BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–2

EBS: an Exact Bayesian Segmentation Algorithm for the analysis of biological data-sets

Alice Cleynen [1]*, Guillem Rigaill [2] and Stéphane Robin [1]

[1]AgroParisTech, 16 rue Claude Bernard, 75231 Paris Cedex 05, France
[2]URGV, INRA-CNRS-Univ. Evry, 2 Rue Gaston Crémieux, 91057 Evry Cedex, France

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: EBS is an R package for the segmentation of biological data-sets (arrayCGH, RNA-seq, etc). It provides, through a Bayesian framework, exact quantities such as the posterior distribution of a change-point position or an efficient ICL criterion for the selection of the total number of change-points. All quantities are computed in quadratic time.

Availability: EBS is available as an R package from CRAN repositories (http://cran.r-project.org/web/packages/EBS)

Contact: alice.cleynen@agroparistech.fr

1 INTRODUCTION

Many biological dataset analyses (CNV or transcript detection, ...) aim at finding some abrupt changes in an ordered signal that is typically observed along the genome, and can therefore be rephrased as change-point detection problems. Most of them do not address crucial questions such as the quality of the segmentation, or the uncertainty on the localisation of breakpoints that are useful when choosing the number of segments, or comparing the segmentation of different profiles.

Among the few that do are the implementation of Bardy and Hartigan's Bayesian approach (`bcp`) that uses MCMC approximation, (Barry and Hartigan, 1993; Erdman and Emerson, 2007), and the frequentist forward-backward algorithm of Guédon (2008) and constrained-HMM framework of Luong *et al* (2012). Those two later approaches compute those useful quantities for fixed values of the segment parameters, and therefore do not account for the uncertainty due to their estimation.

EBS is an implementation of the framework described in Rigaill *et al.* (2010) which derives *exact* posterior probabilities of quantities such as the number of segments, the entropy of a segmentation, or the localisation of the change-points. The general model can be stated as follows: consider data ordered along genomic positions (or probe location) $Y = (Y_1, Y_2, \dots Y_n)$. The whole chromosome is shattered into successive segments r. The observations come from some parametric distribution F of which the parameter depends on the segment: $i \in r \Rightarrow Y_i \sim F(\theta^r)$.

2 AVAILABLE FUNCTIONALITY

Our approach is valid for all models satisfying the following factoriability assumption: if Y denotes the data, m a segmentation and r a segment of m,

$$(H) \quad P(Y,m) = C \prod_{r \in m} a_r P(Y^r|r) \tag{1}$$

where $P(Y^r|r) = \int P(Y^r|\theta_r) P(\theta_r) d\theta_r$.

This condition is satisfied by the Poisson, Normal (Heteroscedastic and Homoscedastic with known variance) and Negative Binomial (with known overdispersion parameter) models which are all included in the package. Normal distribution are dedicated to arrayCGH, whereas Poisson and Negative Binomial are proposed for NGS.

The computation of the quantities of interest rely on the knowledge of $P(Y,K)$ (K being the number of segments) that can be computed in quadratic time as

$$P(Y,K) = \left[\binom{n-1}{K-1} \right]^{-1} \left(A^K\right)_{1,n+1} \tag{2}$$

where $A_{i,j} = P(Y^{[\![i,j[\![})$. (See Proposition 2.2 of Rigaill *et al.* (2010) for proof).

Table 1 gives the list of the functions available in the EBS Package. This section describes and illustrates their use with a continued example.

2.1 Matrix Construction

All quantities of interest can be computed using simple operations on the elements of the matrix A of segment probabilities. The function `EBSegmentation` initializes this matrix with the data according to the user's choice of maximum number of segments and data-model. Each of them is associated with a prior distribution on the parameters (for instance, Gamma for the Poisson model), and the result depends on the value of the hyperparameters. By default `EBSegmentation` proposes to compute and use data-driven values (see EBS Manual for more details), but the user has the possibility of giving his or her own hyperparameters.

The function returns an object of class *EBS* which contains the prior information~~,~~ matrix A, and the two matrices Li and Col in which k^{th} row (respectively column) is the first row (resp. last column) of the k^{th} power of A ($1 \leq k \leq K_{max}$). In other words we have: $Li_{i,j} = P([\![Y_0, Y_j[\![, i)$ and $Col_{i,j} = P([\![Y_i, Y_{n+1}[\![, j)$.

```
> set.seed(1)
> require(EBS)
> x<-c(rnbinom(100,0.4,size=0.95),rnbinom(50,0.1,size=0.95),
rnbinom(75,0.4,size=0.95),rnbinom(125,0.1,size=0.95),
rnbinom(75,0.4,size=0.95))
> out <- EBSegmentation(x, model=3, Kmax=20)
```

2.2 Model Selection

The EBS Package provides two model selection criteria:

- an exact BIC criterion;
- and an exact ICL criterion

These criteria can be called with functions `EBSBIC` and `EBSICL`.

