BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–3

This is a title

Corresponding Author [1,*], Co-Author [2] and Co-Author [2*]

[1]Department of XXXXXXX, Address XXXX etc.
[2]Department of XXXXXXXX, Address XXXX etc.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: Numerous methods have been proposed to find the optimal partition of a signal in a given number of segments, but very few address the question of the quality of the segmentation. In many biological data-sets (CGH-array, NGS, ...), each segment is assumed to be related to a biological event. Thus it is crucial to have precise idea of the uncertainty on the number and position of the segments. EBS is an R package that provides through a Bayesian framework exact quantities such as the posterior distribution of a change-point position, and an efficient ICL criterion for the selection of the total number of change-points. All quantities are computed in a quadratic time.

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

Contact: alice.cleynen@agroparistech.fr

1 INTRODUCTION

Most change-point detection strategies do not address crucial questions like 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 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 Guedon (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.

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 change-point detection 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. The parameter depends on the segment the observation belongs to: $i \in r \Rightarrow Y_i \sim F(\theta^r)$.

*to whom correspondence should be addressed

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

The package includes the Poisson, Normal (Heteroscedastic and Homoscedastic with known variance) and Negative Binomial (with known overdispersion parameter) models that all verify (H). 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 $A_{i,j} = P(Y^r)$ where r stands for the segment $[\![i, j[\![$. (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 own hyperparameters.

The function returns an object of class *EBS* which contains the prior information, matrix A and the two matrices Li and Col 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(125,0.1,size=0.95),rnbinom(25,0.35,size=0.95),
 rnbinom(100,0.1,size=0.95),rnbinom(100,0.35,size=0.95),
 rnbinom(75,0.1,size=0.95))
> out <- EBSegmentation(x, model=3, Kmax=20)
```

2.2 Model Selection

EBS Package provides two model selection criteria:

- an exact BIC criterion

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 optimal value of K
EBSICL	Computes ICL and chooses 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

- an ICL criterion

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

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 :

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

The entropy can be viewed as a penalty term and it is computed in a quadratic time. Even though in this context the Bayesian Information Criterion is exact, it overestimates the number of segments while the ICL performs better (Rigaill *et al.*, 2010). However, the computation of the BIC (through function `EBSPostK`) can be usefull for later analysis such as Bayesian Model Averaging (BMA).

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

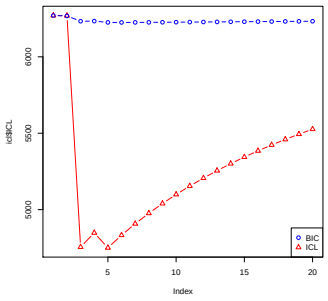

Fig. 1. BIC and ICL criteria as a function of the number of segments

2.3 Change-point location distribution

Once the number of segments is chosen, 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} changepoint 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,
  file="my-segmentation.pdf")
```

Figure 2 shows the output.

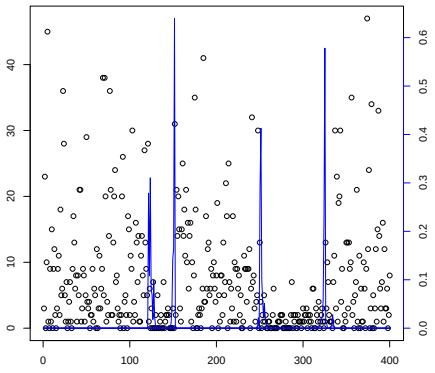

Fig. 2. file *my-segmentation.pdf*, output of function `EBSPlotProba`

3 CONCLUSION

This note introduces the EBS package for the analysis of NGS and CGH-array data. It allows a complete analysis of the segmentation space for biological profiles in an exact Bayesian framework. It provides an efficient criterion for the selection of the number of segments and allows further analysis of variables such as the entropy or the location of a change-point. Future improvements of the package include the analysis of other quantities of interest such as the posterior mean of the signal.

ACKNOWLEDGEMENT

We whish 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

Gudon, 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.