Vignette: NRVATGFLMM

Dossa Houssou Roland G., Lakhal-Chaieb Lajmi, Oualkacha Karim

Introduction

NRVATGFLMM is an R package for performing a family-based association test for rare variants using marginal logistic (NRVAT) and generalized functional linear mixed models (GFLMM) for gene-based association testing of dichotomous traits implying a Gaussian Copula. The latter is employed to model the dependence between relatives.

Models

Consider I families and for $i = 1, \ldots, I$, let n_i be the size of the i^{th} family. The total sample size is $N = \sum_{i=1}^{I} n_i$. For $i = 1, \ldots, I$ and $j = 1, ..., n_i$ let $Y_{ij} \in \{0, 1\}$ be the binary phenotype under investigation for individual j in family i. We begin by specifying the (conditional) marginal distribution of Y_{ij}, denoted $F(y_{ij}|\mathbf{X}_{ij}, \mathbf{G}_{ij})$, where $\mathbf{X}_{ij} = (1, X_{ij1}, \ldots, X_{ijs})^\top$ is a $1 \times (s + 1)$ vector and
$\mathbf{G}_{ij} = (G_{ij1}, \ldots, G_{ijr})^\top$ is a set of genotypes coded as $(0, 1, 2)$ from biallelic variants. Since the response variable is dichotomous, $F(y_{ij}|\mathbf{X}_{ij}, \mathbf{G}_{ij})$ is completely specified by $\mu_{ij} = \mathbf{P}(Y_{ij} = 1|\mathbf{X}_{ij}, \mathbf{G}_{ij})$.

Thus, we relate the binary phenotype Y_{ij} to $\mathbf{X}_{ij}$ and $\mathbf{G}_{ij}$ through a logistic regression model

$$\text{logit}(\mu_{ij}) = \mathbf{X}_{ij}^\top \boldsymbol{\gamma} + \mathbf{G}_{ij}^\top \boldsymbol{\beta}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, I,$$

where $\text{logit}(u) = \log[\frac{u}{1-u}]$, and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_s)^\top$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)^\top$ are sets of regression coefficients. In a matrix notation, one has

$$\text{logit}(\boldsymbol{\mu}_i) = \mathbf{X}_i \boldsymbol{\gamma} + \mathbf{G}_i \boldsymbol{\beta},$$

where $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_{i1}, \ldots, \boldsymbol{\mu}_{in_i})^\top$, $\mathbf{X}_i$ is a $n_i \times (s + 1)$ matrix with the j^{th} row equal to $\mathbf{X}_{ij}$ and $\mathbf{G}_i$ is a $n_i \times r$ matrix with the j^{th} row equal to $\mathbf{G}_{ij}$. The logit was taken element-wise of the entries of $\boldsymbol{\mu}_i$.

In the context of fonctionnal data analysis, we consider I families and for $i = 1, \ldots, I$, n_i be the size of the i^{th} family. The total sample size is $N = \sum_{i=1}^{I} n_i$. For $i = 1, \ldots, I$ and $j = 1, \ldots, n_i$ let $Y_{ij} \in \{0, 1\}$ be a dichotomous trait of interest coded as 1 and 0 denoting respectively, affected and unaffected for individual j in family i. All individuals are sequenced in a genomic region with r genetic variant. Let assume that the physical location of the r genetic variant are ordered, known, and denoted $0 \leq v_1 < \cdots < v_r$, and normalized on the unit region $[0, 1]$. We specify the (conditional) marginal distribution of Y_{ij}, denoted $F(y_{ij}|\mathbf{X}_{ij}, \mathbf{G}_{ij})$, where $\mathbf{X}_{ij} = (1, X_{ij1}, \ldots, X_{ijs})^\top$ is a $1 \times (s + 1)$ covariates vector and $\mathbf{G}_{ij} = (G_{ij}(v_1), \ldots, G_{ij}(v_r))^\top$ is a set of genotypes. We assume that $G_{ij}(v_r) \in \{0, 1, 2\}$ which denotes the number of minor allele of j^{th} individual in family i at the r^{th} variant.