*to whom correspondence should be addressed

Table 1. Functions provided by EBS package

Function Name	Output
EBSegmentation	Initializes matrix A and its K_{max} first powers
EBSBIC	Computes BIC and chooses the optimal value of K
EBSICL	Computes ICL and chooses the optimal value of K
EBSPostK	Returns the posterior probability of the number of segments
EBSDistrib	Computes the distribution of a change-point
EBSPlotProba	Plots distribution of all change-points for a given K

Considering the segmentation as an unobserved variable, we can use the ICL criterion introduced by Biernacki *et al*. (2000) in the context of incomplete data models to select the number of segments. The ICL can be written as $ICL(K) = -\log P(Y,K) + \mathcal{H}(K)$ where the entropy $\mathcal{H}(K)$ is defined as

$$\mathcal{H}(K) = -\log \sum_{m \in \mathcal{M}_K} p(m|Y,K) \log p(m|Y,K) \tag{3}$$

The entropy can be viewed as a penalty term, which favors segmentation with precisely defined change-points location and it is computed in quadratic time. Even though in this context the Bayesian Information Criterion is exact, it tends to overestimate the number of segments while the ICL performs better (Rigaill *et al*., 2010). However, the computation of the BIC (through function `EBSPostK`) can be useful for later carrying analysis such as Bayesian Model Averaging (BMA).

```
> print(bic <- EBSBIC(out)$NbBIC)
[1] 6
> print(icl <- EBSICL(out)$NbICL)
[1] 5
```

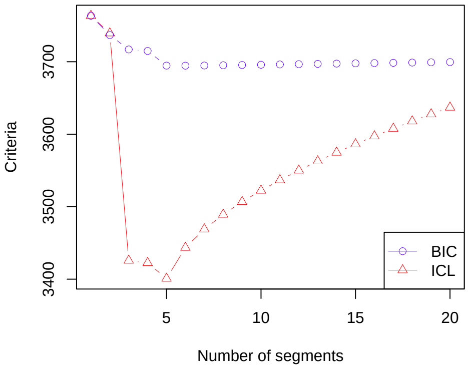

Fig. 1. BIC and ICL criteria as a function of the number of segments

2.3 Change-point location distribution

One might be interested in the distribution of the location of each change-points. Two functions are implemented to address this question. `EBSDistrib` returns the distribution of the k^{th} change-point of a segmentation in K segments. `EBSPlotProba` plots the distribution of all $K-1$ change-points of a segmentation in K segments. The user has the option to plot those distributions on top of the data.

```
> EBSPlotProba(out, icl, data=TRUE)
```

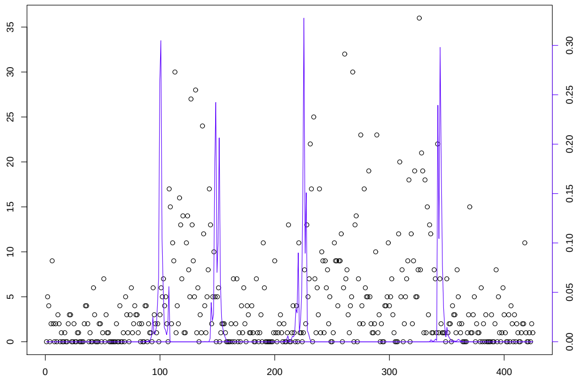

Fig. 2. Output of function `EBSPlotProba`: distribution of the localisation of the change-points on simulated data (right y-axis, blue)

3 CONCLUSION

An exact computation of many powerful quantities is obtained thanks to the exploration of the entire segmentation space in a Bayesian framework adapted to the analysis of NGS and CGH-array data. It provides a useful criterion for the selection of the number of segments, and allows further analysis of variables such as the entropy or the location of change-points. Future improvements to the package may include the calculation of other quantities of interest, such as the posterior mean of the signal, which would provide an estimate of the copy number at a given locus, or of the expression level of a given exon.

ACKNOWLEDGEMENT

We wish to thank Gregory Nuel and Michel Koskas for their help dealing with numerical issues.

REFERENCES

Barry, D., and Hartigan, J. A. (1993) A Bayesian Analysis for Change Point Problems *Journal of the American Statistical Association*, **88** 309-319

Biernacki, C., Celeux, G. and Govaert, G. (2000) Assessing a mixture model for clustering with the integrated completed likelihood, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22** 719-725.

Chandra Erdman and John W. Emerson (2007) **bcp**: An R Package for Performing a Bayesian Analysis of Change Point Problems *Journal of Statistical Software*, **3** 1-13

Guédon, Y. (2008) Exploring the segmentation space for the assessment of multiple change-points models. *Technical report, Preprint INRIA* **6619**

Luong, T. M., Rozenholc, Y. and Nuel, G. (2012) Fast estimation of posterior probabilities in change-point models through a constrained hidden Markov model http://adsabs.harvard.edu/abs/2012arXiv1203.4394L.

Rigaill, G., Lebarbier, E., Robin, S. (2011) Exact posterior distributions over the segmentation space and model selection for multiple change-point detection problems, *Statistics and Computing*, **22-4**, 917-929.