GSOC Tests

Anuraag Srivastava(as4378)

February 22, 2019

Test 1

Run either changepoint::cpt.mean or Fpop::fpop on one of the data sets (vector of logratio values for a given profile.id/chromosome combination) in neuroblatoma$profiles from data(neuroblastoma, package="neuroblastoma"). For one penalty parameter, plot the data as black points and the optimal segment means as horizontal green line segments.

Loading the required packages:

```
data(neuroblastoma, package="neuroblastoma")
options(width=100)
```

Selecting one profile id to continue with the test. This profile is of a children status "relapse" several years after treatment, hence a good candidate for change point detection problem.

```
selected <- data.frame(
  profile.id=paste(c(1)),
  status=c("relapse"))
selected
```

```
##   profile.id  status
## 1          1 relapse
```

Creating a data table for selected profile:

```
rownames(selected) <- selected$profile.id
selected$status.profile <- with(selected, paste(status, profile.id))
some.ids <- rownames(selected)
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.5.2
```

```
someProfiles <- function(all.profiles){
  some <- subset(all.profiles, paste(profile.id) %in% some.ids)
  status.profile <- selected[paste(some$profile.id), "status.profile"]
  some$status.profile <- ifelse(
    is.na(status.profile), paste(some$profile.id), status.profile)
  data.table(some)
}
selected.profiles <- someProfiles(neuroblastoma$profiles)
```

Now, for selected profile there are 24 change-point detection problems (24 chromosomes). Plotting this problem in a grid we get:

```
library(ggplot2)
```

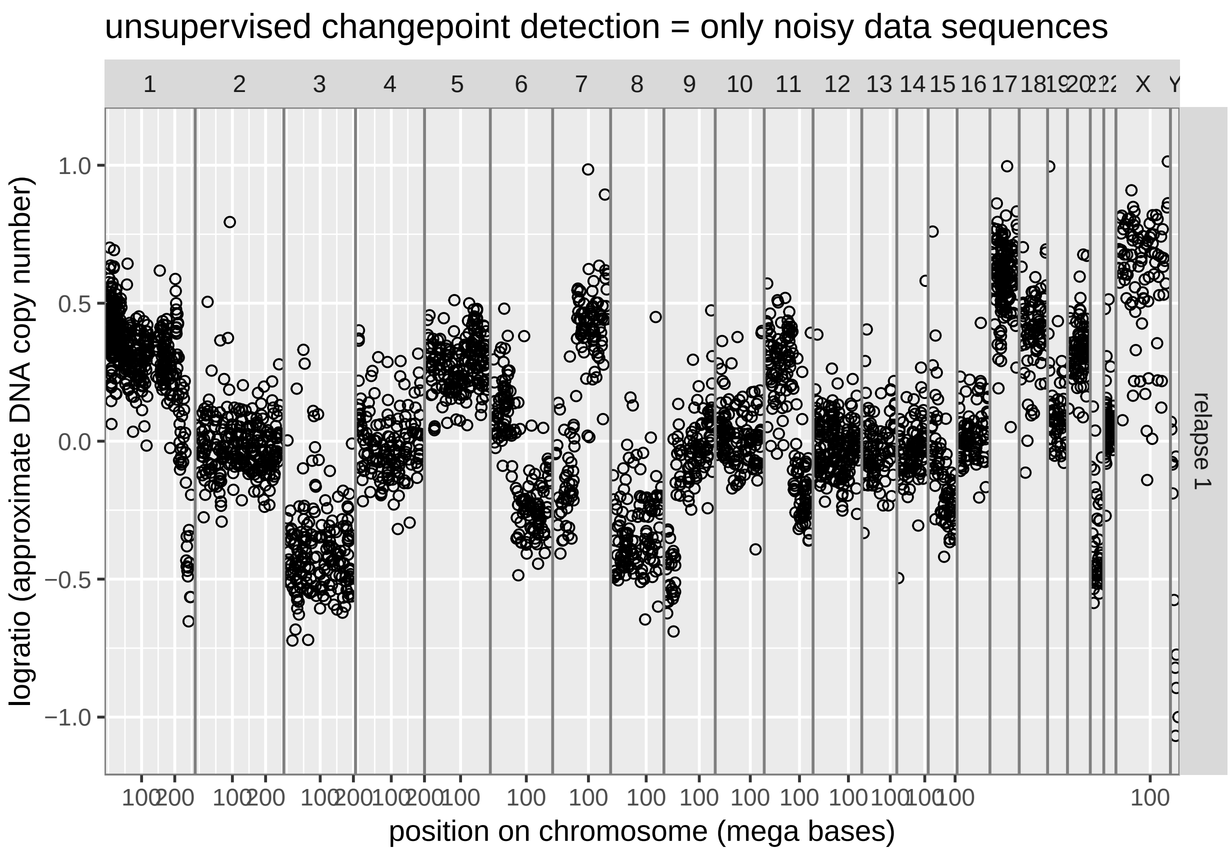
```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
gg.unsupervised <- ggplot()+
  ggtitle("unsupervised changepoint detection = only noisy data sequences")+
  theme(
    panel.margin=grid::unit(0, "lines"),
    panel.border=element_rect(fill=NA, color="grey50")
  )+
  facet_grid(status.profile ~ chromosome, scales="free", space="free_x")+
  geom_point(aes(position/1e6, logratio),
             data=selected.profiles,
             shape=1)+
  scale_x_continuous(
    "position on chromosome (mega bases)",
    breaks=c(100, 200))+
  scale_y_continuous(
    "logratio (approximate DNA copy number)",
    limits=c(-1,1)*1.1)
```

```
## Warning: `panel.margin` is deprecated. Please use `panel.spacing` property instead
```

```
print(gg.unsupervised)
```

```
## Warning: Removed 11 rows containing missing values (geom_point).
```



