biMM: Efficient estimation of genetic variances and covariances for cohorts with high-dimensional phenotype measurements

Matti Pirinen [1,2,3*], Christian Benner [1,3], Pekka Marttinen [4], Marjo-Riitta Järvelin [5,6,7,8], Manuel A. Rivas [9], and Samuli Ripatti [1,3]

November 12, 2016

[1]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. [2]Helsinki Institute for Information Technology HIIT and Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland. [3]Department of Public Health, University of Helsinki, Helsinki, Finland. [4]Helsinki Institute for Information Technology HIIT and Department of Computer Science, Aalto University, Espoo, Finland. [5]Biocenter Oulu, University of Oulu, Oulu, Finland. [6]Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, London, UK. [7]Center for Life Course and Systems Epidemiology, Faculty of Medicine, University of Oulu, Oulu, Finland. [8]Unit of Primary Care, Oulu University Hospital, Oulu, Finland. [9]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA.

Abstract

Genetic research utilizes a decomposition of trait variances and covariances into genetic and environmental parts. Our software package biMM is a computationally efficient implementation of a bivariate linear mixed model for settings where hundreds of traits have been measured on partially overlapping sets of individuals.

Availability: Implementation in R freely available at www.iki.fi/mpirinen.
[*]**Contact:** matti.pirinen@helsinki.fi

1 Introduction

Decomposing phenotypic variance and covariance into genetic and environmental parts is important for designing genetic studies and understanding relationships between traits and diseases. The two main approaches are linear mixed model (LMM) implementations, such as GCTA (Yang *et al.*, 2011), GEMMA (Zhou and Stephens, 2014) or BOLT-REML (Loh *et al.*, 2015), and LD-score regression, implemented in LDSC (Bulik-Sullivan *et al.*, 2015). LMM requires access to the individual-level genotype-phenotype data whereas LDSC only needs output from a genome-wide association study (GWAS) and variant correlations from a reference database, but consequently may be less precise than LMM (Bulik-Sullivan *et al.*, 2015).

We consider settings where individual-level data are available, and hence use LMM. The bivariate LMM for n individuals is $\boldsymbol{Y} = \boldsymbol{G} + \boldsymbol{\varepsilon}$, where $\boldsymbol{Y} = \left(\boldsymbol{Y}_1^T, \boldsymbol{Y}_2^T\right)^T$ is $2n$-vector of mean-centered phenotype values from which the covariates, such as age, sex and principal components of population structure have been regressed out, $\boldsymbol{G} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_G)$ is $2n$-vector of genetic random effects and $\varepsilon \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\varepsilon)$ is $2n$-vector of environmental random effects. The $(2n) \times (2n)$ covariance structures are parameterized by six scalars: genetic variances V_{G1} and V_{G2}, genetic covariance V_{G12}, environmental variances $V_{\varepsilon 1}$ and $V_{\varepsilon 2}$ and environmental covariance $V_{\varepsilon 12}$ as

$$\boldsymbol{\Sigma}_G = \left[\begin{array}{c|c} V_{G1}\boldsymbol{R} & V_{G12}\boldsymbol{R} \\ \hline V_{G12}\boldsymbol{R} & V_{G2}\boldsymbol{R} \end{array}\right] \text{ and } \boldsymbol{\Sigma}_\varepsilon = \left[\begin{array}{c|c} V_{\varepsilon 1}\boldsymbol{I} & V_{\varepsilon 12}\boldsymbol{I} \\ \hline V_{\varepsilon 12}\boldsymbol{I} & V_{\varepsilon 2}\boldsymbol{I} \end{array}\right]$$

expressed as $n \times n$ block matrices. $\boldsymbol{I}$ is the identity matrix and the element i, j of the genetic relationship matrix (GRM) $\boldsymbol{R}$ is

$$\boldsymbol{R}_{ij} = \frac{1}{K} \sum_{k=1}^{K} \left(g_{ik} - 2\widehat{f}_k\right)\left(g_{jk} - 2\widehat{f}_k\right)\left(2\widehat{f}_k\left(1 - \widehat{f}_k\right)\right)^{\alpha},$$

where g_{ik} is the genotype of individual i at variant k, coded as 0, 1 or 2 copies of the minor allele and $\widehat{f}_k$ is the minor allele frequency (MAF). We use the standard scaling of allelic effects determined by $\alpha = -1$.

From this model, an estimate of V_{Gt} approximates additive genetic variance of each trait ($t = 1, 2$) explained by the variants included in the calculation of $\boldsymbol{R}$ and is often used as a lower bound for the (narrow-sense) heritability (detailed assumptions in Yang *et al.* (2015)). An estimate of the genetic correlation $\rho_G = V_{G12}/\sqrt{V_{G1}V_{G2}}$ measures (average) correlation of the allelic effects of the variants on the two traits. Similarly, we can estimate $\rho_\varepsilon = V_{\varepsilon 12}/\sqrt{V_{\varepsilon 1}V_{\varepsilon 2}}$, the correlation in the environmental components between the traits.

A challenge with bivariate LMMs, that operate on an explicit $\boldsymbol{R}$ matrix (e.g. GCTA and GEMMA), is that they require matrix operations cubic in the cohort size for each pair of traits analyzed, which becomes computationally prohibitive for handling hundreds of traits measured on 10,000s of individuals. Our software package biMM speeds up the bivariate LMM analysis (1) by a fast likelihood computation, (2) by reusing matrix decompositions across pairs of traits, and (3) by arranging data to optimize sample overlap between consecutive pairs of traits.

