GEMMA User Manual

Xiang Zhou

December 5, 2012

Contents

1 Introduction

1.1 What is GEMMA

GEMMA is the software implementing the Genome-wide Efficient Mixed Model Association algorithm [6] for a standard linear mixed model and some of its close relatives for genome-wide association studies (GWAS). It fits a standard linear mixed model (LMM) to account for population stratification and sample structure for single SNP association tests [6]. It fits a Bayesian sparse linear mixed model (BSLMM) using Markov chain Monte Carlo (MCMC) for estimating the proportion of variance in phenotypes explained (PVE) by typed genotypes (i.e. chip heritability), predicting phenotypes, and identifying associated SNPs by jointly modeling all SNPs while controlling for population structure [5]. It is computationally efficient for large scale GWAS and uses freely available open-source numerical libraries.

1.2 How to Cite GEMMA

Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics, 44: 821-824, 2012.

Xiang Zhou, Peter Carbonetto and Matthew Stephens. Polygenic modeling with Bayesian sparse linear mixed models. PLOS Genetics. in press. http://arxiv.org/abs/1209.1341.

1.3 Models

1.3.1 Linear Mixed Model

GEMMA can fit a standard linear mixed model in the following form:

$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{x}\beta + \mathbf{u} + \boldsymbol{\epsilon} \quad \mathbf{u} \sim \mathrm{MVN}_n(0, \lambda\tau^{-1}\mathbf{K}) \quad \boldsymbol{\epsilon} \sim \mathrm{MVN}_n(0, \tau^{-1}\mathbf{I}_n)$$

where $\mathbf{y}$ is an n-vector of quantitative traits (or binary disease labels) for n individuals; $\mathbf{W} = (\mathbf{w}_1, \cdots, \mathbf{w}_c)$ is an $n \times c$ matrix of covariates (fixed effects) including a column of 1s; $\boldsymbol{\alpha}$ is a c-vector of the corresponding coefficients including the intercept; $\mathbf{x}$ is an n-vector of marker genotypes; β is the effect size of the marker; $\mathbf{u}$ is an n-vector of random effects; $\boldsymbol{\epsilon}$ is an n-vector of errors; τ^{-1} is the variance of the residual errors; λ is the ratio between the two variance components; $\mathbf{K}$ is a known $n \times n$ relatedness matrix and $\mathbf{I}_n$ is an $n \times n$ identity matrix. MVN_n denotes the n-dimensional multivariate normal distribution.

GEMMA tests the alternative hypothesis $H_1 : \beta \neq 0$ against the null hypothesis $H_0 : \beta = 0$ for each SNP in turn, using one of the three commonly used test statistics (Wald, likelihood ratio or score). GEMMA obtains either the maximum likelihood estimate (MLE) or the restricted maximum likelihood estimate (REML) of λ and β, and outputs the corresponding p value.

1.3.2 Bayesian Sparse Linear Mixed Model

GEMMA can fit a Bayesian sparse linear mixed model in the following form as well as a corresponding probit counterpart:

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\epsilon} \quad \beta_i \sim \pi \mathrm{N}(0, \sigma_a^2\tau^{-1}) + (1-\pi)\delta_0 \quad \mathbf{u} \sim \mathrm{MVN}_n(0, \sigma_b^2\tau^{-1}\mathbf{K}) \quad \boldsymbol{\epsilon} \sim \mathrm{MVN}_n(0, \tau^{-1}\mathbf{I}_n)$$

where $\mathbf{1}_n$ is an n-vector of 1s, μ is a scalar representing the phenotype mean, $\mathbf{X}$ is an $n \times p$ matrix of genotypes measured on n individuals at p genetic markers, $\boldsymbol{\beta}$ is the corresponding p-vector of the genetic marker effects, and other parameters are the same as defined in the standard linear mixed model in the previous section.

In the special case $\mathbf{K} = \mathbf{X}\mathbf{X}^T/p$ (default in GEMMA), the SNP effect sizes can be decomposed into two parts: $\boldsymbol{\alpha}$ that captures the small effects that all SNPs have, and $\boldsymbol{\beta}$ that captures the additional effects of some large effect SNPs. In this case, $\mathbf{u} = \mathbf{X}\boldsymbol{\alpha}$ can be viewed as the combined effect of all small effects, and the total effect size for a given SNP is $\alpha_i + \beta_i$.

There are two important hyper-parameters in the model: PVE, being the proportion of variance in phenotypes explained by the sparse effects ($\mathbf{X}\boldsymbol{\beta}$) and random effects terms ($\mathbf{u}$) together, and PGE, being the proportion of genetic variance explained by the sparse effects terms ($\mathbf{X}\boldsymbol{\beta}$). These two parameters are defined as follows:

$$\mathrm{PVE}(\boldsymbol{\beta}, \mathbf{u}, \tau) := \frac{\mathrm{V}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u})}{\mathrm{V}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) + \tau^{-1}},$$
$$\mathrm{PGE}(\boldsymbol{\beta}, \mathbf{u}) := \frac{\mathrm{V}(\mathbf{X}\boldsymbol{\beta})}{\mathrm{V}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u})},$$

where

$$\mathrm{V}(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

GEMMA uses MCMC to estimate $\boldsymbol{\beta}$, $\mathbf{u}$ and all other hyper-parameters including PVE, PGE and π.

