Yiqiu Shen

Ph.D., Statistics | Data Scientist | Statistician

Los Angeles, CA
yiqiushen2025@gmail.com
www.linkedin.com/in/yiqiu-shen/
github.com/yiqiushen

Summary

Ph.D. statistician and data scientist with a strong foundation in **machine learning**, **statistical modeling**, and **data analysis**. Experienced in applying **Python**, **R**, and **SQL** to solve real-world problems, from wildfire damage modeling using geospatial data to deep learning for medical imaging. Skilled in **communicating** complex analytical findings to both technical and non-technical audiences, with a track record of impactful research and industry collaboration. Proficient at incorporating AI-assisted tools such as **Roo Code** to streamline workflows and boost efficiency.

Education

Aug 2019 – July 2025 **Ph.D. in Statistics**, *University of Southern California*, Los Angeles, CA
GPA: 3.96, Field of Interest: High Dimensional and Robust Statistics, Statistical Learning Theory
Advisor: Dr. Stanislav Minsker

Aug 2016 – May 2019 **B.S. in Applied and Computational Mathematics**, *University of Southern California*, Los Angeles, CA
GPA: 3.92, summa cum laude

Professional Experience

May 2023 – Aug 2023 **Data Scientist Intern**, *Delphire Inc.*, Los Angeles, CA

- Utilized **Pandas, Numpy, and Seaborn** to model and visualize wildfire impact on properties using historical data (1990–2020).
- Created **spatial indexes** using **SQL** (PostgreSQL dialect) and **PostGIS** to integrate Microsoft Building Footprints and Zillow housing data and estimated monetary damage, bringing the query time down to under **1s** for a dataset of **11 mil geoshapes**.
- Operated **FlamMap** and **WindNinja** to simulate wildfire perimeters and assess damage reduction using Sentinel units
- Generated actionable insights that guided leadership decisions and secured additional project funding.

Selected Projects

All codes available on github.com/yiqiushen.

2024 **Ultrasound Image Classification Project**

- Designed a deep learning pipeline using **PyTorch** to classify and segment breast ultrasound images.
- Fine-tuned **ResNet** and **CLIP** models with linear probing and compared against custom CNNs, achieved **83% accuracy**.

2020 **Membership Identification in Large Scaled Sparse Networks**

- Implemented SIMPLE in R and Python to analyze financial data of S&P1000 company networks.
- Improved computation efficiency of both experiment and simulation and reduced running time by 90% using **parallel processing**.

2020 **Pairwise Relationship Identification via Mixed Neural Networks**

- Trained and evaluated a dual-branch **CNN–RNN** model on two benchmark datasets of 1400 + Svirus–host pairs, learning k-mer motifs and their long-range dependencies.
- Sub-sampled 2 kb viral and 5 kb host contigs as paired inputs, capturing motif–motif interactions without full-genome alignment.
- Achieved **87 % accuracy**; results support the hypothesis that co-evolution drives shared word-pattern usage in interacting pairs.

2020 **High Dimensional Classification via Spiked Eigenvalue Theory**

- Designed a high-dimensional classifier that exploits the spiked-eigenvalue structure of 3-mer frequency covariance matrices to separate viral from bacterial contigs.
- Verified the theoretical eigen-spectrum (one dominant component, multiple near-zero spikes) on 77 k+ RefSeq contigs, confirming the Markov-chain assumptions that underlie the model.

Publications

All publications reflect equal collaboration among authors.

2025 **Minimax Supervised Clustering in the Anisotropic Gaussian Mixture Model**
Authors: Minsker, Stanislav, Mohamed Ndaoud, **Shen, Yiqiu**
Journal of Machine Learning Research. To appear

- Derived tight minimax bounds for high-dimensional supervised clustering and showed classical LDA is sub-optimal.
- Proved a fully interpolating least-squares classifier can be optimal, and even robust to covariance corruption; validated via `scikit-learn` simulations.
- Combined cutting-edge theory with practical experimentation, enlightening **model-selection** decisions for segmentation and anomaly-detection tasks in real-world ML pipelines.

2025 **The Impact of Contamination and Correlated Design on the Lasso**
Authors: Minsker, Stanislav, **Shen, Yiqiu**
Probability and Statistics Letters. To appear

- Investigated how **outliers** and **multicollinearity** affect Lasso **feature selection**; extensive simulation study confirms the theory.
- Built a custom, from-scratch Lasso implementation; showcased the ability to **translate new theory into production-ready code** when off-the-shelf libraries are insufficient.
- Demonstrated expertise in developing **reliable, interpretable feature-selection pipelines** that remain trustworthy despite data contamination, a critical skill for production models built on noisy real-world data.

2024 **Concentration and Moment Inequalities for Heavy-Tailed Random Matrices**
Authors: Jirak, Moritz, Minsker, Stanislav, **Shen, Yiqiu**, and Martin Wahl
Submitted to Probability and Related Fields. Minor Revision

- Derived sharp, finite-sample bounds for sums of heavy-tailed random matrices, covering sample covariance and kernel operators.
- Supplied rigorous performance guarantees that stay reliable in the extreme, heavy-tailed scenarios frequently encountered in practice (e.g., financial returns, network-traffic spikes).
- Demonstrated rigorous mathematical thinking and the ability to navigate abstract theoretical concepts while clearly communicating their impact to non-technical audiences.

Skills

Languages	Python, R, MATLAB, SQL (PostgreSQL), Java
ML Frameworks	PyTorch, TensorFlow, Keras
Software	LaTeX, Microsoft Excel, SAS JMP, Mathematica, Tableau, PostGIS, FlamMap
Languages	English (Proficient), Japanese (JLPT N2), Chinese (Native)
Other Skills	Webscraping, Linux, Teaching, Technical Communication, AI-assisted software engineering, Git