For the individual j in family i, we consider $\mathbf{T}_{ij}(v)$, $v \in [0, 1]$ as his/her genetic variant function (GVF). Since the response variable is dichotomous, $F(y_{ij}|\mathbf{X}_{ij}, \mathbf{T}_{ij}(v))$ is completely specified by $\mu_{ij} = \mathbf{P}(Y_{ij} = 1|\mathbf{X}_{ij}, \mathbf{T}_{ij}(v))$.

Thus, we relate the binary phenotype Y_{ij} to the covariates and the genetic variant function through a logistic regression model

$$\text{logit}(\mu_{ij}) = \mathbf{X}_{ij}^\top \boldsymbol{\gamma} + \int_0^1 \mathbf{T}_{ij}(v)^\top \boldsymbol{\beta}(v) dv, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, I,$$

where $\text{logit}(\pi) = e^{\pi}/(1+e^{\pi})$; $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_s)^\top$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)^\top$ are sets of regression coefficients of the genetic variant function at the location u. In a matrix notation, one has

$$\text{logit}(\boldsymbol{\mu}_i) = \mathbf{X}_i\boldsymbol{\gamma} + \int_0^1 \mathbf{T}_i(v)\boldsymbol{\beta}(v)dv,$$

where $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_{i1}, \ldots, \boldsymbol{\mu}_{in_i})^\top$, $\mathbf{X}_i$ is a $n_i \times (s+1)$ matrix with the j^{th} row equal to $\mathbf{X}_{ij}$, and $\mathbf{T}_i(v) = (T_{i1}(v), \ldots, T_{in_i}(v))^\top$ is the genetic variant function. The integral is applied "elementwise" on the inputs of the vector $\mathbf{T}_i(v)\boldsymbol{\beta}(v)$.

We suggested two approaches, as proposed by Jiang et al., (2020) and Zhang et al., (2021) such as (1) smoothing only the genetic effect function $\boldsymbol{\beta}(v)$, which is called beta-smooth only and (2) smoothing both the genetic effect function $\boldsymbol{\beta}(v)$, and the genetic variant function $T_{ij}(v)$.

Beta-Smooth only approach

In this approach, the genetic effect function $\beta(v)$ is assumed to be continuous or smooth with no assumption about the genetic variant function. The integration term $\int_0^1 T_{ij}(v)\beta(v)dv$ is replaced by a summation term $\sum_{l=1}^{r} \mathbf{G}_{ij}(v_l)\beta(v_l)$. The original genotype data $T_{ij} = (G_{ij}(v_1), \ldots, G_{ij}(v_r))^\top$ is directly used here. By expanding the genetic effect function $\beta(v)$, we have that $\beta(v) = (\theta_1(v), \ldots, \theta_{K_\beta}(v))^\top(\beta_1, \ldots, \beta_{K_\beta})$, where *{i}) $\theta_k(v)$, $k = 1, \ldots, K_\beta$, is a series of K_β basis functions of B-Spline and {ii}) $\theta_1(v) = 1$, $\theta_{2p+1}(v) = sin(2\pi pu)$, and $\theta_{2p}(v) = cos(2\pi pu)$, $p = 1, \ldots, (K_\beta - 1)/2$ for Fourier Basis with K_β is taken as a positive odd integer, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{K_\beta})^\top$ is a $K_\beta \times 1$ vector of unknown coefficients. We have then:*

$$logit(\mu_{ij}) = \mathbf{X}_{ij}^\top\boldsymbol{\gamma} + \left[\sum_{l=1}^{r} \mathbf{G}_{ij}(v_l)(\theta_1(v_l), \ldots, \theta_{K_\beta}(v_l))\right](\beta_1, \ldots, \beta_{K_\beta})^\top.$$

