A flexible copula-based approach for the analysis of secondary phenotypes in ascertained samples

Working with SPAC

Karim Oualkacha, Geneviève Lefebvre, Fodé Tounkara & Celia M.T. Greenwood

April 29, 2019

Contents

1 Introduction

This is a unified copula-based framework which tests for secondary phenotypes association and a single SNP in presence of selection bias. The method fits both retrospective and prosepctive likelihoods to handel selection bias under case-contro (CC), extreme-trait (ET) and mulitple-trait (MT) sampling designs. The dependence between primary and secondary phenotypes is modelled via copulas. Thus, SPAC is a robust method for non-normality assumptiom of the secondary trait.

SPAC is an R package that contains methods to perform genetic association of a quantitative secondary phenotype and a single SNP in presence of selection bias. SPAC is a unified copula-based framework fits both retrospective and prosepctive likelihoods to handel selection bias under case-contro (CC), extreme-trait (ET) and mulitple-trait (MT) sampling designs. The dependence between primary and secondary phenotypes is modelled via copulas. Thus, SPAC is a robust method for non-normality assumptiom of the secondary trait; improved power is possible by appropriate modelling of the primary-secondary phenotypes joint distribution via copula models.

The main user-visible function of the package is `SPAC()` function which can be used to analyse single-SNP/Secondary phenotype association in one go.

The main function allows for prospective and retrospective copula-based likelihoods to handle selection bias via the argument *method*:

- pros (default), prospective copula-based likelihood
- retros, retrospective copula-based likelihood

Of note, the retrospective method does not handle covariates while the prosepecetive-based method controls for covariates. The retrospective-based method, *method="retros"*, is the computationally fastest method, and these *p*-values can be used to triage which SNPs of the genome should be re-analyzed with *method="pros"*.

SPAC allows also for primary-secondary dependence modelling using five copulas

- Gaussian, (default);
- Student, Student copula with degree of freedom equals 10;
- Clayton, Clayton copula;
- Gumbel, Gumbel copula;
- Frank, Frank copula.

In order to run any of the examples below, the package needs to be installed and loaded first, of course:

```
library(devtools)
devtools::install()

## Installing SPAC
## '/Library/Frameworks/R.framework/Resources/bin/R' --no-site-file
\
##  --no-environ --no-save --no-restore --quiet CMD INSTALL  \
##  '/Users/KOualkachaUQAM/Dropbox/SPAC'  \
##  --library='/Library/Frameworks/R.framework/Versions/3.4/Resources/library'
\
##  --install-tests
##

# devtools::install_github('KarimOualkacha/SPAC', build_vignettes = TRUE)
devtools::load_all()

## Loading SPAC
## Loading required package:  MASS
## Loading required package:  LaplacesDemon
## Warning:  package 'LaplacesDemon' was built under R version 3.4.4
## Loading required package:  VineCopula
## Warning:  package 'VineCopula' was built under R version 3.4.4
## Loading required package:  copula
## Warning:  package 'copula' was built under R version 3.4.4
##
## Attaching package:  'copula'
## The following object is masked from 'package:LaplacesDemon':
##
##    interval

library(SPAC)
```

2 Loading the input data

Before running an association test with one (or more) of the methods, the following data needs to be present[1]:

- *Phenotype data*; primary and secondary phenpotypes data should be present separatly in the form of an R vector (one value for each individual). It is up to you (as user) to create the vector, for example by reading it from a CSV file using R's `read.csv()` function or like this:

```
  data <- system.file("data","data.RData",
                          package="SPAC")
  load(data)
# data is .RData file contains 3 examples of data sets for CC, ET and MT designs
# y1cc: primary trait under the case-control (CC) design
  head(y1cc)

## [1] 1 1 1 1 1 1

# y2cc: secondary trait under the case-control (CC) design
  head(y2cc)

## [1] 4.5444333 4.6088662 3.4706840 4.3158411 0.5313973 7.2174877
```

- *Covariate data*; this data should be present in the form of a matrix. Like the phenotype data it us up to you to load this data. For example:

[1] The example code in this vignette uses files that are included in the SPAC package, that is why the `package` argument to the `system.file()` function is used.

```
  data <- system.file("data","data.RData",
                            package="SPAC")
  load(data)
# data.file is .RData file contains 3 examples of data sets for CC, ET and MT designs
# matrix of two confounders/covariates: one dichotmous and one continuous covariate
  dim(cov.matCC)

## [1] 1000    2

  head(cov.matCC)

##       conf.1     conf.2
## 79222      1  2.1270291
## 48922      0  2.1488196
## 5552       0  1.9549483
## 98072      1  1.4939171
## 3987       0 -0.5533521
## 3840       1  2.5436760
```

- *Genotype data*; SNP genotype data can be in the form of an R vector (one value for each individual).

```
  data <- system.file("data", "data.RData",
                         package="SPAC")
  load(data)
# data is .RData file contains 3 examples of data sets for CC, ET and MT designs
# SNP data for the CC design: a vector of legnth
  length(markerCC)

## [1] 1000

  head(markerCC)

## [1] 1 1 1 0 0 1
```

3 Analysing a single SNP

With the phenotype data, the covariates and the genotype data loaded it is time for tests of association.

