PSSeg: Parent-Specifc copy number segmentation

M. Pierre-Jean, G. Rigaill, P. Neuvial

September 9, 2014

Abstract

This vignette describes how to use the `jointseg` package to partition bivariate DNA copy number signals from SNP array data into segments of constant parent-specific copy number. We demonstrate the use of the `PSSeg` function of this package for applying two different strategies. Both strategies consist in first identifying a list of candidate change points through a fast (greedy) segmentation method, and then to prune this list is using dynamic programming [1]. The first segmentation method is Recursive Binary Segmentation (RBS, [2]), and the second one is a group fused LARS approach (GFLARS, [7]), for which we ported the original `Matlab` implementation into `R`. We refer to [6] for a more comprehensive performance assessment of such methods and other segmentation methods.

keywords: segmentation, change point model, binary segmentation, dynamic programming, DNA copy number, parent-specific copy number.

Contents

Please see Appendix C for citing `jointseg`.

```
library("jointseg")
```

1 Preparing data to be segmented

`PSSeg` requires normalized copy number signals, in the form of total copy number estimates and allele B fractions for tumor, the (germline) genotype of SNP. Loci are assumed to come from a single chromosome and to be ordered by genomic position.

For illustration, we show of such a data set may be created from real data. We use data from a public SNP array data set, which is distributed in the `acnr` package (on which the `jointseg` package depends).

```
data <- loadCnRegionData(dataSet="GSE29172", tumorFraction=1)
str(data)

## 'data.frame': 192667 obs. of  4 variables:
##  $ c       : num  0.909 0.859 1.304 0.647 0.947 ...
##  $ b       : num  NaN NaN NaN NaN NaN NaN NaN -0.035 NaN NaN ...
##  $ genotype: num  NA NA NA NA NA NA NA 0 NA NA ...
##  $ region  : chr  "(0,1)" "(0,1)" "(0,1)" "(0,1)" ...
```

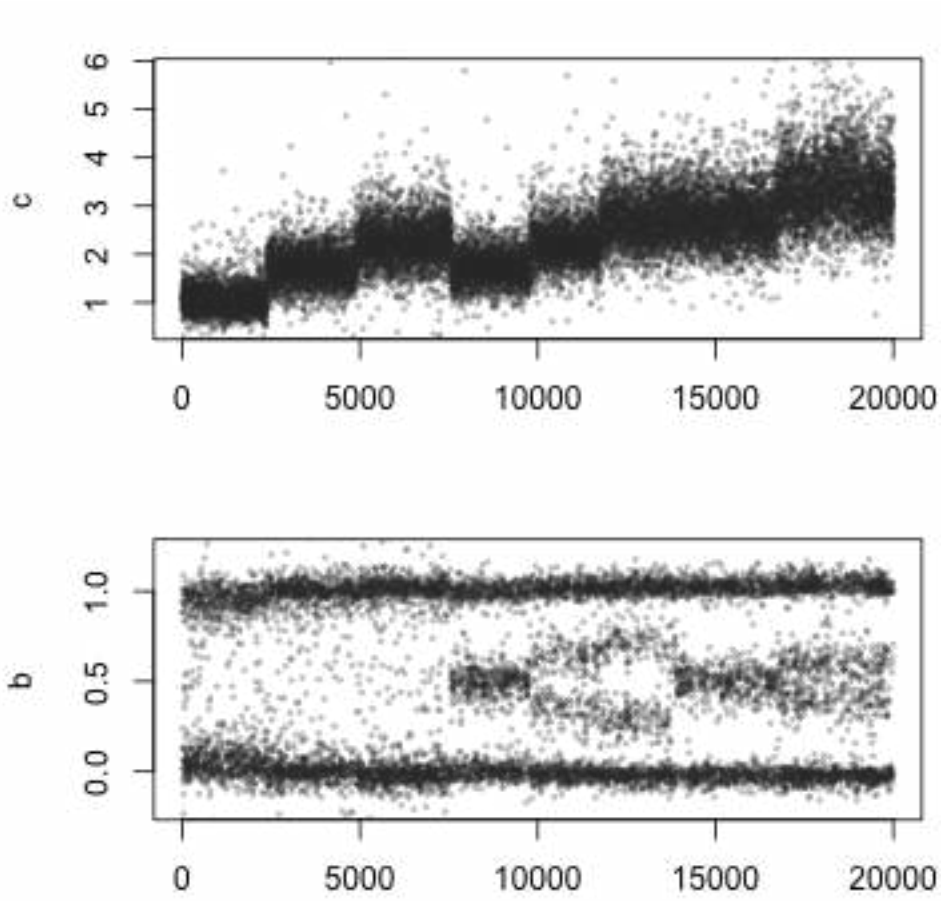
This data set consists of copy number signals from 8 types of genomic regions:

```
table(data[["region"]])

##
## (0,1) (0,2) (0,3) (1,1) (1,2) (1,3) (2,2) (2,3)
## 22615 24135 25405 21539 19048 20903 27924 31098
```

