Gene-based Association Testing of Dichotomous Traits with Generalized Linear Mixed Models for Family Data[1]

Yingda Jiang, Chi-yang Chiu, Daniel E. Weeks, and Ruzong Fan

Georgetown University Medical Center and University of Pittsburgh

January 2020

1 Overview

This document describes a R package to implement family-based additive generalized linear mixed models (GLMM) and generalized functional linear mixed models (GFLMM) for gene-based association testing of dichotomous traits. Section 2 briefly describes the installation of the program. Section 3 describes the data formats. Section 4 explains how to run the program using one example. Section 5 offers explanation of the results and warnings to use the programs. Section 6 provides some suggestions and parameter choices for real data analysis.

The theoretical basis for this program is given in our research papers in **References**. Please refer to the references if you use the program in any published work. In case of suggestions and questions and/or problems, you can contact us via e-mail (chiu@uthsc.edu, weeks@pitt.edu, and rf740@georgetown.edu).

2 Download and Installation

The package is written in R language. Download R codes PedGFLMM.zip. Unzip it to get "PedGFLMM_fixed_model.R", "PedGFLMM_beta_smooth_only.R", "PedGLMM_additive_effect_model.R", and an example file of "Example_PedGFLMM.R". Put the files in a directory you may access. The software needs library(pedigreemm) and library(fda).

3 Data Format

The program needs data.frame in R to define the pedigree structure (typical format used by LINKAGE and PLINK), genotypes, SNP positions, and covariates. The columns in the data.frame must be named as follows:

1. pedigree file which is the same as that of pedgene except for a column named "ID" (Schaid et al. 2013) and has the following 7 variables:

    - ID: identify of each individual.
    - ped: pedigree ID, character or numeric allowed.
    - person: person ID, a unique ID within each pedigree, numeric or character allowed.
    - father: father ID, NA if no father.
    - mother: mother ID, NA if no mother.
    - sex: coded as 1 for male, 2 for female.
    - trait: phenotype, either case-control status coded as 1 for affected and 0 for unaffected. Subjects with missing (NA) will be removed from the analysis.

2. genotype file is a matrix with genotypes for subjects (rows) at each variant position (columns). The first two columns are required to be named "ped" and "person", which are used to match subjects to their data in the pedigree data.frame. The genotypes are coded as 0, 1, 2 for autosomal markers (typically a count of the number of the minor allele).

3. map file provides SNP positions for each SNP. The first column is required for the chromosome number, the second column is for the name of SNPs in the genotype file, and the third column is the position of SNPs in base pairs.

4. covariate file contains covariates. The first two columns are required to be named "ped" and "person", which are used to match subjects to their data in the pedigree data.frame.

4 How to Run the Program

The analysis needs libraries *fda, MASS, Matrix*, and *pedigreemm* in R package. Make sure to install them before running our codes. Open the "Example_PedGFLMM.R" file on an R Console in a PC window. Change the paths leading to the directories of the package "PedGFLMM_fixed_model.R", "PedGFLMM_beta_smooth_only.R", "PedGFLMM_additive_effect_model.R", and the datsets on your computer. Then, you may run the program. The following results are based on the datasets in data.zip by "R x64 3.5.1".

```
> add=PedGLMM_additive_effect_model(ped=Ped, geno = as.matrix(geno),
         covariate = as.matrix(cov), optimizer = "bobyqa", Wald = TRUE)
#fixed-effect model matrix is rank deficient so dropping 39 columns / coefficients
#boundary (singular) fit: see ?isSingular
> add
$`LRT`
[1] 6.076317e-14
$Wald
[1] 0.9999999

> fixed_bsp=PedGFLMM_fixed_model(ped = Ped, geno = as.matrix(geno), ... )
> fixed_fsp=PedGFLMM_fixed_model(ped = Ped, geno = as.matrix(geno), ... )
> fixed_bsp
$`LRT`
[1] 0.0002356555
$Wald
[1] 0.001354728
```

```
> fixed_fsp
$`LRT`
[1] 0.0002147177
$Wald
[1] 0.001233355

> bsmooth_bsp=PedGFLMM_beta_smooth_only(ped = Ped, geno = as.matrix(geno), ... )
> bsmooth_fsp=PedGFLMM_beta_smooth_only(ped = Ped, geno = as.matrix(geno), ... )
> bsmooth_bsp
$`LRT`
[1] 0.0002356555
$Wald
[1] 0.001354729

> bsmooth_fsp
$`LRT`
[1] 0.0002147177
$Wald
[1] 0.001233353
```

5 Explanation of the Results and Warnings

As shown in the Section 4, our program outputs p-values of two tests: likelihood ratio test (LRT) and Wald test. The LRT is conservative and has good power performance. The Wald test can be too conservative and has low power. In real data analysis, we suggest to report the p-values of LRT.

6 Suggestions and Parameters for Real Data Analysis

In this documentation, we present three R functions to perform gene-based association analysis of dichotomous traits using family data. The two PedGFLMM functions, "PedGFLMM_fixed_model.R" and "PedGFLMM_beta_smooth_only.R", usually provide very similar results. In practice, one may use one of them for data analysis. We suggest to use PedGFLMM_fixed_model by either B-spline or Fourier spline basis functions and report the p-values of LRT. If the number of SNPs is not very big, we suggest the following parameters for a data analysis:

```
order  =  4
bbasis = 10
gbasis = 10
fbasis = 11
gfasis = 11
```

If the number of SNPs is large, one should try

```
order  =  4
bbasis = 15
gbasis = 15
fbasis = 16
gfasis = 16
```

or even bigger numbers. The point is that the parameters should be large enough to sufficiently expand information of the genetic data, but can not be too large to decrease the power.

7 References

1. Chiu CY, Yuan F, Zhang BS, Yuan A, Li X, Fang HB, Lange K, Weeks DE, Wilson AF, Bailey-Wilson JE, Lakhal-Chaieb ML, Cook RJ, McMahon FJ, Amos CI, Xiong MM, and Fan

RZ (2019) Pedigree-based linear mixed models for association analysis of quantitative traits with next-generation sequencing data. *Genetic Epidemiology* **43(2)**:189-206.

2. Fan RZ, Wang YF, Mills JL, Wilson AF, Bailey-Wilson JE, and Xiong MM (2013) Functional linear models for association analysis of quantitative traits. *Genetic Epidemiology* **37 (7)**:726-742.

3. Fan RZ, Wang YF, Mills JL, Carter TC, Lobach I, Wilson AF, Bailey-Wilson JE, Weeks DE, and Xiong MM (2014) Generalized functional linear models for case-control association studies. *Genetic Epidemiology* **38 (7)**:622-637.

4. Jiang YD, Chiu CY, Yan Q, Chen W, Gorin MB, Conley YP, Lakhal-Chaieb ML, Cook RJ, Amos CI, Wilson AF, Bailey-Wilson JE, McMahon FJ, Vazquez AI, Yuan A, Zhong XG, Xiong MM, Weeks DE, and Fan RZ (2020) Gene-based association testing of dichotomous traits with generalized linear mixed models for family data.

5. Schaid DJ, McDonnell SK, Sinnwell JP, and Thibodeau SN (2013) Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. *Genetic Epidemiology* **37**:409-418.[2]