Now, fitting the unsupervised change point model using the cpt.mean function and using "SIC0" as the penalty parameter we get:

```
pen.name <- "SIC0"
(unsupervised.models <- selected.profiles[, {
  fit.pelt <- changepoint::cpt.mean(
    logratio, penalty=pen.name, method="PELT")
  end <- fit.pelt@cpts
  before.change <- end[-length(end)]
  after.change <- before.change+1L
  data.table(
    pen.name,
    pen.value=fit.pelt@pen.value,
    changes=list(
    as.integer((position[before.change]+position[after.change])/2)
    ))
}, by=list(profile.id, status.profile, chromosome)])
```

```
##     profile.id status.profile chromosome pen.name pen.value                    changes
##  1:          1      relapse 1          1     SIC0  6.161207                  212809180
##  2:          1      relapse 1          2     SIC0  5.521461
##  3:          1      relapse 1          3     SIC0  5.252273
##  4:          1      relapse 1          4     SIC0  5.036953
##  5:          1      relapse 1          5     SIC0  5.214936
##  6:          1      relapse 1          6     SIC0  5.093750
##  7:          1      relapse 1          7     SIC0  4.859812                   60381530
##  8:          1      relapse 1          8     SIC0  4.836282
##  9:          1      relapse 1          9     SIC0  4.770685
## 10:          1      relapse 1         10     SIC0  4.919981
## 11:          1      relapse 1         11     SIC0  5.043425                   80058339
## 12:          1      relapse 1         12     SIC0  5.303305
## 13:          1      relapse 1         13     SIC0  4.499810
## 14:          1      relapse 1         14     SIC0  4.564348
## 15:          1      relapse 1         15     SIC0  4.369448
## 16:          1      relapse 1         16     SIC0  4.477337
## 17:          1      relapse 1         17     SIC0  5.141664
## 18:          1      relapse 1         18     SIC0  4.343805
## 19:          1      relapse 1         19     SIC0  4.007333
## 20:          1      relapse 1         20     SIC0  4.488636
## 21:          1      relapse 1         21     SIC0  3.713572
## 22:          1      relapse 1         22     SIC0  4.406719
## 23:          1      relapse 1          X     SIC0  4.553877
## 24:          1      relapse 1          Y     SIC0  3.044522  5918094, 9420652,19070635
##     profile.id status.profile chromosome pen.name pen.value                    changes
```

