Bioinformatics, 33(15), 2017, 2405–2407
doi: 10.1093/bioinformatics/btx166
Advance Access Publication Date: 28 March 2017
Applications Note

OXFORD

Genetics and population analysis

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[2,4], Marjo-Riitta Järvelin[5,6,7,8], Manuel A. Rivas[9] and Samuli Ripatti[1,3]

[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 and [9]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on November 16, 2016; revised on February 18, 2017; editorial decision on March 20, 2017; accepted on March 22, 2017

Abstract

Summary: 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 and Implementation: Implementation in R freely available at www.iki.fi/mpirinen.

Contact: matti.pirinen@helsinki.fi

Supplementary information: Supplementary data are available at *Bioinformatics* online.

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 $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \boldsymbol{Y}_2^T)^T$ 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.

The existing bivariate LMM implementations have not been designed for a case where hundreds of traits have been measured on 10 000s of individuals. Our software package biMM combines a fast likelihood computation (similar in speed to GEMMA) with an algorithm that optimizes the sample overlap between consecutive pairs of traits analyzed and therefore efficiently reuses the computationally expensive matrix decompositions. biMM allows user to control how much missing data are tolerated for a single analysis and automatically executes both phenotype imputation and matrix decompositions required to achieve that tolerance.

2 Materials and 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). Similar time complexity is achieved by GEMMA, and efficient algorithms for multivariate LMM have recently been considered also by Furlotte and Eskin (2015) and Casale *et al.* (2015). Our central observation is that a complete sample overlap between two trait pairs means that the same eigendecomposition can be used for both pairs. To fully utilize this observation, we need to keep the eigendecomposition in random access memory (RAM) across the trait pairs and we need to optimize the order of the trait pairs. To our knowledge, neither of these functionalities is available in existing software.

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 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 analyses

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. biMM ran in R-3.3.1 with Intel Math Kernel Library.

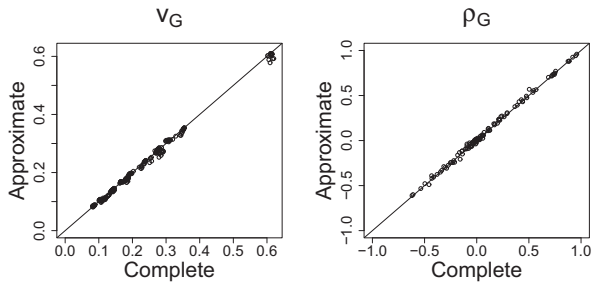

Fig. 1. Comparing estimates for heritability (V_G) and genetic correlation (ρ_G) between an approximate ($t_i = t_d = 200$) and complete ($t_i = t_d = 0$) versions of biMM over 120 pairs of traits

Table 1. Cumulative run time in hours over 120 trait pairs of Figure 1

	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

'Real' is wall clock time. 'CPU' is total CPU time over all cores used.

To assess scaling to larger datasets, we consider 20 000 individuals simulated by HapGen2 (Su *et al.*, 2011) using chromosome 2 of the CEU panel from HapMap3 (International HapMap3 consortium., 2010) with phenotypes generated to have heritabilities between 0.2 and 0.8.

In all examples we used a desktop computer with an Intel Quad-Core i7-3770 CPU @ 3.40 GHz and 16 Gb of RAM.

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.

To assess scaling to larger datasets, we evaluated how much time each additional trait pair would require for a dataset of 20 000 individuals after the eigendecomposition is available and phenotype data are completely observed. The resulting times in CPU seconds are 1.6 for biMM, 75 for GEMMA and 2670 for BOLT-REML. GCTA was unable to run with 16 Gb of RAM. The difference between biMM and GEMMA in this example with no missing data is that biMM holds the eigendecomposition in RAM while GEMMA reads it from a file for each pair of traits. The eigendecomposition itself took 70 CPU minutes with biMM and 450 CPU minutes with GEMMA. Hence, with a desktop computer, an analysis of completely observed or imputed omics data for 500 traits (124 750 trait pairs) measured on 20 000 individuals would take less than 2.5 days with biMM, over 100 days with GEMMA (although with a multivariate analysis strategy GEMMA could finish in 14 days, Supplementary Information) and many years with BOLT-REML.

4 Conclusion

Our freely available biMM software package makes a bivariate linear mixed model analysis of high-dimensional phenotypes on

cohorts of a few tens of thousands of individuals practical using a desktop computer. For even larger cohorts, where explicit matrix decompositions are impractical on current desktop computers, BOLT-REML may be the only available option to analyze a pair of traits, but it cannot utilize sharing of individuals across trait pairs to efficiently analyze tens of thousands of trait pairs.

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.

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.

Conflict of Interest: none declared.

References

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

Casale,F.P. *et al.* (2015) Efficient set tests for the genetic analysis of correlated traits. *Nat. Methods*, **12**, 755–758.

Furlotte,N.A. and Eskin,E. (2015) Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model. *Genetics*, **200**, 59–68.

International HapMap3 consortium. (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.

Loh,P.R. *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**, 191.

Su,Z. *et al.* (2011) HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, **27**, 2304–2305.

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.