Kevin Lin (Post-doctoral researcher in statistics and genomics)

Website: https://linnykos.github.io/ Mobile : (848) 466-6541
Email : kevinL1@wharton.upenn.edu

Research goal: Developing data integration statistical methods to aggregate information from different modalities, with a proclivity towards matrix factorization or networks ideas, in order to investigate biological mechanisms (e.g., epigenetic priming in neurogenesis or therapy resistance in cancer systems respectively) at single-cell resolution.

Education

- **University of Pennsylvania** Philadelphia, PA
  Post-doctoral researcher in Wharton Statistics & Data Science 2020 - Present
  Advisor: Nancy Zhang

- **Carnegie Mellon University** Pittsburgh, PA
  Ph.D. in Statistics & Data Science, Masters in Machine Learning 2014 - 2020
  Thesis advisors: Kathryn Roeder and Jing Lei
  Thesis title: "High-dimensional statistical methods to model heterogeneity in genomic data"

- **Princeton University** Princeton, NJ
  B.S.E. in Operations Research & Financial Engineering 2010 - 2014
  Advisors: Han Liu and Robert Vanderbei
  Certificates in "Statistics and Machine Learning" and "Applications of Computing", graduated with Honors

Preprints

1. **Lin, K.** and Zhang, N. R. (2022). Quantifying common and distinct information in single-cell multimodal data with Tilted-CCA. *bioRxiv preprint*
   bioRxiv: 2022.10.07.511320

2. **Lin, K.**, Qiu, Y., and Roeder, K. (2022). eSVD: Cohort-level differential expression in single-cell RNA-seq data using exponential-family embeddings
   Link: https://linnykos.github.io/papers/cohort_eSVD.pdf

3. **Lin, K.** and Lei, J. (2022). Spectral clustering for heterophilic stochastic block models with time-varying node memberships
   Link: https://linnykos.github.io/papers/dynamicSBM.pdf

4. Guan, P. Y., Lee, J. S., Wang, L., **Lin, K.**, Mei, W., and Jiang, Y. (2022). Destin2: Integrative and cross-modality analysis of single-cell chromatin accessibility data. *bioRxiv preprint*
   bioRxiv: 2022.11.04.515202

Publications (Reverse chronological order)

Note: $(*)$ denotes equal-contribution first authorship.

1. Lei, J. and **Lin, K.** (2022). Bias-adjusted spectral clustering in multi-layer stochastic block models. *Journal of the American Statistical Association*, pages 1–13
   DOI: 10.1080/01621459.2022.2054817, Arxiv: 2003.08222

2. Field, A., Park, C. Y., **Lin, K.**, and Tsvetkov, Y. (2022). Controlled analyses of social biases in Wikipedia bios. In *Proceedings of the ACM Web Conference 2022*, pages 2624–2635
DOI: 10.1145/3485447.3512134, Arxiv: 2101.00078

3. **Lin, K.**, Lei, J., and Roeder, K. (2021a). Exponential-family embedding with application to cell developmental trajectories for single-cell RNA-seq data. *Journal of the American Statistical Association*, 116(534):457–470
DOI: 10.1080/01621459.2021.1886106, Pubmed: 34354320

4. **Lin, K.**, Liu, H., and Roeder, K. (2021b). Covariance-based sample selection for heterogeneous data: Applications to gene expression and autism risk gene detection. *Journal of the American Statistical Association*, 116(533):54–67
DOI: 10.1080/01621459.2020.1738234, Pubmed: 33731968

5. Hyun, S., **Lin, K.**, G'Sell, M., and Tibshirani, R. J. (2021). Post-selection inference for changepoint detection algorithms with application to copy number variation data. *Biometrics*, 77(3):1037–1049
DOI: 10.1111/biom.13422, Pubmed: 33434289

6. Wang, D., Zhao, Z., **Lin, K.**, and Willett, R. (2021). Statistically and computationally efficient changepoint localization in regression settings. *Journal of Machine Learning Research*, 22:248–1
DOI: 10.5555/3546258.3546506, Arxiv: 1906.11364

