Jeremiah Zhe Liu

Language Team, Google Research
Department of Biostatistics, Harvard University

zhl112@mail.harvard.edu
jereliu.info
github.com/jereliu

Education

Harvard University (Boston, MA) *PhD Biostatistics* May 2019

Research Keyword: Bayesian Machine Learning, Ensemble Learning, Uncertainty Quantification, Robust Statistics

GPA: 3.94/4.00.

University of Iowa (Iowa City, IA) *BS Statistics, Mathematics, Minor Computer Science* May 2013

magna cum laude, GPA: 3.96/4.00.

Professional Experience

Google Research
Research Software Engineer 2019-Pres

- Developing statistical solutions to fundemental issues in artificial intelligence (uncertainty quantification and decision making), with application to conversational modeling and recommender systems.
- Work under Google AI Language, in close collaboration with Google Brain.

Research Intern 2018

- Project focus on genomic mutation (i.e. structural variant) detection using deep learning methods . Work under Google Accelerated Science, in close collaboration with Google Brain Genomics.
- Developed a novel neural network module to perform specialized, vision-based processing of gene-sequencing information. Illustrated significant accuracy improvement on mutation type detection tasks.
- Spearheaded the design and implementation of a deep-learning-based system (main architecture: multitask resnet with self-attention) to perform streamlined feature-extraction, mutation site detection and mutation type classification. Illustrated precision and recall improvement over existing structural variant detection tools.

Department of Biostatistics, Harvard University 2019-Pres
Visiting Scientist

- Developing rigorous statistical/machine learning methodology for (1) uncertainty quantification for air pollution exposure assessment, and (2) health effect estimation in large-scale environmental health studies.

Martinos Center for Biomedical Imaging, Mass General Hospital 2017-2019
Graduate Research Fellow / Machine Learning Scientist

- Building reinforcement learning system for automated discovery of novel MRI configurations.
- Partipated in theory development and design of manifold-inspired deep learning architecture for MRI image reconstruction (Nature vol 555).

learnable.ai 2017-2018
Lead Research Engineer

- Designed and supervised the implementation (leading four software engineers) of the company's optical character recognition (OCR) pipeline for processing whole-page mathematical documents.
- Developing a system (leading two research engineers) for joint vision- and language-based understanding and reasoning for high-school geometry questions.
- Provided technical guidance and helped design R& D agenda for classroom video/audio understanding pipeline.

- Other duties include reviewing relevant literature and plan technical solutions, designing and executing R& D agenda, supervising engineer/research progress, and mentoring/management of machine learning engineer interns.

Harvard Clean Air Research Center 2013-2015
Assistant Statistician

- Built spatiotemporal prediction system for heavy-metal air pollutants by integrating information from various sources (air monitoring records, meteorological information, etc) under Random Forrest and Kernel Regression.
- Implemented automated feature selection for GIS features using a combination of measurement error-based weighting and Ridge-type penalization. Conducted stratified cross validation to assess the model's out-of-sample prediction and the influence of prediction error on the risk estimation in second-stage association studies.

Thesis Research

Scalable Bayesian Ensemble Learning with Accurate Predictive Uncertainty, *NeurIPS 2019* 2018-2019
Advisor / Collaborators: Dr. Brent Coull, Dr. John Paisley, & Dr. Marianthi-Anna Kioumourtzoglou

- **Theme**: Spatiotemporally adaptive ensemble learning with accurate uncertainty quantification.
- Proposed a novel ensemble method with spatiotemporally adaptive weights.
- Proposed Bayesian nonparametric machinery to enable model to self-calibrate predictive uncertainty.
- Designed structured VI algorithm to enable scalable and high-quality inference for predictive uncertainty.
- Work applied to optimal aggregation of air pollution predictive models in New England region.

Robust Hypothesis Test for Nonlinear Effect with Gaussian Process, *NeurIPS 2017* 2015-2017
Advisor: Dr. Brent Coull

- **Theme**: Enable classical statistical inference on machine learning models
- Proposed an efficient hypothesis test to detect nonlinear feature effects under Gaussian Process.
- Proposed a cross-validated ensemble estimator for null model to guarantee robust estimation in small sample.
- Work revealed unique connection between model generalizability and the performance of the statistical test.

