Jenna Landy

Updated March 8, 2024

Email: jlandy@g.harvard.edu
Phone: (831) 588-6073
GitHub: github.com/jennalandy
LinkedIn: linkedin.com/in/jenna-landy
Website: www.jennalandy.com/

Research interests

I am interested in both statistical and deep learning models of molecular biology, particularly in the field of cancer genomics. My current research focuses on mutational signatures, which model underlying mutational processes in cancer genomes. I develop new statistical methods to address unanswered questions within this field, drawing knowledge from unsupervised learning, multi-study and ensemble learning, machine learning, and Bayesian modeling and computation. I develop publicly available software to encourage the reproducibility and usability of my work.

I am also passionate about the creation and accessibility of educational resources for statistics and data science. I've taken a leadership role in my department by guiding lab sections and providing tutoring at the masters and PhD levels for probability theory, inference, regression methods, data structures, and algorithms. I have also conducted multiple workshops that demonstrate the effective use of tools such as R Markdown and Quarto in academic research.

Education

Harvard Graduate School of Arts and Sciences Cambridge, MA
PhD in **Biostatistics** (Masters awarded May 2022) August 2020 – May 2025
Advisor: Giovanni Parmigiani *GPA: 3.96*
Selected coursework

- Computer Science and Data Science: Computational Biology and Bioinformatics, Cancer Genome Data Science, Deep Learning in NLP, Data Structures and Algorithms, Decision Theory
- Statistics: Bayesian Statistics and Computation, Probability Theory, Inference, Linear Models, Unsupervised Learning, Statistical Genetics, Causal Inference

California Polytechnic State University, SLO San Luis Obispo, CA
BS in **Statistics**, minor in **Data Science** September 2016 – June 2020
Mentors: Hunter Glanz, Prince Afriye, Brian Granger *GPA: 3.95*

Selected coursework

- Computer Science and Data Science: Distributed Computing, Object Oriented Programming, Machine Learning and Data Science, Database Systems, Ethics in Technology
- Math and Statistics: Multivariate Statistics, Survival Analysis, Calculus, Linear Algebra, Methods of Proof in Mathematics
- Biology: Human Genetics, Cell and Molecular Biology

Publications

Landy, J. M., and Parmigiani, G. (2024). Gridsemble: Selective Ensembling for False Discovery Rates. arXiv preprint arXiv:2401.12865

This paper presents a novel data-driven selective ensembling algorithm for estimating local (fdr) and tail-end (Fdr) false discovery rates. We believe this method will be a useful tool for computing reliable estimates of fdr and for improving replicability in the presence of multiple hypotheses by eliminating the need for an arbitrary choice of method.

Self, B. P., **Landy, J.**, Widmann, J. M., Chen, J., Kerfs, M. (2021, July), ***The Mechanics of SUCCESS: How Non-Cognitive and Affective Factors Relate to Academic Performance in Engineering Mechanics*** Paper presented at 2021 ASEE Virtual Annual Conference Content Access, Virtual Conference. https://strategy.asee.org/37876

This paper investigates how non-cognitive and affective (NCA) competencies (e.g. motivation, grit, belongingness, etc.) can better predict academic success in engineering, as measured by students grades in introductory physics, statics, and dynamics. I performed the analyses and wrote the methods, results, and discussion sections.

Chen, J., **Landy, J. M.**, Scheidt, M., Major, J. C., Ge, J., Chambers, C. E., Grigorian, C., Kerfs, M., Berger, E. J., Godwin, A., Self, B. P., Widmann, J. M. (2020, June), ***Learning in Clusters: Exploring the Association Between Noncognitive and Affective Profiles of Engineering Students and Academic Performance*** Paper presented at 2020 ASEE Virtual Annual Conference Content Access, Virtual Conference. https://peer.asee.org/34901, DOI: 10.18260/1-2–34901

This paper investigates clustering students by their non-cognitive and affective (NCA) competencies and how academic success and retention differs between these groups. I performed the analyses and wrote the results section.

