SEYOON KO

https://kose-y.github.io
kos@ucla.edu, y3kkoseyoon@gmail.com

RESEARCH INTERESTS

- Data-driven discovery in science, particularly in Bioinformatics
- High-dimensional statistical methodology
- Optimization and computational methods in Statistics
- Parallel computing for statistical analysis of large-scale data on GPU and CPU clusters

ACADEMIC POSITIONS

Assistant Adjunct Professor, University of California, Los Angeles *Jul 2023-*
Department of Mathematics

Postdoctoral Scholar, University of California, Los Angeles *Nov 2020-Jun 2023*
Department of Computational Medicine / Biostatistics
Supervisors: Kenneth L. Lange and Hua Zhou

EDUCATION

Ph.D., Statistics *Aug 2020*
Seoul National University, Seoul, Korea
Advisor: Joong-Ho (Johann) Won
Dissertation: Easily parallelizable statistical computing methods and their applications in modern high-performance computing environments

M.S., Computational Science & Technology *Aug 2014*
Seoul National University, Seoul, Korea
Advisor: Sun Kim
Thesis: Flow network model for detection and quantification of gene fusion

B.S., Physics, Mathamatical Sciences, and Computational Sciences *Feb 2012*
Seoul National University, Seoul, Korea *(triple major)*
Cum Laude

HONORS AND AWARDS

QCBio Collaboratory Fellowship *2021-current*
Institute for Quantitative & Computational Biosciences, UCLA
$25,000/yr fellowship for conducting quarterly workshop and consulting for computational biosciences

IDRE Fellowship *2022-23*
Institute for Digital Research & Education, UCLA
$5,000 fellowship for advanced research computing, GPU and many core-based computing architectures, statistical computing and data science and informatics

Postdoctoral Fellowship Program (Fostering Next-generation Researchers) *2020-21*
National Reserach Foundation of Korea
$40,000 award for the most promising postdoctoral researchers to undergo training at overseas research institutes

Best Doctoral Thesis Award of College of Natural Sciences *2020*
Seoul National University

AWS Cloud Credits for Research (formerly AWS Research Grants) *2018-2019*
Amazon Web Services
$20,000 credits for development of easy-to-use cloud and HPC tools for statisticians.
Co-PI

National Scholarship for Science and Engineering *2008-2012*
Korea Student Aid Foundation

PREVIOUS EXPERIENCES

Research Assistant, Seoul National University *Sep 2014-Aug 2020*
Computational Statistics Research Group
High-performance computing for large-scale statistical computation

Visiting Scholar, University of California, Santa Barbara *Jun-Oct 2018*
Department of Statistics and Applied Probability
Analyzed Human Connectome Project (HCP) data with graphical model and inverse covariance estimation
Supervisor: Sang-Yun Oh

NERSC/Computational Research Division, Lawrence Berkeley National Laboratory
Jul-Sep 2015

Analyzed data from Daya Bay Antineutrino Detectors using deep learning on a supercomputer
Supervisor: Sang-Yun Oh

Research Assistant, Seoul National University *Sep 2012-Aug 2014*
Bio and Health Informatics Lab.
Detection and quantification of gene-fusion transcript model using RNA-seq data via optimization

TEACHING EXPERIENCE

Instructor
Program in Computing, Department of Mathematics, UCLA *2023-*
Undergraduate programming courses for students in the non-engineering field

- PIC 16B Python with Applications II (Fall 2023 – 2 lectures, Winter 2024)
  Topics include database queries, interactive visualization, web crawling, webapp development, deep learning libraries, just-in-time compilation, and multithreading.
- PIC 16A Python with Applications I (Winter 2024)
  Topics include Python basics, object-oriented programming, numerical computation, visualization, data wrangling, and machine learning.

Master of Data Science in Health Program, UCLA *2024- (scheduled)*

- BIOSTAT 203C Introduction to Data Science in Python (Spring 2024 (scheduled))
  Scheduled to develop a new course in data science for the new professional degree program.

Instructor, QCBio Collaboratory Workshop *2021-*
Institute for Quantitative and Computational Biosciences, UCLA

Quarterly 3-hour, 3-day workshop series tailored to individuals in the biosciences community who are interested in learning data analysis, programming and statistical techniques. Students can earn credits for BIOINFO 275A/B by taking multiple workshops.

