Machine Learning for large-scale genomics

Sriram Sankararaman

Computer Science, Human Genetics, Computational Medicine
UCLA
Genetic architecture of complex phenotypes

\[ \text{Phenotype} = f(\text{Genotype, Environment}) \]

Visscher et al. AJHG 2012
Growth of Biobanks
Machine Learning for Biobank-scale data

How can we learn about genetic architecture of complex traits and diseases from datasets that contain millions of genomes and thousands of traits?

Statistical

Privacy

Computational

Interpretability
(Narrow-sense) Heritability

\[ h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \]

\[ Y = X \beta + \epsilon \]

Genetic variance component

\[ \beta_m \sim \mathcal{N}(0, \frac{\sigma_g^2}{M}) \]

Environmental variance component

\[ \epsilon_n \sim \mathcal{N}(0, \sigma_e^2) \]
Beyond Heritability

Lee et al. 2012

Loh et al. 2015

Finucane et al. 2015

Bulk-Sullivan et al. 2015
Variance components model

\[ y = \sum_{k=1}^{K} X_k \beta_k + \epsilon \]

\[ \beta_k \sim \mathcal{N}(0, \frac{\sigma_k^2}{M_k} I_{M_k}), \quad k \in \{1, \ldots, K\} \]

\[ \epsilon \sim \mathcal{N}(0, \sigma_e^2 I_N) \]

Goal

Estimate variance components \((\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2, \sigma_e^2)\)
Estimating variance components

Maximum likelihood

\[
(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \hat{\sigma}_K^2, \hat{\sigma}_e^2) = \arg\max (\sigma_1^2, \ldots, \sigma_K^2, \sigma_e^2) \mathcal{L}(\sigma_1^2, \ldots, \sigma_K^2, \sigma_e^2) = \arg\max (\sigma_1^2, \ldots, \sigma_K^2, \sigma_e^2) \mathcal{L}(\sigma_1^2, \ldots, \sigma_K^2, \sigma_e^2) = \arg\max (\sigma_1^2, \ldots, \sigma_K^2, \sigma_e^2) \mathcal{L}(\sigma_1^2, \ldots, \sigma_K^2, \sigma_e^2)
\]

Computationally expensive

Scales as \(O(N^3)\)

Challenging to apply to Biobank-scale data

Zhou and Stephens, Nature Genetics 2012
Loh et al. Nature Genetics 2015
Randomized HE-regression (RHE-mc)

Combines randomization with a method-of-moments estimator

Work with a “sketch” of the genotype

Multiply the genotype matrix with B random vectors

Efficiency depends on B: \( O\left(\frac{MNB}{\log_3(\max(N, M))}\right) \)

Accurate for B as small as 10

Hutchinson 1989
Wu et al. Bioinformatics 2018
Pazokitoroudi et al. RECOMB 2019, Nature Communication 2020
Comparisons of RHE-mc

Relative bias

Runtime (hours)

Sample size (X 1000)

Pazokitoroudi et al. RECOMB 2019, Nature Communication 2020
Insights from Biobank-scale analysis

Heritability

Trait

Per-allele heritability

Minor allele frequency

LD

Height

Minor allele frequency

Low

High

Enrichment

Pazokitoroudi et al. Nature Communication 2020
What is the contribution of non-linear effects?

What is the contribution of environmental interactions?

How are genetic effects shared across traits?
Dominance deviation effects

\[ y = X\beta + D\gamma + \epsilon \]

\[ \beta \sim \mathcal{N}(0, \frac{\sigma_a^2}{M} I_M) \]

\[ \gamma \sim \mathcal{N}(0, \frac{\sigma_d^2}{M} I_M) \]

Heritability

Additive heritability

Dominance heritability

Hivert et al. AJHG (to appear)

Pazokitoroudi et al. AJHG (to appear)
Gene-environment interactions (GxE)

\[ y = X\beta + X \odot E\delta + \epsilon \]

\[ \delta \sim \mathcal{N}(0, \frac{\sigma_{GE}^2}{ML} I) \]

GxE variance component

E = Smoking

GxE heritability
Additive heritability

Kerrin and Marchini AJHG 2020
Pazokitoroudi et al. RECOMB 2021
Novel genetic correlation of coronary artery disease and serum liver enzyme
Scaling Machine Learning to Biobanks

Effectiveness of randomization

Approximate inference

Distributed inference
Promises and challenges
Multi-modal data

- Imaging
- Gene expression
- Genetics
- Biomarkers
- Sensors
- Health outcome
Causal inference

Genetic variant → Exposure → Outcome

Z → ? → Y

Katan et al. Lancet 1986
Cinelli et al. BioRxiv 2020
Generalizability

Martin et al. Nature Genetics 2019

Mostafavi et al. eLife 2020
Acknowledgments

Funding

[Logos of Microsoft Research, Alfred P. Sloan Foundation, Hellman Fellows Fund, NSF, NIH, and The Okawa Foundation]