These regions are coded as (C_1, C_2), where C_1 denotes the minor copy number and C_2 denotes the major copy number, i.e. the smallest and the largest of the two parental copy numbers (see e.g. [4] for more detailed definitions). For example, $(1, 1)$ corresponds to a normal state, $(0, 1)$ to an hemizygous deletion, $(1, 2)$ to a single copy gain and $(0, 2)$ to a copy-neutral LOH (loss of heterowygosity).

```
idxs <- sort(sample(1:nrow(data), 2e4))
plotSeg(data[idxs, ])
```



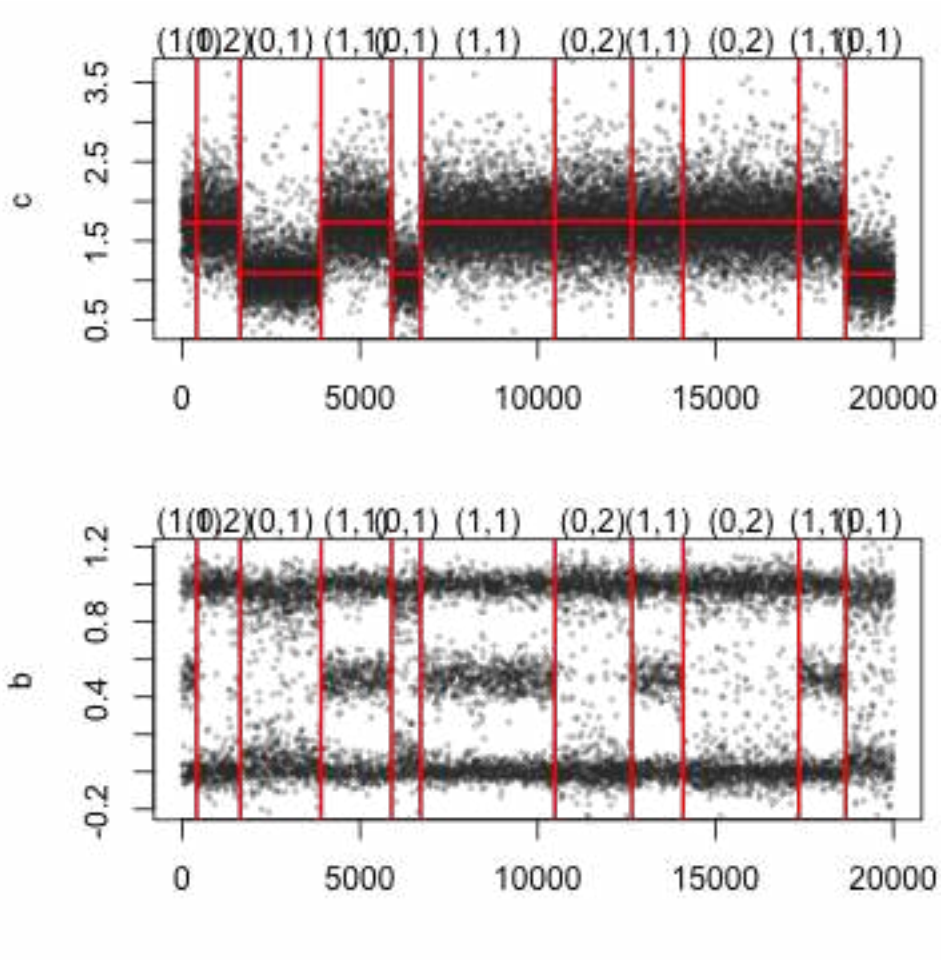
These real data can then be used to create a realistic DNA copy number profile of user-defined length, and harboring a user-defined number of breakpoints. This is done using the `getCopyNumberDataByResampling` function. Breakpoint positions are drawn uniformly) among all possible loci. Between two breakpoints, the copy number state corresponds to one of the types of regions in `data`, and each data point is drawn with replacement from the corresponding true copy number signal from the region. More options are available from the documentation of `getCopyNumberDataByResampling`.

```
K <- 10
bkp <- c(408,1632,3905, 5890,6709, 10481, 12647,14089,17345,18657)
len <- 2e4
sim <- getCopyNumberDataByResampling(len, bkp=bkp, minLength=500, regData=data)
datS <- sim$profile
str(datS)

## 'data.frame': 20000 obs. of  4 variables:
##  $ c       : num  1.33 1.99 2 1.84 1.95 ...
##  $ b       : num  0.958 0.869 0.966 0.442 NaN NaN NaN NaN 0.562 0.964 ...
##  $ genotype: num  1 1 1 0.5 NA NA NA NA 0.5 1 ...
##  $ region  : chr  "(1,1)" "(1,1)" "(1,1)" "(1,1)" ...
```

