Doctoral Dissertation Defense Announcement

Xinran Qi

“Inference with complex-structured data: controlled group-wise variable selection using a Generative Adversarial Network knockoff filter, unconditional reproducibility probability filter, and semi-parametric inference”

Candidate for Doctor of Philosophy in Biostatistics
Division of Biostatistics, Institute for Health and Equity Graduate School of Biomedical Studies
Medical College of Wisconsin

Wednesday, June 9th, 2021 at 1 PM (CST) Public viewing to be held virtually at: https://mcw.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b369e5be-d816-4d8b-93a0-acf400e11c40

Committee in Charge:
Purushottam W. Laud, PhD (Chair)
Aniko Szabo, PhD (Co-advisor)
Tao Wang, PhD (Co-advisor)
Mei-Jie Zhang, PhD
Mingyu Liang, MB, PhD
Graduate Studies:

Design & Analysis of Clinical Biostatistical Computing
Statistical Models & Methods I
Statistical Inference I
Research Seminar
Theory of Statistical Consulting
Statistical Models & Methods II
Introduction to Bayesian Analysis
Statistical Inference II
Statistical Consulting
Reading and Research
Statistical Models and Methods III
Advanced Statistical Computing
Advanced Bayesian Analysis
Into to Probability Model
Intro to Clinical & Translational Science
Applied Survival Analysis
Advanced Statistics I
Optimization
Theory of Functions of a Real Variable II
Ethics & Integrity in Science
Advanced Topics in Mathematical Statistics
Intro to Analysis I
Linear Models I
Theory of Survival Analysis
Intro Translational Bioinformatics
Statistical Genetics
Introduction to Epidemiology
Research Ethics Discussion Series
Doctoral Dissertation
Dissertation

Inference with complex-structured data: controlled group-wise variable selection using a Generative Adversarial Network knockoff filter, unconditional reproducibility probability filter, and semi-parametric inference

Abstract:

We propose a group-wise variable selection method that adjusts for the correlation structure of variables and guarantees the group-wise FDR under a use-specified value. We utilize the Generative Adversarial Network technique to generate negative controls (model-X knockoffs) for the correlated variables so that we can conduct controlled selective inference for complex-structured data.

Whenever statisticians present their proposed variable selection methods, one frequently asked question is that “How can you show that your findings do have some biological importance, or it is statistically replicable”? Usually, the answer is either we can use cross-validation to give an insight on how the method generalizes to an external data; or we can use some gene annotation dictionary, such as g:Profiler, for gene functional enrichment analysis and visualization, biomarkers' (gene, protein, SNP and so on) pathway analysis and identifier conversion. Here we propose an unconditional reproducibility probability filter which could a third answer for that question.

The semi-parametric generalized linear model (SPGLM) proposed by Rathouz and Gao [2009] assumes that the response is from a general exponential family with unspecified reference distribution and can be applied to model the distribution of binomial event-count data with a constant cluster size. We extend SPGLM to model response distributions of binomial data with varying cluster sizes by assuming marginal compatibility. The proposed model combines a non-parametric reference describing the within-cluster dependence structure with a parametric density ratio characterizing the between-group effect. It avoids making parametric assumptions about higher-order dependence and is more parsimonious than non-parametric models. We fit the SPGLM with an Expectation Maximization Newton-Raphson algorithm to the boron acid mouse dataset and compare estimates with existing methods.
Xinran Qi

EDUCATION

Medical College of Wisconsin (MCW) Milwaukee, WI
PhD Candidate in Biostatistics, GPA: 3.9 Aug 2016 - Jun 2021 (expected)
- Advisors: Purushottam (Prakash) W. Laud, Aniko Szabo, Tao Wang

Sun Yat-Sen University (SYSU) Guangzhou, China
Bachelor of Science in Statistics (double major), GPA: 3.8 Aug 2013 - May 2016
- Thesis: Application of Genome-wide Association Study in Pathway Analysis

Bachelor of Science in Chemistry, GPA: 3.8 Aug 2011 - May 2015
- Thesis: Synthesis and Characterization of Mono-nuclear Hepta-coordinate Metal Complex

TECHNICAL SKILLS

Statistical Tools and Programming: R, Python, SAS, Octave, WinBUGS/OpenBUGS, SQL
Dissemination and Collaboration: Microsoft Office Suite, LaTeX, JupyterLab/Notebook, Markdown

PROFESSIONAL PROJECTS AND EXPERIENCE

Research Assistant for Prof. Tao Wang, PhD PI: Prof. Effie W. Petersdorf, MD
Fred Hutchinson Cancer Research Center/DKMS Stiftung Leben Spenden Jan 2020 - Present
Project: Next-Generation Unrelated Donor Hematopoietic Cell Transplantation
- Multi-task learning in 1000 Genomes Project data with a controlled group-wise feature selection algorithm using a GAN knockoff filter to select single nucleotide polymorphisms (SNPs) for phasing multi-allelic human leukocyte antigen (HLA) genes (DMB, DOB, DOA, DMA & DQB1) while adjusting for within-group correlation among each SNPs block.
- Bridge telomere with centromere region for DMB, DOB, DOA, DMA & DQB1 haplotypes using SNPs blocks with linkage disequilibrium based on statistics such as r-square, D prime & Fisher exact test p-value.

