Non-parametric estimation of the age-at-onset distribution from a cross-sectional sample

Soutrik Mandal, Jing Qin, Ruth M. Pfeiffer

Biometrics

July 25, 2022

1 A data example for the proposed EM algorithm

All necessary functions are given in **functions-em.R**. To implement the EM algorithm, the user needs to call the function **func.em**. The arguments in func.em(pc, input.data) are explained below.

pc is a list object in R containing the following inputs:

- K2: Number of categories for T_2 (time from diagnosis to death).
- t1cuts: Cut-points for T_1 (age at diagnosis).
- acuts: Cut-points for A (age at study).
- ageatdx: Column name in data file for age at diagnosis. Note that the entry for non-cases (i.e., healthy subjects) must be NA.
- ageatstudy: Column name in data file for age at study.
- mutation: Column name in data file for mutation status. This column must be coded as: 1 for mutation present, 0 otherwise.

input.data is a data.frame object in R with columns coded according to the description above.

Output

The output from func.em is a list object in R containing the following:

- p0.est: A vector of probabilities p_{0j+}, $j = 1, \ldots, K_1$ for the mutation non-carrier group

$(G = 0)$ where,

$$p_{0j+} = \sum_{k=1}^{K_2} p_{0jk} = \sum_{k=1}^{K_2} P(T_1 = j, T_2 = k|G = 0) \quad (1)$$

- p0.lower: Lower bound of the 95% confidence interval (CI) of p_0.
- p0.upper: Upper bound of the 95% CI of p_0.
- p0.cumulative.lower: Lower bound of the 95% CI of the cumulative sum of p_0.
- p0.cumulative.upper: Upper bound of the 95% CI of the cumulative sum of p_0.
- p1.est: A vector of probabilities p_{1j+}, $j = 1, \ldots, K_1$ for the mutation carrier group $(G = 1)$ where,

$$p_{1j+} = \sum_{k=1}^{K_2} p_{1jk} = \sum_{k=1}^{K_2} P(T_1 = j, T_2 = k|G = 1) \quad (2)$$

- p1.lower: Lower bound of the 95% CI of p_1.
- p1.upper: Upper bound of the 95% CI of p_1.
- p1.cumulative.lower: Lower bound of the 95% CI of the cumulative sum of p_1.
- p1.cumulative.upper: Upper bound of the 95% CI of the cumulative sum of p_1.
- theta.est: Estimate of the genotype prevalence $\theta = P(G = 1)$.
- theta.lower: Lower bound of the 95% CI of θ.
- theta.upper: Upper bound of the 95% CI of θ.

We also provide the file **functions-em-simulation.R** which contains the data generation code used to produce the simulation tables in our paper.

Example:

```
pc= list(K2= 4,
         t1cuts= c(0.3,0.6,1.0),
         acuts= c(0.05,0.15,0.35),
         ageatdx= 'AGEATDX',
         ageatstudy= 'AGEATSTUDY',
```

```
          mutation= 'MUTATION')

input.data= read.csv("data-sim.csv", header=T)

output.em= func.em(pc,input.data)

names(output.em)

output.em$p1.est #to extract the estimates of p1
```