The resulting copy-number profile is plotted below.

```
plotSeg(datS, sim$bkp)
```



2 Preprocessing

We advise the following (typical) preprocessing before segmentation:

1. log-transform total copy numbers in order to stabilize their variance; this step improve segmentation results for all methods.

```
datS$c <- log2(datS$c)-1
```

2. smooth single point outliers as suggested by [5]. This step is controlled by the `dropOutliers` option in the `PSSeg` function, which internally calls the `smooth.CNA` function of the `DNAcopy package`. The default value for this option is `TRUE`.

3. convert allelic ratios to (unimodal) decrease in heterozygosity (d), as initially suggested by [**?**]. This step is performed internally in the `PSSeg` function.

3 PSSeg segmentation using RBS

We can now use the `PSSeg` function to segment signals. The method consists in three steps:

1. run a fast (yet approximate) segmentation on these signals in order to obtain a set of (at most hundreds of) candidate change points. This is done using Recursive Binary Segmentation (RBS) [2];

2. prune the obtained set of change points using dynamic programming [1]
3. select the best number of change points using a model selection criterion proposed by [3]

3.1 Initial segmentation and pruning

```
resRBS <- PSSeg(data=datS, K=2*K, method="RBS", stat=c("c", "d"), profile=TRUE)
```

Note that this is fast:

