MMM v.1.0

A program for analysing a linear mixed model

User Manual

Matti Pirinen

matti.pirinen@iki.fi

July 19, 2012

Contents

1 Overview

MMM is a C-program for analysing a linear mixed model

$$\boldsymbol{Y} = \mu\mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \gamma\boldsymbol{z} + \boldsymbol{\varrho} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$ is the vector of observations on n subjects, μ is the population mean, $\boldsymbol{X} = (x_{ik})$ is the $n \times K$ matrix of covariate values on the subjects, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^T$ collects the (unknown) linear effects of the covariates on the observations $\boldsymbol{Y}$, $\boldsymbol{z} = (z_1, \ldots, z_n)^T$ is a univariate predictor with effect γ and random effects $\boldsymbol{\varrho}$ and $\boldsymbol{\varepsilon}$ are assigned (prior) distributions

$$\boldsymbol{\varrho}|(\eta, \sigma^2) \sim \mathcal{N}(0, \eta\sigma^2\boldsymbol{R}) \text{ and } \boldsymbol{\varepsilon}|(\eta, \sigma^2) \sim \mathcal{N}(0, (1-\eta)\sigma^2\boldsymbol{I}), \tag{1.2}$$

where $\boldsymbol{R}$ is a known positive semi-definite $n \times n$ matrix, $\boldsymbol{I}$ is the $n \times n$ identity matrix, and parameters $\sigma^2 > 0$ and $\eta \in [0, 1]$ determine how the variance is divided between these two components of variance.

MMM computes maximum-likelihood (ML) estimates for $\mu, \boldsymbol{\beta}, \gamma, \sigma^2$ and η as well as significance tests and Bayes factors for γ and η. With binary data (i.e. each $y_i \in \{0, 1\}$) MMM also gives the corresponding ML estimates on the log-odds scale.

MMM is written in such a way that it is efficient to analyse the model

- for many $\boldsymbol{z}$ vectors while keeping $\boldsymbol{R}$, $\boldsymbol{X}$ and $\boldsymbol{Y}$ fixed (e.g. testing genotype-phenotype associations in genome-wide association studies).
- by specifying $\boldsymbol{R}$ in the block-diagonal form (e.g. analysing families or populations by assuming no relatedness between the groups).

1.1 Reference and Licence

Details of the methods implemented in MMM are described in [7] which should be cited in publications that apply MMM. The source code of MMM (written in C) is distributed under the GNU General Public License 3.

1.2 Installing MMM

If compiled versions do not run on your platform, you can compile the programs (MMM and generateR) by following the instructions in the INSTALL file. Note that both programs need GNU Scientific Library (http://www.gnu.org/software/gsl/) and MMM works much faster with LAPACK support than without it.

1.3 Running MMM

There are two ways to run MMM. First, by collecting all information to a parameters file, the program can be run by the command

```
./MMM parameters_file
```

Second, input/output file names can also be specified on the command line. In this case, a parameters file must be designated by the switch '-P_file' and other files can also be defined by their switches. For example, when other parameters than 'gen_file' and 'results_file' are defined in parameters_file, MMM runs with the command

```
./MMM -P_file parameters_file -gen_file chr12.gen -results_file chr12.out
```

2 Input files

Let n be the number of subjects and $K \geq 0$ the number of covariates (excluding the population mean that is always included in the model).
The input data are specified in GROUPs, where a property of grouping is that there are no a priori correlation between the individuals in different GROUPs. Thus, if individuals $i \leq n$ and $j \leq n$ belong to different GROUPs, then $[\boldsymbol{R}]_{ij} = 0$. This means that individuals can be ordered in such a way that $\boldsymbol{R}$ matrix becomes block-diagonal with each block corresponding to one GROUP.
In what follows we suppose that the subjects have been assigned to $g \geq 1$ GROUPs having $n_1, \ldots, n_g$ subjects, respectively, where $n_1 + \ldots + n_g = n$. It is always possible to use just a single GROUP, but this may be very inefficient if the corresponding $\boldsymbol{R}$ matrix is sparse and could be transformed to the block-diagonal form by permuting rows and columns.

2.1 Attributes file

Attributes file (A-file) gives the outcome $\boldsymbol{Y}$ and possible covariates $\boldsymbol{X}$. It is a single ASCII text file whose first line is a header giving names of the columns in the file (names are strings separated by white space). This file must always have at least one column giving outcome $\boldsymbol{Y}$ and if individual exclusion list is applied (see below), then there must also be a column for the individual identifiers. (In principle, outcome and id columns can be the same one.) After the header line, the GROUPs are listed sequentially from 1 to g. Each group i has a title line which starts with 'GROUPi' and then gives the number of subjects belonging to the GROUP, 'n_i'. Then one line per each subject follows specifying the values of the attributes for that subject in the order determined by the header line. For example, with two GROUPs each with 2 individuals, and with id, case-control status, age and sex as attributes:

```
id case age sex

GROUP1 2

ind1 1 57 0
```

```
ind2 0 63 0

GROUP2 2

ind3 0 49 1

ind4 0 52 1
```