Widmann, J., Self, B., Chen, J., Chambers, C., Kusakabe, K., **Landy, J.**, Berger, E., Ge, J., Godwin, J., and Scheidt, M. (2019, July), ***Academic SUCCESS: An Analysis of How Non-Cognitive Profiles Vary by Discipline for Engineering and Computer Science Students***. Paper presented at 2019 Research in Engineering Education Symposium. https://www.sasee.org.za/wp-content/uploads/REES-2019-proceedings.pdf, pages 540 - 548.

The paper looks at how non-cognative and affective competencies differ between students of different years. I administered surveys and wrote the methods section.

Open-Source Software

- **bayesNMF**: an in-progress R package for fitting Bayesian Non-Negative Matrix Factorization (NMF) with a variety of modeling specifications and the option to learn latent rank as part of the Bayesian model.
- **gridsemblefdr**: an R package for estimating local (fdr) and tail-end (Fdr) false discovery rates in large-scale multiple hypothesis testing.
- **easygit**: a minimal JupyterLab extension for basic version control without needing to learn git.
- **jupyterlab-shortcutui**: a JupyterLab extension to edit keyboard shortcuts with a user interface.
- **jupyterlab-git**: an extension to use git/GitHub within the JupyterLab environment.
- **plyto**: an extension to visualize training of ML models in real time within the JupyterLab environment with a corresponding Python package.

Contributed sessions

Women in Statistics and Data Science 2023
Gridsemblefdr: model selection and ensembling for false discovery rates with application to differential expression analysis

Joint Statistical Meetings 2022
Gridsemblefdr: ensembling and hyperparameter optimization for unsupervised learning with application to false discovery rates

Talks, tutorials, and educational materials

Tutorial: Document Creation with Rmd and Quarto 2023
An introduction to using Rmd and Quarto files in quantitative research. Presented at the Dana Farber Data Science workshop series.

Tutorial: Remodel your Rmd 2022
Tips and tricks for experienced quantitative researchers to boost their research workflow with R package development, parameterized reports, Rmd websites, and Rmd customizations. Presented at the Harvard Biostatistics Student Seminar.

Tutorial: Shiny App for Sepsis Prediction 2020
Walking through an intuitive, collaborator-friendly R Shiny user interface to predict hospital readmission due to sepsis in collaboration with Dignity Health. Presented at the Cal Poly honors program senior project poster session.

Textbook: Introduction to Databases and API 2020

A Bookdown and RShiny app introducing statistics students to the use of databases and APIs in R, Python, SAS, and Julia. This work is aimed at lowering technical barriers that keep statistics students from using these tools. All materials are public on GitHub. Worked in a team of two advised by Hunter Glanz, PhD and Rebecca Ottesen, PhD.

Tutorial: JupyterLab Extensions for Enhanced User Experience 2018

Demonstrating three JupyterLab extensions at the Jupytercon 2018 Poster Session.

Industry experience

Defli Diagnostics Baltimore, MD

Title: Data Science Research Intern June - August 2022

Mentor: Laurel Keefer

Project: Data Visualization and Automated Reports R Software Package

Developed and documented an in-house R package for data visualizations, modeling, and automated Rmd reports for use by the Data Science Research team. Followed R tidyverse coding style guide and used Git/GitHub for version control.

Amazon Web Services Seattle, WA

Title: Data Science Intern June - September 2019

Mentor: Steve Loeppky

Project: Public GitHub Notebook Corpus Research Collaboration

Extracted and analyzed all Jupyter Notebooks public on GitHub to understand AWS Sagemaker and Jupyter users, their processes, and their struggles in order to inform big-picture user experience questions. Used the GitHub API and AWS EC2 and S3 instances. All code and results are public on GitHub.

Project Jupyter San Luis Obispo, CA

Title: Software Engineering Intern March - December 2018

Mentor: Brian Granger

Project: Developing JupyterLab extensions to enhance user experience

Contributed to JupyterLab, an open-source interactive development software. Created a visualization toolkit for machine learning and an interface to view and edit keyboard shortcuts in a team with another software intern and a UX intern. Went through design iterations and conducted multiple rounds of user testing for the GitHub extension. Presented projects and conducted user testing at JupyterCon 2018.