1.4 Missing Data

1.4.1 Missing Genotypes

As mentioned before [6], the tricks used in the GEMMA algorithm rely on having complete or imputed genotype data at each SNP. That is, GEMMA requires the user to impute all missing genotypes before association testing. This imputation step is arguably preferable than simply dropping individuals with missing genotypes, since it can improve power to detect associations [1].

Therefore, for fitting both LMM or BSLMM, missing genotypes are required to be imputed first. Otherwise, any SNPs with missingness above a certain threshold (default 5%) will not be analyzed, and missing genotypes for SNPs that do not pass this threshold will be simply replaced with the estimated mean genotype of that SNP. For predictions, though, all SNPs will be used regardless of their missingness. Missing genotypes in the test set will be replaced by the mean genotype in the training set.

1.4.2 Missing Phenotypes

Individuals with missing phenotypes will not be included in the LMM or BSLMM analysis. However, all individuals will be used for calculating the relatedness matrix, so that the resulting relatedness matrix is still an $n \times n$ matrix regardless of how many individuals have missing phenotypes. In addition, predicted values will be calculated for individuals with missing values, based on individuals with non-missing values.

For relatedness matrix calculation, because missingness and minor allele frequency for a given SNP are calculated based on analyzed individuals (i.e. individuals with no missing phenotypes and no missing covariates), if all individuals have missing phenotypes, then no SNP and no individuals will be included in the analysis and the estimated relatedness matrix will be full of "nan"s.

2 Installing and Compiling GEMMA

If you have downloaded a binary executable , no installation is necessary. If you have downloaded the source code, you will need a standard C/C++ compiler such as GNU gcc, as well as the GSL and LAPACK libraries. You will need to change the library paths in the Makefile accordingly. A sample Makefile is provided along with the source code. For details on installing GSL library, please refer to `http://www.gnu.org/s/gsl/`. For details on installing LAPACK library, please refer to `http://www.netlib.org/lapack/`.

If you are interested in fitting BSLMM for a large scale GWAS data set but have limited memory to store the entire genotype matrix, you could compile GEMMA in float precision. A float precision binary executable, named "gemmaf", is available inside the "bin" folder in the source code. To compile a float precision binary by yourself, you can first run "d2f.sh" script inside the "src" folder, and then enable "FORCE_FLOAT" option in the Makefile. The float version could save about half of the memory without appreciable loss of accuracy.

3 Input File Formats

GEMMA requires four input files containing genotypes, phenotypes, relatedness matrix and (optionally) covariates. Genotype and phenotype files can be in two formats, **either both** in the PLINK binary ped format **or both** in the BIMBAM format. Mixing genotype and phenotype files from the two formats (for example, using PLINK files for genotypes and using BIMBAM files for phenotypes) will result in unwanted errors. BIMBAM format is particularly useful for imputed genotypes, as PLINK codes genotypes using 0/1/2, while BIMBAM can accommodate any real values between 0 and 2 (and any real values if paired with "-notsnp" option).

3.1 PLINK Binary PED File Format

GEMMA recognizes the PLINK binary ped file format (`http://pngu.mgh.harvard.edu/~purcell/plink/`) [3] for both genotypes and phenotypes. This format requires three files: *.bed, *.bim and *.fam, all with the same prefix. The *.bed file should be in the default SNP-major mode (beginning with three bytes). One can use the PLINK software to generate binary ped files from standard ped files using the following command:

```
plink --file [file_prefix] --make-bed --out [bedfile_prefix]
```

For the *.fam file, GEMMA only reads the second column (individual id) and the sixth column (phenotype). One can specify a different column as the phenotype column by using "-n [num]", where "-n 1" uses the original sixth column as phenotypes, and "-n 2" uses the seventh column, and so on and so forth.

GEMMA codes alleles as 0/1 according to the plink webpage on binary plink format (`http://pngu.mgh.harvard.edu/~purcell/plink/binary.shtml`). Specifically, the column 5 of the *.bim file is the minor allele and is coded as 1, while the column 6 of the *.bim file is the major allele and is coded as 0. The minor allele in column 5 is therefore the effect allele (notice that GEMMA version 0.92 and before treats the major allele as the effect allele).

GEMMA will read the phenotypes as provided and will recognize either "-9" or "NA" as missing phenotypes. If the phenotypes in the *.fam file are disease status, one is recommended to label controls as 0 and cases as 1, as the results will have better interpretation. For example, the predicted values from a linear BSLMM can be directly interpreted as the probability of being a case. In addition, the probit BSLMM will only recognize 0/1 as control/case labels.