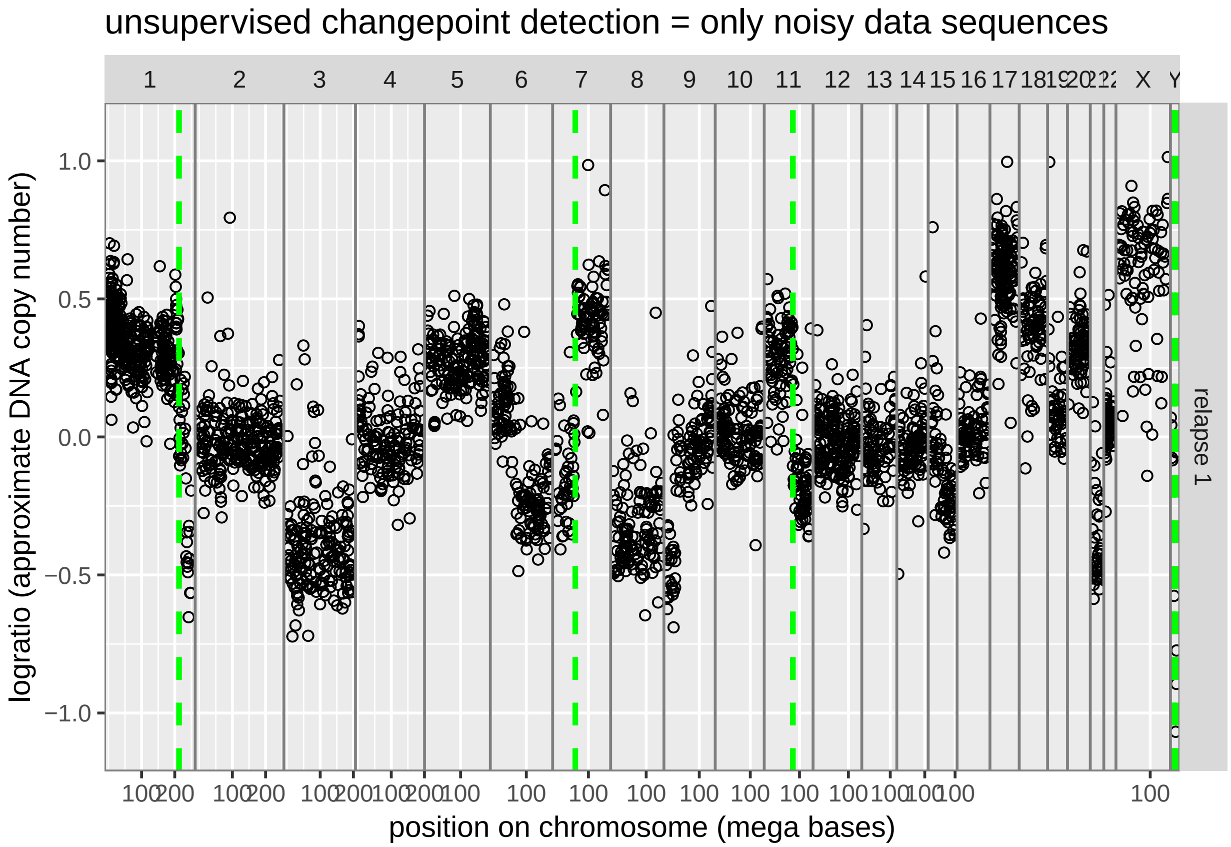
```r
(unsupervised.changes <- unsupervised.models[, data.table(
  change=changes[[1]]
), by=list(profile.id, status.profile, chromosome, pen.name)])
```

```
##    profile.id status.profile chromosome pen.name    change
## 1:          1      relapse 1          1     SIC0 212809180
## 2:          1      relapse 1          7     SIC0  60381530
## 3:          1      relapse 1         11     SIC0  80058339
## 4:          1      relapse 1          Y     SIC0   5918094
## 5:          1      relapse 1          Y     SIC0   9420652
## 6:          1      relapse 1          Y     SIC0  19070635
```

Plotting the data along with optimal segment means (green line segments) we get:

```r
gg.unsupervised+
  theme(legend.box="horizontal")+
  geom_vline(aes(
    xintercept=change/1e6),
    color="green",
    size=1,
    linetype="dashed",
    data=unsupervised.changes)
```

```
## Warning: Removed 11 rows containing missing values (geom_point).
```



Test 2

For two data sets and three penalty parameters, plot the data and optimal models using a ggplot with facet_grid(segments ~ profile.id + chromosome)

This time selecting 2 profiles with status "relapse".

```
selected <- data.frame(
  profile.id=paste(c(1, 4)),
  status=c("relapse", "relapse"))
selected
```

```
##   profile.id  status
## 1          1 relapse
## 2          4 relapse
```

Creating data table for selected profiles:

```
rownames(selected) <- selected$profile.id
selected$status.profile <- with(selected, paste(status, profile.id))
some.ids <- rownames(selected)
library(data.table)
someProfiles <- function(all.profiles){
  some <- subset(all.profiles, paste(profile.id) %in% some.ids)
  status.profile <- selected[paste(some$profile.id), "status.profile"]
  some$status.profile <- ifelse(
    is.na(status.profile), paste(some$profile.id), status.profile)
  data.table(some)
}
selected.profiles <- someProfiles(neuroblastoma$profiles)
```