- Intro to Python (three times in 2022-23)
- Machine Learning with Python (five times in 2021-22, 2023-24)

Guest Lecturer
BIOSTAT M272, UCLA *Mar 15, 2023*
Unsupervised discovery of ancestry informative markers and genetic admixture proportions in biobank-scale data sets

Cloud Administrator, Julia for Data Science and Statistical Computing *Aug 9, 2022*
Short Course at Joint Statistical Meetings 2022
Configured and managed a virtual cluster on Google Cloud running Jupyter Notebooks and Pluto

Instructor / Cloud Administrator, Biomedical Data Science Workshop *Jul 18, 2022*
Lange Symposium Workshop, UCLA
A satellite event of the Lange Symposium. Configured and managed a virtual cluster on Google Cloud for 50 students running Jupyter Notebooks with a ∼10 GB dataset shared among the students.

- Data wangling and visualization with Python
- Easy manipulation of genetic variant data in Julia

Student Lecturer, First Year Graduate Student Seminar *2016-2019*
Department of Statistics, Seoul National University

- Using a cluster computer with Slurm and Rslurm (March 2019)
- Git and GitHub (March 2016, March 2017)

Teaching Assistant *2014-2019*
Department of Statistics, Seoul National University

- Advanced Statistical Computing (Fall 2016, Fall 2019)
  Graduate-level course on convex optimization and computation methods in Statistics. Graded homeworks using Julia. Course managed on GitHub.
- Computational Statistics (Spring 2015, Spring 2016, Spring 2017)
  Senior-level course on computational methods in Statistics. Graded homeworks/exams. Homeworks managed on GitHub.
- Statistical Computing and Lab. (Fall 2014, Fall 2015)
  Freshman programming course in R and C. Developed and managed course materials on GitHub.

TALKS

- Estimation of Genetic Admixture Proportions via Haplotypes. In *American Society of Human Genetics 2023 Annual Meeting*, Washington D.C., November 1–5, 2023. (Poster)
- Intro to Julia: A fast dynamic language for statistical computing and data science. UCLA Institute for Digital Research and Education Early Career Researchers Meeting, October 5, 2023. (Invited)
- Panel discussion on Advanced Research Computing (ARC) applications. UCLA Institute for Digital Research and Education Early Career Researchers Meeting, January 27, 2023. (Invited)
- Unsupervised discovery of ancestry informative markers and genetic admixture proportions in biobank-scale data sets. UCLA Institute for Digital Research and Education Early Career Researchers Meeting, January 6, 2023. (Invited)

- Unsupervised discovery of ancestry informative markers and genetic admixture proportions in biobank-scale data sets. In *American Society of Human Genetics 2022 Annual Meeting*, Los Angeles, California, October 25–29, 2022. (Poster)
- GWAS of longitudinal trajectories at biobank scale. Department of Statistics. Seoul National University, June 14, 2022. (Invited)
- A Distributed Matrix Data Structure and Its Statistical Applications on PyTorch. In Software Workshop of *Lange Symposium 2020*, University of California, Los Angeles, CA, February 20-21, 2020. *Jointly with Johann (Joong-Ho) Won.* [Jupyter Notebook] (Invited)
- Optimal Minimization of the Sum of Three Convex Functions with a Linear Operator. In *The 22nd International Conference on Artificial Intelligence and Statistics*, Okinawa, Japan, April 16–18, 2019. (Poster)
- Multi-GPU distributed statistical computing using deep learning libraries, Department of Statistics and Applied Probability, University of California, Santa Barbara, CA, August 8, 2018. [Invited]
- Fast and accurate reconstruction for susceptibility source separation in QSM. In *Joint Annual Meeting ISMRM-ESMRMB 2018*, Paris, France, June 16-21, 2018. (Poster)
- A feature-splitting distributed algorithm for generalized linear models under generalized and group lasso penalties. In *9th International Conference of the ERCIM Working Group on Computational and Methodological Statistics*, Seville, Spain, December 9-11, 2016. (Oral presentation)

PUBLICATIONS

Peer-reviewed Papers

[1] Chu, B. B., **Ko, S.**, Zhou, J. J., Jensen, A., Zhou, H., Sinsheimer, J. S., Lange, K. (2023). Multivariate Genomewide Association Analysis by Iterative Hard Thresholding. *Bioinformatics*, **39**(4), btad193. [paper][Julia code]

[2] **Ko, S.**, Chu, B. B., Peterson, D., Okenwa, C., Papp, J. C., Alexander, D. H., Sobel, E. M., Zhou, H., Lange, K. (2023). Unsupervised discovery of ancestry informative markers and genetic admixture proportions in biobank-scale data sets. *American Journal of Human Genetics*, **109**(3), pp. 433-445. [paper][Julia code]

