Hanchao Zhang

Website: hanchaozhang.xyz
Github: github.com/Hanchao-Zhang
Address: 180 Madison Ave, 2-31B, New York, NY, 10016
Mobile: 646-206-6662
Email: hz1641@nyu.edu

Education

- **New York Universty, School of Art and Scinece & School of Medicine** — New York, New York
  Doctor of Philosophy - Biostatistics; GPA: 3.7/4.0 — *Aug. 2019 - Now*
  Thesis Advisor: *Prof. Thaddeus Tarpey*
  Research Interest: *Functional Data Analysis, Supverised Learning and Clustering Methods, Optimization Methods*
- **Johns Hopkins University, School of Public Health** — Baltimore, Maryland
  Visiting Student - Biostatistics; — *Sept. 2018 - Jun. 2019*
- **Cornell University, School of Information Sciences & School of Medicine** — New York, New York
  Master of Science - Biostatistics and Data Science; GPA: 4.1/4.3 — *Aug. 2017 - Sept. 2018*
- **Captial University of Economics and Business, School of Finance** — Beijing, China
  Bachelor of Art - International Finance; GPA: 3.8/4.0 (WES Certified) — *Aug. 2013 - Jun. 2017*

Experience

- **Google** — Mountain View, California
  Ph.D. Data Scientist Intern — *Jun. 2022 - Sep. 2022*
  - Developed statistical measurements that quantify cross-questions inter-rater reliability, and detect biased raters with cross-questions information
  - Developed a semi-supervised statistical learning model to performe prediction on the rating and subgroup the bias raters
- **Hedgehog Lab Inc.** — Remote
  Co-founder, Chief Scientist — *Jun. 2021 - Present*
  - Co-founder and Chief Scientist of Hedgehog Lab, a web-based computing language that runs machine learning algorithm
  - Contributed machine learning libraries based on javascript and hedgehogscripts.
- **Johns Hopkins University Bloomberg School of Public Health** — Baltimore, Maryland
  Research Assistant — *Oct. 2018 - Jun. 2019*
  - Performed data cleaning & manipulation over text data; built automatic data pipeline with python to extract features based on NLP (nltk, gensim)
  - Customized machine learning model to make prediction based on multiple data sources (NHANES, Universities & Google Search dataset)
  - Conducted statistical analysis to interpret & optimize model performance based on multiple metrics (ROC, AUC, R-square, confusion matrix)
- **Technical Consulting & Research, Inc.** — New York, New York
  Data Analyst Intern — *Oct. 2018 - Jun. 2019*
  - Performed webscraping with Python (Selenium, BeautifulSoup, Multiprocess) to get 5 k + medical data from National Library
  - Utilized NLP techniques to extract features (location, age, history, condition) from medical script data with Python (nltk, genism, regr)
  - Conducted statistical analysis to interpret & optimize model performance based on multiple metrics (ROC, AUC, R-square, confusion matrix)
- **Weill Cornell Medicine** — New York, New York
  Statistician Intern — *Dec. 2017 - Mar. 2018*
  - Extracted, cleaned & manipulated biomedical data with SQL & R to get clean medical data for statistical research
  - Conducted statistical analysis to generate insight from medical data based on suggestions provided by physicians and biomedical researchers
  - Designed experiments, set up sample size, conducted power analysis, and drafted medical research protocol

Publications and Talks

- **Invited Talk: Optimal Transformations of High-Dimensional Functional Data for Clustering Methods**: Joint Statistical Meeting 2022, Invited Paper Session (JSM 2022)
- **Invited Talk: Functional Data Clustering and Regression Methods**: Columbia University Functional Data Working Group Invited Talk (FDWG 2022)
- **Invited Talk: Optimal Linear Transformations of Functional Data for Clustering Methods**: Joint Statistical Meeting 2021, Invited Paper Session (JSM 2021)
- **Collberative Paper: Validation of EHR medication fill data obtained through electronic linkage with pharmacies**: Saul Blecker, Samrachana Adhikari, **Hanchao Zhang**, etc., Journal of Managed Care and Specialty Pharmacy
- **Collberative Paper: Quantitative evaluation of rejuvenation treatment of nasolabial fold wrinkles by regression model and 3D photography**: Rou-Yu Fang, **Hanchao Zhang**, etc., Journal of Cosmetic Dermatologyh

Projects

- **An Outliers Detection Algorithm for Functional Data**:
  - Developed an outliers detection algorithm for Electroencephalography (EEG) data using a random consecutive window method
  - Applied the developed outliers detection algorithm on the real EEG dataset and improved the AUC from 88% to 96%
  - Built a functional regression model on the data preprocessed by the outliers detection algorithm, and improved prediction accuracy by 12%
- **Association between maternal prenatal stress and human fetal brain development**:
  - Acquired the fMRI and questionnaire data, preprocessed and winsorized the survey data and fMRI data
  - Applied EDA for previewing the data and obtained the association between cortisol value and multiple stress scores
  - Uitilzed unsupervised clustering methods (PCA and T-SNE) to cluster the patients, identify subgroups of the patients for further analysis
- **Covariates Adjustment for Selecting Best Model for Psychosis Improvement**:
  - Utilized the randomized trial data in NIMH with outcome variable quality of life score
  - Built best logistic model with covariates adjustment, evaluate the improvement of accuracy and intention-to-treat parameter
  - Built machine learning models such as random forest, comparing it with the logistic model in the improvement of accuracy
  - Transformed the Model to other datasets representing different populations, evaluating the difference in the intention-to-treat parameter analysis
- **Deep Learning U-Net Model in Segmentation of Brain MR Images**:
  - Utilized the MRI data released for the MICCAI 2012 Grand Challenge on Multi-Atlas Segmentation
  - Established a baseline network for two basic segmentation tasks: brain/non-brain and grey matter/white matter/cerebrospinal fluid
  - Provided the network full 2D axial slices of the MR volumes for training and testing
  - Interpreted the changes that are being made to the network design by exploring intermediate feature maps created in the alternate networks, looking for differences in the organization of features
- **Risk Prediction and Evaluation of the Effect of Myocardial Infarction on Stroke**:
  - Performed data & feature engineering over 1.7 M healthcare data(Strokes) including extraction, manipulation, feature generation & selection
  - Applied Survival Analysis and machine learning methods in computer clusters, including Cox Regression, Accelerated Failure model Survival Random Forest. Selected the Cox Regression without interaction as the best prediction model by Cross- Validation using R (survival, randomForestSRC, and pec) (Dr. Diaz's Lab)
- **Robust Regression for Outliers in Laboratory Data**:
  - Applied three robust regression methods, M-estimation, S-estimation, and MM-estimation, to evaluate the association between urinary PGE-M level and urinary PGD-M level in obese and lean mice, with and without celecoxib by R (MASS, ggplot2); Concluded that M-estimation is slightly better than the other two robust regressions
  - Improved certainty of variance analysis by leveraging each data point with weight and new loss function
- **Transportation and Electronic Health Record (EHR) Database Construction**:
  - Established database schema and populated tables based on electronic health records and New York City transportation data from U.S. Department of transportation on MySQL server
  - Applied Logistic Regression to search association of transportation preference, concluding that number of patients having Benign Essential Hypertension was associated with patients' location in New York City, and Hyperlipidemia and Atrial Fibrillation were associated with patient's transportation preference using R (stat, ggplot2, randomForest)

Skills Summary

- **Programming**: R, Linux, Python(Numpy, Pandas, NLTK), SQL, SAS(Advanced Certified)
- **Languages**: English, Mandarine