7. Lei, J. and **Lin, K.** (2020). Discussion of 'Network cross-validation by edge sampling'. *Biometrika*, 107(2):285–287
DOI: 10.1093/biomet/asaa009

8. An, J.-Y.*, **Lin, K.***, Zhu, L.*, Werling, D. M.*, Dong, S., Brand, H., Wang, H. Z., Zhao, X., Schwartz, G. B., Collins, R. L., Currall, B. B., Dastmalchi, C., Dea, J., Duhn, C., Gilson, M. C., Klei, L., Liang, L., Markenscoff-Papadimitriou, E., Pochareddy, S., Ahituv, N., Buxbaum, J. D., Coon, H., Daly, M. J., Shin Kim, Y., Marth, G. T., Neale, B. M., Quinlan, A. R., Rubenstein, J. L., Sestan, N., State, M. W., Willsey, A. J., Talkowski, M. E., Devlin, B., Roeder, K., and Sanders, S. J. (2018). Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science*, 362(6420)
DOI: 10.1126/science.aat6576, Pubmed: 30545852

9. **Lin, K.**, Sharpnack, J., Rinaldo, A., and Tibshirani, R. J. (2017). A sharp error analysis for the fused lasso, with application to approximate changepoint screening. In *Advances in Neural Information Processing Systems*, pages 6884–6893
DOI: 10.5555/3295222.3295432, Arxiv: 1606.06746

10. Vanderbei, R., **Lin, K.**, Liu, H., and Wang, L. (2016). Revisiting compressed sensing: Exploiting the efficiency of simplex and sparsification methods. *Mathematical Programming Computation*, 8(3):253–269
DOI: 10.1007/s12532-016-0105-y

Articles

1. **Lin, K.** (2017). We, the millennials: The statistical significance of political significance. *Significance*, 14(5):28–33
DOI: 10.1111/j.1740-9713.2017.01073.x

Teaching Experience

- **36-750: Statistical Computing** — Carnegie Mellon University (CMU)
  Guest lecturer for Alexander Reinhart — Fall 2016, Fall 2017, Fall 2018, Fall 2019, Fall 2020
  For PhD students, with lecture "Coding practices: Using R packages and GitHub to develop sustainable codebases." (1 lecture per semester)
  - Lecture about the importance of unit testing, GitHub, and other tools provided by RStudio relevant for developing and maintaining a code-base for statistical projects

- **36-469: Statistical Genomics and High Dimensional Inference** — CMU
  Co-instructor with Kathryn Roeder — Spring 2020
  For upper-level undergraduates & Master students
  - Course about foundational biological questions and how they have been addressed using statistical tools
  - Primarily responsible for designing homeworks that 1) had students analyze genomic datasets to demonstrate the biological and statistical concepts covered in lecture, 2) had students do simulation studies to demonstrate the math principles behind the estimators, and 3) was accessible to a broad audience, as the students had varying degrees of biological, statistical, and coding backgrounds

- **36-490: Undergraduate Research** — CMU
  Data science initiative project fellow under Rebecca Nugent and Peter Freeman — Spring 2019
  For upper-level undergraduates

- **36-350: Statistical Computing** — CMU
  Instructor — Summer 2018
  For entry-level undergraduates
  - Course about the basics of coding in R that would be foundational to future statistical courses in the curriculum
  - Updated more nebulous topics such as unit testing with the intent of having students naturally appreciate the importance of unit testing by debugging an involved algorithm rather than only solving unit testing related exercises

- **36-350: Statistical Computing** — CMU
  Assistant instructor with Ryan J. Tibshirani — Spring 2018
  For entry-level undergraduates

- **36-350: Statistical Computing** — CMU
  Teaching assistant under Peter Freeman — Fall 2017
  For entry-level undergraduates

- **36-350: Statistical Computing** — CMU
  Teaching assistant under Ryan J. Tibshirani — Fall 2016, Fall 2015
  For entry-level undergraduates

- **36-217: Probability Theory and Random Processes** — CMU
  Teaching assistant under Alessandro Rinaldo — Spring 2015
  For entry-level undergraduates