[3] **Ko, S.**, Zhou, H., Zhou, J., and Won, J.-H. (2022). High-Performance Statistical Computing in the Computing Environments of the 2020s. *Statistical Science*, **37**(4), pp. 494–518. [Arxiv preprint] [Python code, PyTorch, dask]

[4] Kim J., Jensen, A., **Ko, S.**, Raghavan, S., Phillips, L. S., Hung, A., Sun, Y., Zhou, H., Reaven, P., Zhou, J. J. (2022), Systematic Heritability and Heritability Enrichment Analysis for Diabetes Complications in UK Biobank and ACCORD Studies, *Diabetes*, **71**(5), pp. 1137–1148.

[5] *__Ko, S.__, *German, C., Jensen, A., Shen, J., Wang, A., Mehrotra, D. V., Sun, Y. V., Sinsheimer, J. S., Zhou, H., Zhou, J. J. (2022), GWAS of longitudinal trajectories at biobank scale, *American Journal of Human Genetics*, **109**(3), pp. 433–445. [Julia code]

[6] Chu, B. B., Sobel, E. M., Wasiolek, R., **Ko, S.**, Sinsheimer, J. S., Zhou, H., Lange, K. (2021), A Fast Data-Driven Method for Genotype Imputation, Phasing, and Local Ancestry Inference: MendelImpute.jl. *Bioinformatics*, **37**(24), pp. 4756–4763.

[7] **Ko, S.**, Li, G. X., Choi, H. and Won, J.-H. (2021). Computationally scalable regression modeling for ultrahigh-dimensional omics data with ParProx. *Briefings in Bioinformatics*, **22**(6), bbab256. [Julia code]

[8] Kim, J., Shen, J., Wang, A., Mehrotra, D. V., **Ko, S.**, Zhou, J. J., Zhou, H. (2021), VCSEL: Prioritizing SNP-Set by Penalized Variance Component Selection, *Annals of Applied Statistics*, **15**(4), pp. 1652–1672.

[9] Ryu, E. K., **Ko, S.**, Won, J.-H. (2020). Splitting with Near-Circulant Linear Systems: Applications to Total Variation CT and PET. *SIAM Journal on Scientific Computing*, **42**(1), pp. B185–B206. [Matlab code]

[10] Zhou, H., Sinsheimer, J. S., Bates, D. M., Chu, B. B., German, C. A., Ji, S. S., Keys, K. L., Kim, J., **Ko, S.**, Mosher, G. D., and Papp, J. C. (2020). OpenMendel: A cooperative programming project for statistical genetics. *Human Genetics*, **139**(1), pp. 61–71. [project homepage]

[11] **Ko, S.**, Yu, D., Won, J.-H. (2019). Easily parallelizable and distributable class of algorithms for structured sparsity, with optimal acceleration. *Journal of Computational and Grahpical Statistics*, **28**(4), pp. 821–833. [Python code, TensorFlow, Docker]

[12] **Ko, S.** and Won, J. H. (2019). Optimal Minimization of the Sum of Three Convex Functions with a Linear Operator. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1185–1194.

[13] Racah, E., **Ko, S.**, Sadowski, P., Bhimji, W., Tull, C., Oh, S.-Y., Baldi, P., and Prabhat (2016). Revealing fundamental physics from the Daya Bay Neutrino Experiment using deep neural networks, In *2016 15th IEEE International Conference on Machine Learning and Applications*, pp. 892–897.

[14] **Ko, S.** and Won, J.-H. (2016). Processing large-scale data with Apache Spark. *Korean Journal of Applied Statistics*, **29**(6), pp. 1077–1094.

[15] **Ko, S.**, Jun, G., and Won, J.-H. (2016). HyperConv: Spatio-spectral classification of hyperspectral images with deep convolutional neural networks. *Korean Journal of Applied Statistics*, **29**(5), pp. 859–872. [Python code, Theano]

[16] Kim, J., Kim, S., **Ko, S.**, In, Y. H., Moon, H. G., Ahn, S. K., Kim, M. K., Lee, M., Hwang, J. H., Ju, Y. S., Kim, J. I., Noh, D. Y., Kim, S., Park, J. H., Rhee, H., Kim, S., and Han, W. (2015). Recurrent fusion transcripts detected by whole transcriptome sequencing of 120 primary breast cancer samples. *Genes, Chromosomes and Cancer*, **54**(11), pp. 681–691.

