Department of Statistics Weekly Seminar Series

Hosted by Dixon Vimalajeewa

Spring 2024

Date: **Wednesday, April 24.**

Speakers: **Dr. Jinghui Li, Genetics Medicine, University of Chicago**

Title: ***Protein-protein interaction is an important trans-regulatory mechanism of the proteome and complex trait***

Abstract:

Recent large-scale proteome data provides us an unprecedented opportunity to investigate the regulatory mechanisms of proteomes and their roles in complex traits. Genetic studies of the proteome have reported extensive associations, especially *trans*- effects. However, it remains unclear what the *trans-mechanisms* are and their importance in complex traits. By analyzing the plasma proteomic data in the UK Biobank (Sun et al. 2023 *Nature*) and RNA-seq datasets (Battle et al. 2014 *Genome Res*; The GTEx Consortium 2020 *Science*), we find that the proportion of heritability explained by *trans*-SNPs is higher in protein than in gene expression (0.90 vs 0.60, p = 1.4e-37). Genes with *trans*-pQTL have a higher probability of being loss of function intolerant (pLI) than genes with *cis*-pQTL (p = 2.2e-14). In addition, by performing colocalization analyses with 50 GWAS traits, we find a higher proportion of *trans*-pQTLs colocalizing with trait loci than *cis*-pQTLs (50.1% vs 34.5%, p < 2.2e-16). These results highlight the important role of *trans-regulations* of proteome. We next investigate the mechanism of pQTLs, and find that most *cis*-

pQTLs and *cis*-eQTLs are shared (true positive rate π_1: 0.69 ~ 0.98). However, most *trans*-pQTLs cannot be replicated in gene expression datasets (π_1: 0.11 ~ 0.24), which indicates prevalent post-translational *trans*-regulations. We observe that the nearby genes of *trans*-pQTLs are enriched in protein-protein interaction (PPI) with their targeting proteins (p = 2.1e-9) but are only weakly enriched in transcription factor activities. We then apply *trans*-PCO (Wang et al. 2022 *bioRxiv*), a multivariate method to map *trans*-QTLs of gene/protein clusters, to identify *trans*-pQTLs of 8,878 PPI clusters. We found a total of 2,346 *trans*-pQTLs (Bonferroni corrected p < 0.05) and 1,326 of them are novel loci not reported by the UK Biobank study. We next conduct a colocalization analysis of *trans*-pQTLs of PPIs with 50 GWAS traits and find a high proportion of *trans*-pQTLs of PPIs colocalizing with trait loci (57.1%). More importantly, the colocalized *trans*-pQTLs of PPIs can help us interpret the mechanisms of GWAS loci. For example, we identified a colocalization between a *trans*-pQTL and GWAS locus of systemic lupus erythematosus near the gene *ITGAM* (PP4 = 0.91). *ITGAM* is a relevant gene to systemic lupus erythematosus (Teruel et al. 2016 *J Autoimmun*), and the protein product of *ITGAM* is a member of the *trans*-target protein complex (CD177, ITGAM, ITGB2), which plays a role in neutrophil signaling and activation. In conclusion, our study shows the crucial role of *trans*-regulation in the proteome and PPI as an important *trans*-regulatory mechanism.

Date: Wednesday, April 17.

Speakers: Dr. Jae Kwang Kim, Department of Biostatistics, Iowa State University

Title: ***Debiased calibration estimation using generalized entropy under selection bias.***

Abstract:

Incorporating the auxiliary information into the survey estimation is a fundamental problem in survey sampling. Calibration weighting is a popular tool for incorporating the auxiliary information. The calibration weighting method of Deville and Sarndal (1992) uses a distance measure between the design weights and the final weights to solve the optimization problem with calibration constraints.

In this paper, we propose a new framework using generalized entropy as the objective function for optimization. Design weights are used in the constraints, rather than in the objective function, to achieve design consistency. The new calibration framework is attractive as it is general and can produce more efficient calibration weights than the classical calibration weights. Furthermore, we identify the optimal choice of the generalized entropy function that achieves the minimum variance among the different choices of the generalized entropy function under the same constraints. Asymptotic properties, such as design consistency and asymptotic normality, are presented rigorously. An extension of the proposed method to doubly robust propensity score estimation will also be presented

Date: Wednesday, April 10.