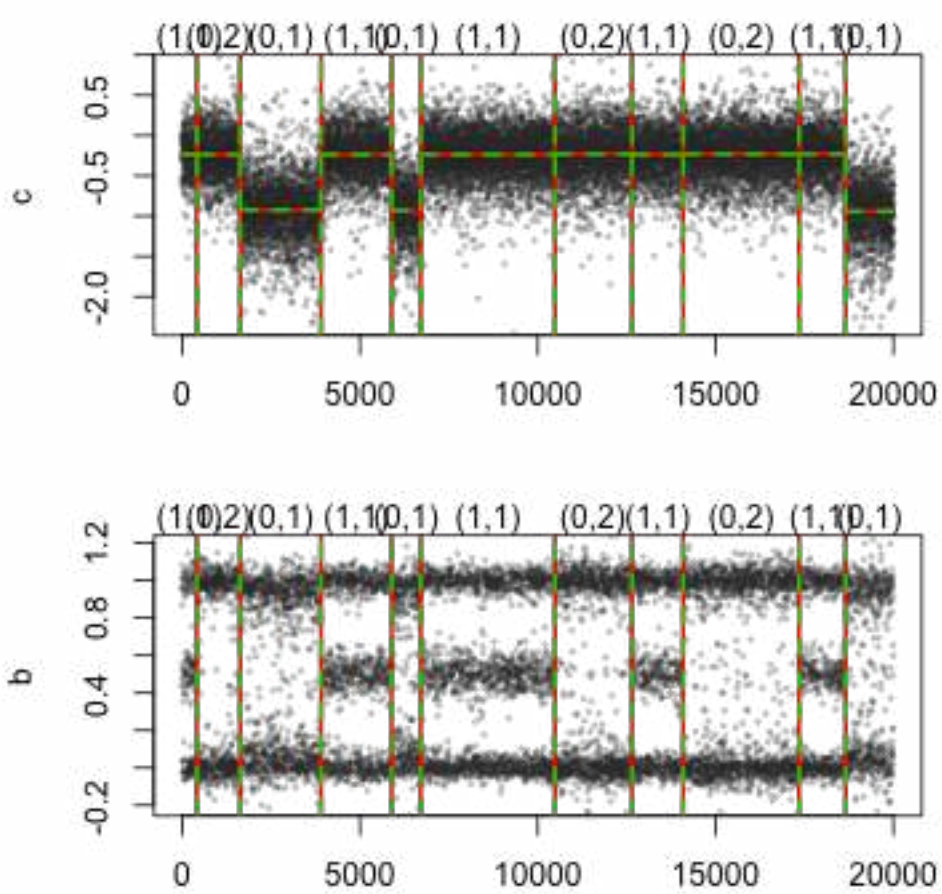
```
resRBS$prof[, "time"]

## segmentation        dpseg
##         0.18         0.00
```

3.2 Plot segmented profile

To plot the PSSeg segmentation results together with the true breakpoints, do :

```
plotSeg(datS, list(true=sim$bkp, est=resRBS$bestBkp))
```



3.3 Results evaluation

The `PSSeg` function returns the original segmentation (by `RBS`), the result of the pruning step, and the best model (among those selected by dynamic programming) according to the criterion proposed by [3].

The quality of the best segmentation can be assessed as follows. The number of true positives (TP) is the number of true change points for which there exists a candidate change point closer than a given tolerance `tol`. The number of false positives is defined as the number of true negatives (all those which are not change points) for which the candidate change points are out of tolerance area and those in tolerance area where there already exists a candidate change point. By construction, $TP \in \{0, 1, \cdots, K\}$ where K is the number of true change points.

```
print(getTpFp(resRBS$bestBkp, sim$bkp, tol=5))

## TP FP
## 10  0
```

Obviously, this performance measure depends on the chosen tolerance:

```
perf <- sapply(0:10, FUN=function(tol) {
  getTpFp(resRBS$bestBkp, sim$bkp, tol=tol,relax = -1)
})
print(perf)

##    [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
## TP    5    7    7    7    8   10   10   10   10    10    10
## FP    5    3    3    3    2    0    0    0    0     0     0
```

4 PSSeg segmentation using GFLARS

We can now use the `PSSeg` function to segment signals with GFLARS method only on heterozygous SNP. The method consists in three steps:

1. run a fast (yet approximate) segmentation on these signals in order to obtain a set of (at most hundreds of) candidate change points. This is done using Group Fused Lars [7];
2. prune the obtained set of change points usign dynamic programming [1]
3. select the best number of change points using a model selection criterion proposed by [3]

4.1 Initial segmentation and pruning

```
resGFL <- PSSeg(data=datS, K=5*K, method="GFLars", stat=c("c", "d"), profile=TRUE)

## Warning:  Missing values detected.  Smoothing will be performed.
```

Note that this is fast due to the low number in the data.

