Regressions on bootstrap UK trees - Comparison SA vs Cluster

S. Le Vu
(Dated: April 14, 2016)

```
detail_knitr <- TRUE
source("functions.R")
```

```
library(ape)
library(phydynR)
```

- restrict or not to cohort of sampling
- add explanatory variables
- categorize age and cd4
- run model and tests

Apply source attribution

List of dated bootstrap trees from LSD

```
 ##- list of lsd trees
 ## filename pattern from LSD changes with LSD version !
 # list.lsd.trees <- list.files(path = "data/LSD", pattern = "result.date", full.names = TRUE) # macbook
   list.lsd.trees <- list.files(path = "data/LSD", pattern = "result_newick_date", full.names = TRUE)
# head(list.lsd.trees)
```

Get CD4s and sample times

```
 ##- read first LSD tree to name
 ##- sampling times and CD4s with tip.labels
 t <- read.tree( list.lsd.trees[2] )
# str(t)
 STFN <- "data/LSD/t000.dates"

 ##- CD4 values
 load("../phylo-uk/data/sub.RData")
 rm(s)
 ## selection of df covariates
# names(df)
 cd4s <- setNames(df$cd4, df$seqindex)[t$tip.label]
 head(cd4s)
```

```
76516 21242 50954 21837 26093  2432
  410   214  1140   166   178   514
```

```
 ##- sampling times
 dates <-( read.table(STFN, skip=1,
                     colClasses=c('character', 'numeric') ) )
# head(dates)
 ##- named vector
 sampleTimes <- setNames( dates[,2], dates[,1] )[t$tip.label]
 head(sampleTimes)
```

```
   76516    21242    50954    21837    26093     2432
2007.953 2003.036 2009.458 2005.786 2002.616 2013.959
```

Parameter for phydynR (todo range of incidence / prevalence)

```
 ##- Maximum height
 MH <- 20
 ##- incidence, prevalence: central scenario # todo: range of values
 ## Yin et al. 2014: 2,820 (95% CrI 1,660-4,780)
 newinf <- 2500 # c(1660, 4780)
 plwhiv <- 43150 / 2 # c(43510 / 2, 43510 / 1.5)
```

SA function

```
sa <- function(lsd_tree){
  W <- phylo.source.attribution.hiv( lsd_tree,
          sampleTimes, # years
          cd4s = cd4s,
          ehi = NA,
          numberPeopleLivingWithHIV = plwhiv,
          numberNewInfectionsPerYear = newinf,
          maxHeight = MH,
          res = 1e3,
          treeErrorTol = Inf)
  return(W)
}
```

Save infector probability files

```
for (i in 1:length(list.lsd.trees)){
  w.fn <- paste("data/phydynR/W0_uk_mh", MH, "_",  i, ".rds", sep = '')
  if(!file.exists(w.fn)){
    tree <- read.tree(file = list.lsd.trees[i])
    W <- sa(lsd_tree = tree)
    saveRDS(W, file = w.fn )
  }
}
```

Load and process patients variables

```
detail_knitr <- TRUE
```

```
##- add individual explanatory variates
##- selection of df covariates
load("../phylo-uk/data/sub.RData")
y <- df[,c("seqindex","patientindex",
           "dob_y", "agediag", "cd4",
           "ydiag", "CHICflag", "ethnicityid")]

y$ethn.bin <- ifelse(y$ethnicityid == "White", "white", "not white")
y$CHICflag <- ifelse(y$CHICflag == "Yes", 1, 0)
y$ethnicityid <- NULL
y <- unfactorDataFrame(y)

## categorize continuous variables
y$agecl <- sapply( y[ , "agediag"] , age2quantile )
y$cd4cl <- sapply( y[ , "cd4"] , cd4toStage )
head(y)
```

```
  seqindex patientindex dob_y agediag cd4 ydiag CHICflag  ethn.bin agecl cd4cl
1    88183            1  1969      27 950  1996        1 not white     2     1
2    56250            2  1955      NA  NA    NA        0 not white    NA    NA
3    41484            3  1977      30 497  2007        1     white     2     3
4    83458            4  1963      32 160  1995        0     white     2     5
5    52521            5  1961      NA  NA    NA        0 not white    NA    NA
6    33345            6  1958      32 256  1990        1     white     2     4

rm(s, df)
```