Research

Selective Ensembling for False Discovery Rates Oct. 2020 – Jan. 2024

Advised by Giovanni Parmigiani, PhD

Gridsemble is a data-driven selective ensembling algorithm for estimating local (fdr) and tail-end (Fdr) false discovery rates in large-scale multiple hypothesis testing. Existing methods for estimating fdr often yield different conclusions, yet the unobservable nature of fdr values prevents the use of traditional model selection. Our method circumvents this challenge by ensembling a subset of methods with weights based on their estimated performances, which are computed on synthetic datasets generated to mimic the observed data while including ground truth. This paper is on arXiv and is currently under review. The corresponding R software package is on GitHub.

Comparative Analysis of Bayesian NMF Models Dec. 2023 –

Advised by Giovanni Parmigiani, PhD

I am developing an R software package, bayesNMF, that implements various model specifications of Bayesian NMF. I also include the option to learn the latent rank as a part of the Bayesian model. I am comparing these models in terms of reconstruction error, correctness of the learned latent factors (in simulation studies), memory usage, and speed.

Graph Neural Networks for PerturbSeq Nov. 2023 -

We are working on predicting changes in gene expressions given a set of perturbations utilizing a graph of known functions from the gene ontology (GO) database. We use a graph neural network to learn perturbation embeddings, which importantly, will allow predictions for combinations of perturbations that were never tested experimentally. We are focusing on the transferability of this model across datasets and making predictions that are robust across cell types and sequencing depths.

Bayesian Causal Inference for Mutational Signatures May 2023 –

Advised by Giovanni Parmigiani, PhD and Nima Hejazi, PhD

We are looking at mutational signatures through the lens of causal inference to answer questions about the causal effects of such exposures on the presence and magnitude of mutational signatures in cancer genomes. Using mutational signatures (or any latent factor) as an outcome in the causal inference framework comes with many challenges.

Deep unfolding Bayesian NMF for Mutational Signatures April 2023 -

Advised by Demba Ba, PhD

Bayesian non-negative matrix factorization is a method used across fields including genomics, audio and signal processing, and neuroscience. The complexity of the posterior of Bayesian NMF requires MCMC methods, such as a Gibbs sampler, or variational inference. We propose a faster solution through deep algorithm unrolling. By designing a neural network where each layer mimics a single iterative update, we are able to improve speed without sacrificing model performance.

Part of Speech-Based Data Augmentation for NMT Oct. – Dec. 2021

Advised by Christopher Tanner, PhD

Data augmentation improves accuracy of ML models for natural language processing tasks, such as neural machine translation (NMT), by increasing the amount and variety of training data. Augmentation approaches for NLP can be applied at the token-level (e.g. contextual replacement) or at the embedding-level (e.g. soft contextual replacement or mixing two sequences by averaging their embeddings with SeqMix). While prior methods keep the semantic meaning of a sentence, a weakness is that they don't maintain syntax. We addressed this by matching POS in word replacement and token mixing, which shows up to a 1 point increase in BLEU. Further, in prior SeqMix methods, the sequences to be mixed are chosen at random, which we address by combining more similar or different sequences. We found that mixing sequences of similar length shows up to a 0.6 point improvement in BLEU. Paper and code are publicly available.

Metabolite Associations for Prostate Cancer Riks Aug. – Dec. 2021

Advised by Lorelei Mucci, PhD and Kathryn Penney, ScD

I investigated connections between metabolites and SNPs used in the prostate cancer PRS. This was a study of all Health Professionals Follow-Up Study (HPFS) participants with both GWAS and metabolomics data from various nested case-control studies within the cohorts.

Studying Underlying Characteristics of Computing and Engineering Student Success (SUCCESS) Feb. 2019 - June 2020

Advised by Jim Widmann, PhD

I was a statistical consultant research assistant on the SUCCESS project, a collaboration between Cal Poly San Luis Obispo and Purdue University. Analyzed survey data on non-cognitive competencies in engineering students as predictors of academic success using R, communicated statistical results to non-statisticians, and contributed to three conference papers.