2.2 R-file

R-file specifies the positive semi-definite covariance matrix $\boldsymbol{R}$. R-file is only needed if U-file and D-file are not available. When an R-file is analysed by MMM the program prints out the eigenvalue decomposition $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$ to U-file and D-file. In the future runs on the same individuals, U-file and D-file may be used directly instead of the original R-file. With large matrices this results in a considerable saving of computation time.
However, for the first run with a given data set an R-file needs to be specified. Its format is a single ASCII text file that gives GROUPs sequentially from 1 to g. Each group i has a title line which starts with 'GROUPi' and then gives the number of subjects belonging to the GROUP, 'n_i' and relatedness-indicator 0 or 1 depending on whether the R-matrix corresponding to this GROUP is identity-matrix or not, respectively. If the relatedness-indicator is 0, then no further lines are needed for this GROUP whereas if the indicator is 1, then the lower diagonal matrix (including the diagonal itself) is given on $n_i(n_i + 1)/2$ consecutive lines. Each line has four elements

```
i j w r
```

where i and j are indexes of the two individuals, w is the precision of the coefficient (e.g. number of SNPs used to calculate genetic relatedness) and r is the actual value r_{ij} of row i and column j (which is same as value r_{ji} of column i and row j) of R-matrix. Indexes i and j must run in the row-wise order, i.e. $(1,1),(2,1),(2,2),(3,1),(3,2),\ldots$. The precision parameters are not used by MMM, but they are useful when several R-matrices are added or subtracted, for example, to make chromosome specific matrices in genetic analyses (see section on 'generateR').
For example, to specify 4×4-matrix

$$\boldsymbol{R} = \begin{pmatrix} 1.05 & 0.52 & 0.0 & 0.0 \\ 0.52 & 1.04 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.00 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.00 \end{pmatrix} \tag{2.1}$$

in block diagonal form with two GROUPs, each with 2 individuals, R-file is

```
GROUP1 2 1

1 1 1000 1.05

2 1 1000 0.52
```

```
2 2 1000 1.04

GROUP2 2 0
```

(Note that for MMM the third column could contain any real numbers and 1000 was chosen arbitrarily for this example.) Naturally, the order of individuals within each GROUP MUST be the same as it is in the corresponding attributes file.
For genetics applications, an R-matrix describing relatedness between individuals can be computed from a set of genotypes by using a program 'generateR' (see section 4 of this manual).

2.3 U-file

U-file contains the eigenvectors of the corresponding $\boldsymbol{R}$ matrix. In principle, the user never needs to write a U-file himself/herself since MMM does the job from a given R-file.
U-file is a single ASCII text file that gives GROUPs sequentially from 1 to g. Each group i has a title line which starts with 'GROUPi' and then gives the number of subjects belonging to the GROUP, 'n_i' and relatedness-indicator 0 or 1 depending on whether the R matrix corresponding to this GROUP is identity-matrix or not, respectively. If relatedness-indicator is 0, then no further lines are needed for this GROUP whereas if the indicator is 1, then $n_i \times n_i$ matrix is given, on n_i consecutive lines. The columns of the matrix are the eigenvectors of the corresponding $\boldsymbol{R}$ matrix, normalized to have a unit length.
For example, to specify the U-file for the $\boldsymbol{R}$-matrix (2.1) in the block diagonal format

```
GROUP1 2 1

 -0.7104980  0.7036992

 -0.7036992 -0.7104980

GROUP2 2 0
```

2.4 D-file

D-file contains the eigenvalues of the corresponding $\boldsymbol{R}$ matrix. In principle, a user never needs to write a D-file himself/herself since MMM does the job from a given R-file.
The format is a single ASCII text file that gives GROUPs sequentially from 1 to g. Each group i has a title line which start with 'GROUPi' and then gives the number of subjects belonging to the GROUP, 'n_i' and relatedness-indicator 0 or 1 depending on whether the R matrix corresponding to this GROUP is identity-matrix or not, respectively. If relatedness-indicator is 0, then no further lines are needed for this GROUP whereas if the indicator is 1, then n_i values are given, on n_i consecutive lines. The values are the eigenvalues of the corresponding $\boldsymbol{R}$ matrix, IN THE SAME ORDER as the eigenvectors in the corresponding U-file. If some of the eigenvalues are negative in D-file, indicating that the $\boldsymbol{R}$ matrix was not numerically positive semi-definite, then MMM turns them

to the user specified minimum treshold (default = 0), to get a positive semi-definite approximation.
For example, the D-file for the $\boldsymbol{R}$-matrix (2.1) in the block diagonal format contains

```
GROUP1 2 1

1.565024

0.524976

GROUP2 2 0
```

2.5 z-file

z-file contains the predictors that are tested by the likelihood ratio test and the Bayesian model comparison methods. An alternative format for reading predictor $\boldsymbol{z}$ is from a gen-file as described below. z-file is a single ASCII text file in which each line corresponds to a single $\boldsymbol{z}$ vector. Each line has $n + 1$ elements, the first of which is a string giving the name of the predictor (e.g. rsids for SNPs in GWAS) and the predictor values for n subjects follow. As opposed to the other input files, GROUPing is not explicit in z-file but, naturally, the individuals must be in the same order in z-file as they are in attributes file. For example, to specify z-file for 4 individuals and three SNPs in a GWAS

```
rs23212 0 1 0 2

rs32322 1.1 2 0.01 1

rs98120 1 1.98 0.01 1
```

If this would be run with the previous examples of A-file and R-file, then at rs23212 0 and 1 would correspond to the individuals in GROUP1 and 0 and 2 to the individuals in GROUP2. Note that the values can be any reals (but with SNP data they are usually between 0 and 2).
It is possible to specify a particular value to denote missing data (see 'missing_z' parameter), and either drop the individuals with missing data from the analysis or set their values to the mean of the non-missing values (see 'impute_missing').

