Colin Pei Cui

456 Snowden Lane, Princeton, NJ 08540
https://cui-colin.github.io/
(415)518-3959 • colstat@gmail.com

Research Interest

I'm interested in statistics, machine learning, and convex optimization. Broadly, I am interested in understanding the mathematical theory of data science.

Education

Rutgers University
MS., [all but M.S. thesis] Statistics
(Incomplete M.S. in statistics, with additional work in mathematical statistics, probability, Bayesian statistics courses. M.S. research topic focused on Bayesian statistics. I compared 11 resampling algorithms for Boltzmann distributions: metropolis, equi-energy sampling, parallel tempering, Hamiltonian, and simulated tempering)

University of California, Davis
B.S. in Statistics, *Cum Laude,* May 2010
Minor: Economics

Relevant Coursework
Probability Theory (Ph.D. course), Statistical Inference (Ph.D. course), Decision Theory (Ph.D. course), Data Mining (Master's course), Bayesian Data Analysis (Ph.D. course), Interpretation of Data (Ph.D. course), Statistical Learning and Nonparametric Estimation (Ph.D. course, Princeton University)

Skills

Software: R, Python, Matlab, Stata, C, C++, Java

Data Science: Extensive training in high-dimensional statistics, and statistical learning theory. Fluent in building statistical models and data visualization in R and Python including packages in caret, glmnet, scikit-learn, NumPy, Pandas, ggplot2, and Plotly. Also, deep learning experience with PyTorch and TensorFlow.

Languages: English, Chinese, French (conversational)

Working papers

Colin Cui. *On Statistical Learning Theory, Oracle inequality, and the Lasso.* Dec. 2020
In this working paper, I highlighted theorems and proofs of high-dimensional statistical learning theory in the past decade. It provided an overview of theory and proof in simple terms to the general audience who may not be familiar to statistics. It also surveys concentration inequalities (which are needed for the proofs), oracle inequality, penalized regressions, and the convergence rate of Lasso. (GitHub link)

Paper contributions

Xiaodong Zhao, et. Al. A Bayesian approach for characterization of soft tissue viscoelasticity in acoustic radiation force imaging. *International Journal for Numerical Methods in Biomedical Engineering,* 32(4):e02741, 2016

In biomedical imaging, acoustic radiation force (ARF) were developed for characterization of the viscoelasticity of soft tissue, which leads to stress distribution of a region of excitation (ROE). To improve estimation of the ROE, we presented a Bayesian inverse formulation. The Bayesian approach formulates the known parameter as a distribution. To make the computation feasible, Gaussian Processes was used as a metamodel to approximate the complex fininte element model. We quantify parameter uncertainty with simulation on the posterior.

Avram Goldberg, et al. *Clinical Outcomes of Scleroderma Patients At High Risk for Pulmonary Hypertension. Analysis of the Pulmonary Hypertension Assessment and Recognition of Outcomes in Scleroderma Registry*. ACR/ARHP Annual Meeting, 2012. (acknowledged)

In this paper, I worked on data visualization with ggplot2 on pulmonary hypertension disease data from Georgetown University Medical School. I performed parametric tests, identified variables that contribute to pulmonary hypertension disease and a patient's survival rate.

Experience

Data Scientist, Ideal (co-founder)**,** Feb. 2018 - Present

- Built statistical machine learning models for NYC start-ups and small businesses
- Served over 50 small firms and individuals in, achieved a first year profit
- Past client partners include NYC professor startup on algorithmic trading in Manhattan

Research Scholar for Prof. Assimina Pelegri, MAE, Rutgers University**,** July 2014 – Jan 2018

- Area of focus: We presented a Bayesian formulation to estimate the stress distribution of region of excitation (ROE) in biomedical imaging. The Bayesian approach formulates the known finite element (FE) model parameter as a probability distribution. To make it computationally feasible, Gaussian Processes was applied as a statistical estimation to solve inverse problem. Our simulation study showed that Bayesian approach to FE improved even in the presence of large uncertainty.
- Selected Project: "A Bayesian approach for characterization of soft tissue viscoelasticity in acoustic radiation force imaging" Xiaodong Zhao, Assimina Pelegri
  Oral Presentation: SIAM Conference (MS10) at Philadelphia, PA (March 2016)

Adjunct Faculty, New Jersey Institute of Technology Sept. 2013 – Feb. 2014

- Taught Math 105 Elementary Statistics and Probability
- Duties include: teaching first year statistics course, grading, reviewing student progress

Research Assistant, Department of Statistics, Rutgers University**,** Jan. 2013 – Jun. 2013

- Advisor: Professor Zhiqiang Tan
- Area of focus: Compared 11 resampling algorithms for Boltzmann distributions including independent metropolis, equi-energy sampling, parallel tempering, and simulated tempering
- Simulated Monte Carlo methods for numerical approximation using statistical software R**.**

Performed stochastic approximation to MCMC algorithm with empirical results in order to prepare manuscript for publication.

- Selected Project: "Resampling Markov chain Monte Carlo algorithms: Basic analysis and empirical comparisons" Zhiqiang Tan (paper)

Research Intern for Dr. Gail Gong (Stanford University) Jun. – Jul. 2010

- Researched in a cohort of four students on estimation in human genetics and disease
- Stimulated family data using R language and compare different ways of estimating the penetrance of disease
- co-generated cancer data, and identified bad alleles and find its probability

Conference

Bayesian Inference Using Gaussian Process Metamodel in Biomedical Imaging (*with* A. Pelegri, and X. Zhao). *Conf. Mathematical Aspect of Material Science,* 2016

Links to selected paper, talk, presentation

1. Paper: *On Statistical Learning Theory, Oracle inequality, and the Lasso*
   https://cui-colin.github.io/main_arxiv.pdf

2. Talk: SIAM Bayesian calibration talk at Philadelphia, PA
   https://github.com/cui-colin/Talks/blob/main/SIAM talk.pdf

3. Project: MCMC using Hamiltonian Dynamics
   https://github.com/cui-colin/Talks/blob/main/HamiltonianMC.pdf

4. Python Data Science Notebook: https://github.com/cui-colin/Python/blob/main/SU2C.ipynb

*A comprehensive list can be found on my webpage.