SPARE R Manual

The R package ‘SPARE’ can be installed and loaded from GitHub using the commands:

```r
library(devtools)
install_github('JasperHof/SPARE')
library(SPARE)
```

The application of SPARE includes two parts: computation of the null model, and running the GWAS.

Computing the null model

After loading the recurrent event data, the ‘null model’ function computes the martingale residuals and the saddlepoint approximation. Recurrent event data should always include:

- Subject ID
- Start time of risk interval
- End time of risk interval (*i.e.* time of recurrence, or time of censoring)
- Recurrence indicator (equals 1 in case of recurrence, 0 in case of censoring)

Optionally, other covariates can be included in the recurrent event data such as age, sex, or principal components. A typical recurrent event data frame looks like:

```
tstart    tstop       Status  subject   age       sex  recurrence
0.000000  7.686297    1       ID1       53.37211  0    1
7.686297  15.412497   1       ID1       53.37211  0    2
15.412497 31.172565   0       ID1       53.37211  0    3
0.000000  12.769121   1       ID2       28.81588  1    1
12.769121 27.858615   1       ID2       28.81588  1    2
27.858615 39.866657   1       ID2       28.81588  1    3
```

Here, ‘tstart’ and ‘tstop’ denote the start- and the end time of the risk interval, ‘Status’ indicates if the subject experienced a recurrence at time ‘tstop’, and ‘recurrence’ indicates for which recurrence the individual is at risk.

Different recurrent event models can be selected as null model. Practical recommendations for selecting a recurrent event model can be found in previous works [1, 6, 3]. Four well-known recurrent event models are the Andersen and Gill model [2], the Prentice, Williams and Petersen Calendar Time (PWP-CT) model, the Prentice, Williams and Petersen Gap Time (PWP-GT) model [5] and the Gap Time-Unrestricted (GT-UR) model [4]. These models can be fitted in R using the coxme function:

```r
library(coxme)
#AG model
fitme = coxme(Surv(tstart, tstop, Status) ~ age + sex + (1|subject),
              data = data)
#PWP-CT model
fitme = coxme(Surv(tstart, tstop, Status) ~ age + sex + strata(recurrence) + (1|subject),
              data = data)
#PWP-GT model
fitme = coxme(Surv(tstop - tstart, Status) ~ age + sex + strata(recurrence) + (1|subject),
              data = data)
#GT-UR model
fitme = coxme(Surv(tstop - tstart, Status) ~ age + sex + (1|subject),
              data = data)
```

Here, ‘data’ is the name of the example data frame. Note that in these examples, an individual frailty is included to correct for within-individual correlated event times.

After fitting a recurrent event model as null model, this model can be used to compute martingale residuals and the saddle-point approximation using the ‘Null_model’ function:

```
obj.null = Null_model(fitme, data, IDs = unique(data$subject))
```

Running the GWAS

After computing the null model, SNPs are tested for association with risk of recurrence using the functions ‘SPARE.bed’ and ‘SPARE.bgen’, respectively. In both cases, the user should provide a character string containing the genotype IDs of the .bed or .bgen file, which is typically included in a sample file.

For .bed files, SPARE can be applied using the ‘SPARE.bed’ function:

```
SPARE.bed(bedfile, gIDs, obj.null, output.file)
```

Here, the ‘bedfile’ should be the name of the .bed file, without the ‘.bed’ extension. The vectors of genotype IDs should be given by gIDs, which are included in the fam file. ‘obj.null’ is the null object computed from the ’Null_model’ function, and ‘output.file’ is the name of the output file.

Likewise, SPARE can be applied for .bgen files using the ‘SPARE.bgen’ function:

```
SPARE.bgen(bgenfile, gIDs, obj.null, output.file)
```

This requires the same input as SPARE.bed, except the name of the .bgen file is required instead of the .bed file. The .bgen file must be accompanied by a .bgi file, which is named ‘<bgenfile>.bgen.bgi’. The SPARE.bgen function uses a directory in which the ‘backingfiles’ of the .bgen files are stored, which are used to read the genotype data. The name of the directory and the backingfiles can be specified in the SPARE.bgen function, and can be removed after the analysis.

Note: The SPARE package relies on the bigsnpr package for loading .bgen files. For formatting the input genotype data, please check the bigsnpr documentation.

It is possible to tune additional parameters for the GWAS, such as threshold for minor allele frequency (0.05 by default) and P value threshold for implementing the saddle-point approximation (P = 0.001 by default). Additional information about the R functions and example code can be found by running ‘?SPARE.bed’ or ‘?SPARE.bgen’ in R.

Approximate log-hazard ratio

The score statistic obtained in SPARE can be related to a log-hazard ratio that would have been obtained in a classical Cox model. This relationship is given by:

$$\alpha = \gamma \cdot \frac{\sum_{i=1}^{n} g_i^2 \cdot \frac{\Lambda_i(\beta_0,0,u_0)}{\theta\Lambda_i(\beta_0,0,u_0)+1}}{\sum_{i=1}^{n} g_i^2}. \tag{1}$$

Here, α is the score statistic in SPARE, γ is the log-hazard ratio, g_i is the (centered) SNP value of individual i, θ is the frailty variance, and Λ_i is the cumulative hazard of individual i. The estimated frailty variance given by the `coxph` or `coxme` functions of the `survival` and `coxme` packages, respectively, are used to compute approximate log-hazard ratios.

References

[1] Leila DAF Amorim and Jianwen Cai. Modelling recurrent events: a tutorial for analysis in epidemiology. *International journal of epidemiology*, 44(1):324–333, 2015.

[2] Per Kragh Andersen and Richard D Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.

[3] Tyler S Kaster, Simone N Vigod, Tara Gomes, Duminda N Wijeysundera, Daniel M Blumberger, and Rinku Sutradhar. A practical overview and decision tool for analyzing recurrent events in mental illness: A review. *Journal of Psychiatric Research*, 137:7–13, 2021.

[4] Patrick J Kelly and Lynette L-Y Lim. Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statistics in medicine*, 19(1):13–33, 2000.

[5] Ross L Prentice, Benjamin J Williams, and Arthur V Peterson. On the regression analysis of multivariate failure time data. *Biometrika*, 68(2):373–379, 1981.

[6] CP Yadav, V Sreenivas, MA Khan, and RM Pandey. An overview of statistical models for recurrent events analysis: a review. *Epidemiology (Sunnyvale)*, 8(4):354, 2018.