2.6 gen-file

This is the format for genotype data used, for example, by the genetic software packages CHIAMO, SNPTEST and IMPUTE[1] and is an alternative way to specify predictor $\boldsymbol{z}$ for MMM. Gen-file is an ASCII text file where each line corresponds to a single SNP. Each line has $5+3n$ entries

[1] http://www.stats.ox.ac.uk/~marchini/software/gwas/gwas.html

```
id1 id2 pos allele_A allele_B p1_AA p1_AB p1_BB ... pn_AA pn_AB pn_BB
```

id1 and id2 are strings giving two names for the SNP (e.g. chromosome and rsid). pos is the physical position on chromosome, allele_A and allele_B give the two alleles at the SNP (among 'A','C','G','T'). For each individual i pi_AA is the probability (given by a genotype calling algorithm) that i is homozygous for allele_A, and similarly for pi_AB and pi_BB. If these three probabilities do not sum to 1, then the remaining probability is assigned to NULL genotype class (i.e. for the event that the calling algorithm did not make a call). There is a possibility to exclude genotypes based on the probability of the NULL class (see 'missing_gen' parameter) and these genotypes can either be dropped from the analysis completely or their genotype can be replaced with the population mean genotype (see 'impute_missing'). For all non-excluded individuals, MMM will use their renormalised average genotype as predictor z in the model, where renormalisation is done over the three non-NULL genotype classes. An example with 2 SNPs on 4 individuals, with NULL class probability 0.02 for the first SNP of the first individual

```
SNP1 rs34 18223 A G 0.9 0.08 0 0 1 0 0 0.99 0.01 0 0 1

SNP2 rs35 19655 A T 1 0 0 1 0 0 0.98 0.02 0 0 1 0
```

Thus the most probable genotypes at SNP1 are AA,AG,AG and GG and for SNP2, AA, AA, AA, and AT, and the $\boldsymbol{z}$ vectors are $(0.0816, 1, 1.01, 2)^T$ for SNP1 and $(0, 0, 0.02, 1)^T$ for SNP2.

2.7 Exclusion file

A single ASCII text file that lists the identifiers of those subjects that will be excluded from the analysis. Ids are listed one per line. The attributes file must have a column that connects the identifiers of the exclusion file to the subjects. For example, to specify that ind2 and ind3 will be dropped from the analysis the exclusion file is simply

```
ind2

ind3
```

Note that if the exclusion file contains identifiers which are not present in the Attributes file, then those ids will be ignored, that is, they cause no exclusions in the analysis. It is always a good practice to check that the number of individuals in the analysis is what it should be after the exclusions are applied (see Output file for individuals).

2.8 Prior file

A single ASCII text file that specifies the prior distributions for the Bayesian analyses in terms of six parameters: $\boldsymbol{m}, \boldsymbol{V}, a, b, r$ and t, by assuming that

$$\begin{aligned} (\boldsymbol{\beta}', \sigma^2) &\sim \text{NIG}(\boldsymbol{m}, \boldsymbol{V}, a, b), \\ \eta &\sim \text{Beta}(r, t), \end{aligned}$$

where $\boldsymbol{\beta}' = (\mu, \boldsymbol{\beta}^T, \gamma)^T$ and NIG$(\boldsymbol{m}, \boldsymbol{V}, a, b)$ is the normal-inverse-gamma distribution with a density function

$$\frac{b^a(\sigma^2)^{-(a+1+K/2)}}{(2\pi)^{K/2}|\boldsymbol{V}|^{1/2}\Gamma(a)} \exp\left(-\frac{1}{2\sigma^2}\left((\boldsymbol{\beta}' - \boldsymbol{m})^T\boldsymbol{V}^{-1}(\boldsymbol{\beta}' - \boldsymbol{m}) + 2b\right)\right).$$

and the density of the beta distribution for $\eta \in [0, 1]$ is

$$\frac{\Gamma(r+t)}{\Gamma(r)\Gamma(t)}\eta^{r-1}(1-\eta)^{t-1}.$$

Thus, $\boldsymbol{m}$ is a $K+2$ dimensional vector of prior expectations for elements of $\boldsymbol{\beta}'$, $\boldsymbol{V}$ is a $(K+2) \times (K+2)$ prior covariance matrix for $\boldsymbol{\beta}'$, and shape parameter a and scale parameter b specify the marginal prior distribution of σ^2 to be Inv-Gamma(a,b). If $\boldsymbol{z}$ is not included in the model, then $\boldsymbol{\beta}' = (\mu, \boldsymbol{\beta}^T)^T$ and $\boldsymbol{m}$ and $\boldsymbol{V}$ are $K+1$ dimensional.

Prior file should always have one line per each of the quantities $\boldsymbol{m}$, a and b. These lines start with tags 'prior_expectation', 'prior_a' and 'prior_b', respectively, which are followed by the corresponding values (dim($\boldsymbol{\beta}'$) values for prior_expectation and single (positive) numbers for prior_a and prior_b).

Matrix $\boldsymbol{V}$ can be specified either on a single line that starts with tag 'prior_variance' and is followed by the diagonal elements of $\boldsymbol{V}$, or by using tag 'prior_variance_matrix' which is followed by the full $\boldsymbol{V}$ matrix given on dim($\boldsymbol{\beta}'$) consecutive rows (each having dim($\boldsymbol{\beta}'$) elements).

The parameters $r \geq 1$ and $t \geq 1$ (in this order) are specified on a same line after the tag 'prior_beta'. If no tag 'prior_beta' is found, then $r = 1$ and $t = 1$ defining a uniform prior on η.

The following two examples specify prior distribution for $K = 1$ covariates with $\boldsymbol{m} = (0, 0, 0)^T$, $\boldsymbol{V} = \text{diag}(10, 10, 0.1)$, $a = 2$, $b = 3$, $r = 1$ and $t = 5$.

Prior file that gives the matrix $\boldsymbol{V}$ in the diagonal form

```
prior_expectation 0 0 0

prior_variance 10 10 0.1

prior_a 2

prior_b 3

prior_beta 1 5
```

and prior file that gives $\boldsymbol{V}$ in the full form

```
prior_expectation 0 0 0

prior_variance_matrix

10 0 0

0 10 0
```

```
0 0 0.1

prior_a 2

prior_b 3

prior_beta 1 5
```

Note that specifying $\boldsymbol{V}$ whose non-diagonal elements are non-zero is possible only by using the full form.
With MMM some R-code is distributed that can be used in choosing appropriate prior parameters also for binary data. See file 'MMM_priors.R'.

