Algorithmic framework for assessing and improving heritability models

Anubhav Kaphle[1], David Balding[1]
[1]Melbourne Integrative Genomics
The University of Melbourne, Parkville, Melbourne

Abstract

There is an ongoing debate regarding the best approach to model how heritability varies across the genome. Different groups have proposed different heritability models, and there is a lack of robust framework to assess which models are most realistic. Recently, Speed et al. proposed a new likelihood approximation for genome-wide regression models based on summary statistics, which provides a statistical framework for assessing heritability models. In this work, we have developed an efficient implementation of the statistical framework proposed by Speed et al., which employs a stochastic gradient descent-based algorithm called ADAM. The software can be used to estimate heritability parameter as well as heritability shared by different SNP categories for any assumed heritability models. We used the software to estimate these parameters for some UK biobank traits as example case. We also estimated heritability contributed by eQTLs and heritability enrichment across tissues using GTEx data. Our method also allows for fine-scale model selection based on huge datasets to identify traits-specific heritability model.

Keywords: Likelihood approximation, stochastic gradient descent, SNP heritability, model selection, UK Biobank, summary statistics

Introduction

Approximate loglikelihood for GWAS summary data is proposed as

$$\mathrm{logl} = \sum_j \frac{1}{u_j}\Big(-\frac{S_j}{2\mathbb{E}[S_j]} - \frac{1}{2}\log(S_j) - \frac{1}{2}\log(2\mathbb{E}[S_j]) - \frac{1}{2}\log(\pi)\Big), \text{ where } u_j = \sum_{l \text{ near } j} r_{jl}^2$$

where S_j is GWAS summary statistic for SNP j, and $\mathbb{E}[S_j]$ is related to heritability model as

$$\mathbb{E}[S_j] = 1 + N\Big(\sum_C \tau_c \sum_{l \text{ near } j,\, l\in c} r_{jl}^2 \times p_l^{1+\alpha} \times A_{lc}\Big)$$

r_{lj}^2 is SNP-SNP squared correlation, τ_c is effect of c SNP category on per-SNP heritability, A_{lc} is the annotation value of SNP l for category c, $p_l = [f_l(1-f_l)]$, f is the minor-allele fraction (MAF), α describes MAF-heritability relationship and can be interpreted as a measure of selection, and N is the sample size used to compute GWAS statistics. Existing heritability models differ by assumptions regarding α and the set of SNP categories used.

Parameters τ_c and α in the model can be estimated using MLE approach.

BLD-LDAK [1] is the most recent improved heritability model that assumes $\alpha = -0.25$ and has 67 different SNP categories mostly related to functional genomic annotations. BLD-LDAK-lite is lower dimensional model and have been shown to produce consistent heritability estimates as the full model.

Materials & Methods

we used ADAM algorithm to estimates parameters in the approximate joint loglikelihood, logl. ADAM [2] stands for Adaptive moment estimation is a stochastic-gradient descent-based algorithm. This algorithm computes adaptive learning rates for each parameter being optimized. Default values suggested by authors were used to estimate the moments of the noisy gradients. Learning rate was set at 0.05.

The implementation is done in Python (Py v3.7.0) programming language.

We use GTEx consortium [4] eQTLs data to analyze heritability contributed by the eQTLs as well as estimate enrichment of heritability across the 49 GTEx tissues based on these eQTLs.

To estimate enrichment across genomic functional region and GTEx tissues and to estimate selection-related parameter α, we used the lite version of the BLD-LDAK model.

Our tool allows to specify different heritability models and can be used to estimate enrichments across different SNP categories.

Results

UK biobank trait	h^2_{SNP} est. (ADAM)	SE	h^2_{SNP} est. (LDAK)	SE
Body Mass Index (BMI)	0.30	0.01	0.28	0.01
Forced vital capacity (FVC)	0.31	0.01	0.30	0.01
Systolic blood pressure (SBP)	0.18	< 0.01	0.17	0.01
Height	0.61	< 0.01	0.59	0.02
Neuroticism score (NEUR)	0.13	0.01	0.12	< 0.01

Table 1: Comparison between estimates of SNP heritability obtained from ADAM implementation and LDAK software that uses second-order Newton-Rhapson optimization method. Heritability model used is the recent BLD-LDAK model. The estimates are greater and more precise compared to LDAK outputs.