Speakers: Dr. Yunju Im, Department of Biostatistics, University of Nebraska Medical Center

Title: ***Bayesian integrative analysis for cancer pathological imaging data with sample selection.***

Abstract:

For identifying the associations between high-dimensional omics variables and a disease outcome, extensive methodological developments have been conducted. Despite significant successes, the existing studies are still often unsatisfactory because of a “lack of information”. In cancer research, a source of information that is broadly available and highly cost-effective comes from pathological images. In this study, we develop an effective Bayesian approach that can identify omics variables that are associated with both the cancer outcome of interest and (manually curated, low-dimensional) pathological imaging features. Improved estimation and selection are expected via this joint variable selection. Additionally, considering that the existing cancer omics studies often cannot afford sample selection and hence may be heterogeneous (that is, there may exist a small number of samples that do not satisfy the same disease model), we design the proposed approach to flexibly distinguish samples that may differ from the majority. Extensive numerical studies, including both simulations and the analysis of TCGA data, demonstrate the effectiveness of the proposed approach.

Date: Wednesday, April 03.

Speakers: Dr. Sanjay Chaudhuri, University of Nebraska-Lincoln

Title: Hamiltonian Monte Carlo In Bayesian Empirical Likelihood Computation

Abstract:

We consider Bayesian empirical likelihood estimation and develop an efficient Hamiltonian Monte Carlo method for sampling from the posterior distribution of the parameters of interest. The proposed method uses hitherto unknown properties of the gradient of the underlying log-empirical likelihood function.
It is seen that these properties hold under minimal assumptions on the parameter space, prior density, and the functions used in the estimating equations determining the empirical likelihood. We overcome major challenges posed by complex, non-convex boundaries of the support routinely observed for empirical likelihood which prevents efficient implementation of traditional Markov chain Monte Carlo methods like random walk Metropolis-Hastings, etc. with or without parallel tempering. Our method employs a finite number of estimating equations and observations but produces valid semi-parametric inference for a large class of statistical models including mixed effects models, generalized linear models, hierarchical Bayes models, etc. A simulation study confirms that our proposed method converges quickly and draws samples from the posterior support efficiently. We further illustrate its utility by analyzing a discrete data set in small area estimation.

Date: Wednesday, March 06.

Speakers: *Ms. Ankona Banerjee and Mr. Kenneth Jhune Nobleza, Center for Epidemiology & Population Health Baylor College of Medicine, Houston.*

Title: *Decoding Health Data: A Deep Dive into NHANES and Survey Analysis Techniques*

Abstract:

The topic of this presentation is survey data analysis, with a primary focus on the National Health and Nutrition Examination Survey (NHANES). In this introductory presentation, the speakers will cover the unique aspects of this nationally representative survey data, elaborate the difference between the complex survey designs and simple random sampling, explore the structures of complex survey designs, and provide an overview of the NHANES content. The presentation also includes step-by-step instructions on the basics of conducting an analysis using NHANES. Additionally, the speakers will highlight the importance of data privacy and offer practical tips to avoid errors. This lighthearted and easy-to-understand presentation aims to provide researchers and public health professionals who are interested in using data obtained through complex survey design with valuable hands-on information on how to get started. While a working knowledge of R is beneficial, it is not necessary to make the most of this seminar.

Date: Wednesday, February 21.

Speaker: **Dr. Ran Dai, Department of Biostatistics at the University of Nebraska Medical Center (UNMC).**

Title: **Controlling FDR in selecting group-level simultaneous signals from multiple data sources with application to the National Covid Collaborative Cohort data**

Abstract:

One challenge in exploratory association studies using observational data is that the signals are potentially weak; and the features have complex correlation structures. False discovery rate (FDR) controlling procedures can provide important statistical guarantees for replicability in risk factor identification in exploratory research. In the recently established National COVID Collaborative Cohort (N3C), electronic health record (EHR) data on the same set of candidate features are independently collected in multiple different sites, offering opportunities to identify signals by combining information from different sources. This paper presents a general knockoff-based variable selection algorithm to identify mutual signals from unions of group-level conditional independence tests with exact FDR control guarantees under finite sample settings. This algorithm can work with general regression settings, allowing heterogeneity of both the predictors and the outcomes across multiple data sources. We demonstrate the performance of this method with extensive numerical studies and an application to the N3C data

Date: Wednesday, February 07.

