Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

Comparing binsegRcpp with other implementations of binary segmentation

Toby Dylan Hocking
Northern Arizona University

Abstract

Binary segmentation is a classic algorithm for detecting changepoints in sequential data. In theory it should be extremely fast for N data and K segments, $O(NK)$ in the worst case, and $O(N \log K)$ in the best case. In practice existing implementations such as ruptures (Python module) and changepoint (R package) are much slower, and in fact sometimes do not return a correct result. We present the R package **binsegRcpp**, which provides an efficient C++ implementation of binary segmentation, and always returns correct results. We discuss some advanced C++ coding techniques which were used to avoid repetition, for efficiency, and for readability. We additionally include detailed theoretical and empirical comparisons with other implementations of binary segmentation in R packages **fpop** and **wbs**.

Keywords: JSS, style guide, comma-separated, not capitalized, R.

1. Introduction: Count data regression in R

The introduction is in principle "as usual". However, it should usually embed both the implemented *methods* and the *software* into the respective relevant literature. For the latter both competing and complementary software should be discussed (within the same software environment and beyond), bringing out relative (dis)advantages. All software mentioned should be properly `\cite{}`d. (See also Appendix B for more details on BibTeX.)

For writing about software JSS requires authors to use the markup `\proglang{}` (programming languages and large programmable systems), `\pkg{}` (software packages), `\code{}` (functions, commands, arguments, etc.). If there is such markup in (sub)section titles (as above), a plain text version has to be provided in the LaTeX command as well. Below we also illustrate how abbrevations should be introduced and citation commands can be employed. See the LaTeX code for more details.

TODO (Standard Template Library containers, virtual classes, static variables, function pointers, templates, macros).

TODO `fpop::multiBinSeg` (Maidstone, Hocking, Rigaill, and Fearnhead 2017).

TODO `wbs::sbs` (Baranowski and Fryzlewicz 2019).

TODO `changepoint::cpt.mean(method="BinSeg")` (Killick and Eckley 2014; Killick, Haynes, and Eckley 2022).

TODO `ruptures.Binseg` (Truong, Oudre, and Vayatis 2020).

2. Models and software

The basic Poisson regression model for count data is a special case of the GLM framework McCullagh and Nelder (1989). It describes the dependence of a count response variable y_i ($i = 1, \ldots, n$) by assuming a Poisson distribution $y_i \sim \text{Pois}(\mu_i)$. The dependence of the conditional mean $\mathsf{E}[y_i \,|\, x_i] = \mu_i$ on the regressors x_i is then specified via a log link and a linear predictor

$$\log(\mu_i) \quad = \quad x_i^\top \beta, \tag{1}$$

where the regression coefficients β are estimated by maximum likelihood (ML) using the iterative weighted least squares (IWLS) algorithm.

> Note that around the `{equation}` above there should be no spaces (avoided in the LaTeX code by `%` lines) so that "normal" spacing is used and not a new paragraph started.

R provides a very flexible implementation of the general GLM framework in the function `glm()` (Chambers and Hastie 1992) in the **stats** package. Its most important arguments are

```
glm(formula, data, subset, na.action, weights, offset,
  family = gaussian, start = NULL, control = glm.control(...),
  model = TRUE, y = TRUE, x = FALSE, ...)
```

where `formula` plus `data` is the now standard way of specifying regression relationships in R/S introduced in Chambers and Hastie (1992). The remaining arguments in the first line (`subset`, `na.action`, `weights`, and `offset`) are also standard for setting up formula-based regression models in R/S. The arguments in the second line control aspects specific to GLMs while the arguments in the last line specify which components are returned in the fitted model object (of class '`glm`' which inherits from '`lm`'). For further arguments to `glm()` (including alternative specifications of starting values) see `?glm`. For estimating a Poisson model `family = poisson` has to be specified.