2.9 Parameters file

Parameters file collects the information that is needed for running MMM. Its format is an ASCII text file where each line contains two items, a 'tag' and a 'value' (Table I), except for the tag 'covariates' for which there can be several values on the same line. Possible tags and their meanings are

- 'n_groups' is g, the number of groups.
- 'outcome' specifies the name of the column in A-file that is used as outcome $\boldsymbol{Y}$.
- 'covariates' specifies K, the number of covariates used in the analysis, and lists the names of those covariates as given on the header line of A-file.
- 'id' specifies the name of the column in A-file that is used to identify the individuals in exclusion file.
- if 'bayesian'=1, BFs for γ's and η are computed.
- if 'logOR'=1 or 2, effects and their standard errors are transformed to the log-odds scale, otherwise (0) effects are on the linear scale. If this is 1 or 2 then the outcome must be 0 or 1 for every individual. Option 2 is to be used for GWAS application as it gives more accurate effects for the tested z-predictor WHEN z is a genotype coded 0,1,2. With other predictors than genotypes, use 1. See Additional notes for details.
- if 'mean_center'=1, each $\boldsymbol{z}$ is mean centered. This is important for accuracy if the effects are measured on the log-odds scale.
- if 'check_geno'=1, each z_i is checked to lie within $[0, 2]$ (before possible mean centering). This is an extra check for valid SNP data and has an effect only if data is read from a z-file.
- 'tol' gives the tolerance for maximisation algorithm.
- 'min_d' gives the lower bound for accepted eigenvalues of $\boldsymbol{R}$. Eigenvalues less than min_d are set equal to min_d in the analysis.

- if 'fixed_eta' is defined, then η parameter in the model is fixed to that value and not estimated by maximum-likelihood. Results in reductions in both the running time and the accuracy of the results. Does not have an effect on the Bayesian model comparisons.
- if 'missing_z' is defined, then all elements of $\boldsymbol{z}$ that have that value in z-file are considered missing.
- if 'missing_gen' is defined, then all genotypes (in gen-file) whose non-NULL probability is less than this value are considered missing.
- if 'impute_missing'=1, then missing data (in z-file or gen-file) is set to the mean of the non-missing data. If value is 0, then all subjects with missing data are dropped from the analysis. This is only possible when R-file is specified and bayesian=0, and it may be slow for large GROUPs since matrix decompositions may need to be done anew at each $\boldsymbol{z}$.
- if 'save_U_D'=1 and R-file is specified, then MMM outputs the corresponding decomposition $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$ to U and D-files.
- max-parameters control the accuracy of maximisation. Higher values give higher accuracy, and the default values seem to work well.
- 'n_integr_intervals' is the number of intervals to which [0,1] is divided when numerical integration over η is done for computing Bayes factors.

2.10 Command line parameters

All 10 parameters in Table I ending with '_file' can also be read from the command line by using switch '-TAG' followed by the path to the file (e.g. -A_file attr.txt to read Attributes file from file 'attr.txt'). If any of these paths are defined on the command line, then the same TAG cannot be defined in the parameters file and the parameters file must be defined on the command line with the switch '-P_file'. (See subsection Running MMM.)

3 Output files

3.1 Output file for individuals

Gives the list of individuals in the Attributes file (using 'id' column as labels) with information on the GROUP of the individual and whether the individual was excluded from the analysis. By default, the name of this file is 'MMM_inds.out'.

Table I: Parameter file

TAG	VALUE	required?
A_file	path to Attributes file	always
R_file	path to R-file	either R or U-file
U_file	path to U-file	either R or U-file
D_file	path to D-file	with U-file
z_file	path to z-file	
gen_file	path to gen-file	
E_file	path to Exclusion file	
results_file	path to results file	
inds_out_file	path to output file for individuals	
n_groups	positive integer	always
outcome	string	always
covariates	integer, 0 (default) + list of covariate names	
id	string	with E_file
bayesian	0 (default) or 1	
prior_file	path to prior-file	if bayesian=1
logOR	0 (default) or 1 or 2	
mean_center	0 (default) or 1	
check_geno	0 (default) or 1	
tol	positive real, 1e-9 (default)	
min_d	positive real, 0 (default)	
fixed_eta	[0,1]	
missing_z	real	
missing_gen	[0,1], 0.1 (default)	
impute_missing	0 (default) or 1	
save_U_D	1 (default) or 0	
max_iterations	positive int, 100 (default)	
max_eta_iterations	positive int, 100 (default)	
max_eta_intervals	positive int, 10 (default)	
max_eta_divisions	positive int, 10 (default)	
n_integr_intervals	positive int, 100 (default)	

3.2 U and D-files

If 'save_U_D'=1 then MMM writes out the eigenvalue decomposition of the $\boldsymbol{R}$ matrix AFTER the exclusion list was applied to files with suffixes _U.out and _D.out added to the name of R_file. Note that if some individuals are excluded from the analysis then U and D files do not correspond to the full $\boldsymbol{R}$ matrix, but instead these files should be stored and interpreted together with the corresponding individual output file that tells which individuals and in which order were included in the decomposed matrix. In the future runs U and D files can be used together with the original attributes file and z-file or gen-file AS LONG AS the same exclusion list (E-file) is provided as when U and D files were computed.