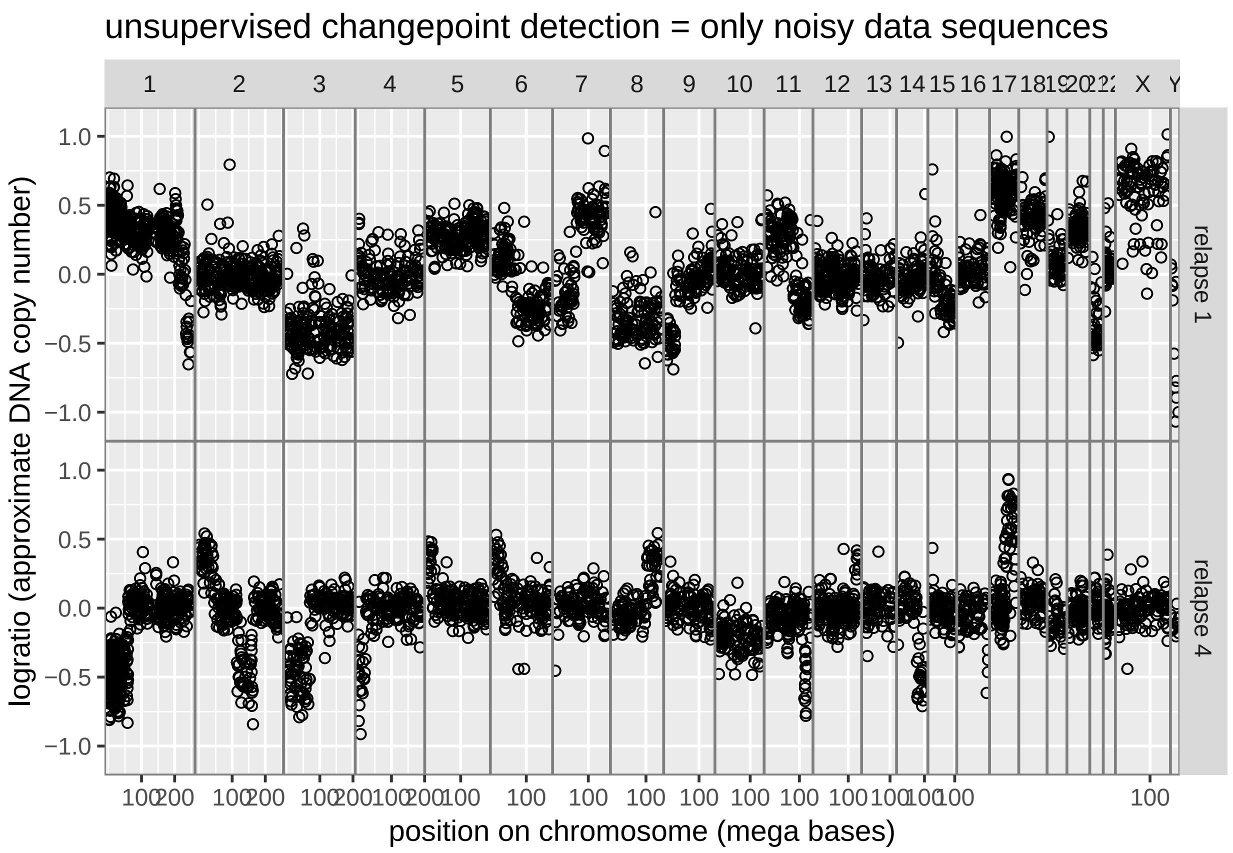
Printing the problems (24 * 2 = 48):

```
library(ggplot2)
gg.unsupervised <- ggplot()+
  ggtitle("unsupervised changepoint detection = only noisy data sequences")+
  theme(
    panel.margin=grid::unit(0, "lines"),
    panel.border=element_rect(fill=NA, color="grey50")
  )+
  facet_grid(status.profile ~ chromosome, scales="free", space="free_x")+
  geom_point(aes(position/1e6, logratio),
             data=selected.profiles,
             shape=1)+
  scale_x_continuous(
    "position on chromosome (mega bases)",
    breaks=c(100, 200))+
  scale_y_continuous(
    "logratio (approximate DNA copy number)",
    limits=c(-1,1)*1.1)
```

```
## Warning: `panel.margin` is deprecated. Please use `panel.spacing` property instead
```

```
print(gg.unsupervised)
```

```
## Warning: Removed 11 rows containing missing values (geom_point).
```



