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 !
 if( any(grep("MacBook", Sys.info())) ){
 list.lsd.trees <- list.files(path = "data/LSD", pattern = "result.date", full.names = TRUE)
 } else {
 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)

81625 41073 80125 85494 92828 83878
  420   884   600   950   309   323

 ##- 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)

   81625    41073    80125    85494    92828    83878
2003.285 2002.118 2004.951 2009.871 2011.704 2003.532
```

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, list of W, list of clusters

Get list of infector probs

```
## list of infector prob files
list.W0 <- list.files("data/phydynR", pattern = 'mh20', full.names = TRUE)
## order
list.W0 <- list.W0[order(nchar(list.W0), list.W0)]
```

Set depth in time or cohort. Applied for both outdegree and cluster size determination

```
thr_year <- Inf
```

Calculate outdegrees by bootstrap

```
if(FALSE){
#### for m bootstrap
### function: input filename of W
outdegree <- function(w.fn, t = thr_year){
  W <- readRDS(w.fn)

  ## restrict to cohort sampled within thr_year years
  cohort <- names(sampleTimes[sampleTimes > (max(sampleTimes) - t ) ] )
  i <- which( (W$donor %in% cohort) & (W$recip %in% cohort ) )
  WW <- list( donor = W$donor[i] , recip = W$recip[i], infectorProbability = W$infectorProbability[i] )

  ## calculate outdegrees
  out <- aggregate(
    x = list(outdegree = WW$infectorProbability),
    by = list(patient = WW$donor),
    FUN = function(x) sum(x, na.rm = T) )
  return(out)
}

list.outdegree <- lapply(list.W0, outdegree)
}
```

Get cluster list

```
l_bs_uk <- readRDS( file = "data/listUK_ucsd_clus.rds")
```

Pruning: cluster size and membership are recomputed as time restriction exclude patients

```
##- function to prune cluster according to a threshold of sampling time to control for cohort effect
##- recalculate size and cluster membership
## or use ydiag ? which is different (median diff # 2.5 years)
## depends on clustering algorithm ?

prune.clus <- function(a, t = thr_year){
  ## subset df by sampling times
  cohort <- names(sampleTimes[sampleTimes > (max(sampleTimes) - t ) ] )
  aa <- a[a[,"id"] %in% cohort,]
  if(identical(aa,a)){
    print('do nothing')
    } else {
  ## for each clusterID, re-calculate size and binclus membership
  for (i in unique(aa[,"ClusterID"])){
    aa[ aa[,"ClusterID"] == i, "size" ] <- nrow(aa[ aa[,"ClusterID"] == i, ])
  }
  aa[,"binclus"] <- ifelse(aa[,"size"] < 2, 0, 1 )
    }
  return(aa)
}

list.clus.pruned <- lapply(l_bs_uk, function(x){
  lapply(x, prune.clus)
})

Error in lapply(l_bs_uk, function(x) {:  object 'l_bs_uk' not found
```

Load patients variables

```
##- 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)
```

Load pre-computed list of clusters and outdegrees merged with patients variables

```
#### variables
if(FALSE){
cluster <- lapply(list.clus.pruned, function(u){
  lapply(u, function(x) {
    merge(x, y,
        by.x = "id", by.y = "seqindex",
        all.x = T, sort = FALSE)
})
  })

od <- lapply(list.outdegree, function(x){
  merge(x, y,
        by.x = "patient", by.y = "seqindex",
        all.x = T, sort = FALSE)
})
}
```

```
###- save and read
# saveRDS(cluster, file = "data/list_cluster_uk_bs_thr_demo.rds")
# saveRDS(od, file = "data/list_outdegree_uk_bs_demo.rds")
cluster <- readRDS(file = "data/list_cluster_uk_bs_thr_demo.rds")
od <- readRDS(file = "data/list_outdegree_uk_bs_demo.rds")

list.total <- c("SA" = list(od), "Cluster" = cluster)
names(list.total)

[1] "SA"           "Cluster.0.01" "Cluster.0.02" "Cluster.0.05"
```

Regression models

Function to summarize regression results

```r
source("test_fn_compare.reg.sum.bs.R")
compare.reg.bs

function (ls, reg, model, alpha = 0.05, ...)
{
    coef <- lapply(ls, function(x) {
        lapply(x, function(x) {
            if ("size" %in% names(x)) {
                full.model <- sub("y", "scale(size)", model)
            }
            else if ("outdegree" %in% names(x)) {
                full.model <- sub("y", "scale(outdegree)", model)
            }
            else stop("cannot find y")
            coef(summary(reg(formula = full.model, data = x,
                ...)))
        })
    })
    pvalue <- lapply(coef, function(x) {
        sapply(x, function(x) {
            identity(x[, 4])
        })
    })
    sum.signif <- sapply(pvalue, function(x) {
        apply(x, 1, function(x) sum(x < alpha)/length(x))
    })
    param <- lapply(coef, function(x) {
        sapply(x, function(x) {
            identity(x[, 1])
        })
    })
    mean.parms <- signif(sapply(param, function(x) {
        apply(x, 1, mean)
    }), 2)
    if (identical(reg, lm)) {
        r2 <- lapply(ls, function(x) {
            sapply(x, function(x) {
                if ("size" %in% names(x)) {
                  full.model <- sub("y", "scale(size)", model)
                }
                else if ("outdegree" %in% names(x)) {
                  full.model <- sub("y", "scale(outdegree)",
                    model)
                }
                else stop("cannot find y")
                summary(reg(full.model, data = x))$r.squared
            })
        })
        mean.r2 <- signif(sapply(r2, function(x) {
            mean(x)
        }), 3)
        return(list(model = model, `mean parameter` = mean.parms,
            `signif pvalue` = sum.signif, `mean r.squared` = mean.r2))
    }
    else {
        return(list(model = model, `mean parameter` = mean.parms,
            `signif pvalue` = sum.signif))
    }
}
```

List of models