3.3 Results file

If the results file is not specified, then it will be formed by adding the suffix '_res.out' to the name of 'z_file' or 'gen_file' if one of them is specified, or to the name of the outcome variable if neither of the predictor files is used.
All ML-estimates are from the full linear mixed model, (with or without η fixed depending on the parameters file), excepting columns '_eta0' which are from the standard linear model i.e. $\eta = 0$. Columns in the results file include some subset of the following

- 'z_id' from z_file or 'snp_id1', 'snp_id2', 'pos', 'allele_0', 'allele_1' from gen_file are simply copied to the results file with allele 0 in results file corresponding to allele A in gen file and allele 1 corresponding to allele B.

- 'n_included' is the number of subjects included in the analysis. If missingness is properly modelled ('impute_missing'=0) then this may vary among predictors. Otherwise this is the same for all predictors.

- 'n_missing' is the number of individuals who have been excluded from the analysis because their predictor value is missing according to the given tresholds. This is non-zero only if 'impute_missing'=0, and in that case 'n_included'+'n_missing' equals the number of individuals after exclusions in E-file have been done.

- 'n_imputed' is the number of individuals that were set to carry the population mean predictor value (see parameter 'impute_missing').

- 'error' is one of 0=succesfull; 1=initialisation failed (e.g. covariates/predictor are colinear as with a monomorphic SNP in a GWAS); 2=maximisation did not converge (try to increase iteration parameters); 3=error reading the predictor from z_file/gen_file.

- 'pop_mean_est' and 'pop_mean_se' are the estimate and standard error of the population mean parameter μ under the linear mixed model. If 'logOR=1 or 2' then these are on the log-odds scale.

- 'covar_?_est' and 'covar_?_se' are the estimate and standard error of the corresponding covariate ($\boldsymbol{\beta}$ coefficients). If 'logOR=1 or 2' then these are on the log-odds scale.

- 'var_est' and 'var_se' are the ML estimate and standard error of the variance parameter σ^2.
- 'eta_est' and 'eta_se' are the ML estimate and standard error of the variance parameter η.
- 'loglkhood' is the maximum log-likelihood (base e) of the full model.
- 'z_est_eta0' and 'z_se_eta0' are the ML estimate and its standard error of the predictor's effect γ under the model where $\eta = 0$, i.e., under the standard linear model. If 'logOR=1 or 2' then these are on the log-odds scale. If gen-file is used, then the effects are reported for the allele B which is called allele 1 in MMM output. In case-control GWAS, where either gen-file is used or z codes for genotype (0,1,2), use 'logOR=2' to get more accurate estimates than with 'logOR=1'.
- 'chisq_eta0' and 'pval_eta0' are the test statistic and the corresponding p-value from the likelihood-ratio test for the predictor's effect γ under the model $\eta = 0$.
- 'z_est', 'z_se', 'chisq' and 'pval' are the corresponding quantitites from the full linear mixed model.
- 'log10BF_eta0' and 'log10BF' are the log-10 of the Bayes factors for the predictor's effect γ, under the restricted model $\eta = 0$ and the full model $\eta \sim \text{Beta}(r, t)$, respectively.
- 'lr_stat_eta' is the likelihood ratio statistic between models $\eta > 0$ and $\eta = 0$ and is only written to the output file if no predictor file is read in.
- 'pval_score_eta' is a p-value for the hypothesis $\eta = 0$ by using a score test (see subsection "Testing whether $\eta = 0$") and is only written to the output file if no predictor file is read in.
- 'log10BF_eta' is the log-10 of the Bayes factor comparing the full model to the model with $\eta = 0$. Only written to the output file if no predictor is read in. Otherwise this is written to the standrd output before any predictor is analysed.

  The following are printed only when the input is read from a gen-file.

- '00','01','11' and 'NULL' are the sums of the probabilities of the genotype classes 0/0, 0/1, 1/1 and NULL, respectively, over all individuals who were both not excluded (in E-file) and did not have missing/imputed data at this SNP. Allele 0 corresponds to 'allele_A' in gen-file and allele 1 corresponds to 'allele_B'.
- 'freq_1' is the allele frequency of the allele 1 among non-excluded, non-missing genotypes.
- 'var_info_nonmissing' and 'var_info_all' are information measures on the genotypes as described below.

Information measures. The idea is to quantify how much information there is about the individual genotypes in the gen-file relative to the information about the genotypes based only on the estimated population allele frequencies. This is an extension of the info measure used in the software packages SNPTEST and IMPUTE [4] to the setting where there are non-zero probabilities for the NULL genotype class[2].

Let G_i denote the number of copies of allele 1 (allele B in gen-file) that the individual i carries. We model $G_i = X_i I_E + S(1 - I_E)$, where I_E is the indicator of the event that the genotype is sampled from the individual specific probability distribution and its complement is that the genotype is sampled from the population. Event E occurs with probability $1-p_N^{(i)}$, where $p_N^{(i)}$ is the probability of the NULL genotype class for individual i in gen-file. Thus, with probability $1 - p_N^{(i)}$, G_i equals the value of the random variable

$$X_i = \begin{cases} 0, & \text{with probability } p_{AA}^{(i)}, \\ 1, & \text{with probability } p_{AB}^{(i)}, \\ 2, & \text{with probability } p_{BB}^{(i)}, \end{cases}$$

where $p_{AA}^{(i)}, p_{AB}^{(i)}$ and $p_{BB}^{(i)}$ are the genotype probabilities from the gen-file that have been renormalised[3] to sum to one. Under the complementary event, which occurs with probability $p_N^{(i)}$, G_i equals the value of a random variable $S \sim \text{Bin}(2, f_B)$ which models a genotype sampled from the population under Hardy-Weinberg equilibrium, given the observed allele frequency f_B of the allele B among the non-excluded, non-missing genotypes at this SNP.

