Processing ExaML boostrap trees - April 2016

S. Le Vu
(Dated: April 8, 2016)

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

```
library(ape)

Warning:  package 'ape' was built under R version 3.2.3
```

- Starting with 100 bootstrap trees
- Transformed into edge lists of distances
- As inputs of UCSD software at thresholds c("0.01", "0.02", "0.05")
- Obtain a list of 3 * 100 cluster assignements

```
cl2 <- readRDS( file = "data/ucsd_results/list.hivclust.rds" )
```

```
  ## number of different cluster of size > 1
  sapply(aa, function(x){
    summary(sapply(x, function(x) dim(x)[1]))
  })

        0.01 0.02 0.05
Min.     974 1274 1153
1st Qu. 1039 1358 1401
Median  1068 1374 1430
Mean    1070 1377 1417
3rd Qu. 1093 1396 1449
Max.    1226 1462 1503
```

Number of clusters (n > 1) increases because at larger threshold, more patients are included. See below how it changes when everybody is in.

```
  ## stats of mean size
  sapply(aa, function(x){
    summary(sapply(x, function(x) mean(x$Freq)))
  })

         0.01  0.02  0.05
Min.    2.679 3.349 4.816
1st Qu. 2.864 3.551 5.449
Median  2.907 3.637 5.693
Mean    2.915 3.646 5.788
3rd Qu. 2.955 3.704 5.923
Max.    3.129 4.064 8.174
```

Mean cluster size logically increases, with some variation between bootstraps
After having merged patients data with cluster assignments,

```
# saveRDS(listUKclus, file = "data/listUKclus.rds")
listUKclus <- readRDS( file = "data/listUKclus.rds")
```

```
## number of different clusters (counting size 1)
sapply(listUKclus, function(x){
  summary(sapply(x, function(x) {
    length(unique(x$ClusterID) )
  }))
})
```

```
         0.01 0.02 0.05
Min.     9554 7685 3892
1st Qu. 10030 8380 5216
Median  10140 8563 5482
Mean    10110 8517 5407
3rd Qu. 10220 8671 5696
Max.    10530 9171 6466
```

Now cluster number decreases

```
## proportion of cluster membership
sapply(listUKclus, function(x){
  summary(sapply(x, function(x) sum(x$binclus) / length(x$binclus)))
})
```

```
          0.01   0.02   0.05
Min.    0.2145 0.3508 0.5912
1st Qu. 0.2454 0.3999 0.6515
Median  0.2538 0.4089 0.6656
Mean    0.2567 0.4130 0.6720
3rd Qu. 0.2647 0.4268 0.6885
Max.    0.3154 0.4884 0.7748
```

```
## stats of mean size
sapply(listUKclus, function(x){
  summary(sapply(x, function(x) mean(x$size)))
})
```

```
         0.01  0.02   0.05
Min.    1.861 4.540  13.31
1st Qu. 2.271 5.825  25.85
Median  2.447 6.262  38.95
Mean    2.513 6.356  82.70
3rd Qu. 2.731 6.751  76.39
Max.    3.679 8.801 840.10
```

Linear regressions

```
lapply(models, function(x) {reg.sum.bs(ls = listUKclus, reg = lm, model = x)
  })
```

```
[[1]]
[[1]]$model
[1] "model0"

[[1]]$`mean parameter`
                    0.01      0.02      0.05
```

```
(Intercept)     1.7e-15 -5.0e-16 -1.6e-15
scale(agediag) -2.0e-02 -1.6e-02  7.7e-03

[[1]]$`signif pvalue`
               0.01 0.02 0.05
(Intercept)    0.00 0.00 0.00
scale(agediag) 0.63 0.39 0.24

[[1]]$`mean r.squared`
    0.01     0.02     0.05
0.000444 0.000281 0.000208


[[2]]
[[2]]$model
[1] "model1"

[[2]]$`mean parameter`
                   0.01   0.02    0.05
(Intercept)      0.0032 0.0039 0.00071
scale(sqrt(cd4)) 0.0440 0.0540 0.02100

[[2]]$`signif pvalue`
                 0.01 0.02 0.05
(Intercept)         0    0 0.00
scale(sqrt(cd4))    1    1 0.54

[[2]]$`mean r.squared`
    0.01     0.02     0.05
0.001930 0.002890 0.000693


[[3]]
[[3]]$model
[1] "model2"

[[3]]$`mean parameter`
                         0.01    0.02   0.05
(Intercept)           -0.0047 -0.0084 -0.058
factor(ethn.bin)white  0.0058  0.0100  0.071

[[3]]$`signif pvalue`
                      0.01 0.02 0.05
(Intercept)              0    0 0.70
factor(ethn.bin)white    0    0 0.74

[[3]]$`mean r.squared`
    0.01     0.02     0.05
2.99e-05 2.68e-05 9.21e-04


[[4]]
[[4]]$model
[1] "model3"

[[4]]$`mean parameter`
                     0.01    0.02   0.05
```

```
(Intercept)         0.007  0.0021  0.016
factor(CHICflag)No -0.034 -0.0100 -0.077

[[4]]$`signif pvalue`
                  0.01 0.02 0.05
(Intercept)       0.00    0 0.26
factor(CHICflag)No 0.31    0 0.85

[[4]]$`mean r.squared`
    0.01     0.02     0.05
2.30e-04 3.05e-05 1.15e-03


[[5]]
[[5]]$model
[1] "model4"

[[5]]$`mean parameter`
                          0.01    0.02   0.05
(Intercept)            0.0024 -0.0057 -0.035
scale(agediag)        -0.0150 -0.0110  0.013
scale(sqrt(cd4))       0.0410  0.0520  0.021
factor(ethn.bin)white  0.0045  0.0080  0.062
factor(CHICflag)No    -0.0150  0.0170 -0.079

[[5]]$`signif pvalue`
                     0.01 0.02 0.05
(Intercept)          0.00 0.00 0.30
scale(agediag)       0.31 0.07 0.33
scale(sqrt(cd4))     1.00 1.00 0.54
factor(ethn.bin)white 0.00 0.00 0.66
factor(CHICflag)No   0.02 0.00 0.78

[[5]]$`mean r.squared`
   0.01    0.02    0.05
0.00231 0.00309 0.00307
```

- Age: negative effect - fading out as threshold increases - low R2
- CD4: positive effect - 100% significant except high threshold (54%)
- Ethnicity: positive effect (whites in larger clusters) only significant at high threshold
- CHIC: positive effect (CHIC in large clusters) - not always significant
- Full model: Only CD4 would show a constantly significant effect over all bootstrap trees up to a high threshold. Ethnicity and CHIC come out at high threshold. Overall small R2

Logistic regression

```
reg.sum.bs(ls = listUKclus, reg = glm, model = logit_model_uk, family = binomial(link = "logit"))

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

$`mean parameter`
                        0.01   0.02  0.05
(Intercept)          -1.100 -0.350 0.670
scale(agediag)       -0.034 -0.013 0.023
```

```
scale(sqrt(cd4))       0.230  0.240 0.250
factor(ethn.bin)white -0.022 -0.062 0.029
factor(CHICflag)No     0.280  0.300 0.290

$`signif pvalue`
                     0.01 0.02 0.05
(Intercept)          1.00 0.99 1.00
scale(agediag)       0.26 0.01 0.07
scale(sqrt(cd4))     1.00 1.00 1.00
factor(ethn.bin)white 0.02 0.08 0.03
factor(CHICflag)No   1.00 1.00 1.00
```

Only CD4 and CHIC show a constant effect