GEMMA User Manual V0.90

Xiang Zhou

March 25th, 2012

Contents

1 Introduction

1.1 What is GEMMA

GEMMA is the software implementing the Genome-wide Efficient Mixed Model Association algorithm [5], which tests for association in genome-wide association studies (GWAS) using a standard linear mixed model to account for population stratification and sample structure. The software calculates Wald [5] or likelihood ratio [5] or score test [1] statistics and p values, and is computationally efficient for large GWAS. The current implementation of GEMMA uses freely available open-source numerical libraries, and can handle approximately 23,000 individuals on a machine with 64Gb memory, in double precision. One can easily modify the code to use float precision for even larger samples.

1.2 The Model

GEMMA considers a standard linear mixed model:

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

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

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 and score). The software can provide either maximum likelihood estimate (MLE) or restricted maximum likelihood estimate (REML) of λ and β, and output the corresponding p value.

1.3 Missing Data

As mentioned in the paper [5], the tricks used in 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 drop individuals with missing genotypes, since it can improve power to detect associations [3]. In the current implementation, 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 the analyzed SNPs will be simply replaced with the estimated mean genotype of that SNP.

2 Installing and Compiling GEMMA

If you have downloaded an executable binary, no installation is necessary. If you downloaded the source code, you will need a standard C/C++ compiler such as GNU gcc, as well as 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/`.

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 in the PLINK binary ped format or in the BIMBAM format which is useful for imputed genotypes.

3.1 PLINK Binary PED File Format

GEMMA recognizes the PLINK binary ped file format (`http://pngu.mgh.harvard.edu/~purcell/plink/`) [4] for both genotypes and phenotypes. This format requires three files: *.bed, *.bim and *.fam, all with the same 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 phenotype column by using "-n [num]", where "-n 1" uses the original sixth column and "-n 2" uses the seventh column as phenotypes, so on and so forth.

3.2 BIMBAM File Formats

GEMMA also recognizes BIMBAM file format (`http://stephenslab.uchicago.edu/software.html`) [3], which is useful for imputed genotypes. BIMBAM format consists of three files, a mean genotype file, a phenotype file, and an optional SNP annotation file.

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, and the rest 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
```

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. Missing phenotype 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 to the phenotype file, and specify a different column for association tests by using "-n [num]", where "-n 1" uses the original first column and "-n 2" uses the second column as phenotypes, 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
```

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 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 format, 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 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 file does not provide individual id, the second format only works with 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)

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 1s 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 GEMMA source code recently, you will find an example folder containing a small GWAS example dataset. This data set comes from Global Wheat program of the International Maize and Wheat Improvement Center (CIMMYT), and is directly downloaded from the supplementary information link of the original research paper [2].

The data set consists of 599 wheat lines and 1279 Diversity Array Technology (DArT) markers. Both genotype and phenotype files are in BIMBAM format, and the relatedness matrix file is in $n \times n$ matrix format. The phenotype file (wheatdata.pheno.txt) contains average grain yield performance for these wheat lines, the genotype file (wheatdata.geno.txt) contains DArT markers taking values of either 0 or 1 (where allele types are pseudo-coded as A and C), and the relatedness matrix file (wheatdata.kin.txt) contains a 599×599 relatedness matrix estimated using the Browse application of the International Crop Information System (ICIS) [2]. This relatedness matrix file is included for completeness, but is not needed for analysis using GEMMA.

A demo.txt file inside the folder shows detailed steps on how to use GEMMA to estimate relatedness matrix from genotypes, and on how to perform a genome-wide analysis thereafter. The results from GEMMA are available inside the result folder.

4.2 Estimate Relatedness Matrix from Genotypes

GEMMA provides two ways to estimate the relatedness matrix from genotypes, by 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 1s. Then the two relatedness matrix GEMMA can calculate are as follows:

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

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

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 "-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.

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.

4.3 Association Tests

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

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

where "-fa [num]" option specifies which frequentist test to use, i.e. "-fa 1" performs Wald test, "-fa 2" performs likelihood ratio test, "-fa 3" performs score test, and "-fa 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.

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.

5 Output Files

All output files will be in an output folder in the current directory.

5.1 Log File: prefix.log.txt

The log file contains detailed information of the run parameters, estimated $\hat{\lambda}$ in the null model, and total computation time. $\hat{\lambda}$ can be used to estimate the proportion of phenotypic variation (PVE) explained by typed genotypes (details will be available in a separate paper). If we denote s_g as the mean of the n diagonal elements of the relatedness matrix $\mathbf{G}$, then the estimated PVE takes the following form:

$$\hat{\text{PVE}} = \frac{s_g \hat{\lambda}}{s_g \hat{\lambda} + 1}$$

5.2 Relatedness Matrix File: prefix.*XX.txt

This output file contains estimated relatedness matrix in the $n \times n$ matrix format, where prefix.cXX.txt stores the centered relatedness matrix and prefix.sXX.txt stores the standardized relatedness matrix. Typically, the diagonal elements of the prefix.cXX.txt matrix are around 0.3, and the diagonal elements of the prefix.sXX.txt matrix are around 1.0.

5.3 Association Tests Result File: prefix.assoc.txt

This output file contains association test results for all tested SNPs.

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, where missing genotype value is represented by "NA"
- -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
- -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)

Relatedness Matrix Calculation Options

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

Association Tests Options

- -fa [num] specify frequentist analysis choice (default 1; valid value 1-4: 1 Wald test, 2 likelihood ratio test, 3 score test, 4 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)

References

[1] Mark Abney, Carole Ober, and Mary Sara McPeek. Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: Fasting serum-insulin level in the hutterites. *American Journal of Human Genetics*, 70:920–934, 2002.

[2] Gustavo de los Campos, Hugo Naya, Daniel Gianola, Jos Crossa, Andrs Legarra, Eduardo Manfredi, Kent Weigel, and Jos Miguel Cotes. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 182:375–385, 2009.

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

[4] 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. *American Journal of Human Genetics*, 81:559–575, 2007.

[5] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed model analysis for association studies. *Nature Genetics, in press*, 2012.