For each individual i let v_i be the variance of his/her genotype distribution

$$v_i = \text{Var}(G_i) = \left(1 - p_N^{(i)}\right) \text{Var}(X_i) + p_N^{(i)} \text{Var}(S) + p_N^{(i)} \left(1 - p_N^{(i)}\right) (\text{E}(X_i) - \text{E}(S))^2,$$

and define $v_{HW} = \text{Var}(S) = 2f_B(1 - f_B)$ as the variance of a genotype sampled from the population under Hardy-Weinberg equilibrium. The info measure is defined by

$$\text{var\_info} = 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{v_i}{v_{HW}} ,$$

where the sum is either over the individuals who do not have missing genotypes according to the defined thresholds ('var_info_nonmissing') or over all individuals not mentioned in the E-file ('var_info_all').

The value of the info measure is 1 if and only if all genotypes are determined without uncertainty (every $v_i = 0$) and the info decreases by $\frac{1}{n}$ per each genotype whose NULL probability is 1 or whose variance (for some other reason) equals the variance under Hardy-Weinberg equilibrium. In theory this info measure can also be negative (e.g. if all individuals have $p_{AA} = p_{AB} = p_{BB} = \frac{1}{3}$).

[2]Thanks to Gavin Band for ideas and discussions.

[3]renormalisation has an effect only if the NULL class probability $p_N^{(i)} > 0$

4 generateR for computing R-matrix

An application software '*generateR*' is provided for computing an $\boldsymbol{R}$ matrix from a set of variables formatted either as a z-file or as a gen-file. For example, *generateR* can use a genome-wide panel of SNPs to compute a genetic relatedness matrix between individuals. Suppose that an input file (either z-file or gen-file) with L lines each with n columns is provided. *generateR* outputs matrix $\boldsymbol{R}$ whose element (i, j) is

$$\boldsymbol{R}_{ij} = \frac{1}{|M_{ij}|} \sum_{\ell \in M_{ij}} \frac{(z_{\ell i} - 2\hat{p}_\ell)(z_{\ell j} - 2\hat{p}_\ell)}{2\hat{p}_\ell(1 - \hat{p}_\ell)},$$

where M_{ij} is the set of loci ℓ where neither i nor j has missing data, $\boldsymbol{z}_\ell$ is the z-vector corresponding the line ℓ of the input file and $\hat{p}_\ell$ is the sample estimate of the frequency of allele 1 at locus ℓ (gen-file) or half of the sample mean of $\boldsymbol{z}_\ell$ vector (z-file). For gen-files $\boldsymbol{z}_\ell$ corresponds to the expected genotypes after possible renormalisation, for more details see the section about gen-file. Note also that if z-file contains haploid data, i.e., values are 0 or 1, then the denominator should be changed to $2\hat{p}_\ell(1 - 2\hat{p}_\ell)$ in the source code.
generateR is run from the command line with options given as arguments. The possible options and their values are

- -n, the number of individuals. One less than the number of columns in z-file or $(g - 5)/3$ where g is the number of columns of gen-file.
- -z_file, the path to z-file (or '-' for stdin).
- -gen_file, the path to gen-file (or '-' for stdin).
- -out_file, the path to output R-file.
- -out_matrix_file, the path to output matrix file.
- -exclusion_file, the path to the file that contains the ids in the z-file (1st col) or gen-file (2nd col) of the lines that will be excluded from the computation.
- -missing_z, the value that labels the missing data in z-file.
- -missing_gen, the threshold that defines missingness in gen-file. A genotype is considered missing if non-NULL probability is less than this value. Default 0.10.
- -maf, minor allele frequency threshold. If $\min\{\hat{p}_\ell, 1 - \hat{p}_\ell\}$ is less than this value then z_ℓ is excluded from the calculations. Default 0.0.
- -missing, missingness threshold. If proportion of individuals with missing data is over this value then z_ℓ is excluded from the calculations. Default 1.0.
- -info, info threshold. If var_info for a SNP in the gen-file is less than this value, then the SNP is excluded. Default 0.0.

For example, *generateR* could be run by either of the following commands

```
./generateR -n 1000 -gen_file in.gen -out_file R.out -maf 0.01
   -missing 0.02 -missing_gen 0.1 -info 0.8

./generateR -n 1000 -z_file in.z -out_file R.out -maf 0.01
   -missing 0.02 -missing_z -9
```

Note that you need to add a header line to the output matrix of *generateR* before it can be read by MMM.
generateR also prints out the individuals who have more than 0.02 missing data and the pairs of individuals whose relatedness is above 0.90. These threshold values can only be changed at the source code.

4.1 Output files

generateR prints out the matrix information in R-file format when -out_file is specified. Such a file contains one line per each pair of individuals, including a line where an individual is paired with itself:

```
ind1 ind2 N r
```

Here the indexes run from 1 to n and ind1 $\geq$ ind2. N is the number of non-missing data points on which the relatedness estimate r is based on and can vary between pairs of individuals. This format is useful when several matrices need to be combined because in N column it contains the information on how different r estimates should be weighted. There are the following options to operate on R-files:

- -add, paths to two R-files which are combined.
- -subtract, paths to two R-files of which the second is subtracted from the first one.
- -to_matrix, path to R-file that is transformed to a matrix file.

All these commands must be accompanied with -n option and with output file given by -out_list or -out_matrix (or both). One of the input files can be read from the standard input by using '-' as the file name.
There is also an option to print the matrix as full symmetrix matrix by using switch -out_matrix. It is possible to use both -out_file and -out_matrix in the same run.