> As the synopsis above is a code listing that is not meant to be executed, one can use either the dedicated `{Code}` environment or a simple `{verbatim}` environment for this. Again, spaces before and after should be avoided.
>
> Finally, there might be a reference to a `{table}` such as Table 1. Usually, these are placed at the top of the page (`[t!]`), centered (`\centering`), with a caption below the table, column headers and captions in sentence style, and if possible avoiding vertical lines.

package	binsegRcpp	changepoint	wbs	fpop	ruptures
function	binseg	cpt.mean	sbs	multiBinSeg	Binseg
version	2022.3.29	2.2.3	1.4	2019.8.26	1.1.6
weights	yes	no	no	no	no
max segs	yes	yes	no	yes	yes
dim	one	one	one	multi	multi
correct	yes	no	yes	yes	no
losses	C++	C	L2	L2	Python
storage	STL multiset	arrays	recursion	heap	LRU cache
space	$O(S)$	$O(S^2)$	$O(S)$	$O(S)$	$O(S)$
cumsum	yes	yes	yes	yes	no
best	$O(N \log N)$	$O(N^3)$	$O(N \log N)$	$O(N \log N)$	$> O(N \log N)$
worst	$O(N^2)$	$O(N^3)$	$O(N^2)$	$O(N^2)$	$O(N^2)$
CV	yes	no	no	no	no
params	yes	largest	no	no	no

Table 1: TODO

case/splits:	one		best/equal		worst/unequal	
operation:	insert	argmin	insert	argmin	insert	argmin
list	$O(1)$	$O(n)$	$O(S \log S)$	$O(S)$	$O(S)$	$O(S)$
heap/multiset	$O(\log n)$	$O(1)$	$O(S)$	$O(S^2)$	$O(S)$	$O(S)$
Find new split			$O(N \log S)$		$O(NS)$	

Table 2: container properties

Loss computation. We run each algorithm on `N.data` points up to `max.segments = max.changes+1`. `binsegRcpp::binseg` result is a list which contains a data table with `max.segments` rows and column `loss` that is the square loss. `changepoint::cpt.mean` result is a list of class `cpt.range` with method `logLik` which returns the square loss of one of the models. `wbs::sbs` result is a list which contains a data frame with `N.data-1` rows and `CUSUM` and `min.th` columns TODO.`fpop::multiBinSeg` result is a list with element `J.est`, which is a vector of `max.changes` square loss decrease values. `ruptures.Binseg.predict` result is a vector of segment ends for one model size, which can be passed to `sum_of_costs` method to compute the square loss.

Three packages have implemented the normal change in mean and variance model. `binsegRcpp::binseg` loss values are the normal negative log likelihood (NLL). The `changepoint::logLik` function returns two times the NLL. The `ruptures.Binseg` loss is related via this equation,

$$\text{NLL} = (\text{rupturesLoss} + N[1 + \log(2\pi)])/2 \tag{2}$$

3. Illustrations

For a simple illustration of basic Poisson and NB count regression the `quine` data from the **MASS** package is used. This provides the number of `Days` that children were absent from school in Australia in a particular year, along with several covariates that can be employed