Smoothness of the genetic effect function and the genetic variant function

Here, we suppose that both the genetic effect function $\beta(v)$ and the GVF, $T_{ij}(v)$ are expanded by a series of basis functions of either B-Spline or Fourier. In this approach, the estimation of the GVF for each subject, $T_{ij}(v)$, is required. To obtain the latter, we rely on an ordinary least squares smoother (see (Jiang et al., (2020) for more details). Indeed, following these works, the ordinary least squares smoother method assumes that the subject genetic variant function is smooth. This means that $T_{ij}(v) = a(v) + \epsilon$, where $a(v)$ is an unknown smooth function and ϵ is an error term; i.e. for the positions of the observed genotypes, $v_l, l = 1, \ldots, r$, one has $T_{ij}(v_l) = a(v_l) + \epsilon_l$. The function $a(.)$ can then be approximated using smoothing techniques and the r observed genotypes of each subject within ordinary least squares regression model. More precisely, for the positions of the observed genotypes, $v_l, l = 1, \ldots, r$, the approximation of $a(.)$ can formally be expressed as

$$a(v_l) = \sum_{k}^{K_a} c_k\phi_k(v_l) = \mathbf{c}^\top\phi(v_l),$$

where the vector $\mathbf{c}$ of length K_a contains the coefficients c_k's and $\phi(v) = (\phi_1(v), \ldots, \phi_{K_a}(v))^\top$ is a column vector of the basis functions. Thus, a linear smoother of $a(.)$ is obtained by determining the coefficients of the expansion c_k when minimizing the following least squares criterion $||\mathbf{G}_{ij} - \mathbf{c}\Upsilon||^2$, which yields to $\hat{\mathbf{c}} = [\Upsilon^\top\Upsilon]^{-1}\Upsilon^\top\mathbf{G}_{ij}$, where Υ represents the $r \times K_a$ matrix carrying the values $\Upsilon_{l\,k} = \phi_k(v_l),\ l \in 1, \ldots, r$.

Finally, $T_{ij}(v)$ can be estimated as follows

$$\hat{T}_{ij}(v) = (G_{ij}(v_1), \ldots, G_{ij}(v_r))\ \Upsilon[\Upsilon^\top\Upsilon]^{-1}\phi(v).$$

Hence,

$$logit(\mu_{ij}) = \mathbf{X}_{ij}^{\top}\boldsymbol{\gamma} + \left[(G_{ij}(v_1), ..., G_{ij}((v_r))\ \Upsilon[\Upsilon^{\top}\Upsilon]^{-1}\int_0^1 \phi(v)\ \boldsymbol{\theta}^{\top}(v)\, dv\right](\beta_1, \ldots, \beta_{K_\beta})^{\top}.$$

There is available codes in statistical packages in R and Matlab for calculating the terms $\Upsilon[\Upsilon^{\top}\Upsilon]^{-1}$ *and* $\int_0^1 \phi(v)\ \boldsymbol{\theta}^{\top}(v)\, dv$. *For more details, see (Jiang et al., (2020)).*

Installation

The ***NRVATGFLMM*** *package is publicly available at github.com/houssoudossa/NRVATGFLMM.*

The package is written in R language, and can be installed as follows

1- Install the devtools package by starting up R and running this command

```
install.packages("devtools")
```

2- Load the devtools library to make its commands available

```
library(devtools)
```

3- Install the ***NRVATGFLMM*** *R package from the github repository via this command*

```
install_github("houssoudossa/NRVATGFLMM")
```

The main three functions of ***NRVATGFLMM*** *package requires several input files, namely, Pedigree, Genotypes, Map, Covariates and Kinship matrix (kin2) files. These input files are described next.*

Description of the input files

Pedigree file

The pedigree file is in the same format as that used by the PedGFLMM R package and has the following columns:

