Generalized Functional Linear Models for Gene-based Case-Control Association Studies

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

1 Overview

This document describes a R package to implement generalized functional linear models for testing association between a dichotomous trait and multiple genetic variants in a genetic region while adjusting for covariates (Fan et al. 2014). 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 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 (fanr@mail.nih.gov).

2 Download and Installation

Download R codes of "GFLM_fixed_model.R" and "GFLM_beta_smooth_only.R", and example files of "Example_GFLM.R" and "Example_GFLM_multiple_gene_analysis.R" from GFLM_web.zip. Also, download "FLM_FPCA.R" and "FLM_FPCA_no_position.R" from FLM_web.zip, which is a R package to implement the models for functional linear models for association analysis of quantitative traits (Fan et al. 2013). 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 package needs libraries fda, MASS, Matrix, and globaltest in R package. Make sure to install them before running our codes. Open the "Example_GFLM.R" file on an R Console in a PC window. Change the paths leading to the directory of the package “GFLM_beta_smooth_only.R”, “GFLM_fixed_model.R”, “FLM_FPCA.R”, and “FLM_FPCA_no_position.R” 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”.

```
> gflm_fixed_model(pheno, mode = "Additive", geno, pos, order, bbasis, fbasis,
                   gbasis, covariate, base = "bspline", interaction = FALSE)
$LRT
[1] 0.4160104
$Chisq
[1] 0.4160104
$Rao
[1] 0.6044884
$gt
[1] 0.2761322
```

```
> gflm_fixed_model(pheno, mode = "Additive", geno, pos, order, bbasis, fbasis,
                   gfasis, covariate, base = "fspline", interaction = FALSE)
$LRT
[1] 0.3945785
$Chisq
[1] 0.3945785
$Rao
[1] 0.548139
$gt
[1] 0.9690488
```

```
> gflm_beta_smooth_only(pheno, mode = "Additive", geno, pos, order, bbasis,
                        covariate, base = "bspline", interaction = FALSE)
$LRT
[1] 0.4160104
$Chisq
[1] 0.4160104
$Rao
[1] 0.6044884
$gt
[1] 0.8995719

> gflm_beta_smooth_only(pheno, mode = "Additive", geno, pos, order, fbasis,
                        covariate, base = "fspline", interaction = FALSE)
$LRT
[1] 0.3945785
$Chisq
[1] 0.3945785
$Rao
[1] 0.548139
$gt
[1] 0.8450543

flm_fpca_no_position(pheno, mode = "Additive", geno, covariates = covariate,
                     kz = 20, kb = 10, smooth.cov=FALSE, family = "binomial")
$LRT
[1] 0.4160104
$Chisq
```

```
[1] 0.4160104

$Rao

[1] 0.6044884

$gt

[1] 0.7533608


> fpca = flm_fpca(pheno, mode = "Additive", geno, covariates = covariate, pos,

                  kz = 20, kb = 10, smooth.cov=FALSE, family = "binomial")

$LRT

[1] 0.2999991

$Chisq

[1] 0.2999991

$Rao

[1] 0.4842508

$gt

[1] 0.5426588
```

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_GFLM_multiple_gene_analysis.R" file on an R Console in a PC window. Change the paths leading to the directories of the package "GFLM_fixed_model.R", "GFLM_beta_smooth_only.R", "FLM_FPCA.R", "FLM_FPCA_no_position.R", and the datasets on your computer.

Then, you may get one csv file named "y_mode=Additive_order=4_bbasis=10_fbasis=11.csv" after running "Example_GFLM_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 can output 4 p-values based on likelihood ratio test (LRT), Chisq, Rao's efficient score test (Rao), and global test (gt). The LRT is the same as χ^2, which inflates type I error rates (Fan et al. 2014). Rao and gt have conservative and accurate type I error rates (Fan et al. 2014). If you use the R codes to analyze your data, we recommend to report the p-values of Rao and gt.

5 Suggestions and Parameters for Real Data Analysis

In this documentation, we present four R functions to perform gene-based association analysis of quantitative traits. In practice, one may use one of them for data analysis. We suggest to use gflm_fixed_model by either B-spline or Fourier spline basis functions. We also suggest the following parameters for a data analysis:

```
order  =  4

bbasis = 10

gbasis = 11

fbasis = 10

gfasis = 11
```

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. 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*, 38:622-637.