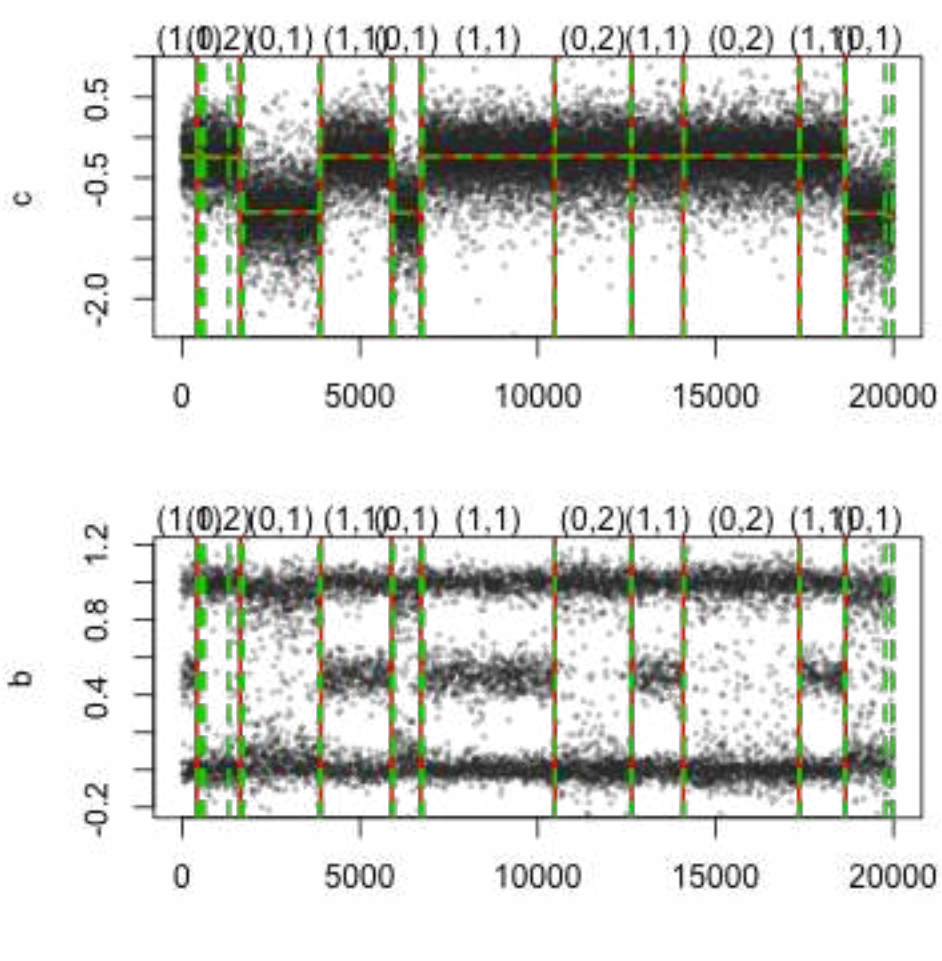
```
resGFL$prof[, "time"]

## segmentation        dpseg
##          0.3          0.0
```

4.2 Plot segmented profile

To plot the PSSeg segmentation results together with the true breakpoints, do :

```
plotSeg(datS, list(true=sim$bkp, est=resGFL$bestBkp))
```



4.3 Results evaluation

The `PSSeg` function returns the original segmentation (by `GFLARS`), the result of the pruning step, and the best model (among those selected by dynamic programming) according to the criterion proposed by [3].

```
print(getTpFp(resGFL$bestBkp, sim$bkp, tol=15))

## TP FP
## 10 14
```

Obviously, this performance measure depends on the chosen tolerance:

```
perf <- sapply(0:20, FUN=function(tol) {
  getTpFp(resGFL$bestBkp, sim$bkp, tol=tol)
})
print(perf)

##     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16]
```

```
## TP    1    5    6    8    8    9    9    9    9     9    10    10    10    10    10    10
## FP   23   19   18   16   16   15   15   15   15    15    14    14    14    14    14    14
##    [,17] [,18] [,19] [,20] [,21]
## TP    10    10    10    10    10
## FP    14    14    14    14    14
```

A Acknowledgements

This package has been documented using the inlinedocs package.