For prediction problems, one is recommended to list all individuals in the file, but label those individuals in the test set as missing. This will facilitate the use of the prediction function implemented in GEMMA.

3.2 BIMBAM File Formats

GEMMA also recognizes BIMBAM file format (`http://stephenslab.uchicago.edu/software.html`) [1], which is particularly useful for imputed genotypes as well as for general covariates other than SNPs. BIMBAM format consists of three files, a mean genotype file, a phenotype file, and an optional SNP annotation file. We explain these files in detail below.

3.2.1 Mean Genotype File

This file contains genotype information. The first column is SNP id, the second and third columns are allele types with minor allele first, and the remaining columns are the posterior/imputed mean genotypes of different individuals numbered between 0 and 2. An example mean genotype file with two SNPs and three individuals is as follows:

```
rs1, A, T, 0.02, 0.80, 1.50
rs2, G, C, 0.98, 0.04, 1.00
```

GEMMA codes alleles exactly as provided in the mean genotype file, and ignores the allele types in the second and third columns. Therefore, the minor allele is the effect allele only if one codes minor allele as 1 and major allele as 0.

One can use the following bash command (in one line) to generate BIMBAM mean genotype file from IMPUTE genotype files (`http://www.stats.ox.ac.uk/~marchini/software/gwas/file_format.html`)[2]:

```
cat [impute filename] | awk -v s=[number of snps]
'{ printf $2 "," $4 "," $5; for(i=1; i<=s; i++) printf "," $(i*3+3)*2+$(i*3+4); printf "\n" }'
> [bimbam filename]
```

Notice that one may need to manually input the two quote symbols ' . Depending on the terminal, a direct copy/paste of the above line may result in "-bash: syntax error near unexpected token '(' " errors.

Finally, the mean genotype file can accommodate values other than SNP genotypes. One can use the "-notsnp" option to disable the minor allele frequency cutoff and to use any numerical values as covariates.

3.2.2 Phenotype File

This file contains phenotype information. Each line is a number indicating the phenotype value for each individual in turn, in the same order as in the mean genotype file. Notice that only numeric values are allowed and characters will not be recognized by the software. Missing phenotype information is denoted as NA. The number of rows should be equal to the number of individuals

in the mean genotype file. An example phenotype file with five individuals and one phenotype is as follows:

```
1.2
NA
2.7
-0.2
3.3
```

One can include multiple phenotypes as multiple columns in the phenotype file, and specify a different column for association tests by using "-n [num]", where "-n 1" uses the original first column as phenotypes, and "-n 2" uses the second column, and so on and so forth. An example phenotype file with five individuals and three phenotypes is as follows:

```
1.2  -0.3  -1.5
NA    1.5   0.3
2.7   1.1   NA
-0.2 -0.7   0.8
3.3   2.4   2.1
```

For binary traits, one is recommended to label controls as 0 and cases as 1, as the results will have better interpretation. For example, the predicted values from a linear BSLMM can be directly interpreted as the probability of being a case. In addition, the probit BSLMM will only recognize 0/1 as control/case labels.

For prediction problems, one is recommended to list all individuals in the file, but label those individuals in the test set as missing. This will facilitate the use of the prediction function implemented in GEMMA.

3.2.3 SNP Annotation File (optional)

This file contains SNP information. The first column is SNP id, the second column is its base-pair position, and the third column is its chromosome number. The rows are not required to be in the same order of the mean genotype file, but must contain all SNPs in that file. An example annotation file with four SNPs is as follows:

```
rs1, 1200, 1
rs2, 1000, 1
rs3, 3320, 1
rs4, 5430, 1
```

If an annotation file is not provided, the SNP information columns in the output file for association tests will have "-9" as missing values.

3.3 Relatedness Matrix File Formats

GEMMA, as a linear mixed model software, requires a relatedness matrix file in addition to both genotype and phenotype files. GEMMA takes relatedness file in two formats. The first format is a $n \times n$ matrix, where each row and each column corresponds to individuals in the same order as in the *.fam file or in the mean genotype file, and ith row and jth column is a number indicating the relatedness value between ith and jth individuals. An example relatedness matrix file with three individuals is as follows:

```
0.3345  -0.0227   0.0103
-0.0227  0.3032  -0.0253
0.0103  -0.0253   0.3531
```

The second relatedness matrix format is a three column "id id value" format, where the first two columns show two individual id numbers, and the third column shows the relatedness value between these two individuals. Individual ids are not required to be in the same order as in the *.fam file, and relatedness values not listed in the relatedness matrix file will be considered as 0. An example relatedness matrix file with the same three individuals above is shown below:

```
id1  id1  0.3345
id1  id2  -0.0227
id1  id3  0.0103
id2  id2  0.3032
id2  id3  -0.0253
id3  id3  0.3531
```

As BIMBAM mean genotype files do not provide individual id, the second format only works with the PLINK binary ped format. One can use "-km [num]" to choose which format to use, i.e. use "-km 1" or "-km 2" to accompany PLINK binary ped format, and use "-km 1" to accompany BIMBAM format.

