Hanchao Zhang

Chair of CSSA at NYU Grossman School of Medicine
Website: hanchaozhang.xyz
Address: 180 Madison Ave, 5-31C, New York, NY, 10016
Mobile: (+1) - 646 - 206 - 6662
Email: hanchao.zhang@nyu.edu

Education

- **Columbia University, School of Art and Science** — New York, New York
  Visiting Ph.D. Student - Statistics Department through IUDC; — *Jun. 2023 - Present*
  Research : *Stochastic Process*

- **New York University, Grossman School of Medicine & School of Art and Science** — New York, New York
  Doctor of Philosophy - Biostatistics; — *Aug. 2019 - Present*
  Thesis Advisor: *Professor Thaddeus Tarpey (Chair, and Program Director)*
  Research Interest: *Manifold Learning, Self-Consistency, Functional Data, Supervised Learning and Clustering Methods in High-Dimensional Data*
  Awards: *Special MacCracken Awards 2023 for Research Outstanding*

- **Johns Hopkins University, School of Public Health** — Baltimore, Maryland
  Visiting Student - Biostatistics, Applied Mathematics and Statistics; — *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. Research Scientist Intern — *Jun. 2022 - Sep. 2022*
  - Developed statistical model that quantifies cross-questions inter-rater reliability and detects unreliable raters with cross-questions information based on linear mixed effect model with repetitive measurements
  - Developed a semi-supervised statistical model to estimate latent variables representing unreliable raters. The model improves the prediction accuracy of rating by 10.25% compared to the linear mixed effect model

- **Hedgehog Lab Inc. (https://hlab.app/)** — Remote
  Co-founder, Chief Scientist — *Jun. 2021 - Jun. 2022*
  - 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 predictions 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 web scraping with Python (Selenium, BeautifulSoup, Multiprocess) to get 5k + 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)

Publications & Talks

- **Methodology Paper: K-Tensors: Clustering Positive Semi-Definite Matrices**: **Hanchao Zhang**, Thaddeus Tarpey., arXiv
- **Methodology Paper: Principal and Self-Consistency for Positive Semi-Definite Matrices on Riemannian Manifold**: **Hanchao Zhang**, Thaddeus Tarpey., Working Paper
- **Collberative Paper: Low muscle mass is associated with a higher risk of all–cause and cardiovascular disease–specific mortality in cancer survivors**: Dongyu Zhang, **Hanchao Zhang**, etc., Nutrition
- **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
- **Invited Talk: Clustering Positive Semi-Definite Matrices: A Metric Learning Approach to Diseases Subtyping**: International Biometric Society, Invited Paper Session (ENAR 2023)
- **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 (FDAWG 2022)
- **Invited Talk: Optimal Linear Transformations of Functional Data for Clustering Methods**: Joint Statistical Meeting 2021, Invited Paper Session (JSM 2021)

Projects

- **Principal and Self-Consistent Positive Semi-Definite Manifold**:
  - Developed and defined a novel self-consistent submanifold framework tailored for positive semi-definite manifold data
  - Investigated and established the statistical properties governing this submanifold, showcasing its principal and self-consistency properties
  - Defined a novel distance metric for positive semi-definite matrices, contributing to the advancement of distance-based analysis on manifolds
- **K-Tensors: A Self-Consistent Algorithm for Clustering Positive Semi-Definite Matrices**:
  - Developed a self-consistent clustering algorithm tailored specifically for positive semi-definite matrix-valued data
  - Extended the clustering algorithm to seamlessly accommodate manifold data, enhancing its applicability and robustness on manifold learning
  - Improved the clustering algorithm's capabilities in detecting the true data generation process
- **Self-Consistent Convexity-Based Clustering Algorithm for Bregman Distance**:
  - Developed a self-consistent support hyperplane approach for convexity-based clustering as a weighted maximum volume problem
  - Conducted analysis of the intricate relationship between convexity-based clustering and clustering algorithms based on the Bregman distance
- **FairElasticnet: A Framework for Fairness Variable Selection**:
  - Developed the FairElasticnet model, a novel and impactful approach for fairness variable selection in prediction models
  - Applied augmentation methods and ADMM optimizer to simplify and optimize ℓ-1 and ℓ-2 problems within the FairElasticnet framework
- **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
  - Utilized unsupervised clustering methods (PCA and T-SNE) to cluster the patients and identify subgroups of the patients for further 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 the U.S. Department of Transportation on the MySQL server
  - Applied Logistic Regression to search association of transportation preference, concluding that number of patients having Benign Essential Hypertension was associated with the patient's location in New York City, and Hyperlipidemia and Atrial Fibrillation were associated with the patient's transportation preference using R (stat, ggplot2, randomForest)

Skills Summary

- **Programming**: Python, R, Linux, SQL, SAS (Advanced Certified)
- **Languages**: English, Mandarin