Now, fitting the unsupervised change point model using the cpt.mean function and using 3 penalty paramters "SIC0, BIC0 and AIC0" we get:

```
pen.name <- "SIC0"
(unsupervised.models <- selected.profiles[, {
  fit.pelt <- changepoint::cpt.mean(
    logratio, penalty=pen.name, method="PELT")
  end <- fit.pelt@cpts
  before.change <- end[-length(end)]
  after.change <- before.change+1L
  data.table(
    pen.name,
    pen.value=fit.pelt@pen.value,
    changes=list(
    as.integer((position[before.change]+position[after.change])/2)
    ))
}, by=list(profile.id, status.profile, chromosome)])
```

```
##      profile.id status.profile chromosome pen.name pen.value                         changes
##   1:          4      relapse 4          1     SIC0  6.059123                        59792500
##   2:          4      relapse 4          2     SIC0  5.455321                        45164625
##   3:          4      relapse 4          3     SIC0  5.141664                        69953902
##   4:          4      relapse 4          4     SIC0  4.983607
##   5:          4      relapse 4          5     SIC0  5.247024
##   6:          4      relapse 4          6     SIC0  4.976734
##  7:          4 relapse 4               7     SICO  4.753590
##  8:          4 relapse 4               8     SICO  4.779123
##  9:          4 relapse 4               9     SICO  4.779123
## 10:          4 relapse 4              10     SICO  4.890349
## 11:          4 relapse 4              11     SICO  4.990433
## 12:          4 relapse 4              12     SICO  5.247024
## 13:          4 relapse 4              13     SICO  4.442651
## 14:          4 relapse 4              14     SICO  4.330733                   76603452
## 15:          4 relapse 4              15     SICO  4.406719
## 16:          4 relapse 4              16     SICO  4.330733
## 17:          4 relapse 4              17     SICO  5.030438                   41646489
## 18:          4 relapse 4              18     SICO  4.330733
## 19:          4 relapse 4              19     SICO  3.784190
## 20:          4 relapse 4              20     SICO  4.488636
## 21:          4 relapse 4              21     SICO  3.761200
## 22:          4 relapse 4              22     SICO  4.276666
## 23:          4 relapse 4               X     SICO  4.770685
## 24:          4 relapse 4               Y     SICO  2.484907
## 25:          1 relapse 1               1     SICO  6.161207                  212809180
## 26:          1 relapse 1               2     SICO  5.521461
## 27:          1 relapse 1               3     SICO  5.252273
## 28:          1 relapse 1               4     SICO  5.036953
## 29:          1 relapse 1               5     SICO  5.214936
## 30:          1 relapse 1               6     SICO  5.093750
## 31:          1 relapse 1               7     SICO  4.859812                   60381530
## 32:          1 relapse 1               8     SICO  4.836282
## 33:          1 relapse 1               9     SICO  4.770685
## 34:          1 relapse 1              10     SICO  4.919981
## 35:          1 relapse 1              11     SICO  5.043425                   80058339
## 36:          1 relapse 1              12     SICO  5.303305
## 37:          1 relapse 1              13     SICO  4.499810
## 38:          1 relapse 1              14     SICO  4.564348
## 39:          1 relapse 1              15     SICO  4.369448
## 40:          1 relapse 1              16     SICO  4.477337
## 41:          1 relapse 1              17     SICO  5.141664
## 42:          1 relapse 1              18     SICO  4.343805
## 43:          1 relapse 1              19     SICO  4.007333
## 44:          1 relapse 1              20     SICO  4.488636
## 45:          1 relapse 1              21     SICO  3.713572
## 46:          1 relapse 1              22     SICO  4.406719
## 47:          1 relapse 1               X     SICO  4.553877
## 48:          1 relapse 1               Y     SICO  3.044522  5918094, 9420652,19070635
##     profile.id status.profile chromosome pen.name pen.value                    changes
```

```
pen.name <- "BICO"
(unsupervised.models1 <- selected.profiles[, {
  fit.pelt <- changepoint::cpt.mean(
    logratio, penalty=pen.name, method="PELT")
  end <- fit.pelt@cpts
  before.change <- end[-length(end)]