3.4 Covariates File Format (optional)

One can provide a covariates file if needed for fitting LMM if necessary. GEMMA fits a linear mixed model with an intercept term if no covariates file is provided, but does not internally provide an intercept term if a covariates file is available. Therefore, if one has covariates other than the intercept and wants to adjust for those covariates ($\mathbf{W}$) simultaneously, one should provide GEMMA with a covariates file containing an intercept term explicitly. The covariates file is similar to the above BIMBAM multiple phenotype file, and must contain a column of 1's if one wants to include an intercept. An example covariates file with five individuals and three covariates (the first column is the intercept) is as follows:

```
1  1  -1.5
1  2   0.3
1  2   0.6
1  1  -0.8
1  1   2.0
```

4 Running GEMMA

4.1 A Small GWAS Example Dataset

If you downloaded the GEMMA source code recently, you will find an "example" folder containing a small GWAS example dataset. This data set comes from the heterogeneous stock mice data, kindly provided by Wellcome Trust Centre for Human Genetics on the public domain `http://mus.well.ox.ac.uk/mouse/HS/`, with detailed described in [4].

The data set consists of 1904 individuals from 85 families, all descended from eight inbred progenitor strains. We selected one particular phenotype from this data set, the percentage of CD8+ cells, with measurements in 1410 individuals. The phenotypes were already corrected for sex, age, body weight, season and year effects by the original study, and we further quantile normalized the phenotypes to a standard normal distribution. In addition, we obtained a total of 12,226 autosomal SNPs, with missing genotypes replaced by the mean genotype of that SNP in the family. Genotype and phenotype files are in both BIMBAM and PLINK binary formats.

For demonstration purpose, we randomly divided the 85 families into two sets, where each set contained roughly half of the individuals (i.e. *inter-family* split) as in [5]. We also created artificial binary phenotypes from the quantitative phenotypes, by assigning the half individuals with higher quantitative values to 1 and the other half to 0, as in [5]. Therefore, the phenotype files contain five columns of phenotypes. The first column contains the quantitative phenotypes for all individuals. The second column contains quantitative phenotypes for individuals in the training set. The third column contains quantitative phenotypes for individuals in the test set. The fourth and fifth columns contain binary phenotypes for individuals in the training set and test set, respectively.

A demo.txt file inside the same folder shows detailed steps on how to use GEMMA to estimate the relatedness matrix from genotypes, how to perform association tests using the standard linear mixed model, how to fit the Bayesian sparse linear mixed model and how to obtain predicted values using the output files from fitting BSLMM. The output results from GEMMA for all the examples are also available inside the "result" subfolder.

4.2 Estimate Relatedness Matrix from Genotypes

4.2.1 Basic Usage

The basic usages to calculate an estimated relatedness matrix with either the PLINK binary ped format or the BIMBAM format are:

```
./gemma -bfile [prefix] -gk [num] -o [prefix]
./gemma -g [filename] -p [filename] -gk [num] -o [prefix]
```

where the "-gk [num]" option specifies which relatedness matrix to estimate, i.e. "-gk 1" calculates the centered relatedness matrix while "-gk 2" calculates the standardized relatedness matrix; "-bfile [prefix]" specifies PLINK binary ped file prefix; "-g [filename]" specifies BIMBAM mean genotype file name; "-p [filename]" specifies BIMBAM phenotype file name; "-o [prefix]" specifies output file prefix.

4.2.2 Detailed Information

GEMMA provides two ways to estimate the relatedness matrix from genotypes, using either the centered genotypes or the standardized genotypes. We denote $\mathbf{X}$ as the $n \times p$ matrix of genotypes, $\mathbf{x}_i$ as its ith column representing genotypes of ith SNP, $\bar{x}_i$ as the sample mean and v_{x_i} as the sample variance of ith SNP, and $\mathbf{1}_n$ as a $n \times 1$ vector of 1's. Then the two relatedness matrices GEMMA can calculate are as follows:

$$G_c = \frac{1}{p}\sum_{i=1}^{p}(\mathbf{x}_i - \mathbf{1}_n\bar{x}_i)(\mathbf{x}_i - \mathbf{1}_n\bar{x}_i)^T,$$
$$G_s = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{v_{x_i}}(\mathbf{x}_i - \mathbf{1}_n\bar{x}_i)(\mathbf{x}_i - \mathbf{1}_n\bar{x}_i)^T.$$

Which of the two relatedness matrix to choose will largely depend on the underlying genetic architecture of the given trait. Specifically, if SNPs with lower minor allele frequency tend to have larger effects (which is inversely proportional to its genotype variance), then the standardized genotype matrix is preferred. If the SNP effect size does not depend on its minor allele frequency, then the centered genotype matrix is preferred. In our previous experience based on a limited examples, we typically find the centered genotype matrix provides better control for population structure in lower organisms, and the two matrices seem to perform similarly in humans.

