MINGXUAN CAI

Department of Mathematics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
(+852) 5324 2681 || mcaiad@ust.hk
Home page: https://mxcai.github.io/

EDUCATION

Ph.D., Statistics, 09/2018 – 6/2022 (expected)
Department of Mathematics, The Hong Kong University of Science and Technology
Advisor: Prof. Can Yang
M.Phil., Statistics, 09/2016 – 06/2018
Department of Mathematics, Hong Kong Baptist University
B.S., Statistics and Operation Research, 09/2012 – 06/2016
Department of Mathematics, Hong Kong Baptist University

HONORS

- **Postgraduate Research Excellence Award**, School of Science, HKUST, 2020. This award is present to research postgraduate students with outstanding research achievements.
- **Din-Yu Hsieh Teaching Award**, Department of Mathematics, HKUST, 2021. This award is in appreciation of the excellent teaching performance of teaching assistants.

RESEARCH INTEREST

- Statistical machine learning with application in genomics data
- Variance components and mixed-effects models
- Scalable algorithms in statistical machine learning

PUBLICATIONS

- **Cai, M.***, Xiao, J.*, Zhang, S.*, Wan, X., Zhao, H., Chen, G., Yang, C. (2021). A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. *The American Journal of Human Genetics*, 108(4): 632-655. We present a unified statistical framework (XPA) to improve the prediction accuracy of human traits using multi-ancestry genetic data. Paired with innovations in data structure and algorithm design, our framework is highly scalable, with both computational cost and memory storage linear to the sample size and number of predictors. In practice, XPA can analyze 3 million variants from 430K samples with only 385 Gb memory usage in 54.5 hours. In a Chinese cohort, our method achieves 7.3%-198.0% accuracy gain for height prediction in terms of R^2 compared to existing methods.
- Xiao, J.*, **Cai, M.***, Hu, X., Wan, X., Chen, G., Yang, C. (2022). XPXP: Improving polygenic prediction by cross-population and cross-phenotype analysis. *Bioinformatics*, in press. We propose a cross-population and cross-phenotype (XPXP) method to construct accurate PRSs by leveraging biobank-scale datasets in European populations and multiple GWASs of genetically correlated phenotypes while allowing for incorporation of population-specific and phenotype-specific effects. We showed that the height PRSs constructed by XPXP achieved 12% and 18% improvement over the runner-up method in terms of predictive R^2 in East Asian and African populations, respectively. We

also showed that XPXP substantially improved the stratification ability in identifying individuals at high genetic risk of Type 2 Diabetes.

- **Cai, M.**, Chen, L., Liu, J., Yang, C. (2020). IGREX for quantifying the impact of genetically regulated expression on phenotypes. *NAR Genomics and Bioinformatics*, 2(1): lqaa010. Many genetic variants affect phenotypes by regulating the gene expression level. We develop a statistical model, IGREX, to quantify the impact of genetically regulated expression on various human traits and inform trait-relevant tissue types. Efficient parameter expanded EM (PX-EM) algorithm and Method of Moments are adopted to optimize computational efficiency.
- **Cai, M.**, Dai, M., Ming, J., Peng, H., Liu, J., Yang, C. (2019). BIVAS: A scalable Bayesian method for bi-level variable selection. *Journal of Computational and Graphical Statistics*, 29(1): 40-52. We develop a scalable Bayesian method for bi-level variable selection. Our approach combines variational inference and multi-thread computing to improve computational efficiency in large- scale dataset. In the application of genetic data, bivas can analyze large GWAS datasets and achieves almost the same estimation accuracy as MCMC based algorithm while using only 1% of its computational time.
- Ming, J., Dai, M., **Cai, M.**, Wan, X., Liu, J., Yang, C. (2018). LSMM: A statistical approach to integrating functional annotations with genome-wide association studies. *Bioinformatics*, 34(16): 2788-2796. We propose a latent sparse mixed model (LSMM) to integrate functional annotations with GWAS data. Not only does it increase the statistical power of identifying risk variants, but also offers more biological insights by detecting relevant functional annotations
- Dai, M., Ming, J., **Cai, M.**, Liu, J., Yang, C., Wan, X., Xu, Z. (2017). IGESS: A statistical approach to integrating individual level genotype data and summary statistics in genome wide association studies. *Bioinformatics*, 33(18): 2882-2889. We propose a statistical approach, IGESS, to increase statistical power of identifying risk variants and improving accuracy of risk prediction by integrating individual level genotype data and summary statistics. An efficient algorithm based on variational inference is developed to handle genome-wide analysis.

* represents equal contributions.

CONFERENCE PRESENTATIONS

- "BIVAS: A scalable Bayesian method for bi-level variable selection", Joint Meeting of 10th Asian Regional Section (ARS) of the International Association for Statistical Computing (IASC) and the NZ Statistical Association (NZSA). Dec 2017

STATISTICAL SOFTWARES

- **Softwares available on my GitHub page: https://github.com/mxcai**
- **VCM**: An efficiently implemented R package for variance components estimation. The software provides 3 algorithms for fitting the variance components model, including the Expectation-Maximization algorithm, the Minorization-Maximization algorithm, and the Method of Moments. Available at https://github.com/mxcai/VCM.
- **XPASS**: An R package for constructing genetic prediction of polygenic traits by leveraging cross-population datasets from GWAS summary data. Available at https://github.com/mxcai/ XPASS.
- **iGREX**: An R package for the IGREX model that provides efficient quantification of the impact of genetically regulated expression on complex traits and diseases. Available at https:// github.com/mxcai/iGREX.
- **bivas**: An R package for a scalable Bayesian bi-level variable selection model. Available at https://github.com/mxcai/bivas.

TEACHING EXPERIENCE

- Calculus II, Teaching Assistant at HKUST, Spring 2021.
- Machine Learning and its Applications, Teaching Assistant at HKUST, Spring 2020.
- Statistical Machine Learning (PG), Teaching Assistant at HKUST, Spring 2020.
- Applied Statistics, Teaching Assistant at HKSUT, Spring 2020.
- Statistical Machine Learning (UG), Teaching Assistant at HKUST, Fall 2019, Fall 2020.
- Sampling, Teaching Assistant at HKUST, Spring 2018.
- Mathematics for Personal Financial Management, Teaching Assistant at HKBU, Spring 2017.
- Discrete Mathematics, Teaching Assistant at HKBU, Fall 2017.