Pleiotropy Analysis of Quantitative Traits at Gene Level by Multivariate Functional Linear Models

Ruzong Fan and Yifan Wang, NICHD/NIH, December 2014

1 Overview

This document describes a R package to implement the multivariate functional linear models for pleiotropy analysis of quantitative traits at gene level. 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 paper 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 R codes "MFLM_fixed_model.R", "MFLM_beta_smooth_only.R", "MFLM_FPCA.R", and "MFLM_FPCA_no_position.R", and example files of "Example_MFLM.R" and "Example_MFLM_multiple_gene_analysis.R" from MFLM_web.zip. Plus, you will need datasets from data.zip to run the examples. Put the files in a directory you may access.

3 How to Run the Program

3.1 One Gene Analysis

The analysis needs libraries fda, MASS, and Matrix in R package. Make sure to install them before running our codes. Open the "Example_MFLM.R" file on an R Console in a PC window. Change the paths leading to the directories of the package "MFLM_fixed_model.R", "MFLM_beta_smooth_only.R",

"MFLM_FPCA.R", "MFLM_FPCA_no_position.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 i386 3.1.2".

```
> mflm_fixed_model(pheno, mode = "Additive", geno, pos, order, bbasis,
         fbasis, gbasis, covariate, base = "bspline", interaction = FALSE)
$Pillai
[1] 0.6251404
$Wilks
[1] 0.6266297
$Hotelling_Lawley
[1] 0.628154
......
> mflm_fixed_model(pheno, mode = "Additive", geno, pos, order, bbasis,
         fbasis, gfasis, covariate, base = "fspline", interaction = FALSE)
$Pillai
[1] 0.4798107
$Wilks
[1] 0.4771207
$Hotelling_Lawley
[1] 0.4745115
......
> mflm_beta_smooth_only(pheno, mode = "Additive", geno, pos, order, bbasis,
                        covariate, base = "bspline", interaction = FALSE)
$Pillai
[1] 0.6511031
$Wilks
[1] 0.6486651
```

```
$Hotelling_Lawley
[1] 0.6462962
......
> mflm_beta_smooth_only(pheno, mode = "Additive", geno, pos, order, fbasis,
                        covariate, base = "fspline", interaction = FALSE)
$Pillai
[1] 0.4798107
$Wilks
[1] 0.4771207
$Hotelling_Lawley
[1] 0.4745115
.......
mflm_fpca_no_position(pheno, mode = "Additive", geno, covariates = covariate,
                     kz = 20, kb = 10, smooth.cov=FALSE)
$Pillai
[1] 0.6511031
$Wilks
[1] 0.6486651
$Hotelling_Lawley
[1] 0.6462962
.......
> mflm_fpca(pheno, mode = "Additive", geno, covariates = covariate, pos,
           kz = 20, kb = 10, smooth.cov=FALSE)
$Pillai
[1] 0.4789084
$Wilks
```

```
[1] 0.4762317
$Hotelling_Lawley
[1] 0.4736356
......
```

3.2 Multiple Gene Analysis

The analysis needs libraries fda, MASS, and Matrix in R package. Make sure to install them before running our codes. Open the "Example_MFLM_multiple_gene_analysis.R" file on an R Console in a PC window. Change the paths leading to the directories of the package "MFLM_fixed_model.R", "MFLM_beta_smooth_only.R", "MFLM_FPCA.R", "MFLM_FPCA_no_position.R", and the datasets on your computer.

Then, you may get one csv file named "y_mode=Additive_order=4_bbasis=15_fbasis=25.csv" after running "Example_MFLM_multiple_gene_analysis.R" file. Note that only two genes are analyzed, but you may add more for multiple gene analysis.

4 Explanation of the Results and Warnings

As shown in the Section 3, our program outputs p-values of three types of approximate F-distribution tests based on Pillai-Bartlett trace, Hotelling-Lawley trace, and Wilks's Lambda. In addition, it also outputs p-values of three types of approximate F-distribution tests based on Roy's maximum root, and spherical F-test as well as its corrected versions.

The approximate F-distribution tests based on Pillai-Bartlett trace, Hotelling-Lawley trace, and Wilks's Lambda have conservative and accurate type I error rates (Wang et al. 2014). 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 approximate F-distribution tests based on Roy's maximum root, and spherical F-test as well as its corrected versions can inflate type I error rates. We do not recommend to use them in real

data analysis.

5 Suggestions and Parameters for Real Data Analysis

In this documentation, we present four R functions to perform gene-based pleiotropy analysis of multiple quantitative traits. In practice, one may use one of them for data analysis. We suggest to use mflm_fixed_model by either B-spline or Fourier spline basis functions and report the p-values of approximate F-distribution tests based on Pillai-Bartlett trace. We also suggest the following parameters for a data analysis:

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

6 References

1. 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.

2. Wang YF, Liu AY, Mills JL, Wilson AF, Bailey-Wilson JE, Xiong MM, Wu CO, and Fan RZ (2014) Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. *Genetic Epidemiology*, in reviews.