- *ID: identity of each individual.*
- *ped: pedigree ID, character or numeric are allowed.*
- *person: person ID, a unique ID within each pedigree, numeric or character allowed.*
- *father: father ID,* 0 *if no father.*
- *mother: mother ID,* 0 *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.*

The first 6 rows of the example pedigree file are the following:

```
#>   ID ped person father mother sex trait
#> 1  1   1      1      0      0   1     0
#> 2  2   1      2      0      0   2     0
#> 3  3   1      3      1      2   2     0
#> 4  4   2      1      0      0   1     1
#> 5  5   2      2      0      0   2     0
#> 6  6   2      3      1      2   2     0
```

Genotype file

The genotype file is a matrix with genotypes for subjects (rows) for several variant positions (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 file. The genotypes are coded as $0, 1, 2$ *for autosomal markers (typically a count of the number of the minor allele). The following output example shows the first* 8 *SNPs genotypes (columns) of the* 100 *genotypes considered in the genotypes file.*

```
#>   ped person V3 V4 V5 V6 V7 V8 V9 V10
#> 1   1      1  0  0  0  0  0  0  0   0
#> 2   1      2  0  0  0  1  0  0  0   0
#> 3   1      3  0  0  0  1  0  0  0   0
#> 4   2      1  0  0  0  0  0  0  0   0
#> 5   2      2  0  0  0  0  0  0  0   0
#> 6   2      3  0  0  0  0  0  0  0   0
```

Map file

The 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.

Below, we have the first 6 rows of the example map file:

```
#>   chr       snp pos
#> 1   1 rs6681049   1
#> 2   1 rs4074137   2
#> 3   1 rs7540009   3
#> 4   1 rs1891905   4
#> 5   1 rs9729550   5
#> 6   1 rs3813196   6
```

Covariate file

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

The first 6 rows of the covariates file example are stated below:

```
#>   ped person Intercept        X1 X2
#> 1   1      1         1 0.1266321  0
#> 2   1      2         1 0.6589364  1
#> 3   1      3         1 0.4829154  0
#> 4   2      1         1 0.3401634  1
#> 5   2      2         1 0.8287422  0
#> 6   2      3         1 0.4277390  1
```

Kinship matrix

The kinship matrix (kin2) is a positive semi-definite relationship matrix (e.g. kinship matrix in genetic association studies). Its entries give the probability of sharing identically by descent %(IBD) at the level of the whole genome between subjects. For instance, $\mathbf{\Phi}_{jj} = 0.5$ *for all subjects* j*,* $\mathbf{\Phi}_{jk} = 0.25$ *if* j *and* k *are siblings or if one of them is a parent of the other,* $\mathbf{\Phi}_{jk} = 0.125$ *if one of them is a grand-parent of the other and* $\mathbf{\Phi}_{jk} = 0.0625$ *if* j *and* k *are cousins. The rownames and colnames of this block matrices must at least include all samples as specified in the ped column of the pedigree file. Below, we have the first 6 rows of the kinship matrix example (first* 6 *columns) which show two independent families of two parents and one child.*

```
#>      1/1  1/2  1/3  2/1  2/2  2/3
#> 1/1 0.50 0.00 0.25 0.00 0.00 0.00
#> 1/2 0.00 0.50 0.25 0.00 0.00 0.00
#> 1/3 0.25 0.25 0.50 0.00 0.00 0.00
#> 2/1 0.00 0.00 0.00 0.50 0.00 0.25
#> 2/2 0.00 0.00 0.00 0.00 0.50 0.25
#> 2/3 0.00 0.00 0.00 0.25 0.25 0.50
```

Running NRVATGFLMM

NRVATGFLMM *consists of three main functions, which implement the association tests described in Dossa et al (2022) (chapter 3) and Dossa et al (2023) (chapter 4) such as:*