3.1 Using SPAC for a case-control design

This is the simplest way to run SPAC for a single-SNP/secondary phenotype association test under the CC design:

```
SNPresults <- SPAC(y1 = y1cc,
                   y2 = y2cc,
                   G = markerCC,
                   covariates = as.matrix(cov.matCC),
                   link = "probit",
                   copfit = "Gaussian",
                   method = "pros",
                   Design = "CC",
                   prev = 0.1)

## Starting association analysis of the SNP...

SNPresults

## $intercept.SNP.SecP
## [1] 0.63826541 0.07032754
##
## $SNP.SecP
## [1] 0.01637158 0.05628887
##
## $P.value.SecP
## [1] 0.7711666
##
## $intercept.SNP.PrP
## [1] -2.24385138  0.09328779
##
## $SNP.PrP
## [1] 0.17561800 0.06694801
##
## $P.value.PrP
## [1] 0.008710817
##
## $alpha
##
## 0.4744384
##
## $tau
##       tau
## 0.3146976
##
## $df2
##
```

```
## 5004.195
##
## $AIC
## [1] 3885.296
```

- Here the `prev` is a scalar between 0 and 1, specifies the primary phenotype prevalance. It is needed for the `method = "prosp"` and `Design = "CC"` or `Design = "MT"`. Default is `prev = NULL`. If it is not specified, it will be estimated form the data.

- The `link` is a character specifies the link function to be used for modelling the marginal distribution of the binary primary phenotype for the CC and MT designs. The available link functions are `link=c("probit","logit","cloglog")`. The defaut is `link = "probit"`, which th liabiltiy latent model.

- The `copfit` is a character that selects the copula model to use for modelling priamry-secondary phenptypes dependence. Can be one of the following:
  - `copfit = "Gaussian"`, (default)
  - `copfit = "Student"`, Student copula with degree of freedom equals 10
  - `copfit = "Clayton"`, Clayton copula
  - `copfit = "Gumbel"`, Gumbel copula
  - `copfit = "Frank"`, Frank copula

3.2 Using SPAC for an extrem-trait (ET) design

The next code run SPAC for a single-SNP/secondary phenotype association test under the ET design:

```
SNPresults.ET <- SPAC(y1 = y1et,
                   y2 = y2et,
                   G = markeret,
                   covariates = as.matrix(cov.matCC),
                   copfit = "Gaussian",
                   method = "pros",
                   Design = "ET",
                   cutoffs = c(-1.498953,2.174215)
                   )

## Starting association analysis of the SNP...

SNPresults.ET
```

```
## $intercept.SNP.SecP
## [1] 1.02536920 0.09166647
##
## $SNP.SecP
## [1] 0.06940319
##
## $P.value.SecP
## [1] 0.5538218
##
## $intercept.SNP.PrP
## [1] 0.42018677 0.05895765
##
## $SNP.PrP
## [1] 0.08110449 0.04060297
##
## $P.value.PrP
## [1] 0.04577073
##
## $alpha
##
## 0.7291998
##
## $tau
##      tau
## 0.520215
##
## $df2
##
## 41.31134
##
## $AIC
## [1] 5257.224
```

The `cutoffs` is a vector or scalar, depending on the sampling mechanism design `Design = "ET"` or `Design = "MT"`. For the ET design the `cutoffs` is a 2×1 vector, `cutoffs = c(ylb,yub)`, with

- ylb is the lower primary trait threshold
- yub is the upper primary trait threshold

Of note, the ET design does not need a `link` argument since it fits a liability thresohold model for the primary phenotype.

3.3 Using SPAC for an multiple-trait (MT) design

The next code run SPAC for a single-SNP/secondary phenotype association test under the MT design. Here we use `method = "Clayton"` as a copula model to model the joint distribution of the primary-secondary phenotypes:

```
SNPresults.MT <- SPAC(y1 = y1mt,
                   y2 = y2mt,
                   G = markermt,
                   covariates = as.matrix(cov.matMT),
                   link = "probit",
                   copfit = "Clayton",
                   method = "pros",
                   Design = "MT",
                   cutoffs = 1.865465,
                   prev = 0.1)

## Starting association analysis of the SNP...
## Warning in sqrt(diag(mvar)):  NaNs produced

SNPresults.MT

## $intercept.SNP.SecP
## [1] 0.66932760 0.07369333
##
## $SNP.SecP
## [1] 0.05814342
##
## $P.value.SecP
## [1] 0.6116331
##
## $intercept.SNP.PrP
## [1] -2.2577354  0.1104643
##
## $SNP.PrP
## [1] 0.15285477 0.06968403
##
## $P.value.PrP
## [1] 0.02826844
##
## $alpha
##
## 1.225416
##
## $tau
##       tau
## 0.3799249
##
## $df2
##
```

```
## 50.97831
##
## $AIC
## [1] 3757.415
```

In the MT design, the `cutoffs` is a scalar, `cutoffs = y2ub`, with `y2ub` is the upper secondary trait threshold

References

[1] Fodé Tounkara, Geneviève Lefebvre, Celia MT Greenwood and Karim Oualkacha (2019). *A flexible copula-based approach for the analysis of secondary phenotypes in ascertained samples.* Statistics in Medicine. 1-32. Under revision.