  after.change <- before.change+1L
  data.table(
    pen.name,
    pen.value=fit.pelt@pen.value,
    changes=list(
    as.integer((position[before.change]+position[after.change])/2)
    ))
}, by=list(profile.id, status.profile, chromosome)])
```

```
##     profile.id status.profile chromosome pen.name pen.value                   changes
##  1:          4      relapse 4          1     BIC0  6.059123                  59792500
##  2:          4      relapse 4          2     BIC0  5.455321                  45164625
##  3:          4      relapse 4          3     BIC0  5.141664                  69953902
##  4:          4      relapse 4          4     BIC0  4.983607
##  5:          4      relapse 4          5     BIC0  5.247024
##  6:          4      relapse 4          6     BIC0  4.976734
##  7:          4      relapse 4          7     BIC0  4.753590
##  8:          4      relapse 4          8     BIC0  4.779123
##  9:          4      relapse 4          9     BIC0  4.779123
## 10:          4      relapse 4         10     BIC0  4.890349
## 11:          4      relapse 4         11     BIC0  4.990433
## 12:          4      relapse 4         12     BIC0  5.247024
## 13:          4      relapse 4         13     BIC0  4.442651
## 14:          4      relapse 4         14     BIC0  4.330733                  76603452
## 15:          4      relapse 4         15     BIC0  4.406719
## 16:          4      relapse 4         16     BIC0  4.330733
## 17:          4      relapse 4         17     BIC0  5.030438                  41646489
## 18:          4      relapse 4         18     BIC0  4.330733
## 19:          4      relapse 4         19     BIC0  3.784190
## 20:          4      relapse 4         20     BIC0  4.488636
## 21:          4      relapse 4         21     BIC0  3.761200
## 22:          4      relapse 4         22     BIC0  4.276666
## 23:          4      relapse 4          X     BIC0  4.770685
## 24:          4      relapse 4          Y     BIC0  2.484907
## 25:          1      relapse 1          1     BIC0  6.161207                 212809180
## 26:          1      relapse 1          2     BIC0  5.521461
## 27:          1      relapse 1          3     BIC0  5.252273
## 28:          1      relapse 1          4     BIC0  5.036953
## 29:          1      relapse 1          5     BIC0  5.214936
## 30:          1      relapse 1          6     BIC0  5.093750
## 31:          1      relapse 1          7     BIC0  4.859812                  60381530
## 32:          1      relapse 1          8     BIC0  4.836282
## 33:          1      relapse 1          9     BIC0  4.770685
## 34:          1      relapse 1         10     BIC0  4.919981
## 35:          1      relapse 1         11     BIC0  5.043425                  80058339
## 36:          1      relapse 1         12     BIC0  5.303305
## 37:          1      relapse 1         13     BIC0  4.499810
## 38:          1      relapse 1         14     BIC0  4.564348
## 39:          1      relapse 1         15     BIC0  4.369448
## 40:          1      relapse 1         16     BIC0  4.477337
## 41:          1      relapse 1         17     BIC0  5.141664
## 42:          1      relapse 1         18     BIC0  4.343805
## 43:          1      relapse 1         19     BIC0  4.007333
## 44:          1      relapse 1         20     BIC0  4.488636
## 45:          1      relapse 1         21     BIC0  3.713572
## 46:          1      relapse 1         22     BIC0  4.406719
## 47:          1      relapse 1          X     BIC0  4.553877
## 48:          1      relapse 1          Y     BIC0  3.044522  5918094, 9420652,19070635
##     profile.id status.profile chromosome pen.name pen.value                    changes
```

```
unsupervised.models <- rbind(unsupervised.models, unsupervised.models1)


pen.name <- "AIC0"
(unsupervised.models2 <- selected.profiles[, {
  fit.pelt <- changepoint::cpt.mean(
    logratio, penalty=pen.name, method="PELT")
  end <- fit.pelt@cpts
  before.change <- end[-length(end)]
  after.change <- before.change+1L
  data.table(
    pen.name,
    pen.value=fit.pelt@pen.value,
    changes=list(
    as.integer((position[before.change]+position[after.change])/2)
    ))
}, by=list(profile.id, status.profile, chromosome)])
```

```