```
model1 <- "y ~ factor(agecl)"
model1c <- "y ~ scale(agediag)"
model1i <- "y ~ factor(agecl) + factor(cd4cl)"
model2 <- "y ~ factor(cd4cl)"
model3 <- "y ~ factor(ethn.bin)"
model4 <- "y ~ factor(CHICflag)"
model5 <- "y ~ factor(agecl) + factor(cd4cl) + factor(agecl)*factor(cd4cl) "
model6 <- "y ~ scale(agediag) + scale(sqrt(cd4)) + factor(ethn.bin) + factor(CHICflag)"
```

Age only

```
test <- compare.reg.bs(ls = list.total, reg = lm, model = model1, alpha = 0.05)
test

$model
[1] "y ~ factor(agecl)"

$`mean parameter`
                   SA Cluster.0.01 Cluster.0.02 Cluster.0.05
(Intercept)     0.095        0.041        0.042       0.0160
factor(agecl)2 -0.090       -0.042       -0.052      -0.0420
factor(agecl)3 -0.076       -0.049       -0.044      -0.0079
factor(agecl)4 -0.180       -0.068       -0.065      -0.0054

$`signif pvalue`
                 SA Cluster.0.01 Cluster.0.02 Cluster.0.05
(Intercept)    0.10         0.61         0.69         0.10
factor(agecl)2 0.04         0.25         0.55         0.36
factor(agecl)3 0.03         0.42         0.29         0.07
factor(agecl)4 0.34         0.75         0.78         0.07

$`mean r.squared`
          SA Cluster.0.01 Cluster.0.02 Cluster.0.05
    0.005630     0.000633     0.000581     0.000543
```

Interpretation:

- negative effect on both OD and cluster size (decrease with age)
- detected more frequently in cluster size at low thresholds

Adding CD4 to age

```
test2 <- compare.reg.bs(ls = list.total, reg = lm, model = model1i, alpha = 0.05)
test2
```

```
$model
[1] "y ~ factor(agecl) + factor(cd4cl)"

$`mean parameter`
                   SA Cluster.0.01 Cluster.0.02 Cluster.0.05
(Intercept)     0.060       0.0630       0.0630       0.0370
factor(agecl)2 -0.088      -0.0400      -0.0510      -0.0370
factor(agecl)3 -0.030      -0.0410      -0.0380      -0.0012
factor(agecl)4 -0.140      -0.0490      -0.0440       0.0067
factor(cd4cl)2  0.110       0.0160       0.0370      -0.0030
factor(cd4cl)3  0.063      -0.0032       0.0045      -0.0250
factor(cd4cl)4 -0.062      -0.0200      -0.0320      -0.0490
factor(cd4cl)5 -0.220      -0.1400      -0.1600      -0.0540

$`signif pvalue`
                 SA Cluster.0.01 Cluster.0.02 Cluster.0.05
(Intercept)    0.02         0.58         0.62         0.16
factor(agecl)2 0.06         0.20         0.46         0.22
factor(agecl)3 0.01         0.26         0.12         0.05
factor(agecl)4 0.21         0.39         0.27         0.08
factor(cd4cl)2 0.13         0.00         0.03         0.03
factor(cd4cl)3 0.04         0.00         0.00         0.10
factor(cd4cl)4 0.04         0.00         0.01         0.32
factor(cd4cl)5 0.31         1.00         1.00         0.42

$`mean r.squared`
          SA Cluster.0.01 Cluster.0.02 Cluster.0.05
     0.01940      0.00355      0.00470      0.00154
```

Interpretation:

- Again, negative effect of age on both OD and cluster size
- Only effect detected for CD4 < 200 vs CD4 > 700, always significant for cluster size models and 31% of SA model

Results of model with continuous age and CD4 + ethnicity and CHIC

```
test6 <- compare.reg.bs(ls = list.total, reg = lm, model = model6, alpha = 0.05)
test6

$model
[1] "y ~ scale(agediag) + scale(sqrt(cd4)) + factor(ethn.bin) + factor(CHICflag)"

$`mean parameter`
                            SA Cluster.0.01 Cluster.0.02 Cluster.0.05
(Intercept)            0.00025      -0.0130        0.011       -0.110
scale(agediag)        -0.04400      -0.0150       -0.011        0.013
scale(sqrt(cd4))       0.07400       0.0410        0.052        0.021
factor(ethn.bin)white -0.00160       0.0045        0.008        0.062
factor(CHICflag)1     -0.00620       0.0150       -0.017        0.079