- **46-921 & 46-923: Financial Data Analysis I and II** CMU
  Teaching assistant under Chad Schafer Spring 2014
  For Master business students

- **ORF 350: Analysis of Big Data** Princeton University
  Course designer with Han Liu Spring 2014, Spring 2013, Spring 2012
  For upper-level undergraduates
  - Course about the four main categories of machine learning for big data (regression, classification, dimension-reduction, and clustering) and their relation to the maximum-likelihood principle
  - Primarily responsible for designing homeworks that had students 1) analyze real-life datasets to demonstrate the effectiveness of methods taught in lecture, and 2) perform simulation studies to demonstrate the math concepts taught in lecture

Mentoring Experience

I've had the pleasure and opportunity to mentor undergraduates during my Ph.D. These experiences were a great way for me to contribute directly to the community and to guide undergraduates in applying their coursework knowledge in more unstructured settings, as well as get hands-on mentoring experience.

- **Taewan Kim** University of Chicago
  Masters in Statistics Summer 2020-Present
  Working on: Dependency diagnostic: Visually understanding pairwise variable relationships for single-cell RNA-seq data
  - Mentored starting when he was an undergraduate student (CMU, senior in Statistics & Data Science) after his interest in genomics after taking my course "36-469: Statistical Genomics and High Dimensional Inference"
  - Guided him through our research project while he was a Masters student, and our paper is currently in preparation for submission

- **Julie Kim, Sophia Wen, Jae Won Yoon, Wanhe Zhao** CMU
  Senior undergraduates in Statistics & Data Science Spring 2019
  Title: Utilizing infant EEG brain patterns to predict childhood ADHD
  - Mentored as a Data science initiative project fellow for "36-490: Undergraduate Research" in collaboration with Cassie Eng and Anna Fisher (CMU Psychology), organized by Peter Freeman
  - Yielded poster presentation at Meeting of the Minds (CMU, 2019)

- **Grace Cao, Steve Kim, Eric Shi, Theo Yannekis** CMU
  Senior undergraduates in Statistics & Data Science Spring 2019
  Title: Do streamlined books improve young students' reading comprehension?
  - Mentored as a Data science initiative project fellow "36-490: Undergraduate Research" in collaboration with Cassie Eng and Anna Fisher (CMU Psychology), organized by Peter Freeman

- Yielded poster presentation at Meeting of the Minds (CMU, 2019)

- **Amy Tian** Princeton
  Senior undergraduate in Operations Research & Financial Engineering Spring 2017
  Title: A high-dimensional visualization system with applications to portfolio selection
  - Mentored for the undergraduate senior thesis research, organized by Han Liu
  - Yielding thesis (146 pages) for completion of Bachelor of Science in Engineering degree

- **Mark Aksen** Princeton
  Senior undergraduate in Mathematics Spring 2017
  Title: A study of functional connectivity for schizophrenia using a Gaussian graphical model
  - Mentored for independent work as part of the Program in Applied & Computational Mathematics, organized by Han Liu
  - Yielded talk presentation for Program in Applied & Computational Mathematics (Princeton, 2017)

- **Felix Xiao** Princeton
  Senior undergraduate in Operations Research & Financial Engineering Spring 2016
  Title: Approaches to brain parcellation using energy statistics and graph partitioning
  - Mentored for the undergraduate senior thesis research, organized by Han Liu
  - Yielded thesis (92 pages) for completion of Bachelor of Science in Engineering degree

Honors and Awards

- **PhD TAs of the year** Carnegie Mellon University
  For the Spring 2020 semester May 2020
  1 of 2 total recipients

- **Honorable mention in student paper competition** American Statistical Association
  For "Dependency diagnostic: Visually understanding pairwise variable relationships" January 2018
  For ASA section: Statistical Computing and Statistical Graphics

- **Winner of Statistical excellence for early-career writing** Significance magazine
  For article "We, the millennials: The statistical significance of political significance" June 2017
  Competition held jointly with the Young Statisticians Section of Royal Statistical Society