##     profile.id status.profile chromosome pen.name pen.value
##  1:          4      relapse 4          1     AIC0         2
##  2:          4      relapse 4          2     AIC0         2
##  3:          4      relapse 4          3     AIC0         2
##  4:          4      relapse 4          4     AIC0         2
##  5:          4      relapse 4          5     AIC0         2
##  6:          4      relapse 4          6     AIC0         2
##  7:          4      relapse 4          7     AIC0         2
##  8:          4      relapse 4          8     AIC0         2
##  9:          4      relapse 4          9     AIC0         2
## 10:          4      relapse 4         10     AIC0         2
## 11:          4      relapse 4         11     AIC0         2
## 12:          4      relapse 4         12     AIC0         2
## 13:          4      relapse 4         13     AIC0         2
## 14:          4      relapse 4         14     AIC0         2
## 15:          4      relapse 4         15     AIC0         2
## 16:          4      relapse 4         16     AIC0         2
## 17:          4      relapse 4         17     AIC0         2
## 18:          4      relapse 4         18     AIC0         2
## 19:          4      relapse 4         19     AIC0         2
## 20:          4      relapse 4         20     AIC0         2
## 21:          4      relapse 4         21     AIC0         2
## 22:          4      relapse 4         22     AIC0         2
## 23:          4      relapse 4          X     AIC0         2
## 24:          4      relapse 4          Y     AIC0         2
## 25:          1      relapse 1          1     AIC0         2
## 26:          1      relapse 1          2     AIC0         2
## 27:          1      relapse 1          3     AIC0         2
## 28:          1      relapse 1          4     AIC0         2
## 29:          1      relapse 1          5     AIC0         2
## 30:          1      relapse 1          6     AIC0         2
## 31:          1      relapse 1          7     AIC0         2
## 32:          1      relapse 1          8     AIC0         2
## 33:          1      relapse 1          9     AIC0         2
## 34:          1      relapse 1         10     AIC0         2
## 35:          1      relapse 1         11     AIC0         2
## 36:          1      relapse 1         12     AIC0         2
## 37:          1      relapse 1         13     AIC0         2
## 38:          1      relapse 1         14     AIC0         2
## 39:          1      relapse 1         15     AIC0         2
## 40:          1      relapse 1         16     AIC0         2
## 41:          1      relapse 1         17     AIC0         2
## 42:          1      relapse 1         18     AIC0         2
## 43:          1      relapse 1         19     AIC0         2
## 44:          1      relapse 1         20     AIC0         2
## 45:          1      relapse 1         21     AIC0         2
## 46:          1      relapse 1         22     AIC0         2
## 47:          1      relapse 1          X     AIC0         2
## 48:          1      relapse 1          Y     AIC0         2
##     profile.id status.profile chromosome pen.name pen.value
##                                                        changes
##  1:                                                   59792500
##  2:                               45164625,114042111,163323003
##  3:                                                   69953902
##  4:                                                   22554400
##  5:
##  6:
##  7:
##  8:                                                  102756380
##  9:
## 10:
## 11:
## 12:
## 13:
## 14:                                                   76603452
## 15:
## 16:
## 17:                                                   41646489
## 18:
## 19:
## 20:
## 21:
## 22:
## 23:
## 24:
## 25:                                                  212809180
## 26:
## 27:
## 28:
## 29:
## 30:                                                   68487187
## 31:                                                   60381530
## 32:
## 33:                                                   28103371
## 34:
## 35:                                                   80058339
## 36:
## 37:
## 38:
```

```
## 39:
## 40:
## 41:
## 42:
## 43:
## 44:
## 45:
## 46:
## 47:
## 48:  2182240, 3006400, 5918094, 9420652,19070635,21291814,...
##                                                        changes
```