In the current implementation, only polymorphic SNPs that have missingness below 5% (use "-miss [num]" to change, e.g. "-miss 0.1" changes the threshold to 10%) and minor allele frequency above 1% (use "-maf [num]" to change, e.g. "-maf 0.05" changes the threshold to 5%) will be used to estimate the relatedness matrix. Both missingness and minor allele frequency of a given SNP are calculated based on analyzed individuals (i.e. individuals with no missing phenotypes and no missing covariates). Therefore, if all individuals have missing phenotypes, no SNP will be analyzed and the output matrix will be full of "nan"s.

4.2.3 Output Files

There will be two output files, both inside an output folder in the current directory. The prefix.log.txt file contains some detailed information about the running parameters and computation time, while the prefix.cXX.txt or prefix.sXX.txt contains a $n \times n$ matrix of estimated relatedness

matrix.

4.3 Association Tests with a Standard Linear Mixed Model

4.3.1 Basic Usage

The basic usages for association analysis with either the PLINK binary ped format or the BIMBAM format are:

```
./gemma -bfile [prefix] -k [filename] -lmm [num] -o [prefix]
./gemma -g [filename] -p [filename] -a [filename] -k [filename] -lmm [num] -o [prefix]
```

where the "-lmm [num]" option specifies which frequentist test to use, i.e. "-lmm 1" performs Wald test, "-lmm 2" performs likelihood ratio test, "-lmm 3" performs score test, and "-lmm 4" performs all the three tests; "-bfile [prefix]" specifies PLINK binary ped file prefix; "-g [filename]" specifies BIMBAM mean genotype file name; "-p [filename]" specifies BIMBAM phenotype file name; "-a [filename]" (optional) specifies BIMBAM SNP annotation file name; "-k [filename]" specifies relatedness matrix file name; "-o [prefix]" specifies output file prefix.

4.3.2 Detailed Information

The algorithm calculates either REML or MLE estimate of λ in the evaluation interval from 1×10^{-5} (corresponding to almost pure environmental effect) to 1×10^{5} (corresponding to almost pure genetic effect). Although unnecessary in most cases, one can expand or reduce this evaluation interval by specifying "-lmin" and "-lmax" (e.g. "-lmin 0.01 -lmax 100" specifies an interval from 0.01 to 100). The log-scale evaluation interval is further divided into 10 equally spaced regions, and optimization is carried out in each region where the first derivatives change sign. Although also unnecessary in most cases, one can increase or decrease the number of regions by using "-region" (e.g. "-region 100" uses 100 equally spaced regions on the log-scale), which may yield more stable or faster performance, respectively.

For binary traits, one can label controls as 0 and cases as 1, and follow our previous approaches to fit the data with a linear mixed model by treating the binary case control labels as quantitative traits [6, 5]. This approach can be justified partly by recognizing the linear model as a first order Taylor approximation to a generalized linear model, and partly by the robustness of the linear model to model misspecification [5].

Again, in the current implementation, only polymorphic SNPs that have missingness below 5% (use "-miss [num]" to change, e.g. "-miss 0.1" changes the threshold to 10%) and minor allele frequency above 1% (use "-maf [num]" to change, e.g. "-maf 0.05" changes the threshold to 5%) will be analyzed. Both missingness and minor allele frequency of a given SNP are calculated based on analyzed individuals (i.e. individuals with no missing phenotypes and no missing covariates).

4.3.3 Output Files

There will be two output files, both inside an output folder in the current directory. The prefix.log.txt file contains some detailed information about the running parameters and computation time. In addition, prefix.log.txt contains PVE estimate and its standard error in the null linear mixed model.

The prefix.assoc.txt contains the results. An example file with a few SNPs is shown below:

```
chr rs ps n_miss beta se l_remle p_wald
1 rs3683945 3197400 0 -7.788676e-02 6.193497e-02 4.317971e+00 2.087605e-01
1 rs3707673 3407393 0 -6.654296e-02 6.210229e-02 4.316122e+00 2.841257e-01
1 rs6269442 3492195 0 -5.344258e-02 5.377462e-02 4.323589e+00 3.204786e-01
1 rs6336442 3580634 0 -6.770167e-02 6.209261e-02 4.315691e+00 2.757527e-01
1 rs13475700 4098402 0 -5.659112e-02 7.175371e-02 4.340122e+00 4.304285e-01
```

The eight columns are: chromosome numbers, snp ids, base pair positions on the chromosome, number of missing values for a given snp, beta estimates, standard errors for beta, remle estimates for lambda, and p values from Wald test.