[17] Kim, Y., Kang, Y. S., Lee, N. Y., Kim, K. Y., Hwang, Y. J., Kim, H. W., Rhyu, I. J., Her, S., Jung, M. K., Kim, S., Lee, C. J., **Ko, S.**, Kowall, N. W., Lee, S. B., Lee, J., and Ryu, H. (2015). Uvrag targeting by Mir125a and Mir351 modulates autophagy associated with Ewsr1 deficiency. *Autophagy*, **11**(5), pp. 796–811.

Working Papers and Preprints

[I] **Ko, S.**, Suchard, M., and Holbrook, A. (2024+). Analyzing millions of SARS-CoV-2 cases with spatiotemporal Hawkes processes.

[II] **Ko, S.**, Zhou, H., Sobel, E., and Lange, K. (2024+). Accurate estimation of genetic admixture proportions via haplotype modeling.

[III] **Ko, S.**, Zhou, H., Zhou, J., and Won, J.-H. (2024+). DistStat.jl: Towards unified programming for high-performance statistical computing environments in Julia. Under revision. [Arxiv preprint] [Julia code]

Conference Abstracts

[i] **Ko, S.**, Sobel, E., Zhou, H., and Lange, K., Unsupervised Learning of Ancestry Informative Markers and Genetic Admixture Coefficients From Biobank Data. In *Annual Meeting of American Society of Human Genetics 2022*, Los Angeles, CA, USA, October 25-29, 2022.

[ii] **Ko, S.**, Lee, J., Won, J.-H., and Lee, J., Fast and accurate reconstruction for susceptibility source separation in QSM. In *Joint Annual Meeting ISMRM-ESMRMB 2018*, Paris, France, June 16-21, 2018.

[iii] Bhimji, W., Racah, E., **Ko, S.**, Sadowski, P., Tull, C., and Oh, S.-Y., Exploring Raw HEP Data using Deep Neural Networks at NERSC. In *38th International Conference on High Energy Physics*, Chicago, IL, USA, August 3-10, 2017.

[iv] **Ko, S.**, Yu, D., and Won, J.-H., A feature-splitting distributed algorithm for generalized linear models under generalized and group lasso penalties. In *9th International Conference of the ERCIM Working Group on Computational and Methodological Statistics*, Seville, Spain, December 9-11, 2016.

[v] Lee, H. B., Han, W., **Ko, S.**, Kim, M. S., Lim. S., Lee, K. M., Kang, Y. J., Han, J. H., Kim, Y., Yoo, T. K., Moon, H. G., Noh, D. Y., and Kim, S., Identification of ESR splice variants associated with prognosis in estrogen receptor positive breast cancer. In *Thirth-Eighth Annual CTRC-AACR San Antonio Breast Cancer Symposium*, San Antonio, TX, USA, December 8-12, 2015.

PROFESSIONAL SERVICE

Invited reviewer: *Journal of Computational and Graphical Statistics, Journal of the Royal Statistical Society: Series C, Annals of Applied Statistics, Computational Statistics and Data Analysis, Bioinformatics*
Reviewer: *Journal of Open Source Software*

OPEN SOURCE CONTRIBUTIONS

- `MPI.jl`: Added CUDA-aware MPI support to make HPC programming easier in a hybrid CPU-GPU environment in Julia. [CUDA support announcement with acknowledgement] [code examples]
- `OpenMendel` ecosystem: Julia project for statistical genomics. [project homepage]
  - Mainly contributed to `TrajGWAS.jl` and utilities for handling compressed storage of SNP data: `SnpArrays.jl` (PLINK 1 BED format), `PGENFiles.jl` (PLINK 2 PGEN format), and `BGEN.jl` (Oxford BGEN format).
  - Maintaining 20+ packages from the group

COMPUTER SKILLS

- Programming languages: C, C++, Julia, Python, Matlab, and R
- Linux operating system
- Deep learning and machine learning packages: TensorFlow, PyTorch, and JAX
- Developed I/O packages for genetic file-formats including PLINK 1 BED, PLINK 2 PGEN, and Oxford BGEN
- Developed packages involving MPI and CUDA
- High-performance computing on supercomputers and on a cloud
- Managed JupyterHub on shared workstations and a cloud via Kubernetes for research and teaching
- Working knowledge of Dask, Docker, Apache Spark, Hadoop/MapReduce