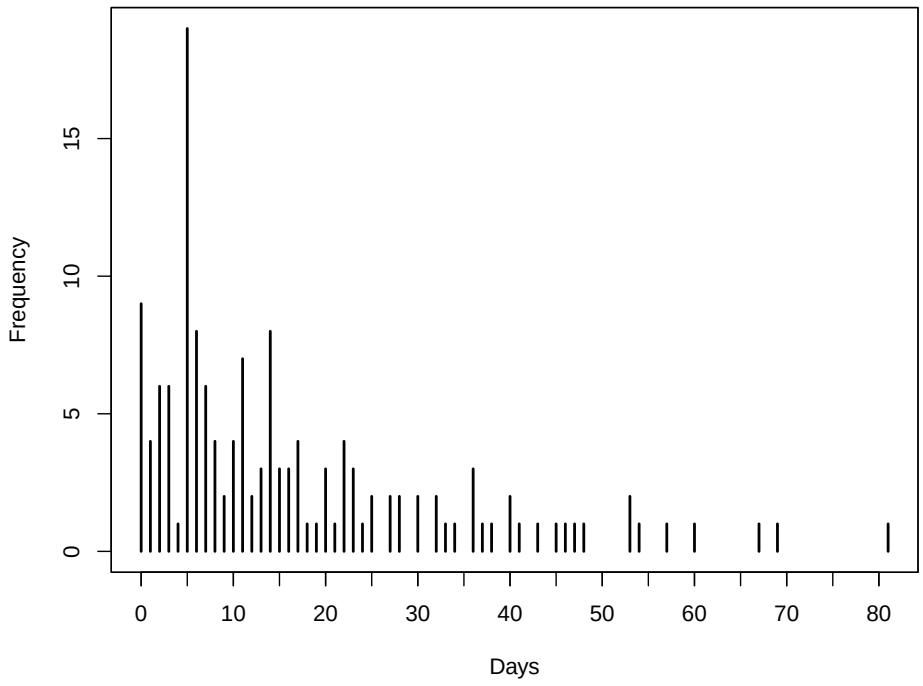

Figure 1: Frequency distribution for number of days absent from school.

as regressors. The data can be loaded by

```
R> data("quine", package = "MASS")
```

and a basic frequency distribution of the response variable is displayed in Figure 1.

> For code input and output, the style files provide dedicated environments. Either the "agnostic" `{CodeInput}` and `{CodeOutput}` can be used or, equivalently, the environments `{Sinput}` and `{Soutput}` as produced by `Sweave()` or **knitr** when using the `render_sweave()` hook. Please make sure that all code is properly spaced, e.g., using `y = a + b * x` and *not* `y=a+b*x`. Moreover, code input should use "the usual" command prompt in the respective software system. For R code, the prompt `"R> "` should be used with `"+ "` as the continuation prompt. Generally, comments within the code chunks should be avoided – and made in the regular LaTeX text instead. Finally, empty lines before and after code input/output should be avoided (see above).

As a first model for the `quine` data, we fit the basic Poisson regression model. (Note that JSS prefers when the second line of code is indented by two spaces.)

```
R> m_pois <- glm(Days ~ (Eth + Sex + Age + Lrn)^2, data = quine,
+    family = poisson)
```

To account for potential overdispersion we also consider a negative binomial GLM.

```
R> library("MASS")
R> m_nbin <- glm.nb(Days ~ (Eth + Sex + Age + Lrn)^2, data = quine)
```

In a comparison with the BIC the latter model is clearly preferred.

```
R> BIC(m_pois, m_nbin)

       df      BIC
m_pois 18 2046.851
m_nbin 19 1157.235
```

Hence, the full summary of that model is shown below.

```
R> summary(m_nbin)

Call:
glm.nb(formula = Days ~ (Eth + Sex + Age + Lrn)^2, data = quine,
    init.theta = 1.60364105, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0857  -0.8306  -0.2620   0.4282   2.0898

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.00155    0.33709   8.904  < 2e-16 ***
EthN        -0.24591    0.39135  -0.628  0.52977
SexM        -0.77181    0.38021  -2.030  0.04236 *
AgeF1       -0.02546    0.41615  -0.061  0.95121
AgeF2       -0.54884    0.54393  -1.009  0.31296
AgeF3       -0.25735    0.40558  -0.635  0.52574
LrnSL        0.38919    0.48421   0.804  0.42153
EthN:SexM    0.36240    0.29430   1.231  0.21818
EthN:AgeF1  -0.70000    0.43646  -1.604  0.10876
EthN:AgeF2  -1.23283    0.42962  -2.870  0.00411 **
EthN:AgeF3   0.04721    0.44883   0.105  0.91622
EthN:LrnSL   0.06847    0.34040   0.201  0.84059
SexM:AgeF1   0.02257    0.47360   0.048  0.96198
SexM:AgeF2   1.55330    0.51325   3.026  0.00247 **
SexM:AgeF3   1.25227    0.45539   2.750  0.00596 **
SexM:LrnSL   0.07187    0.40805   0.176  0.86019
AgeF1:LrnSL -0.43101    0.47948  -0.899  0.36870
AgeF2:LrnSL  0.52074    0.48567   1.072  0.28363
AgeF3:LrnSL       NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.6036) family taken to be 1)

    Null deviance: 235.23  on 145  degrees of freedom
```

```
Residual deviance: 167.53  on 128  degrees of freedom
AIC: 1100.5

Number of Fisher Scoring iterations: 1


              Theta:  1.604
          Std. Err.:  0.214

 2 x log-likelihood:  -1062.546
```

4. Summary and discussion

- As usual ...