4.4 Fit a Bayesian Sparse Linear Mixed Model

4.4.1 Basic Usage

The basic usages for fitting a BSLMM with either the PLINK binary ped format or the BIMBAM format are:

```
./gemma -bfile [prefix] -bslmm [num] -o [prefix]
./gemma -g [filename] -p [filename] -a [filename] -bslmm [num] -o [prefix]
```

where the "-bslmm [num]" option specifies which model to fit, i.e. "-bslmm 1" fits a standard linear BSLMM, "-bslmm 2" fits a ridge regression/GBLUP, and "-bslmm 3" fits a probit BSLMM; "-bfile [prefix]" specifies PLINK binary ped file prefix; "-g [filename]" specifies BIMBAM mean genotype file name; "-p [filename]" specifies BIMBAM phenotype file name; "-a [filename]" (optional) specifies BIMBAM SNP annotation file name; "-o [prefix]" specifies output file prefix.

4.4.2 Detailed Information

In default, GEMMA does not require the user to provide a relatedness matrix explicitly. It internally calculates and uses the centered relatedness matrix, which has the nice interpretation that each effect size β_i follows a mixture of two normal distributions *a priori*. Of course, one can choose to supply a relatedness matrix by using the "-k [filename]" option. In addition, GEMMA does not

take covariates file when fitting BSLMM. However, one can use the BIMBAM mean genotype file to store these covariates and use "-notsnp" option to use them.

The option "-bslmm 1" fits a linear BSLMM using MCMC, "-bslmm 2" fits a ridge regression/GBLUP with standard non-MCMC method, and "-bslmm 3" fits a probit BSLMM using MCMC. Therefore, option "-bslmm 2" is much faster than the other two options, and option "-bslmm 1" is faster than "-bslmm 3". For MCMC based methods, one can use "-w [num]" to choose the number of burn-in iterations that will be discarded, and "-s [num]" to choose the number of sampling iterations that will be saved. In addition, one can use "-smax [num]" to choose the number of maximum SNPs to include in the model (i.e. SNPs that have addition effects), which may also be needed for the probit BSLMM because of its heavier computationAL burden. It is up to the users to decide these values for their own data sets, in order to balance computation time and computation accuracy.

For binary traits, one can label controls as 0 and cases as 1, and follow our previous approach to fit the data with a linear BSLMM by treating the binary case control labels as quantitative traits [5]. This approach can be justified by recognizing the linear model as a first order Taylor approximation to a generalized linear model. One can of course choose to fit a probit BSLMM, but in our limited experience, we do not find appreciable prediction accuracy gains in using the probit BSLMM over the linear BSLMM for binary traits (see a briefly discussion in [5]). This of course could be different for a different data set.

The genotypes, phenotypes (except for the probit BSLMM), as well as the relatedness matrix will be centered when fitting the models. The estimated values in the output files are thus for these centered values. Therefore, proper prediction will require genotype means and phenotype means from the individuals in the training set, and one should always use the same phenotype column, with individuals in the test set labeled as missing, to fit the BSLMM and to obtain predicted values described in the next section.

Again, in the current implementation, only polymorphic SNPs that have missingness below 5% (use "-miss [num]" to change, e.g. "-miss 0.1" changes the threshold to 10%) and minor allele frequency above 1% (use "-maf [num]" to change, e.g. "-maf 0.05" changes the threshold to 5%) will be analyzed. Both missingness and minor allele frequency of a given SNP are calculated based on analyzed individuals (i.e. individuals with no missing phenotypes and no missing covariates).

4.4.3 Output Files

There will be five output files, all inside an output folder in the current directory. The prefix.log.txt file contains some detailed information about the running parameters and computation time. In addition, prefix.log.txt contains PVE estimate and its standard error in the null linear mixed model (not the BSLMM).

The prefix.hyp.txt contains the posterior samples for the hyper-parameters (h, PVE, ρ, PGE,

π and $|\gamma|$), for every 10th iteration. An example file with a few SNPs is shown below:

```
h   pve   rho   pge   pi   n_gamma
4.777635e-01 5.829042e-01 4.181280e-01 4.327976e-01 2.106763e-03 25
5.278073e-01 5.667885e-01 3.339020e-01 4.411859e-01 2.084355e-03 26
5.278073e-01 5.667885e-01 3.339020e-01 4.411859e-01 2.084355e-03 26
6.361674e-01 6.461678e-01 3.130355e-01 3.659850e-01 2.188401e-03 25
5.479237e-01 6.228036e-01 3.231856e-01 4.326231e-01 2.164183e-03 27
```

The prefix.param.txt contains the posterior mean estimates for the effect size parameters (α, $\beta|\gamma == 1$ and γ). An example file with a few SNPs is shown below:

```
chr rs ps n_miss alpha beta gamma
1 rs3683945 3197400 0 -7.314495e-05 0.000000e+00 0.000000e+00
1 rs3707673 3407393 0 -7.314495e-05 0.000000e+00 0.000000e+00
1 rs6269442 3492195 0 -3.412974e-04 0.000000e+00 0.000000e+00
1 rs6336442 3580634 0 -8.051198e-05 0.000000e+00 0.000000e+00
1 rs13475700 4098402 0 -1.200246e-03 0.000000e+00 0.000000e+00
```

Notice that the beta column contains the posterior mean estimate for $\beta_i|\gamma_i == 1$ rather than β_i. Therefore, the effect size estimate for the additional effect is $\beta_i\gamma_i$, and in the special case $\mathbf{K} = \mathbf{X}\mathbf{X}^T/p$, the total effect size estimate is $\alpha_i + \beta_i\gamma_i$.

The prefix.bv.txt contains a column of breeding value estimates $\hat{\mathbf{u}}$. Individuals with missing phenotypes will have values of "NA".

The prefix.gamma.txt contains the posterior samples for the gamma, for every 10th iteration. Each row lists the SNPs included in the model in that iteration, and those SNPs are represented by their row numbers (+1) in the prefix.param.txt file.

4.5 Predict Phenotypes Using Output from BSLMM

4.5.1 Basic Usage

The basic usages for association analysis with either the PLINK binary ped format or the BIMBAM format are:

```
./gemma -bfile [prefix] -epm [filename] -emu [filename] -ebv [filename] -k [filename]
-predict [num] -o [prefix]
./gemma -g [filename] -p [filename] -epm [filename] -emu [filename] -ebv [filename]
-k [filename] -predict [num] -o [prefix]
```

where the "-predict [num]" option specifies where the predicted values need additional transformation with the normal cumulative distribution function (CDF), i.e. "-predict 1" obtains predicted

values, "-predict 2" obtains predicted values and then transform them using the normal CDF to probability scale; "-bfile [prefix]" specifies PLINK binary ped file prefix; "-g [filename]" specifies BIMBAM mean genotype file name; "-p [filename]" specifies BIMBAM phenotype file name; "-epm [filename]" specifies the output estimated parameter file (i.e. prefix.param.txt file from BSLMM); "-emu [filename]" specifies the output log file which contains the estimated mean (i.e. prefix.log.txt file from BSLMM); "-ebv [filename]" specifies the output estimated breeding value file (i.e. prefix.bv.txt file from BSLMM); "-k [filename]" specifies relatedness matrix file name; "-o [prefix]" specifies output file prefix.

4.5.2 Detailed Information

GEMMA will obtain predicted values for individuals with missing phenotype, and this process will require genotype means and phenotype means from the individuals in the training set. Therefore, one should always use the same phenotype column as used in fitting BSLMM. There are two ways to obtain predicted values: use $\hat{\mathbf{u}}$ and $\hat{\boldsymbol{\beta}}$, or use $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. We note that $\hat{\mathbf{u}}$ and $\hat{\boldsymbol{\alpha}}$ are estimated in slightly different ways, and so even in the special case $\mathbf{K} = \mathbf{X}\mathbf{X}^T/p$, $\hat{\mathbf{u}}$ may not equal to $\mathbf{X}\hat{\boldsymbol{\alpha}}$. However, in this special case, these two approaches typically give similar results based on our previous experience. Therefore, if one used the default matrix in fitting the BSLMM, then it may not be necessary to supply "-ebv [filename]" and "-k [filename]" options, and GEMMA can use only the estimated parameter file and log file to obtain predicted values by the second approach. But of course, one can choose to use the first approach which is more formal, and when do so, one needs to calculate the centered matrix based on the same phenotype column used in BSLMM (i.e. to use only SNPs that were used in the fitting). On the other hand, if one did not use the default matrix in fitting the BSLMM, then one needs to supply the same relatedness matrix here again.

The option "-predict 2" should only be used when a probit BSLMM was used to fit the data. In particular, for binary phenotypes, if one fitted the linear BSLMM then one should use the option"-predict 1", and use option "-predict 2" only if one fitted the data with the probit BSLMM.

Here, unlike in previous sections, all SNPs that have estimated effect sizes will be used to obtain predicted values, regardless of their minor allele frequency and missingness. SNPs with missing values will be imputed by the mean genotype of that SNP in the training data set.

4.5.3 Output Files

There will be two output files, both inside an output folder in the current directory. The prefix.log.txt file contains some detailed information about the running parameters and computation time, while the prefix.prdt.txt contains a column of predicted values for all individuals. In particular, individuals with missing phenotypes will have predicted values, while individuals with non-missing phenotypes will have "NA"s.

5 Questions and Answers

1. Q: I tried to run GEMMA to calculate the relatedness matrix, but it looked like 0 SNPs were analyzed, and the output matrix only contained "nan"s.
   A: This can happen when all individuals have missing pheotypes (e.g. "-9" in .fam file). Please refer to the "Missing Phenotypes" section for details.

