

How to use the findSegments function to fit a piecewise constant curve

Wolfgang Huber

May 22, 2005

Contents

1	Introduction	1
2	How to choose the number of segments?	4
3	Some more testing	4

1 Introduction

The problem of segmenting a series of numbers into piecewise constant segments occurs in multiple application areas. Two examples are

- arrayCGH data, where the segments correspond to regions of copy number gain, loss, or no change.
- tiling microarray data for transcriptomics, where the segments correspond to transcripts. Here we assume that the probe effects (which lead to different fluorescence intensities even for the same mRNA abundance) have been normalized away, so that all probes for one transcript have the same fluorescence (in expectation).

To demonstrate and verify the correctness of the algorithm, let's generate simulated data:

```
> genData = function(lenx, nrCP, stddev = 0.1) {  
+   x = numeric(lenx)  
+   cp = c(1, sort(sample(1:floor(lenx/15), nrCP - 1) * 15),  
+         lenx + 1)
```

```

+     s = 0
+     for (j in 2:length(cp)) {
+         sel = cp[j - 1]:(cp[j] - 1)
+         s = (0.5 + runif(1)) * sign(rnorm(1)) + s
+         x[sel] <- rnorm(length(sel), mean = s, sd = stddev)
+     }
+     return(list(x = x, cp = cp[-1]))
+ }

> lenx = 1000
> nrcp = 10
> gd = genData(lenx, nrcp)
> plot(gd$x, pch = ".")
> abline(v = gd$cp, col = "red")

```

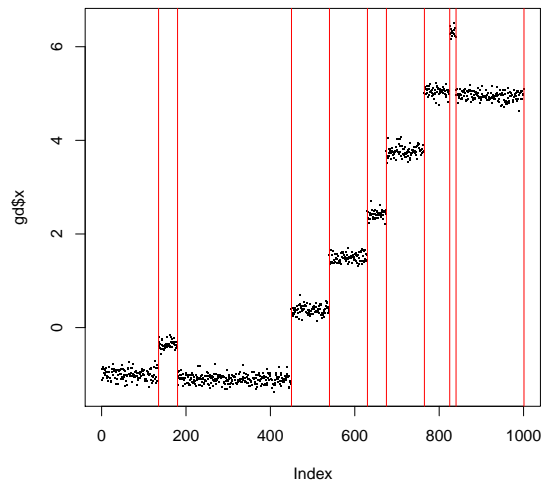


Figure 1: A simulated data example with 10 change points, shown with red vertical lines

The result is shown in Figure 1. We can use the function *findSegments* to reconstruct the changepoints from the x data alone.

```

> library(tilingArray)

```

```

> library(tilingArray)
> maxk = 500
> maxcp = 12
> seg = findSegments(gd$x, maxk = maxk, maxcp = maxcp)
> seg

$J
[1]      -Inf -0.414019  1.202624  1.792334  2.500388  2.879470  3.019222
[8]  3.480042  3.751572  4.642962  4.650403  4.662729

$th
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
[1,] 1001    0    0    0    0    0    0    0    0    0    0    0
[2,]  501 1001    0    0    0    0    0    0    0    0    0    0
[3,]  450  675 1001    0    0    0    0    0    0    0    0    0
[4,]  450  630  765 1001    0    0    0    0    0    0    0    0
[5,]  450  540  675  765 1001    0    0    0    0    0    0    0
[6,]  450  540  630  675  765 1001    0    0    0    0    0    0
[7,]  181  450  540  630  675  765 1001    0    0    0    0    0
[8,]  450  540  630  675  765  825  840 1001    0    0    0    0
[9,]  181  450  540  630  675  765  825  840 1001    0    0    0
[10,]  135  180  450  540  630  675  765  825  840 1001    0    0
[11,]  135  180  444  450  540  630  675  765  825  840 1001    0
[12,]  135  180  450  540  630  675  699  710  765  825  840 1001

attr(,"class")
[1] "segmentation" "list"

> gd$cp

[1]  135  180  450  540  630  675  765  825  840 1001

```

We see that the 10-th row of the matrix `seg$th` exactly reconstructs the change points `gd$cp` that were used in the simulation.

The parameters `maxcp` and `maxk` are the maximum number of segments and the maximum length per segment. The algorithm finds for each value of k from 1 to `maxcp` the best segmentation under the restriction that no individual segment be longer than `maxk`. In the original paper of Picard et al. [1] and in their software, `maxk` is implicitly set to the number of data points `length(x)`. I have introduced this parameter to reduce the algorithm's complexity. The complexity of Picard's software is `length(x)*length(x)`

in memory and `length(x)*length(x)*maxcp` in time, the complexity of the *findSegments* function is `length(x)*maxk` in memory and `length(x)*maxk*maxcp` in time. As I am envisaging applications with `length(x) $\approx 10^5$` and `maxk ≈ 250` , the difference can be substantial.

2 How to choose the number of segments?

Need to assess goodness of fit (contained in `gd$J`) together with model complexity (i.e. the number of change points). Details will follow ...

3 Some more testing

Here is a little for-loop that generates data using random parameters and checks whether *findSegments* can reconstruct them. The purpose of this is for checking the validity of the code.

```
> set.seed(4711)
> for (i in 1:20) {
+   gd = genData(lenx, nrcp)
+   seg = findSegments(gd$x, maxk = maxk, maxcp = maxcp)
+   stopifnot(seg$th[nrcp, 1:nrcp] == gd$cp)
+ }
```

References

- [1] A statistical approach for CGH microarray data analysis. Franck Picard, Stephane Robin, Marc Lavielle, Christian Vaisse, Gilles Celeux, Jean-Jacques Daudin. Rapport de recherche No. 5139, Mars 2004, *Institut National de Recherche en Informatique et en Automatique (INRIA)*, ISSN 0249-6399. http://www.inapg.fr/ens_rech/mathinfo/recherche/mathematique/outil.html