging ERP across longitudinal and spatial repetitions in order to increase the signal-to-noise ratio of the ERP function. Features are then extracted from the averaged ERP and analyzed using simple statistical methods, ignoring additional information that may be found in the collapsed dimensions. In this talk, I discuss methodology for preserving and analyzing the lost dimensions of ERP data. In particular, I focus on multidimensional functional principal components analysis (MD-FPCA), a two-step procedure used to summarize important characteristics across all three dimensions of the ERP data structure into an interpretable, low-dimensional form. MD-FPCA is applied to a study on neural correlates of visual implicit learning in young children with autism spectrum disorder (ASD). Application of the proposed methods reveal meaningful trends and substructures in the implicit learning processes of ASD children when compared to typically developing controls. Results indicate proposed methodology effectively preserves important information contained within the multiple dimensions of ERP data.

Bio:Dr. Kyle Hasenstab recently earned his PhD in Statistics from the University of California, Los Angeles where he researched methods for analyzing data from EEG experiments to study implicit learning in children with autism spectrum disorder. He has worked as a postdoctoral fellow at the Centers for Disease Control and Prevention in their Chronic Viral Diseases Branch --and is currently working as a statistician for AT&T.

03/01/2017

Inference of High-dimensional, Non-sparse and Strongly Dependent Gaussian Observations

David Azriel, PhD

Abstract: Motivated by a data set obtained from brain imaging, we study inference of high-dimensional observations without assuming a sparse parameter space. Our approach starts from computing the z-scores at each cortical voxel. The result is a large strongly-dependent vector of observations, assumed to be Gaussian. We study two issues: first, we investigate the empirical distribution of this vector and its possible departure from a standard normal distribution. Second, we study inference of linear-projections of this vector. Our analysis shows that the global null hypothesis (when there is no dependence between the response and the measurements) is not likely to be true. Furthermore, we find that the effect is widespread (non-sparse) but not large enough to be significant anywhere.

Bio:Dr. David Azriel is a senior lecturer in statistics at the Technion - Israel Institute of Technology since 2015. Previously, he was a postdoc with Larry Brown at Wharton at the University of Pennsylvania. He completed his PhD thesis at the Hebrew University in Jerusalem in 2012. His research interests are in high dimensional data, model selection and optimal clinical trial design.

02/15/2017

Fast Estimation of Regression Parameters in a Broken-Stick Model for Longitudinal Data

Bin Nan, PhD

Abstract: Estimation of change-point locations in the broken-stick model has significant applications in modeling important biological phenomena. In this talk, Dr. Nan will present a computationally economical likelihood-based approach for estimating change-point(s) efficiently in both cross-sectional and longitudinal settings. The method, based on local smoothing in a shrinking neighborhood of each change-point, is shown via simulations to be computationally more viable than existing methods that rely on search procedures, with dramatic gains in the multiple change-point case. The proposed estimates are shown to have root-n consistency and asymptotic normality--in particular, they are asymptotically efficient in the cross-sectional setting--allowing us to provide meaningful statistical inference. As the primary and motivating longitudinal application, a two change-point broken-stick model appears to be a good fit to the Michigan Bone Health and Metabolism Study cohort data to describe patterns of change in log estradiol levels, before and after the final menstrual period. A plant growth dataset in the cross-sectional setting is also illustrated. This is a joint work with Rito Das, Mouli Banerjee, and Huiyong Zheng.

Bio: Dr. Bin Nan is Professor of Biostatistics and Statistics at the University of Michigan. He received his Ph.D. in Biostatistics from the University of Washington in 2001 and joined the faculty at the University of Michigan in the same year. Dr. Nan's research interests are in various areas of statistics and biostatistics including semiparametric inference, failure time and survival analysis, longitudinal data, missing data and two-phase sampling designs, and high-dimensional data analysis. He is collaborating in many studies in areas of epidemiology, bioinformatics, and brain imaging, particularly in cancer, HIV, women's health, and neurodegenerative diseases. He is Fellow of the American Statistical Association and Fellow of the Institute of Mathematical Statistics.

02/09/2017

Recurrent Event Data Analysis with Intermittently Observed Time-varying Covariates

Chiung-Yu Huang, PhD

Abstract: Although recurrent event data analysis is a rapidly evolving area of research, rigorous studies on modeling and estimation of the effects of time-varying covariates on the risk of recurrent events have been lacking. Existing methods for analyzing recurrent event data usually require that the covariate processes are observed throughout the entire follow-up period. However, covariates are often observed periodically rather than continuously. We propose a novel semiparametric estimator for the regression parameters in the popular proportional rate model. The proposed estimator is based on an estimated score function where we kernel smooth the mean covariate process. We show that the proposed semiparametric estimator is asymptotically unbiased, normally distributed and derive the asymptotic variance. Simulation studies are conducted to compare the performance of the proposed estimator and the simple methods carrying forward the last covariates. The different methods are applied to an observational study designed to assess the effect of Group A streptococcus (GAS) on pharyngitis among school children in India.

Bio: Dr. Huang is Associate Professor of Oncology and Biostatistics at the Johns Hopkins University. Her main area of research is in general biostatistics methodology and its application to the biomedical sciences. She has extensive experience in the statistical analysis of survival outcomes, recurrent events, competing risks, longitudinal measurements, missing data, biased sampling, and design and monitoring of clinical trials.

01/24/2017

Optimally Combining Outcomes to Improve Prediction

David Benkeser, MPH, PhD

Abstract: In many studies, multiple instruments are used to measure different facets of an unmeasured outcome of interest. For example, in studies of childhood development, children are administered tests in several areas and researchers combine these test scores into a univariate measure of neurocognitive development. Researchers are interested in predicting this development score based on household and environment characteristics early in life in order to identify children at high risk for neurocognitive delays. We propose a method for estimating the combined measure that maximizes predictive performance. Our approach allows modern machine learning techniques to be used to predict the combined outcome using potentially high-dimensional covariate information. In spite of the highly adaptive nature of the procedure, we nevertheless obtain valid estimates of the prediction algorithm’s performance for predicting the combined outcome as well as confidence intervals about these estimates. We illustrate the methodology using longitudinal cohort studies of early childhood development.

Bio: David Benkeser, PhD, MPH is a post-doctoral researcher under Mark Van der Laan in the Division of Biostatistics at the University of California, Berkeley where he works on developing methods for machine learning, causal inference, and the integration of the two fields. He obtained his PhD from the Department of Biostatistics at the University of Washington where his research focused on causal inference in complex longitudinal settings with applications in preventive vaccine efficacy trials for infectious diseases**.**