1. *NRVAT: association test implementing our marginal logistic model.*
2. *CBGF_Beta_Only: association test implementing generalized functional linear mixed model (CBGF), by assuming that the genetic effect function is continuous/smooth.*
3. *CBGF_Fixed: association test implementing generalized functional linear mixed model (CBGF), by assuming that the genetic variant function (GVF) and the genetic effect function are continuous/smooth.*

After package installation, one can access help document for each of these three functions, including a running example. For instance, to access the help documentation for the NRVAT function, one can proceed as follows

```
library(NRVATGFLMM)
?NRVAT
```

Getting the example Data

Before starting, we first load the example data.

```
library(NRVATGFLMM)
data(Ped)
data(geno)
data(cov)
data(snpPos)
data(kin2)
```

NRVAT

The ***NRVAT*** *function implements a region-based association test using our NRVAT approach. Figure below shows the results of NRVAT model when using the input presented above with all the kernel matrices including the estimation of covariates coefficients and the heritability parameter.*

```
h.s=0.2
Result_NRVAT=NRVAT(heritability=h.s, a1=1, b1=25, Ped=Ped, covariates=cov, geno=geno, kin2=kin2)

Result_NRVAT
  Intercept    gam1      gam2       h.S_estim  NRVAT_S_obsL
1 -2.256907   1.379476  1.194767   0.21       2606.676
  NRVAT_p.valueL   NRVAT_S_obsQ  NRVAT_p.valueQ
1  0.5681522       166698.3      0.5570728
  NRVAT_S_obsIB   NRVAT_p.valueIB  NRVAT_S_obsG
1  0.2516299       0.6139498        77.21667
  NRVAT_p.valueG   NRVAT_S_obsP  NRVAT_p.valueP
1  0.5264557       1615371092    0.5557433
```

In the two functions CBGF_Beta_Only and CBGF_Fixed, we fixed, by default, the number of B-spline basis

functions as $K = K_a = K_\beta = 10$, and the number of Fourier basis functions to be $K = K_a = K_\beta = 11$. The order of the basis functions was fixed as 4.

CBGF_Beta_Only (which we named CBGF-1 in the main article)

*The **CBGF_Beta_Only** function implements a region-based association test using our "beta smooth only" generalized functional linear mixed model (CBGF), by assuming that the genetic effect function is continuous/smooth. This can be done using either B-spline or Fourier basis functions, and the order of the basis functions need to be specified by the user. Figure below shows the results of CBGF-1 model when using the input presented above under B-spline basis functions (Bs) and Fourier basis functions (Fs) with all the kernel matrices, including the estimation of covariates coefficients and the heritability parameter.*

```
h.s=0.2
beta_basis_BO_BS=10
beta_basis_BO_FS=11
order = 4

Result_CBGF1=CBGF_Beta_Only(heritability=h.s, a1=1, b1=25, Ped=Ped, covariates=cov,
                            geno=geno, kin2=kin2, pos = snpPos, order=order,
                            beta_basis_BO_BS=beta_basis_BO_BS, base_BO_BS = "bspline",
                            beta_basis_BO_FS=beta_basis_BO_FS, base_BO_FS = "fspline")

Result_CBGF1
  Intercept   gam1      gam2       h.S_estim   StatL.Beta_Bs   p.v_obsbL.Beta_Bs StatQ.Beta_Bs
1 -2.256907   1.379476  1.194767   0.21        159.252         0.2814301         694.3374
  p.v_obsbQ.Beta_Bs  StatG.Beta_Bs  p.v_obsbG.Beta_Bs  StatP.Beta_Bs   p.v_obsbP.Beta_Bs
1  0.7866799         39.26447        0.04244662        40768.38        0.8968183
  StatL.Beta_Fs    p.v_obsbL.Beta_Fs StatQ.Beta_Fs  p.v_obsbQ.Beta_Fs  StatG.Beta_Fs
1  3073.721         0.3518492         202584.2      0.8588227          101.2937
  p.v_obsbG.Beta_Fs  StatP.Beta_Fs    p.v_obsbP.Beta_Fs
1  0.1463586         23836470         0.8643131
```