Computational details

If necessary or useful, information about certain computational details such as version numbers, operating systems, or compilers could be included in an unnumbered section. Also, auxiliary packages (say, for visualizations, maps, tables, ...) that are not cited in the main text can be credited here.

The results in this paper were obtained using R 4.1.3 with the **MASS** 7.3.55 package. R itself and all packages used are available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/.

Acknowledgments

All acknowledgments (note the AE spelling) should be collected in this unnumbered section before the references. It may contain the usual information about funding and feedback from colleagues/reviewers/etc. Furthermore, information such as relative contributions of the authors may be added here (if any).

References

Baranowski R, Fryzlewicz P (2019). *wbs: Wild Binary Segmentation for Multiple Change-Point Detection*. R package version 1.4, URL https://CRAN.R-project.org/package=wbs.

Chambers JM, Hastie TJ (eds.) (1992). *Statistical Models in S*. Chapman & Hall, London.

Killick R, Eckley IA (2014). "changepoint: An R Package for Changepoint Analysis." *Journal of Statistical Software*, **58**(3), 1–19. URL https://www.jstatsoft.org/v58/i03/.

Killick R, Haynes K, Eckley IA (2022). *changepoint: An R package for changepoint analysis*. R package version 2.2.3, URL https://CRAN.R-project.org/package=changepoint.

Maidstone R, Hocking T, Rigaill G, Fearnhead P (2017). "On optimal multiple changepoint algorithms for large data." *Statistics and Computing*, **27**. URL https://link.springer.com/article/10.1007/s11222-016-9636-3.

McCullagh P, Nelder JA (1989). *Generalized Linear Models*. 2nd edition. Chapman & Hall, London. doi:10.1007/978-1-4899-3242-6.

Truong C, Oudre L, Vayatis N (2020). "Selective review of offline change point detection methods." *Signal Processing*, **167**, 107299. ISSN 0165-1684. doi:https://doi.org/10.1016/j.sigpro.2019.107299. URL https://www.sciencedirect.com/science/article/pii/S0165168419303494.

A. More technical details

Appendices can be included after the bibliography (with a page break). Each section within the appendix should have a proper section title (rather than just *Appendix*).

For more technical style details, please check out JSS's style FAQ at https://www.jstatsoft.org/pages/view/style#frequently-asked-questions which includes the following topics:

- Title vs. sentence case.
- Graphics formatting.
- Naming conventions.
- Turning JSS manuscripts into R package vignettes.
- Trouble shooting.
- Many other potentially helpful details...

B. Using BibTeX

References need to be provided in a BibTeX file (`.bib`). All references should be made with `\cite`, `\citet`, `\citep`, `\citealp` etc. (and never hard-coded). This commands yield different formats of author-year citations and allow to include additional details (e.g., pages, chapters, ...) in brackets. In case you are not familiar with these commands see the JSS style FAQ for details.

Cleaning up BibTeX files is a somewhat tedious task – especially when acquiring the entries automatically from mixed online sources. However, it is important that informations are complete and presented in a consistent style to avoid confusions. JSS requires the following format.

- JSS-specific markup (`\proglang`, `\pkg`, `\code`) should be used in the references.
- Titles should be in title case.
- Journal titles should not be abbreviated and in title case.
- DOIs should be included where available.
- Software should be properly cited as well. For R packages `citation("pkgname")` typically provides a good starting point.

Affiliation:

Achim Zeileis
Journal of Statistical Software
and
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstr. 15
6020 Innsbruck, Austria
E-mail: Achim.Zeileis@R-project.org
URL: https://www.zeileis.org/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd