BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 2 2011, pages 161–166
doi:10.1093/bioinformatics/btq647

Genome analysis

Advance Access publication November 18, 2010

Efficient change point detection for genomic sequences of continuous measurements

Vito M. R. Muggeo* and Giada Adelfio

Dipartimento di Scienze Statistiche e Matematiche 'Vianelli', Università di Palermo, viale delle Scienze, 90128 Palermo, Italy

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Knowing the exact locations of multiple change points in genomic sequences serves several biological needs, for instance when data represent aCGH profiles and it is of interest to identify possibly damaged genes involved in cancer and other diseases. Only a few of the currently available methods deal explicitly with estimation of the number and location of change points, and moreover these methods may be somewhat vulnerable to deviations of model assumptions usually employed.

Results: We present a computationally efficient method to obtain estimates of the number and location of the change points. The method is based on a simple transformation of data and it provides results quite robust to model misspecifications. The efficiency of the method guarantees moderate computational times regardless of the series length and the number of change points.

Availability: The methods described in this article are implemented in the new R package cumSeg available from the Comprehensive R Archive Network at http://CRAN.R-project.org/package=cumSeg.

Contact: vito.muggeo@unipa.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 9, 2010; revised on November 10, 2010; accepted on November 13, 2010

1 INTRODUCTION

Statistical analysis in biological research is often faced with outcomes expressed by specific measurements for the gene expression levels. When these observed responses are ordered by their location on the genome, the values form clouds with different observed means, supposedly reflecting different mean levels. By exploiting differences in the observed responses, the statistical analysis of these sequences aims at identifying chromosomal regions with 'abnormal' (increased or decreased) mean levels; changes in the mean level at a particular genome location may be interesting for certain goals, such as assessing the status not only of specific loci (like tumor suppressors and oncogenes) but also any structural genomic alteration, resulting in genomic imbalances, involved in pathogenesis or in various observed abnormal phenotypes. For instance, detecting chromosomal aberrations is crucial for the diagnosis of some diseases, including mental retardation and cancer (e.g. Albertson and Pinkel, 2003; Redon *et al.*, 2006; Veltman *et al.*, 2003) often associated to aneuploidies, deletions, duplications and/or amplifications. In this respect, the output of the array comparative genomic hybridization (aCGH) experiment is a fluorescence color ratio understood to represent the DNA copy numbers of the actual tumor cells plus noise generated by the normal cells and by the experiment; aCGH analysis aims at finding abrupt changes in the mean of the fluorescence color ratios (or usually their logarithms) to detect chromosomal aberrations. The analysis of aCGH data concerns one of the most important applications involving change point detection, but more generally the analysis of genomic sequences to detect abrupt changes is an important step in several areas of the biological research: recombination of viruses (Halpern, 2000), characterization of complete transcriptomes via high-density DNA tiling microarrays (Huber *et al.*, 2006a) or investigation of DNA sequences in general, see Karkin and Brendel (1993), for a wider discussion.

Several authors have addressed the problem using different approaches within a likelihood-based or Bayesian paradigms. Algorithms based on segmentation techniques include works by Jong *et al.* (2003), Myers *et al.* (2004), Olshen *et al.* (2004), Wang *et al.* (2005), Picard *et al.* (2005) and Lipson *et al.* (2005); hidden Markov models have been presented in Fridlyand *et al.* (2004). When mean levels are of main interest, rather than the change points, Bayesian approaches have been discussed in Barry and Hartigan (1993) and Erdman and Emerson (2008), while Eilers and Menezes (2004), Huang *et al.* (2005) and Tibshirani and Wand (2008) have used L_1-penalties to deal with abrupt changes; smoothing techniques have been applied by Hupé *et al.* (2004). Although comprehensive reviews and comparisons of some of the aforementioned methods are given by Lai *et al.* (2005) and Willenbrock and Fridlyand (2005), the focus of the aforementioned procedures is into 'calling gains and losses', namely classification into damaged and non-damaged genes. In general, little emphasis is given to the problem of estimating number and locations of the change points, especially when the stochastic terms are not simple Gaussian independent identically distributed (*iid*). Typical statistical analyses assume *iid* Gaussian noise, but there is no guarantee that this is always true.

In this article, we present a new approach to detect and estimate change points in genomic sequences. While associated mean levels can be obtained, we are specifically interested in the change points. The main goal is to provide a procedure which works adequately well in different and real scenarios where the true error structure is usually unknown. The method is conceptually simple and it appears to be quite robust to departures from standard model assumptions, such as autocorrelation, non-normality and heteroscedasticity. Moreover,

*To whom correspondence should be addressed.

it can be straightforwardly generalized to include interprobe distance, quality of measurements and multivariate and binary responses.