CBGF_Fixed (which we named CBGF-2 in the main article)

*The **CBGF_Fixed** function carries out a region-based association test using an expansion of the genetic variant function (GVF) and the genetic effect function in our generalized functional linear mixed model (CBGF). This expansion can be done using either B-spline or Fourier basis functions, and the order of the basis functions need to be specified by the user. Figure below shows the results of CBGF-2 model when using the input presented above under B-spline basis functions (Bs) and Fourier basis functions (Fs) with all the kernel matrices including the estimation of covariates coefficients and the heritability parameter.*

```
h.s=0.2
beta_basis_F_BS = geno_basis_F_BS = 10
beta_basis_F_FS = geno_basis_F_FS = 11
order = 4

Result_CBGF2=CBGF_Fixed(heritability=0.2, a1=1, b1=25, Ped=Ped, covariates=cov,
                        geno=geno, kin2=kin2, pos = snpPos, order=order,
                        beta_basis_F_BS = beta_basis_F_BS, geno_basis_F_BS = geno_basis_F_BS,
                        base_F_BS = "bspline", beta_basis_F_FS = beta_basis_F_FS,
                        geno_basis_F_FS = geno_basis_F_FS, base_F_FS ="fspline")

Result_CBGF2
  Intercept  gam1      gam2       h.S_estim  StatL.Fixed_Bs   p.v_obsbL.Fixed_Bs  StatQ.Fixed_Bs
1 -2.256907  1.379476  1.194767   0.21       0.0227694        0.2849979           0.0455453
```

```
  p.v_obsbQ.Fixed_Bs  StatG.Fixed_Bs  p.v_obsbG.Fixed_Bs StatP.Fixed_Bs  p.v_obsbP.Fixed_Bs
1  0.2852501          0.004554381     0.2849006          0.4560379       0.2875183
  StatL.Fixed_Fs  p.v_obsbL.Fixed_Fs  StatQ.Fixed_Fs   p.v_obsbQ.Fixed_Fs  StatG.Fixed_Fs
1  0.5144763       0.2394277          1.033001         0.2426641           0.093562
  p.v_obsbG.Fixed_Fs  StatP.Fixed_Fs    p.v_obsbP.Fixed_Fs
1  0.2384471          11.80829          0.2746678
```

References

Breslow, N. E. and Clayton, D. G. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association ***88,*** *9–25 (1993).*

Chen, Han and Huffman, Jennifer E and Brody, Jennifer A and Wang, Chaolong and Lee, Seunggeun and Li, Zilin and Gogarten, Stephanie M and Sofer, Tamar and Bielak, Lawrence F and Bis, Joshua C and others. Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies The American Journal of Human Genetics ***104,*** *260–274 (2019).*

Jiang, Yingda and Chiu, Chi-Yang and Yan, Qi and Chen, Wei and Gorin, Michael B and Conley, Yvette P and Lakhal-Chaieb, M'Hamed Lajmi and Cook, Richard J and Amos, Christopher I and Wilson, Alexander F and others. Gene-Based Association Testing of Dichotomous Traits With Generalized Functional Linear Mixed Models Using Extended Pedigrees: Applications to Age-Related Macular Degeneration. Journal of the American Statistical Association 1–15 (2020).

Lakhal-Chaieb, Lajmi and Oualkacha, Karim and Richards, Brent J and Greenwood, Celia MT. A rare variant association test in family-based designs and non-normal quantitative traits. Statistics in medicine ***35,*** *905–921 (2016).*

Zhang, Bingsong and Chiu, Chi-Yang and Yuan, Fang and Sang, Tian and Cook, Richard J and Wilson, Alexander F and Bailey-Wilson, Joan E and Chew, Emily Y and Xiong, Momiao and Fan, Ruzong. Gene-based analysis of bi-variate survival traits via functional regressions with applications to eye diseases. Genetic Epidemiology (2021).