5 Additional notes

5.1 Reading data from the standard input

It is possible to read (at most) one input file for MMM from the standard input. This is done by specifying the corresponding filename as '-' in the parameters file or on the

command line. For example, to read gen-file from stdin, the parameters file should have a line

```
gen_file -
```

or a command line could be

```
./MMM -P_file parameters_file -gen_file -
```

By default, when z/gen-file is read from stdin, the results will be written to '-_res.out' which can be changed via the tag 'results_file'. When R-file is read from stdin, the decomposition is written to '-_U.out' and '-_D.out' unless a line 'save_U_D 0' is found in the parameters file.
You can also read one of the input files from the standard input for the program 'generateR' by using '-' as the file name.

5.2 GLS approximation

A generalised least squares (GLS) approximation to the full linear mixed model is achieved by first running MMM without z-file/gen-file to get a maximum likelihood estimate for the parameter η without the predictors, and then using 'fixed_eta' tag in the parameter file to fix η to that value also with z-file/gen-file. This procedure results in a reduction of the running time when many predictors are tested, but in some situations may be noticeably less accurate than the full mixed model analysis [7]. In GWAS, where the individual genetic effects are small, the GLS approximation is usually very good.

5.3 Accuracy of log-odds estimates

When estimating log-odds you should ALWAYS set 'mean_center 1' in the parmeters file. The linear model approximation to the logistic regression model works best when the outcome is well balanced, say proportion of cases is between 30% and 70%, the effect on the log-odds scale is not large, say the absolute value are less than 0.4 (odds-ratios are less than 1.5), and the tested predictors are not very rare. The effect sizes on log-odds scale are also distorted if the model includes covariates with large effects. However, for case-control GWAS where the tested predictor codes for a genotype (0,1,2), accurate log-odds estimate for the genetic effect results even in less balanced settings by using 'logOR 2' instead of 'logOR 1' in the parameters file. This does not affect the effect size estimates of the covariates nor the estimated p-values of the predictor. For details, see [7] where 'logOR 1' is called the First order approximation (FOA) and 'logOR 2' is called the GWAS approximation.

5.4 Zero p-values

MMM outputs chi-square values computed as twice the difference of the log-likelihoods between the full model and the null model as well as the corresponding p-values by

assuming that under the null hypothesis the chi-squares have the chi-square distribution with 1 df. If a p-value is less than 10^{-16} then MMM sets it to 0. More accurate p-values in those cases can be computed for example in the R software package by command

```
pchisq(chisq,df=1,lower=FALSE)
```

5.5 Problems when $\eta = 1$ and standard errors are NA

If the ML-estimate of η is close to the boundary value of 1, MMM may not be able to produce standard errors of the parameter estimates due to singularity of the information matrix. However, the chi-square values and p-values are still valid as they come from likelihood-ratios, not from the estimated standard errors.

In these cases one should make sure that no eigenvalues are zero, since then the model is ill-defined at $\eta = 1$. This can be achieved by setting min_d to some small positive number such as 0.01. If any eigenvalue is smaller than 0.01, MMM restricts $\eta < 0.99$ to avoid problems.

To get standard errors you may try setting η to fixed value say 0.99 ('fixed_eta 0.99' in parameters file) and checking whether that changes chisqs and z_est compared to your current results. If the results are not changed, then you can use those SEs.

You can also estimate 'z_se' using 'z_est' and 'chisq' simply by

$$\text{z\_se} = \sqrt{(\text{z\_est})^2/\text{chisq}}.$$

This is like asking that if this 'chisq' came from a Wald's test and I know the point estimate 'z_est' what is the corresponding standard error.

In genetic studies the most common reason for getting $\eta = 1$ is that the R-matrix either includes duplicated individuals or it has been computed from a panel of bad quality genetic variants that spuriously strongly predict the phenotype. For example, if cases and controls are genotyped separately and careful quality control has not been applied, then there may be some SNPs that have failed in only one of the collections and including those in the matrix calculations tend to increase η and also reduce power to see real genetic effects.

Even if there are the above mentioned reasons why $\eta = 1$ is suspicious in genetic studies there is not necessarily anything wrong with data sets where $\eta = 1$ (as long as the eigenvalues are not near 0 in the same time).

5.6 Controlling for population and family structure

When the mixed model is used to control for relatedness structure in genetic association studies then the relatedness matrix would ideally be computed for each variant separately by leaving the tested variant and its close neighbours out from the matrix [5, 6]. The reason is that if the tested variant is also in the matrix, then we lose a bit of power to see the true effect at that locus. A practical way to implement this is to compute a separate matrix for each chromosome that would contain the (relevant) variants from the other

chromosomes (but see also the solution of [6]). With really large data sets it may be feasible to do the matrix computations only once, in which case one can just use a single genome-wide matrix, with a slight loss of power at some of the tested variants.

When the purpose of the matrix is to describe genome-wide relatedness, then one should (probably) thin the variants to a less dependent set so that any one region of the genome does not have a disproportionally strong effect on the matrix. For example, in [8] we used the same set of SNPs for the matrix calculation as we used for investigating population structure via principal components analysis. (Good quality, common SNPs, not in strong LD with each other and lying outside of the special regions such as MHC.)

When putative associations have been identified by the mixed model it is important to check whether including a few leading principal components of the genetic population structure as covariates in the model noticeably reduces their signals. If that is the case then the variant is geographically highly differentiated and thus it is hard to rule out a spurious association. If a suitable panel of PCA SNPs has been used to compute an R-matrix (via, e.g., 'generateR'), then the leading principal components can be extracted from the first columns of the corresponding U-matrix. For example, in Unix PCs 1-3 are extracted by the command

```
awk 'NR>1 {print $1,$2,$3}' U-filename > PCs1-3.txt
```

Then they can be attached to an attributes file and read in as covariates in MMM.

5.7 Testing whether $\eta = 0$ (region-based tests in genetics)

When no predictor file (z-file or gen-file) is specified, MMM outputs the ML-estimates for the covariates and the variance parameters together with three quantities for assessing evidence that $\eta > 0$.