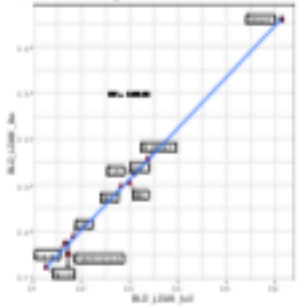

Fig 2: Comparison of the heritability estimates obtained from lite and full BLD-LDAK heritability models using our implementation. The estimates are consistent re-confirming results from [1]

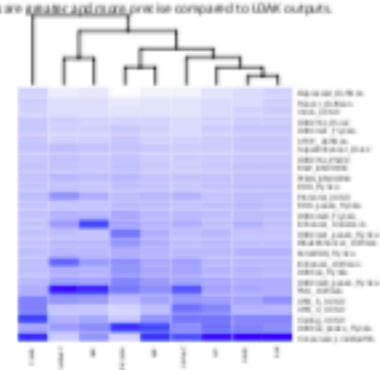

Fig 1: Heritability enrichment estimates across functional genomics categories. We use BLD-LDAK lite model to estimate enrichments.

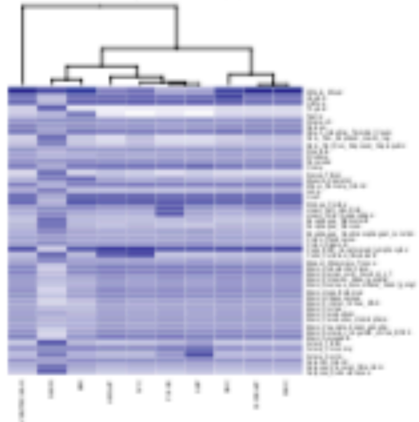

Figure 3: Heritability enrichment estimates across 49 GTEx tissues. Assumed heritability model here is BLD-LDAK-lite. We use BLD-LDAK lite model to estimate enrichment.

Trait	α estimated	LDAK α estimate
BMI	-0.20	-0.15
FVC	-0.25	-0.30
SBP	-0.35	-0.30
IMPEDENCE	-0.35	-0.20
NEUR	-0.30	-0.25

Table 2: α estimates for traits using profile likelihood-based approach. Assumed heritability model here is BLD-LDAK-lite. We chose 41 different α values ranging from $(-1, 1)$ in step size of 0.05. Then, the best possible value or mode α value for our model is the one with maximum value for the likelihood function estimated using best optimized $\hat{\tau}_\alpha$. $\hat{\tau}_\alpha = \arg\max_\tau LL_\alpha(\tau)$ and $\hat{\alpha} = \arg\max_\alpha LL_\alpha(\hat{\tau}_\alpha)$. The estimates.

Results

Trait	SNP-her	se	eQTL-her	se	Enrich	se
BMI	0.30	0.00	0.10	0.00	1.01	0.00
FVC	0.31	0.01	0.12	0.00	1.27	0.05
SBP	0.18	0.00	0.07	0.00	1.21	0.00
HEIGHT	0.61	0.00	0.27	0.00	1.35	0.00
NEUR	0.13	0.00	0.04	0.00	1.00	0.00
WBC	0.34	0.00	0.17	0.00	1.53	0.00
RBC	0.28	0.00	0.13	0.00	1.43	0.01
PLATELET	0.34	0.00	0.17	0.00	1.53	0.00
PULSE	0.17	0.00	0.07	0.00	1.26	0.01
VENTRICULAR	0.16	0.01	0.09	0.01	1.70	0.12

Table 3: Enrichment of heritability contributed by GTEx eQTLs. PLATELET - Platelet count, WBC – White Blood cell count, RBC - red cell count, VENTRICULAR - Ventricular rate, PULSE - Pulse rate. Heritability model used is BLD-LDAK.

Conclusions

- We present an implementation of approximate loglikelihood for GWAS summary data using first-order stochastic optimization algorithm ADAM.
- We present examples of how we obtain consistent estimates for heritability parameters using our implementation that are more precise.
- We estimate heritability explained by GTEX eQTLs as well as compute enrichment of heritability across functional genomic regions and GTEx tissues.
- We also estimate selection-related parameter α for some of the UK biobank traits using profile-likelihood based approach and re-confirm that α varies across traits.
- Our framework allows to evaluate different heritability models.

Acknowledgement

AK acknowledges Melbourne Research Scholarship (MRS) for supporting his PhD research. We acknowledge Prof. Doug Speed, Aarhus University, Denmark for his valuable support and advice for some of the presented work.

We acknowledge Neale lab (http://www.nealelab.is/uk-biobank) for making GWAS summary data available for UK biobank traits.

References

[1] Doug Speed, John Holmes, and David J Balding. Evaluating and improving heritability models using summary statistics. *Nature Genetics*, 52(4):458–462, 2020.
[2] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.
[3] www.ldak.org
[4] https://gtexportal.org/home/