Teaching experience

Teaching Assistant, Biostatistics Department, Harvard

BST 221: Introduction to Data Structures and Algorithms 2022

Taught lab section, hosted weekly office hours, and graded assignments for a PhD level course in developing algorithms, proving correctness, and analyzing runtime.

Average student rating: 5/5

BST 201: Introduction to Statistical Methods 2021

Taught weekly lab section, hosted weekly office hours, graded all assignments and exams for a Masters level course covering basic statistical techniques.

Average student rating: 4.8/5

Tutor, Biostatistics Department, Harvard 2021 –
Tutored one-on-one and group sessions for the Biostatistics Ph.D. Core Coursework: Probability, Inference, Methods, Data Structure, and Algorithms. Prepared lessons, answered questions, and provided resources.

Teaching Assistant, Statistics Department, Cal Poly
STAT 427: Mathematical Statistics 2020
Hosted weekly office hours and graded assignments for an upper-division course on the theory of hypothesis testing and its applications.

STAT 426: Estimation and Sampling Theory 2020
Hosted weekly office hours and graded assignments for an upper-division course on properties of statistics obtained from samples and asymptotics.

STAT 425: Probability Theory 2019
Hosted weekly office hours, assisted in lab section, and graded assignments for an upper-division course on the rigorous development of probability theory.

Competitions

DataFest Hackathon, UCLA 2018, 2019, 2020
ASA sponsored annual data science hackathon. In 2019, my team won the data visualization category out of 80 student teams.

RShiny App Competition, Cal Poly Statistics Department 2020
Built an RShiny web app for Medium data science article recommendations and visualizing article popularity through web scraping and topic modeling. I won the competition and was awarded funding to attend the RStudio conference in January 2020.

Honors & Awards

Biostatistics Department, TH Chan School of Public Health
Certificate of Distinction in Teaching 2021-2022
Robert B. Reed Prize for Excellence in Biostatistics for the highest score on the written qualifying exam 2021

California Polytechnic State University, San Luis Obispo
Summa Cum Laude 2020
Graduated with Honors (Cal Poly Honors Program) 2020
Award for Contribution to the Objective and Public Image of the College of Science and Mathematics 2020
Academic Merit Award (Department of Statistics) 2020
Humanitarian Award (Department of Statistics) 2020
Dean's List and President's Honors List 2016-2020
Merit-based full-ride Frost scholarship (Cal Poly Frost Fund) 2016-2020

Skills

Mathematics and Statistics
Probability Theory, Inference, Regression, Bayesian Methods and Computation, Unsupervised Learning, Statistical Genetics, Calculus, Linear Algebra

Machine Learning and Deep Learning
Autoencoders, Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Graph Neural Networks (GNN), Generative Adversarial Networks (GAN), Diffusion Models, Model-Based Deep Learning, Deep Algorithm Unrolling

Programming
Object-Oriented and Functional Programming, Data Manipulation (Pandas, R), Simulation, Databases, APIs, Machine Learning (PyTorch, Scikit-Learn, R), Debugging, Visualization (matplotlib, ggplot, RShiny), Cluster Resources, Bash, Software Development, Git/GitHub

Languages
Python (PyTorch), R (tidyverse), Java, JavaScript (TypeScript, React), MongoDB, Hadoop, PySpark, SQL (JDBC), SAS, Stata

Service and outreach

Stat Start 2022
Volunteered teaching courses on basic statistics, R, and data analysis for Stat Start, a computational summer program for high school students.

Cal Poly College of Science and Math Student Council 2017 – 2020
This is a committee of student leaders and the Dean that serves as a line of communication between students and faculty in the college. I initiated a college-wide peer mentoring group, volunteered at the college's annual research conference, and contributed to a campaign to have faculty discuss campus climate in their classrooms. As president in the 2018 – 2019 academic year, I planned events, scheduled and ran meetings, and organized guest speakers from across campus.

Professional Memberships

American Statistical Association Student member 2017 –
Caucus for Women in Statistics Student member 2023 –