2 Methods

2.1 Reusing eigendecomposition

Once an eigendecomposition of $\boldsymbol{R}$ is available, our biMM algorithm drops the time complexity from cubic to quadratic for a trait pair and from cubic to linear for a single evaluation of the likelihood function (Supplementary Information). A crucial observation is that a complete sample overlap between two trait pairs means that the same eigendecomposition can be used for both pairs.

2.2 Ordering pairs, imputing and dropping values

We order the trait pairs in such a way that the consecutive pairs have a large sample overlap. biMM further allows imputing at most t_i missing values and/or dropping at most t_d

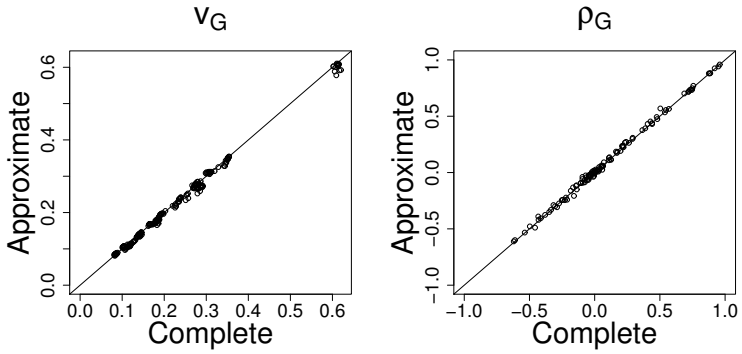

Figure 1: Comparing estimates for heritability (V_G) and genetic correlation (ρ_G) between an approximate ($t_i = t_d = 200$) and complete ($t_i = t_d = 0$) versions of biMM over 120 pairs of traits.

	biMM approx	biMM compl	GEMMA	BOLT-REML	GCTA
Real (h)	0.05	0.49	2.76	19.20	21.39
CPU (h)	0.07	1.49	2.76	19.20	21.39

Table 1: Cumulative run time over 120 trait pairs of Fig. 1. 'Real' is wall clock time. 'CPU' is total CPU time over all cores used. We used an Intel Quad-Core i7-3770 CPU @ 3.40GHz. biMM ran in R-3.3.1 with Intel Math Kernel Libarary.

non-missing values for a trait pair to make it match the available eigendecomposition (Supplementary Information). Only when this is not possible for any remaining pair does biMM a new eigendecomposition. Algorithmically, given user-specified t_i and t_d, biMM finds an ordering that results in a small number of total eigendecompositions. This is an instance of the shortest Hamiltonian path problem that we tackle by a greedy heuristic (Supplementary Information).

2.3 Example analysis

We consider data from the Northern Finland Birth Cohort 1966 (NFBC1966) (Rantakallio *et al.*, 1969) with 16 traits having sample sizes between 4736 and 5025 individuals (Supplementary Table S1) and preprocessed by Tukiainen *et al.* (2014). We analyzed all 120 pairs of traits using both the complete ($t_i = t_d = 0$) and an approximate versions ($t_i = t_d = 200$) of biMM and compared with GCTA 1.25.3, GEMMA 0.94.1 and BOLT-REML 2.2 with their default parameters.

3 Results

Figure 1 shows that the complete and approximate versions of biMM are very similar across the 120 pairs of traits. Table 1 shows that the approximate version is much faster than either the complete version or any other software package tested. Detailed results are in Supplementary Figures S1-S4. In short, biMM and GEMMA gave essentially the same results and they were also similar to the results from GCTA and BOLT-REML.

Funding

This work was supported by the Academy of Finland [257654 and 288509 to M.P.; 286607 and 294015 to P.M.; 251217 and 255847 to S.R.]. S.R. was supported by EU FP7 projects ENGAGE (201413) and BioSHaRE (261433), the Finnish Foundation for Cardiovascular Research, Biocentrum Helsinki and the Sigrid Juselius Foundation. NFBC1966 received financial support from University of Oulu Grant no. 65354, Oulu University Hospital Grant no. 2/97, 8/97, Ministry of Health and Social Affairs Grant no. 23/251/97, 160/97, 190/97, National Institute for Health and Welfare, Helsinki Grant no. 54121, Regional Institute of Occupational Health, Oulu, Finland Grant no. 50621, 54231.

Acknowledgements

This study made use of NFBC1966 data. We thank the late professor Paula Rantakallio (launch of NFBC1966), the participants in the 31yrs study and the NFBC project center.

References

Bulik-Sullivan B. et al. (2015) An atlas of genetic correlations across human diseases and traits. *Nat Gen* **47**, 1236-1241.

Loh PR. et al. (2015) Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat Gen* **47**, 1385-1392.

Rantakallio P. et al. (1969) Groups at risk in low birth weight infants and perinatal mortality. *Acta Paediatr Scand* **193**, (Suppl 193): 191.

Tukiainen T. et al. (2014) Chromosome X-wide association study identifies loci for fasting insulin and height and evidence for incomplete dosage compensation. *PLoS Genet*, **10**, e1004127.

Yang J. et al. (2011) GCTA: a tool for Genome-wide Complex Trait Analysis. *Am J Hum Genet*, **88**, 76-82.

Yang J. et al. (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Gen* **47**, 1114-1120.

Zhou X. and Stephens M. (2014) Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods* **11**, 407-409.