Research Assistant for Prof. Tao Wang, PhD PI: Prof. Effie W. Petersdorf, MD
NIH Project Number: 5U01AI069197-14 NIH Grant Project: Hematopoietic Stem Cell & Cord Blood Transplantation Jan 2018 - Aug 2020
- Gene hunting in MHC & KIR genetic regions using a model-free knockoff filter for high-dimensional, controlled group-wise feature selection with a GAN generator to validate functional SNPs variants and to examine genotype associations with transplant outcomes.
- Proposed an unconditional association test with multiple testing adjustment for reproducible selective inference which denoised signals from SNPs with sparse frequency and determined rank order of immunogenetic factors that strongly predicted survival & post-transplant complications.
- Designed a new determinant of post-transplant outcomes that corrected for genetic sub-structure in a better way than patient/donor’s self-identified ethnicity.
- Applied a two-step group Lasso to survival analysis in Cox regression model context, in order to select significant SNPs to predict survival status, treatment related mortality, relapse, acute/chronic GVHD & other transplant outcomes while adjusting for confounding covariates such as HLA match.

Research Assistant for Prof. Tao Wang, PhD PI: Prof. Mary Horowitz, MD, MS
NIH/NCI Project Number: 5U24CA76518-20 Center for International Blood and Marrow Transplant Research (CIBMTR) Aug 2020 - Present
NIH/NCI Grant Project: A Data Resource for Analyzing Blood & Marrow Transplants
- Collaborate with clinicians & conduct multi-variate survival analysis and model inference for bone marrow transplantation studies.

Data Analyst Director: Prof. Aniko Szabo, PhD
MCW Biostatistics Consulting Service
- Project 1: Significance of Corpora Amylacea in Prostate Tissues
  - Assessed whether corpora amylacea in benign glands of pathology biopsies & other specimens were associated with concurrent cancer based on a study design from Urology Service.
  - Explored whether corpora amylacea were related to patient’s age, race, body mass index, serum prostate-specific antigen, presence of glandular atrophy & chronic/acute inflammation.
- Project 2: Post-allo Hematopoietic Stem Cell Transplantation Relapse Survival in Multiple Myeloma
  - Conducted uni/multi-variate Cox regression analysis for post-relapse survival in patients with multiple myeloma.
- Project 3: Prior Concussion & Physical Conditions Conferring Risk for Sports Related Concussions
  - Ascertained potential risk factors that predisposed athletes to sports related concussions.
- Project 4: Fine Wire Electromyography (EMG) & Surface EMG Measures of Muscle Activation
  - Investigated & compared distribution of fine wire EMG measurements with surface EMG measurements of muscle activation for each exercise task & muscle.
- Project 5: In-hospital Mortality & Post-transplant Complications in Patients with Amyloid Light-chain (AL) Amyloidosis: A Nationwide Inpatient Sample (NIS) Analysis
  - Based on NIS data base, created analytical dataset under scope of comparing in-hospital mortality & post-transplant complications in patients with AL amyloidosis/multiple myeloma using ICD-9 codes for diagnoses & procedures.
  - Constructed descriptive summary tables of demographic characteristics, stem cell sources, mortalities, complications, Charlson Comorbidity Index for AL amyloidosis/multiple myeloma.
- Project 6: Efficacy of Small Bowel Obstruction (SBO) Protocols Using Gastrografin & Omnipaque
  - Evaluated & compared two SBO protocols in terms of length of stay, mortality, operative rate, complication rate, time to resolution, small bowel resection rate between Mayo Clinic (using Gastrografin) & MCW Froedtert hospital (using Omnipaque).

Research Assistant for Prof. Aniko Szabo, PhD
Project: Semi-parametric Models for Clustered Binary & Ordinal Data
- Developed a family of semi-parametric density ratio models that described entire distributions of correlated binary & ordinal count data with varying cluster sizes.
- Formulated a semi-parametric relative risk model for clustered binary count data which combined a non-parametric baseline describing within-cluster dependence structure with a parametric relative risk characterizing between-group effect and applied it to developmental toxicology datasets.

Teaching Assistant for Prof. Ruta Brazauskas, PhD
Division of Biostatistics, MCW
- Graded students’ homework for two courses (Statistics for Basic Sciences & Introduction to Biostatistics).

PUBLICATIONS

Manuscripts in Preparation

JSM Proceedings