HiCseg: Package for two-dimensional segmentation of HiC data

C. Lévy-Leduc, M. Delattre, T. Mary-Huard, S. Robin

June 10, 2014

This vignette explains how to use the package HiCseg which is dedicated to the two-dimensional segmentation of HiC data. More precisely, the goal of this package is to provide the boundaries of cis-interacting regions in such data by using a dynamic programming algorithm. For further details on the statistical model and on the implementation we refer the reader to [1].

After having installed the package in R, the package has to be loaded by using the following instruction:

```
> library(HiCseg)
```

The package HiCseg contains a function called `HiCseg_linC_R` which makes the link between the C language and R. The usage of this function is the following:

```
result = HiCseg_linkC_R(size_mat, nb_change_max, distrib, mat_data, model)
```

where the arguments are:

- `size_mat`: Size of the data matrix
- `nb_change_max`: Maximal number of change-points
- `distrib`: Distribution of the data: "B" is for Negative Binomial distribution, "P" is for the Poisson distribution and "G" is for the Gaussian distribution.
- `mat_data`: Matrix of data
- `model`: Type of model: "D" for block-diagonal and "Dplus" for the extended block-diagonal model.

and where the output `result` is a list of three attributes:

- `t_hat`: Contains the estimated change-points
- `J`: Values of the log-likelihood for different number of change-points up to some constants

- `t_est_mat`: It gives the matrix of the estimated change-points for different number of change-points: in the first line when there is no change-point, in the second line when there is one change-point, in the third line when there are two change-points....

More precisely, `mat_data` has to be a matrix which can be loaded in the R environment thanks to a `.Rdata` object. For example, we can load the matrix of observations provided with the package which is a toy example:

```
> data(matrix)
> dim=dim(matrix)
> n=dim[1]
> image(1:n,1:n,matrix,xlab="",ylab="")
```

It is a symmetric matrix of size 200×200 with block boundaries on the x-axis at: 40, 80, 120, 160 and 201 (by convention). HiCseg allows you to automatically recovers these boundaries thanks to the `HiCseg_linkC_R` function.

```
> result = HiCseg_linkC_R(200, 10, "G", matrix, "D")
```

In this line of command:

- The first argument is 200 since this is the size of the data matrix.
- The second argument is 10 which means that the maximal number of change-points the algorithm is going to look for is 10.
- The third argument is "G" because we decided to model the observations as reasliza-tions of Gaussian random variables. It could be replaced by "P" or "B" if we have to deal with integer values. Note that "P" means Poisson distribution and "B" Negative Binomial distribution.
- The fourth argument is `matrix` here because this is the matrix of observations in which we want to find blocks.
- The last argument is "D" because we decided to fit a block-diagonal model which assumes that the observations are realizations of random variables having their mean which changes along the diagonal blocks but which is constant outside the diagonal blocks. Here the argument "Dplus" could also be used if we want to make the assumption that the mean is not constant anymore outside the diagonal blocks but constant in bands outside the diagonal blocks. For further details on the differences between these two modelings we refer the user to [1].

```
> result$t_est_mat
```

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,]  200    0    0    0    0    0    0    0    0     0
 [2,]   39  200    0    0    0    0    0    0    0     0
 [3,]   39  159  200    0    0    0    0    0    0     0
 [4,]   39  119  159  200    0    0    0    0    0     0
 [5,]   39   79  119  159  200    0    0    0    0     0
 [6,]   39   41   79  119  159  200    0    0    0     0
 [7,]   39   75   77   79  119  159  200    0    0     0
 [8,]   39   73   75   77   79  119  159  200    0     0
 [9,]   39   71   73   75   77   79  119  159  200     0
[10,]   39   41   71   73   75   77   79  119  159   200
```

```
> result$J

 [1] -38499.61 -32349.69 -27856.21 -23552.36 -20031.73
 [6] -20339.39 -20626.68 -20862.68 -21125.81 -21396.09
```

```
> result$t_hat

 [1]  39  79 119 159 200   0   0   0   0   0
```

```
> plot(result$J,type="o",xlab="K",ylab="Log-likelihood")
```

`result$t_est_mat` provides the matrix of the estimated change-points for different number of change-points: in the first line when there is no change-point (1 segment which ends at $n = 200$), in the second line when there is one change-point (here: 39) or two segments, in the third line when there are two change-points (39 and 159) or three segments...

`result$J` provides the values of the log-likelihood (up to some constants) for $K = 1, \ldots, 10$.

`result$t_hat` gives the list of the $\hat{K}$ change-points where $\hat{K}$ is the value of K at which J is maximized (here $\hat{K} = 5$) that is gives the values of change-points found at line 5 of the matrix `result$t_est_mat`. The values that we get correspond to the end of each diagonal blocks.

In order to plot the blocks thus obtained one can use the following lines

```
> image(1:n,1:n,matrix,xlab="",ylab="")
> t_hat=c(1,result$t_hat[result$t_hat!=0]+1)
> for (i in 1:(length(t_hat)-1))
+ {
+   lines(c(t_hat[i],t_hat[i]),c(t_hat[i],(t_hat[(i+1)]-1)))
+   lines(c(t_hat[(i+1)]-1,t_hat[(i+1)]-1),c(t_hat[i],t_hat[(i+1)]-1))
+   lines(c(t_hat[i],t_hat[(i+1)]-1),c(t_hat[i],t_hat[i]))
+   lines(c(t_hat[i],t_hat[(i+1)]-1),c(t_hat[(i+1)]-1,t_hat[(i+1)]-1))
+ }
```

Hereafter, we also provide some information about the R session

```
> sessionInfo()

R version 2.14.1 (2011-12-22)
Platform: x86_64-pc-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=fr_FR.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=fr_FR.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=fr_FR.UTF-8    LC_MESSAGES=fr_FR.UTF-8
 [7] LC_PAPER=C                 LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=fr_FR.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets
[6] methods   base

other attached packages:
[1] HiCseg_1.1

loaded via a namespace (and not attached):
[1] tools_2.14.1
```

References

[1] C.˜Lévy-Leduc, M.˜Delattre, T.˜Mary-Huard, and S.˜Robin. Two-dimensional segmentation for analyzing hic data. Technical report, AgroParisTech/INRA MIA 518, 2014. Submitted to ECCB'14 and available from http://www.agroparistech.fr/mia/levyleduc/.