Speaker: Dr. Rakheon Kim, Department of Statistical Sciences, Baylor University, Waco, Texas.

Title: Positive-definite thresholding estimators of covariance matrices with zeros

Abstract:

A positive definite estimator of a covariance matrix with zero entries provides a valid covariance matrix that can be used an an input in almost any area of multivariate statistical analysis. However, most current approaches do not yet guarantee positive definiteness or deal with the asymptotic efficiency of the covariance estimator. Focusing on the classical setting when the number of Gaussian variables is fixed and the sample size increases, we construct a positive definite and asymptotically efficient estimator by the iterative conditional fitting algorithm [Chaudhuri, 2007] when the location of the zero entries is known. If the location of the zero entries is unknown, we further construct a positive definite thresholding estimator by

combining the iterative conditional fitting algorithm with thresholding. We prove our thresholding estimator is asymptotically efficient with probability tending to one. In simulation studies, we show our estimator more closely matches the true covariance and more correctly identifies the non-zero entries than competing estimators. We apply our estimator to a neuroimaging study of Huntington disease to detect non-zero correlations among brain regional volumes. Such correlations are timely for ongoing treatment studies to inform how different brain regions are likely to be affected by these treatments.

Date: Wednesday, January 31.

Speaker: Dr. Blaine Johnson, Department of Agronomy, UNL.

Title: Reflections on a Career: Education and Teaching within the Disciplines of Plant Breeding and Statistics.

Abstract:

A career that has spanned over 35 plus years has included teaching formal courses as well as providing many informal educational and mentoring opportunities in both academia and industry. This seminar is a reflection on approaches used during that career to convey sometime complex and technical information within the disciplines of plant breeding and statistics to audiences having a wide range of diversity in technical knowledge and experience. The seminar beings with some general philosophies on teaching and education, then moves to specific examples of the delivery of sometimes highly technical knowledge and information on plant breeding and statistics to (a) an audience very little technical knowledge, (b) an audience of intermediate technical knowledge, and (c) an audience with a high degree of technical knowledge of either or both disciplines. The examples are real world, originating from seminars/presentations, workshops, and formal courses taught in both academia and industry, but always with a strong industry perspective.

Fall 2023

Date: Wednesday, November 08.

Speaker: Dr. Jesús Arroyo, Texas A&M University, College Station, Texas.

Title: Joint spectral clustering in multilayer network data

Abstract:

Modern network datasets are often composed of multiple layers, such as different views, time-varying observations or independent sample units. These data require flexible and tractable models and methods capable of aggregating information across the networks. To that end, this talk considers the community detection problem under the multilayer degree-corrected stochastic blockmodel. We establish the identifiability of the model and propose a spectral clustering algorithm. Our theoretical results demonstrate that the misclustering error rate improves exponentially with multiple network realizations, even in the presence of significant layer heterogeneity. The methodology is illustrated in simulations and a case study of US airport data. This is joint work with Joshua Agterberg (University of Pennsylvania) and Zachary Lubberts (University of Virginia).

Date: Wednesday, October 18.

Speaker: Dr. Rebecca Killick, Lancaster University, UK.

Title: Modelling nonstationarity through changepoints in environmental processes

Abstract:

It is widely acknowledged that the assumption of stationarity of model parameters over time in environmental processes if flawed. The simplest break from stationarity of model parameters is a piecewise stationary model - where each piece is a standard statistical model, but the parameter estimates across pieces vary. This talk will introduce the audience to changepoint models, the challenges in fitting them, and some substantive applications in the inference for data around tree growth and soil monitoring.

Date: Monday, October 9.