6 Options

File I/O Related Options

- -bfile [prefix] specify input plink binary file prefix; require .fam, .bim and .bed files
- -g [filename] specify input bimbam mean genotype file name
- -p [filename] specify input bimbam phenotype file name
- -n [num] specify phenotype column in the phenotype file (default 1)
- -a [filename] specify input bimbam SNPs annotation file name (optional)
- -k [filename] specify input kinship/relatedness matrix file name
- -km [num] specify input kinship/relatedness file type (default 1; valid value 1 or 2).
- -c [filename] specify input covariates file name (optional); an intercept term is needed in the covariates file
- -epm [filename] specify input estimated parameter file name
- -en [n1] [n2] [n3] [n4] specify values for the input estimated parameter file (with a header) (default 2 5 6 7 when no -ebv -k files, and 2 0 6 7 when -ebv and -k files are supplied; n1: rs column number; n2: estimated alpha column number (0 to ignore); n3: estimated beta column number (0 to ignore); n4: estimated gamma column number (0 to ignore).)
- -ebv [filename] specify input estimated random effect (breeding value) file name
- -emu [filename] specify input log file name containing estimated mean
- -mu [filename] specify estimated mean value directly, instead of using -emu file
- -snps [filename] specify input snps file name to only analyze a certain set of snps; contains a column of snp ids
- -pace [num] specify terminal display update pace (default 100000).
- -o [prefix] specify output file prefix (default "result")

SNP Quality Control Options

- -miss [num] specify missingness threshold (default 0.05)
- -maf [num] specify minor allele frequency threshold (default 0.01)

- -notsnp [num] minor allele frequency cutoff is not used and so all real values can be used as covariates

Relatedness Matrix Calculation Options

- -gk [num] specify which type of kinship/relatedness matrix to generate (default 1; valid value 1-2; 1: centered matrix; 2: standardized matrix.)

Linear Mixed Model Options

- -lmm [num] specify frequentist analysis choice (default 1; valid value 1-4; 1: Wald test; 2: likelihood ratio test; 3: score test; 4: all 1-3.)
- -lmin [num] specify minimal value for lambda (default 1e-5)
- -lmax [num] specify maximum value for lambda (default 1e+5)
- -region [num] specify the number of regions used to evaluate lambda (default 10)

Bayesian Sparse Linear Mixed Model Options

- -bslmm [num] specify analysis choice (default 1; valid value 1-3; 1: linear BSLMM; 2: ridge regression/GBLUP; 3: probit BSLMM.)
- -hmin [num] specify minimum value for h (default 0)
- -hmax [num] specify maximum value for h (default 1)
- -rmin [num] specify minimum value for rho (default 0)
- -rmax [num] specify maximum value for rho (default 1)
- -pmin [num] specify minimum value for log10(pi) (default log10(1/p), where p is the number of analyzed SNPs)
- -pmax [num] specify maximum value for log10(pi) (default log10(1))
- -smin [num] specify minimum value for —gamma— (default 0)
- -smax [num] specify maximum value for —gamma— (default 300)
- -gmean [num] specify the mean for the geometric distribution (default: 2000)
- -hscale [num] specify the step size scale for the proposal distribution of h (value between 0 and 1, default min(10/sqrt(n),1))
- -rscale [num] specify the step size scale for the proposal distribution of rho (value between 0 and 1, default min(10/sqrt(n),1))

- -pscale [num] specify the step size scale for the proposal distribution of log10(pi) (value between 0 and 1, default min(5/sqrt(n),1))
- -w [num] specify burn-in steps (default 100,000)
- -s [num] specify sampling steps (default 1,000,000)
- -rpace [num] specify recording pace, record one state in every [num] steps (default 10)
- -wpace [num] specify writing pace, write values down in every [num] recorded steps (default 1000)
- -seed [num] specify random seed (a random seed is generated by default)
- -mh [num] specify number of MH steps in each iteration (default 10; requires 0/1 phenotypes and -bslmm 3 option)

Prediction Options

- -predict [num] specify prediction options (default 1; valid value 1-2; 1: predict for individuals with missing phenotypes; 2: predict for individuals with missing phenotypes, and convert the predicted values using normal CDF.)

References

[1] Yongtao Guan and Matthew Stephens. Practical issues in imputation-based association mapping. *PLoS Genetics*, 4:e1000279, 2008.

[2] Bryan N. Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5:e1000529, 2009.

[3] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham. PLINK: a toolset for whole-genome association and population-based linkage analysis. *The American Journal of Human Genetics*, 81:559–575, 2007.

[4] William Valdar, Leah C. Solberg, Dominique Gauguier, Stephanie Burnett, Paul Klenerman, William O. Cookson, Martin S. Taylor, J Nicholas P. Rawlins, Richard Mott, and Jonathan Flint. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genetics*, 38:879–887, 2006.

[5] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modelling with Bayesian sparse linear mixed models. *PLOS Genetics*, in press:http://arixv.org/abs/1209.1341, 2013.

[6] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44:821–824, 2012.