B Session information

```
sessionInfo()

## R version 3.1.1 (2014-07-10)
## Platform: x86_64-apple-darwin10.8.0 (64-bit)
##
## locale:
## [1] fr_FR.UTF-8/fr_FR.UTF-8/fr_FR.UTF-8/C/fr_FR.UTF-8/fr_FR.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  base
##
## other attached packages:
## [1] jointseg_0.6.1    acnr_0.2.0        R.utils_1.33.0    R.oo_1.18.2
## [5] R.methodsS3_1.6.2 knitr_1.6.10
##
## loaded via a namespace (and not attached):
## [1] DNAcopy_1.38.1     evaluate_0.5.5     formatR_1.0        highr_0.3.1
## [5] matrixStats_0.10.1 methods_3.1.1      stringr_0.6.2      tools_3.1.1
```

C Citing `jointseg`

```
citation("jointseg")

##
## To cite package 'jointseg' in publications use:
##
##   Morgane Pierre-Jean and Guillem Rigaill and Pierre Neuvial (2014). jointseg:
##   Joint segmentation of multivariate (copy number) signals. R package version
##   0.6.1.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{pierre-jean14jointseg,
##     title = {jointseg: Joint segmentation of multivariate (copy number) signals},
##     author = {Morgane Pierre-Jean and Guillem Rigaill and Pierre Neuvial},
##     year = {2014},
##     note = {R package version 0.6.1},
```

```
##   }
##
##   Morgane Pierre-Jean and Guillem Rigaill and Pierre Neuvial (2014). A
##   performance evaluation framework of DNA copy number analysis methods in cancer
##   studies; application to SNP array data segmentation methods. arXiv preprint
##   arXiv:1402.7203.
##
## A BibTeX entry for LaTeX users is
##
##   @TechReport{pierre-jean14performance,
##     title = {A performance evaluation framework of {DNA} copy number analysis methods in cancer studi
##     author = {Morgane Pierre-Jean and Guillem Rigaill and Pierre Neuvial},
##     year = {2014},
##     institution = {Laboratoire de Math\'ematiques et Mod\'elisation d'\'Evry, Universit\'e d'\'Evry V
##     note = {arXiv preprint http://arxiv.org/abs/1402.7203},
##   }
```

References

[1] Richard Bellman. On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*, 4(6):284, 1961.

[2] S. Gey and E. Lebarbier. Using cart to detect multiple change points in the mean for large sample. Technical report, Statistics for Systems Biology research group, 2008.

[3] E. Lebarbier. Detecting multiple change-points in the mean of gaussian process by model selection. *Signal processing*, 85(4):717–736, 2005.

[4] Pierre Neuvial, Henrik Bengtsson, and Terence P Speed. Statistical analysis of single nucleotide polymorphism microarrays in cancer studies. In *Handbook of Statistical Bioinformatics*, Springer Handbooks of Computational Statistics. Springer, 1st edition, March 2011.

[5] A B Olshen, E S Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, 2004.

[6] Morgane Pierre-Jean, Guillem J Rigaill, and Pierre Neuvial. Performance evaluation of DNA copy number segmentation methods. *Briefings in Bioinformatics*, to appear.

[7] J.-P. Vert and K. Bleakley. Fast detection of multiple change-points shared by many signals using group LARS. *Advances in Neural Information Processing Systems*, 23:2343–2351, 2010.