Speaker: *Dr. Aurobindo Ghosh, Lee Kong Chian School of Business, Singapore Management University.*

Title: *Fractile Regression in Finance: Nonparametric Regression with a New Perspective*

Abstract:

Fractile Graphical Analysis was proposed by Prashanta Chandra Mahalanobis (Mahalanobis, 1960) in a series of papers and seminars as a method for comparing two distributions at two different points (of time or space) controlling for the rank of a covariate through fractile groups. Mahalanobis used a heuristic method of approximating the standard error of the dependent variable using fractile graphs from two independently selected "interpenetrating sub-samples." We highlight the potential and revisit this under-utilized technique of fractile graphical analysis with a historical perspective. We evaluate a new non-parametric regression method called Fractile Regression where we condition on the ranks of the covariate, and compare it with existing regression techniques. We apply this method to compare financial distributions like private equity returns and mutual fund inflow distributions after conditioning on size and returns respectively.

Date: Wednesday, October 4.

Speaker: *Dr. Jian Cao, Department of Statistics, University of Huston.*

Title: Linear-Cost Vecchia Approximation of Multivariate Normal Probabilities.

Abstract:

Multivariate normal (MVN) probabilities are needed in the model estimation and posterior inference for several important extensions of the classic Gaussian process (GP). They are analytically intractable and need to be evaluated through Monte-Carlo-based numerical integration. We discover that the dominant complexity for the classic separation-of-variable (SOV) and minimax exponential tilting (MET) methods for the numerical evaluation of MVN probabilities comes from the computation of the conditional mean and variance of the integration variables. With this realization, we propose to restructure the SOV and MET algorithms and use the Vecchia approximation to reduce the per-sample complexity from O(n^2) to O(n), where n is the dimension of the MVN probability. Furthermore, MET significantly improves the estimation accuracy compared with SOV but additionally requires solving a non-linear system of 2n parameters for its proposal density, which has O(n^3) complexity and is frequently more expensive than the Monte Carlo (MC) sampling that is supposed to be the main computation component. We propose a new parameterization for the non-linear system that utilizes the sparse inverse Cholesky factor of the Vecchia approximation, achieving a linear complexity for finding the proposal density. We also reduce the complexities of both the univariate reordering used in MET and SOV and sampling truncated MVN (TMVN) distributions by one order of magnitude. Overall, we achieve linear-complexity for estimating MVN probabilities and sampling TMVN distributions, both at the same convergence/acceptance rate as MET, the currently most accurate method for estimating MVN probabilities. We use a groundwater tetrachloroethylene concentration dataset with four thousand observed locations and more than twenty thousand censored locations to demonstrate the scalability of our method for estimating and making posterior inference for a partially censored Gaussian process model.

Date: Wednesday, September 13.

Speaker: *Dr. Tianjing Zhao, Department of Animal Science, University of Nebraska Lincoln.*

Title: **Solving emerging challenges in statistical genetics and genomics: new data, large data, and sharing data.**

Abstract:

With the development of high-throughput sequencing, whole-genome analysis, such as genomic prediction and genome-wide association studies (GWAS), plays an important role in animal, plant, and human studies. As the amount and diversity of omics data continue to grow, several challenges arise for the linear mixed model. First, there is a need to extend mixed models to incorporate multiple sequential layers of data as one connected network (e.g., the regulatory cascades). Second, due to increasing concerns about data privacy, there is a need to adopt mixed models for encrypted data, enabling the sharing of confidential data in genome-to-phenome analyses. Also, there is a need to address computational costs for large data analysis. We proposed new methods to solve these challenges.

Date: **Wednesday, August 30.**

Speaker: *Dr. Pavel N. Krivitsky, senior lecturer in Mathematics and Statistics, University of New South Wales.*

Title: **A Tale of Two Datasets: Representativeness and Generalisability of Inference for Samples**

of Networks.

Abstract:

Two large, heterogeneous samples of small networks of within-household contacts in Belgium were collected using two different but complementary sampling designs: one smaller but with all contacts in each household observed, the other larger and more

representative but recording contacts of only one person per household. We wish to combine their strengths to learn the social forces that shape household contact formation and facilitate simulation for prediction of disease spread, while generalising to the population of households in the region.

To accomplish this, we describe a flexible framework for specifying multi-network models in the exponential family class and identify the requirements for inference and prediction under this framework to be consistent, identifiable, and generalisable, even when data are incomplete; explore how these requirements may be violated in practice; and develop a suite of quantitative and graphical diagnostics for detecting violations and suggesting improvements to candidate models. We report on the effects of network size, geography, and household roles on household contact patterns (activity, heterogeneity in activity, and triadic closure).