# CBMAT: Copula Based Multivariate Association Test for bivariate mixed phenotypes

Julien St-Pierre & Karim Oualkacha

March 9, 2021

## Contents

## 1 Introduction

CBMAT is an R package that contains methods to perform region-based genetic association of a bivariate phenotype. The dependence between phenotypes is modelled via copulas. Thus, CBMAT is a robust method for non-normality assumption of the traits. It also allows for testing association with a mixed discrete/continuous bivariate phenotype.

The main user-visible function of the package is `CBMAT()` function which can be used to analyse genetic region/bivariate phenotype association in one go. CBMAT allows also for phenotype dependence modelling using four copulas

- Gaussian copula;
- Clayton copula;
- Gumbel copula;
- Frank copula.

CBMAT can be downloaded at https://github.com/julstpierre/CBMAT and can be installed using `devtools` package:

```
# development version from GitHub
if (!requireNamespace("devtools")) install.packages("devtools")
devtools::install_github("julstpierre/CBMAT")
```

## 2 Input Data

Before running an association test with CBMAT, the following data is needed:

- *Phenotypes*: phenotypes data should be present separately in the form of an R vector (one value for each individual). For example, here we show the first 6 entries of the phenotypes in the simulated data set `data_mixed`:

```
  #Load and attach data
  data(data_mixed,package = "CBMAT")
  attach(data_mixed)

  #One binary phenotype
  head(y.bin)

## [1] 0 0 1 0 0 0

  #Two continuous phenotypes
  head(y.gauss)

## [1]  0.1688371 -0.5772787  6.5811132  3.1596381  3.5511046  2.5926228

  head(y.Gamma)

## [1] 2.54060075 1.12152345 4.98853560 0.07337192 2.34025750 0.33434987
```

- *Covariates*: covariates data should be present in the form of an $n \times k$ matrix, where $n$ represents the number of subjects and $k$ the number of covariates, including the intercept. For example, here we show the first 6 entries of the covariate matrix included in `data_mixed`:

```
  dim(x)

## [1] 503   3

  head(x)
```

```
##      [,1] [,2]        [,3]
## [1,]    1    0 -0.35173586
## [2,]    1    0 -1.10067195
## [3,]    1    1  1.33908052
## [4,]    1    1 -1.03802577
## [5,]    1    1 -0.50020571
## [6,]    1    1 -0.01141293
```

- *Genotype data*: SNPs genotype must be in the form of an $n \times p$ matrix, where $n$ represents the number of subjects and $p$ the number of SNPs in the region of interest. For example, here we show the the first 6 entries of the first 10 SNPs of the genotype matrix included in `data_mixed`:

```
  dim(G)

## [1] 503  30

  head(G[,1:10])

##      V1 V2 V3 V4 V5 V6 V7 V8 V9 V10
## [1,]  0  0  0  0  0  0  0  0  0   0
## [2,]  0  0  0  0  0  0  0  0  0   0
## [3,]  0  0  0  0  0  0  0  0  0   0
## [4,]  0  0  0  0  0  0  0  0  0   0
## [5,]  0  0  0  0  0  0  0  0  0   0
## [6,]  0  0  0  0  0  1  0  0  0   0
```

# 3 Running CBMAT

If CBMAT has been successfully installed, you can load it in an R session using

```
  library(CBMAT)
```

Here we provide two simple examples of performing score tests while fitting generalized linear models (glms) under the null hypothesis of no association.

## 3.1 Using CBMAT for a bivariate continous phenotype

When both traits are continuous, one can run CBMAT for a region-based association test with the following code:

```
cont.score <- CBMAT(y1=y.gauss,
                    fam1="gaussian()",
                    y2=y.Gamma,
                    fam2="Gamma(link=log)",
                    x=x,
                    G=G,
                    copfit=c("Gaussian","Clayton","Frank","Gumbel"),
                    weight=FALSE,
                    weight.para1=1,
                    weight.para2=25,
                    pval.method="min")

## Starting association analysis...

cont.score

## $p.value
## [1] 0.5554563
##
## $alpha
## [1] 0.2672714
##
## $tau
## [1] 0.172244
##
## $gamma.y1
## [1] 1.773083 1.600274 1.815735
##
## $gamma.y2
## [1] 0.2501279 1.0248267 0.7646350
##
## $cop
## [1] "Gaussian"
```

- `fam1` and `fam2` are characters specifying the error distributions and link functions to be used in each marginal model.
- `copfit` is a character vector that specifies the copula model(s) to use for modelling phenotypes dependence. The default option is to select between the Gaussian, Clayton, Frank and Gumbel copulas, based on AIC of the different models.
- `weight` is logical variable indication if weights should be used to increase power for rare variants.
- `weight.para1` and `weight.para2` are parameters of beta distribution used to simulate weights.

- `pval.method` is a character that specifies which method should be used to calculate p-value of score test. See [1] for further details. Can be one of the following:
  - `pval.method = "min"`, optimal p-value (default)
  - `pval.method = "Fischer"`, Fisher's method
  - `pval.method = "MFKM"`, MFKM method

### 3.2 Using CBMAT for a mixed discrete-continuous bivariate phenotype

If one trait is discrete, it must be entered as the first phenotype, using a probit link function:

```
mixed.score <-CBMAT(y1=y.bin,
                    fam1="binomial(link=probit)",
                    y2=y.Gamma,
                    fam2="Gamma(link=log)",
                    x=x,
                    G=G,
                    copfit=c("Gaussian","Clayton","Frank","Gumbel"),
                    weight=FALSE,
                    weight.para1=1,
                    weight.para2=25,
                    pval.method="min")
```

```
## Starting association analysis...
```

```
mixed.score
```

```
## $p.value
## [1] 0.4540166
##
## $alpha
## [1] 0.3429957
##
## $tau
## [1] 0.1463919
##
## $gamma.y1
## [1] -1.991215  1.481467  1.881062
##
## $gamma.y2
## [1] 0.2565104 1.0136100 0.7674400
##
## $cop
## [1] "Clayton"
```

# References

[1] Jianping Sun, Karim Oualkacha, Celia M.T. Greenwood, Lajmi Lakhal-CHaieb. *Multivariate Association Test for Rare Variants Controlling for Cryptic and Family Relatedness*. Canadian Journal of Statistics, vol. 47, no. 1, Mar. 2019, pp. 90-107.