- **Teaching assistant excellence award recipient** Carnegie Mellon University
  For article 36-350: "Statistical Computing" in Fall 2017 May 2017
  1 of 5 total recipients

- **Award recipient of Kenneth H. Condit Prize** Princeton University
  For excellence in service to department May 2014

Talks and posters

Invited talks:

- **UCLA Department of Statistics: Seminar Series** (Remote)
  Exponential-family embedding for single-cell data with applications to developmental trajectories 2021

- **Joint Statistical Meetings** (Remote)
  Exponential-family embedding with application to cell developmental trajectories for single-cell data 2020
  For session: Analysis of single-cell RNA-seq data
  Delivered jointly with Kathryn Roeder

- **StatScale Seminar** (Remote)
  Time-varying stochastic block models, with application for dynamics of gene co-expression networks 2021

Talks:

- **Joint Statistical Meetings** Washington DC
  Tilted-CCA: Quantifying common and distinct information in multiomic single-cell data 2022
  For session: Novel approaches for omics and multi-omics analysis

- **Symposium on Data Science and Statistics** Pittsburgh, PA
  Spectral clustering for multi-layer stochastic block models: Analysis of dynamic heterophilic networks 2022
  For session: Time analyses

- **Joint Statistical Meetings** (Remote)
  Time-varying stochastic block models via kernel smoothing, with application to RNA-seq data 2020
  For session: Statistical methods in gene expression data analysis I

- **Joint Statistical Meetings** Denver, Colorado
  Exponential-family embedding with application to cell developmental trajectories for single-cell data 2019
  For session: Statistical methods for single-cell genomics

- **Joint Statistical Meetings** Vancouver, Canada
  Dependency diagnostic: Visually understanding pairwise variable relationships 2018
  For session: A mixed bag of graphical delights

- **Joint Statistical Meetings** Baltimore, MD
  Hypothesis testing for simultaneous variable clustering and correlation network estimation 2017
  For session: Selected topics on hypothesis testing and statistical inference

- **Joint Statistical Meetings** Chicago, IL
  Longitudinal Gaussian graphical model for autism risk gene detection 2016
  For session: Network and graphical models for analysis of genomic data

- **Modeling and Optimization: Theory and Applications** Bethlehem, PA
  Optimization for compressed sensing: New insights and alternatives 2014
  For session: Algorithms for big data

Posters:

- **American Society of Human Genetics** (Remote)
  Exponential-family embedding with application to cell developmental trajectories for single-cell data 2020
- **Conference on Neural Information Processing Systems** Long Beach, CA
  A sharp error analysis for the fused lasso, with application to approximate changepoint screening 2017
- **American Society of Human Genetics** Baltimore, MD
  Gaussian graphical model integrating microarray and sequencing data for autism risk gene detection 2015

Professional Service

In addition to being research leaders and educators, professors are also important members of the department's community. Hence, I value mental health training since I strive to help students experiencing emotional turbulence when juggling classes, research, and personal growth.

- Certified by Mental Health First Aid USA (Fall 2020)
- Certified by CMU's Eberly Center's Future Faculty Program, which included two observed lectures in two different semesters (Fall 2019 to Summer 2020)
- Certified with Gatekeeper certificate by the QPR's (Question, Persuade, Refer) suicide prevention program (February 2020)
- Founder and organizer for "Statistical Inference" reading group for PhD students in the Statistics & Data Science department (2017-2018)
- Founding member of Carnegie Mellon University's Statistics and Data Science department's Wellness Network (2018-2020)
- Association member of American Statistical Association, and American Society of Human Genetics
- Reviewer for:
  - Annals of Applied Statistics
  - Annals of Statistics
  - Biometrika
  - Electronic Journal of Statistics
  - IEEE Transactions of Network Science and Engineering
  - Journal of Molecular Biology
  - Journal of American Statistical Association
  - Nature Neuroscience (as a code reviewer)
  - PLOS Genetics
  - Statistical Sinica
  - Statistics and Probability Letters
  - Statistics in Medicine
  - Technometrics