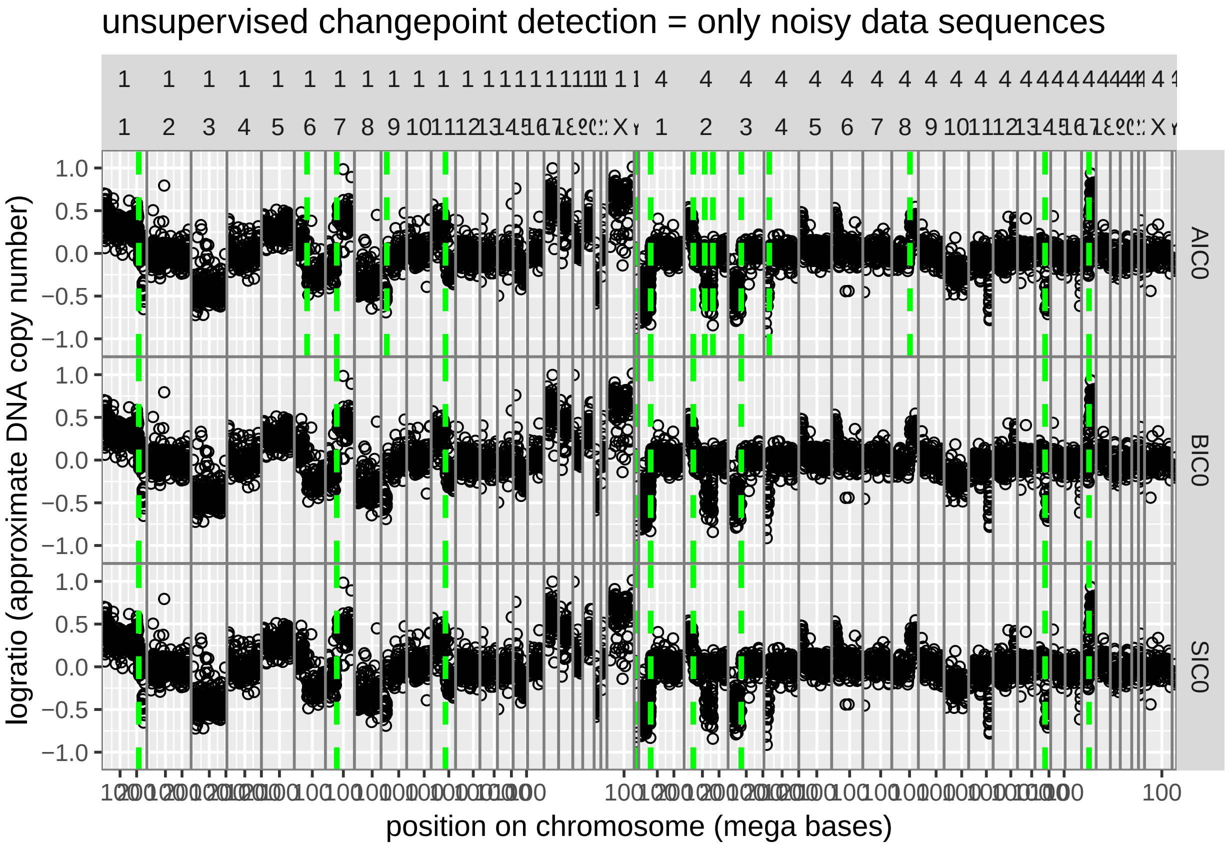
```
unsupervised.models <- rbind(unsupervised.models, unsupervised.models2)


(unsupervised.changes <- unsupervised.models[, data.table(
  change=changes[[1]]
), by=list(profile.id, status.profile, chromosome, pen.name)])
```

```
##     profile.id status.profile chromosome pen.name    change
##  1:          4      relapse 4          1     SIC0  59792500
##  2:          4      relapse 4          2     SIC0  45164625
##  3:          4      relapse 4          3     SIC0  69953902
##  4:          4      relapse 4         14     SIC0  76603452
##  5:          4      relapse 4         17     SIC0  41646489
##  6:          1      relapse 1          1     SIC0 212809180
##  7:          1      relapse 1          7     SIC0  60381530
##  8:          1      relapse 1         11     SIC0  80058339
##  9:          1      relapse 1          Y     SIC0   5918094
## 10:          1      relapse 1          Y     SIC0   9420652
## 11:          1      relapse 1          Y     SIC0  19070635
## 12:          4      relapse 4          1     BIC0  59792500
## 13:          4      relapse 4          2     BIC0  45164625
## 14:          4      relapse 4          3     BIC0  69953902
## 15:          4      relapse 4         14     BIC0  76603452
## 16:          4      relapse 4         17     BIC0  41646489
## 17:          1      relapse 1          1     BIC0 212809180
## 18:          1      relapse 1          7     BIC0  60381530
## 19:          1      relapse 1         11     BIC0  80058339
## 20:          1      relapse 1          Y     BIC0   5918094
## 21:          1      relapse 1          Y     BIC0   9420652
## 22:          1      relapse 1          Y     BIC0  19070635
## 23:          4      relapse 4          1     AIC0  59792500
## 24:          4      relapse 4          2     AIC0  45164625
## 25:          4      relapse 4          2     AIC0 114042111
## 26:          4      relapse 4          2     AIC0 163323003
## 27:          4      relapse 4          3     AIC0  69953902
## 28:          4      relapse 4          4     AIC0  22554400
## 29:          4      relapse 4          8     AIC0 102756380
## 30:          4      relapse 4         14     AIC0  76603452
## 31:          4      relapse 4         17     AIC0  41646489
## 32:          1      relapse 1          1     AIC0 212809180
## 33:          1      relapse 1          6     AIC0  68487187
## 34:          1      relapse 1          7     AIC0  60381530
## 35:          1      relapse 1          9     AIC0  28103371
## 36:          1      relapse 1         11     AIC0  80058339
## 37:          1      relapse 1          Y     AIC0   2182240
## 38:          1      relapse 1          Y     AIC0   3006400
## 39:          1      relapse 1          Y     AIC0   5918094
## 40:          1      relapse 1          Y     AIC0   9420652
## 41:          1      relapse 1          Y     AIC0  19070635
## 42:          1      relapse 1          Y     AIC0  21291814
## 43:          1      relapse 1          Y     AIC0  23715526
##     profile.id status.profile chromosome pen.name    change
```

Plotting the data along with optimal segment means (green line segments) we get:

```
gg.unsupervised+
  facet_grid(pen.name ~ profile.id + chromosome, scales="free", space="free_x")+
  theme(legend.box="horizontal")+
  geom_vline(aes(
    xintercept=change/1e6),
    color="green",
    size=1,
    linetype="dashed",
    data=unsupervised.changes)
```

```
## Warning: Removed 33 rows containing missing values (geom_point).
```


unsupervised changepoint detection = only noisy data sequences
logratio (approximate DNA copy number)
position on chromosome (mega bases)
AIC0
BIC0
SIC0
X