Gene Level Meta-analysis of Quantitative Pleiotropic Traits with Multivariate Functional Linear Models

Ruzong Fan, NICHD/NIH, 2016

1 Overview

This document describes a R package to implement the multivariate functional linear models for gene level meta-analysis of quantitative pleiotropic traits (Chiu et al. 2016). Section 2 briefly describes the installation of the program. Section 3 explains how to run the program using one example. Section 4 offers explanation of the results and warnings to use the programs. Section 5 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 reference 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 (fanr@mail.nih.gov).

2 Download and Installation

The package is written in R. Download the following files "MetaMFLM_fixed_model.R", "MetaMFLM_beta_smooth_only.R", "MetaMFLM_FPCA.R", "MetaMFLM_FPCA_no_position.R", "MetaMFLM_additive_effect_model.R", an example files of "MetaSKAT_Example_for_MetaMFLM.R", and "MetaSKAT_Example_cov_functions.R" from MFLM_meta.zip. Put the files in a directory you may access.

3 How to Run the Program

The analysis needs R libraries of fda, MASS, Matrix, and MetaSKAT in R package. The genotype datasets are from the package MetaSKAT. Make sure to install them before running our codes. Open the "MetaSKAT_Example_for_MetaMFLM.R" file on an R Console in a PC window. Please change

the paths leading to directory of "MetaMFLM_fixed_model.R", "MetaMFLM_beta_smooth_only.R", "MetaMFLM_FPCA.R", "MetaMFLM_FPCA_no_position.R", and "MetaSKAT_Example_cov_functions.R" on your computer. Then you may run "MetaSKAT_Example_for_MetaMFLM.R". Please note that the following codes

```
A = matrix(c(1,0.6,-0.35,0.6,1,-0.45,-0.35,-0.45,1),3,3,byrow=T)

pheno = covariate = geno = pos = list(length = L)

covariate[[1]] = Generate_Covariates_1 ( length(y.list[[1]]) )
covariate[[2]] = Generate_Covariates_2 ( length(y.list[[2]]) )
covariate[[3]] = Generate_Covariates_3 ( length(y.list[[3]]) )

X1 = Get_Alpha0_Q1(covariate[[1]])
X2 = Get_Alpha0_Q2(covariate[[2]])
X3 = Get_Alpha0_Q3(covariate[[3]])

pheno[[1]] = X1 + mvrnorm(n = length(y.list[[1]]), mu=rep(0,3), Sigma = A)
pheno[[2]] = X2 + mvrnorm(n = length(y.list[[2]]), mu=rep(0,3), Sigma = A)
pheno[[3]] = X3 + mvrnorm(n = length(y.list[[3]]), mu=rep(0,3), Sigma = A)
```

will generate 3 random samples of covariate and quantitative trait values. Therefore, the results will be different from time to time. On April 13, 2016, I got the following results

```
>MetaMFLM_beta_smooth_only(L, is.homo = TRUE, pheno, mode = "Additive", geno, pos,
       order, bbasis, covariate, base = "bspline", interaction = FALSE)
$Pillai
[1] 0.2190323
```

```
$Wilks
[1] 0.2192848
$Hotelling_Lawley
[1] 0.2195375
$Roy
[1] 0.03609096
$Spherical
[1] 0.339629
$Spherical_GG
[1] 0.3563375
$Spherical_HF
[1] 0.3563116
 ......
> MetaMFLM_add_effect(L, is.homo = FALSE, pheno, mode = "Additive", geno, covariate)
$Pillai
[1] 0.8081166
$Wilks
[1] 0.808422
$Hotelling_Lawley
[1] 0.8087312
$Roy
[1] 0.2927468
$Spherical
[1] 0.986425
$Spherical_GG
[1] 0.9648986
```

```
$Spherical_HF
[1] 0.9649502


> ### The following function is not working since geno[[k]] have different number
      of columns ###
> MetaMFLM_add_effect(L, is.homo = TRUE, pheno, mode = "Additive", geno, covariate)
Error in rbind(U, geno[[k]]) :
  number of columns of matrices must match (see arg 2)
```

To make "MetaMFLM_add_effect(L, is.homo = **TRUE**, y, mode = "Additive", geno, covariate)" to run, one needs that each individual of the L studies is sequenced at the same variants. However, the function "MetaMFLM_add_effect(L, is.homo = **FALSE**, pheno, mode = "Additive", geno, covariate)" can analyze different genotype data among multiple studies, i.e., individuals of different studies may be genotyped at different genetic markers. The details are provided in Chiu et al. (2016).

4 Explanation of the Results and Warnings

As shown in the Section 3, our program can output 3 p-values of approximate F-distribution tests based on Pillai-Bartlett trace, Hotelling-Lawley trace, and Wilks's Lambda. In addition, the program outputs 4 p-values of Roy's maximum root, spherical, spherical_GG, and Spherical_HF F-tests. If you use the R codes to analyze your data, we recommend to report the p-values of approximate F-distribution tests based on Pillai-Bartlett trace, Hotelling-Lawley trace, and Wilks's Lambda. The rest four tests can inflate the type I errors and we do not recommend to use them.

5 Suggestions and Parameters for Real Data Analysis

In practice, one may use one of them for data analysis. Please use one of MetaMFLM_fixed_model.R and MetaMFLM_beta_smooth_only.R by either B-spline or Fourier spline basis functions. We also suggest the following parameters for a data analysis:

```
order  =  4
bbasis = 15
gbasis = 15
fbasis = 25
```

The two functions, "MetaMFLM_FPCA.R" and "MetaMFLM_FPCA_no_position.R", are based on functional principal component analysis. We do find that they provide correct type I error rates and similar power levels as MetaMFLM_beta_smooth_only and MetaMFLM_fixed_model, although the results are not presented in Chiu et al. (2016).

6 References

Chiu CY, Jung JS, Chen W, Weeks DE, Ren H, Boehnke M, Amos CI, Liu AY, Mills JL, Lee MLT, Xiong MM and Fan RZ (2016) Meta-analysis of quantitative pleiotropic traits at gene level with multivariate functional linear models.

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**:726-742.

Fan RZ, Wang YF, Boehnke M, Chen W, Li Y, Ren HB, Lobach I, and Xiong MM (2015) Gene level meta-analysis of quantitative traits by functional linear models. *Genetics* **200**:1089-1104.

Wang YF, Liu AY, Mills JL, Boehnke M, Wilson AF, Bailey-Wilson JE, Xiong MM, Wu CO, and Fan RZ (2015) Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. *Genetic Epidemiology* **39 (4)**:259-275.