1. Likelihood ratio statistic ('lr_stat_eta') is $2\log(L_1/L_0)$ where L_1 and L_0 are the values of the maximised likelihoods of the full model and the model where $\eta = 0$, respectively. The true null distribution of LR-statistic is complex [1], but in some cases it seems to be well approximated by $0.5\delta_0 + 0.5\chi^2_1$, i.e., a 50:50 mixture of the point mass at zero and a chi-square variate with one degree of freedom. However, it is not clear when this approximation is particularly good/bad, so it should be used with caution.

2. A score test p-value ('pval_score_eta') is computed from a score statistic:

$$\sum_{i=1}^{n}(d_i - 1)\frac{(\widetilde{\boldsymbol{Y}}_i - \widetilde{\boldsymbol{X}}_i\widehat{\boldsymbol{\beta}})^2}{\widehat{\sigma^2}}, \tag{5.1}$$

   where $\widetilde{\boldsymbol{Y}} = \boldsymbol{U}^T\boldsymbol{Y}$ and $\widetilde{\boldsymbol{X}} = \boldsymbol{U}^T\boldsymbol{X}$ are transformed data and covariates, $\boldsymbol{U}$ is the matrix of eigenvectors of $\boldsymbol{R}$, $(d_i)_{i=1}^n$ are the corresponding eigenvalues and the ML-estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma^2}$ come from the null model where $\eta = 0$.

Under the null model $\eta = 0$, this score statistic (5.1) is distributed as a mixture

$$\sum_{i=1}^{n}(d_i - 1)\chi^2_{1,i},$$

where each $\chi^2_{1,i}$ is an independent draw from the central chi-square distribution with one degree of freedom. MMM implements the p-value computations from this distribution by using Davies method [2] as recently implemented in the R-package CompQuadForm [3]. Previously, similar approach has been used by [9].

3. The log10 of the Bayes factor between the models $\eta \sim \text{Beta}(r, t)$ and $\eta = 0$ is given by 'log10BF_eta'.

5.8 Problems with large $\boldsymbol{R}$ matrix (>20,000 × 20,000)

If the dimension of $\boldsymbol{R}$ matrix is much larger than 20,000, then the current version of LAPACK (3.3.0) seems not to be able to handle it and the program exits with a segementation fault. This is likely to be a known problem with LAPACK, see bug0020 in LAPACK errata (http://www.netlib.org/lapack/Errata/).
A way around this would be to do the decomposition of the matrix by some other software than LAPACK (but which?!?) and then input the decomposed matrix through U and D-files to MMM. MMM also has ability to do the decomposition using GSL instead of LAPACK (set macro USE_LAPACK to 0 in the source code), but this approach takes over a week for matrices of this size and is thus not practical. A simple way around the problem is to split the data into subparts of smaller dimension, if there is a reasonable way to do that in the context considered.

5.9 Diagonal $\boldsymbol{R}$ matrix

When $\boldsymbol{R}$ matrix is diagonal, then there is no information about η and as a consequence MMM outputs standard errors as NA except for the column z_se_eta0, which in this situation can be used as the correct standard error also for the full model.

5.10 PLINK-files

To transform other genetic data formats into gen files you may use GTOOL, freely available at www.stats.ox.ac.uk/~marchini/software/gwas/gtool.html

5.11 Generating R-matrix

You may use additional program 'generateR' to generate an R-matrix from either z-file or gen-file formatted set of variables, see section 4 of this manual. With genetic data careful quality control should be applied to the panel of genetic variants BEFORE computing $\boldsymbol{R}$. See "Problems when $\eta = 1$" above.

5.12 MMM in use

MMM has been used in

- IMSGC&WTCCC2. (2011) Genetic Risk and a Primary Role for Cell-mediated Immune Mechanisms in Multiple Sclerosis. Nature. 476: 214-219.

Acknowledgements

Thanks to Gavin Band, Le Si Quang and Chris Spencer for their comments and suggestions about the program and Dan Davison for help with the matrix decompositions.
This work was carried out at the Wellcome Trust Centre for Human Genetics, University of Oxford as part of the Wellcome Trust Case Control Consortium 2 project (WTCCC2). Funding for WTCCC2 came from the Wellcome Trust [085475/B/08/Z and 085475/Z/08/Z] and this work was also funded by the Wellcome Trust core grant for Wellcome Trust Centre for Human Genetics [090532/Z/09/Z].

References

[1] Crainiceanu CM and Ruppert D. (2004).*Likelihood ratio tests in linear mixed models with one variance component.* J R Statist Soc B. 66, 165-185.

[2] Davies DB. (1980). *Algorithm AS 155: The distribution of a linear combination of* χ^2 *random variables.* J R Statist Soc C. 29, 323-333.

[3] Duchesne P and Lafaye de Micheaux P. (2010). *Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods.* Computational Statistics and Data Analysis. 54, 858-862.

[4] Marchini J and Howie B. (2010). *Genotype imputation for genome-wide association studies.* Nat Rev Genet. 11, 499-511.

[5] Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI and Heckerman D. (2011) *FaST linear mixed models for genome-wide association studies.* Nat Methods. 8, 833-835.

[6] Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E and Heckerman D. (2012) *Improved linear mixed models for genome-wide association studies.* Nat Methods. 9, 525-526.

[7] Pirinen M, Donnelly P and Spencer C. (2012). *Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies.* Ann Appl Stat. (in press)

[8] IMSGC&WTCCC2. (2011). *Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis.* Nature. 476: 214-219.

[9] Wu MC, Lee S, Cai T, Li Y, Boehnke M and Lin X. (2011). *Rare-variant association testing for sequencing data with the sequence kernel association test.* Am J Hum Genet. 89, 82-93.