Technical Skills

- **Analysis & Modelling:** Python (tensorflow, pytorch, pyMC3), R, Matlab
- **Graphics & Documents:** ggplot2, OpenGL, Shiny, ArcGIS, LaTeX
- **High Performance Computing:** C (CUDA, OpenCL, OpenMP)
- **Software Development:** Python, C++, Java, Bash

Open Source Software

cabernet: Calibrated Bayesian Ensemble Regression Network (*in progress*) 2019

- A TensorFlow Probability implementation of Bayesian nonparametric ensemble method.
- Developed a modularized variational inference program that allows flexible mixture of various variational families (e.g. decoupled sparse Gaussian process) to achieve high-quality inference for Gaussian process in near $O(n)$ time.
- Implemented a model zoo of statistical and neural ensemble methods, including cross-validated stacking, generalized additive ensemble, and mixture density network (MDN).

GURLS_MKL: Fast Multiple Kernel Learning Library for GURLS Package 2015

- Independently developed multiple kernel learning functionality for *Grand Unified Regularized Least Squares* (GURLS), an state-of-art supervised-learning package developed at MIT

• Extended fast Proximal Forward-Backward Splitting (PFBS) optimization algorithm to allow memory-efficient iteration update with parallel support. Derived boundary conditions on algorithm parameters to guarantee model convergence.

GPU-Accelerated Sampling for Bayesian Normal Conditional Autoregressive Models 2012

• Designed and implemented parallel algorithms in OpenCL for new model computation strategy proposed by Cowles et al.(2012) for Bayesian Normal CAR model.
• Implementation incorporated into R package *CARrampsOcl*.

Mentorship Experience

Wenying Deng, Doctoral Candidate in Biostatistics, Harvard University 2018-Pres.

• **Project 1**: A Bootstrap Test for Nonlinear Interaction using Cross-valided Kernel Ensemble. *arXiv:1811.11025*
• **Project 2**: Scalable Variable Selection with Theoretical Guarantee using Variational Neural Networks. *In Progress*

Professional Service

Referee, NeurIPS 2019, ICLR 2020

Publications

Machine Learning, Theory & Method

Liu JZ, Coull B. *Robust Hypothesis Test for Nonlinear Effect with Gaussian Processes*. Advances in Neural Information Processing Systems 30 (NeurIPS 2017)

Liu JZ, Paisley J, Kioumourtzoglou M, Coull B. *Adaptive and Calibrated Ensemble Learning with Tail-free Process*. Bayesian Nonparametrics workshop, NeurIPS 2018.

Liu JZ, Paisley J, Kioumourtzoglou M, Coull B. *Accurate Uncertainty Estimation and Decomposition in Ensemble Learning*. Advances in Neural Information Processing Systems 32 (NeurIPS 2019)

Machine Learning, Application

Zhu B, **Liu JZ**, Rosen B, Rosen M *Image reconstruction by domain transform manifold learning*. Nature Vol 555, (22 March 2018) doi:10.1038/nature25988

Zhu B, **Liu J**, Koonjoo N, Rosen B, and Rosen M *AUTOmated pulse SEQuence generation (AUTOSEQ) using Bayesian reinforcement learning in an MRI physics simulation environment.* Joint Annual Meeting ISMRM-ESMRMB 2018

Liu JZ, Lee J, Lin P, Valeri L, Christiani D, Bellinger D, Wright R, Mazumdar M, Coull B *A Robust Hypothesis Test for Continuous Nonlinear Interactions in Nutrition-Environment Studies: A Cross-validated Ensemble Approach*. Journal of the American Statistical Association. *In Submission* (Distinguished Paper Award, ENAR 2019)

Deng W, **Liu JZ**, E Lake, B Coull. ***CVEK:*** *Robust Nonlinear Effect Estimation and Testing with Gaussian Process Ensemble*. Journal of Statistical Software. *arXiv:1811.11025*

Public Health & Biomedicine

Hswen Y, Brownstein J, **Liu JZ**, Hawkins J *Use of a Digital Health Application for Influenza Surveillance in China*. American Journal of Public Health, 2017; e1 DOI: 10.2105/AJPH.2017.303767

Wang Z, Zheng Y, Zhao B, Zhang Y, **Liu Z**, Xu J, Chen Y, Yang Z, Wang F, Wang H, He J, Zhang R, Abliz Z. *Human Metabolic Responses to Chronic Environmental Polycyclic Aromatic Hydrocarbon Exposure by a Metabolomic Approach*. Journal of Proteome Research, 2015, 14 (6), pp 2583 - 2593

Liu Z, Zhang J, Zhao B, et al. *Population-based reference for birth weight for gestational age in northern China*. Early Human Development 2014;90(4):177-87.

Honors & Awards

IMS Hannan Travel Award, Institute of the Mathematical Statistics, 2019

ENAR Distinguished Paper Award, International Biometric Society, 2019

Certificates of Distinction and Excellence in Teaching, Harvard Derek Bok Center for Teaching and Learning, 2018

Phi Beta Kappa, Alpha of Iowa Chapter, CLAS, University of Iowa, 2012