$`signif pvalue`
                        SA Cluster.0.01 Cluster.0.02 Cluster.0.05
(Intercept)           0.00         0.00         0.00         0.83
scale(agediag)        0.05         0.31         0.07         0.33
scale(sqrt(cd4))      0.65         1.00         1.00         0.54
```

```
factor(ethn.bin)white 0.01         0.00         0.00         0.66
factor(CHICflag)1     0.01         0.02         0.00         0.78

$`mean r.squared`
          SA Cluster.0.01 Cluster.0.02 Cluster.0.05
     0.01090      0.00231      0.00309      0.00307
```

Interpretation:

- No effect for ethnicity and CHIC, except for high cluster threshold where everything pops out ???
- CD4 is frequently associated with dependent variable, especially for cluster size

Results of model with factorized age and CD4 plus interactions

```
test5 <- compare.reg.bs(ls = list.total, reg = lm, model = model5, alpha = 0.05)
test5

$model
[1] "y ~ factor(agecl) + factor(cd4cl) + factor(agecl)*factor(cd4cl) "

$`mean parameter`
                                  SA Cluster.0.01 Cluster.0.02 Cluster.0.05
(Intercept)                    0.100       0.0660       0.0590       0.0520
factor(agecl)2                -0.120      -0.0470      -0.0520      -0.0320
factor(agecl)3                -0.140      -0.0590      -0.0600      -0.0410
factor(agecl)4                -0.180      -0.0330       0.0089      -0.0220
factor(cd4cl)2                 0.140      -0.0390       0.0047      -0.0570
factor(cd4cl)3                -0.074       0.0490       0.0560      -0.0017
factor(cd4cl)4                -0.210      -0.0091      -0.0086      -0.0650
factor(cd4cl)5                -0.052      -0.2100      -0.2300      -0.0950
factor(agecl)2:factor(cd4cl)2 -0.054       0.0660       0.0240       0.0490
factor(agecl)3:factor(cd4cl)2  0.015       0.0930       0.1000       0.1000
factor(agecl)4:factor(cd4cl)2 -0.075       0.0460      -0.0067       0.0550
factor(agecl)2:factor(cd4cl)3  0.120      -0.0580      -0.0350      -0.0630
factor(agecl)3:factor(cd4cl)3  0.240      -0.0300      -0.0250       0.0150
factor(agecl)4:factor(cd4cl)3  0.170      -0.1300      -0.1700      -0.0360
factor(agecl)2:factor(cd4cl)4  0.200       0.0018      -0.0082       0.0049
factor(agecl)3:factor(cd4cl)4  0.190      -0.0340      -0.0240       0.0420
factor(agecl)4:factor(cd4cl)4  0.170      -0.0140      -0.0730       0.0200
factor(agecl)2:factor(cd4cl)5 -0.230       0.0640       0.0710      -0.0015
factor(agecl)3:factor(cd4cl)5 -0.031       0.0950       0.0900       0.0370
factor(agecl)4:factor(cd4cl)5 -0.230       0.0650       0.0500       0.1000

$`signif pvalue`
                                SA Cluster.0.01 Cluster.0.02 Cluster.0.05
(Intercept)                   0.00         0.09         0.01         0.14
factor(agecl)2                0.00         0.00         0.00         0.06
factor(agecl)3                0.02         0.01         0.00         0.06
factor(agecl)4                0.02         0.01         0.00         0.05
factor(cd4cl)2                0.08         0.03         0.00         0.15
factor(cd4cl)3                0.00         0.03         0.00         0.04
factor(cd4cl)4                0.00         0.00         0.00         0.14
factor(cd4cl)5                0.00         0.93         0.99         0.27
factor(agecl)2:factor(cd4cl)2 0.03         0.01         0.00         0.07
factor(agecl)3:factor(cd4cl)2 0.01         0.07         0.04         0.16
factor(agecl)4:factor(cd4cl)2 0.00         0.04         0.00         0.06
factor(agecl)2:factor(cd4cl)3 0.00         0.03         0.00         0.11
factor(agecl)3:factor(cd4cl)3 0.01         0.01         0.00         0.02
factor(agecl)4:factor(cd4cl)3 0.01         0.16         0.34         0.08
factor(agecl)2:factor(cd4cl)4 0.00         0.00         0.00         0.01
factor(agecl)3:factor(cd4cl)4 0.02         0.00         0.00         0.03
factor(agecl)4:factor(cd4cl)4 0.00         0.00         0.00         0.03
factor(agecl)2:factor(cd4cl)5 0.00         0.00         0.00         0.01
factor(agecl)3:factor(cd4cl)5 0.01         0.01         0.00         0.02
factor(agecl)4:factor(cd4cl)5 0.02         0.00         0.00         0.10

$`mean r.squared`
          SA Cluster.0.01 Cluster.0.02 Cluster.0.05
     0.02920      0.00465      0.00574      0.00286
```

Interpretation:

- R2 increased... a little
- cluster still see some effect of CD4 $<$ 200