The article is organized as follows. Section 2 describes the model estimation and selection, while Section 3 presents results from a simulation study aimed at comparing the proposed approach with respect to its competitors in selecting the right number of change points. Section 4 reports analyses of two historical datasets, and finally Section 5 is devoted to discussion and future work.

2 METHODS

2.1 The piecewise constant model

Let $\{(x_i, y_i)\}_i^n$ be the observed data: y_i is the outcome, e.g. the log fluorescence ratio, and x_i represents the genome location for $i = 1, 2, \dots, n$ genes/clones. As usual in statistical modeling, we assume the datum is the sum of the signal and noise, i.e. $y_i = \mu_i + \epsilon_i$. To deal with possible change points in the sequence, we further assume the signal μ_i is approximated by a piecewise constant (or mean-shift) regression function with $K_0 + 1$ segments,

$$y_i = \beta_1 + \delta_1 I(x_i > \psi_1) + \cdots + \delta_{K_0} I(x_i > \psi_{K_0}) + \epsilon_i \tag{1}$$

where ϵ_i is noise which, depending on its 'strength', distorts the signal μ_i and complicates change point detection. In Equation (1), $I(\cdot)$ is the indicator function being equal to one when its argument is true, and the ψs represent the K_0 change points on the genome, (e.g. the locations of the changes), β_1 is the mean level (e.g. the log2 ratio) for $x_i < \psi_1$ and the δs are the differences in the mean levels at the change points; for instance $\beta_1 + \delta_1$ is the mean level when $\psi_1 \le x_i < \psi_2$. Typically it is assumed $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, but in our framework we only need that the noise is zero mean, $E[\epsilon_i] = 0$; no assumption on the error density or its higher moments is requested. Notice although in model (1) we have used the simple index variable $i = 1, 2, \dots, n$ denoting the genome index, any continuous variable may be employed. In fact, this restriction is not essential in our framework (see Section 5), but we will assume $x_i = i$ throughout the article for the sake of simplicity.

The basic statistical problem with model (1) consists in identification of the number of change points (K_0). Estimation of their locations (the ψ_ks) and the mean levels assumed in between (the regression coefficients β_1 and δ) may also be of interest, but we do not deal specifically with them in the present article.

2.2 Model estimation: preliminary

We first discuss estimation assuming fixed and known number of change points, K_0. Taking the cumulative sums of both side of Equation (1) yields

$$z_i = \beta_1 x_i + \delta_1 (x_i - \psi_1)_+ + \cdots + \delta_{K_0} (x_i - \psi_{K_0})_+ + \eta_i \tag{2}$$

where for each k and i, $z_i = \sum_j^i y_j$, $x_i = i = \sum_j^i 1$, $\eta_i = \sum_j^i \epsilon_j$, and $(x_i - \psi_k)_+ = \sum_j^i I(x_j > \psi_k) = (x_i - \psi_k) I(x_i > \psi_k)$. Model (2) has the same parameters of Equation (1), but different responses, covariates and errors. More importantly, Equation (2) underlies a different relationship: unlike (1), it assumes a *piecewise linear* or *segmented* relationship, namely the regression function *continuous* at the change points ψ_k. Figure 1 illustrates a simple example: the top panel reports the 'original' data and the underlying piecewise constant signal, while the bottom plot shows the cumulative sums of the data and the corresponding piecewise linear signal induced. The change points are the same and the mean levels of the former plot correspond to the slopes of the latter; as a consequence, parameter estimates of model (1) may come from estimation of model (2).

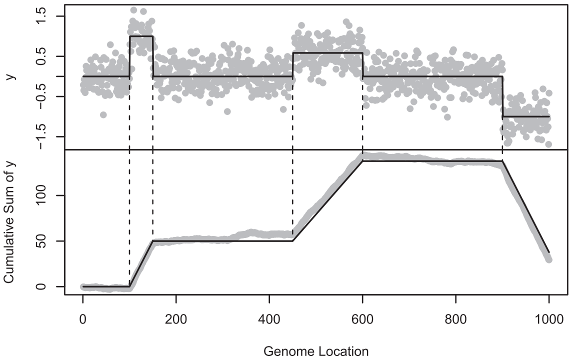

Fig. 1. An example of original measurements (upper panel) and transformed data (bottom panel) with corresponding underlying signals superimposed. These data have been simulated according to Huang *et al.* (2005).

The reason to consider continuous change points [i.e. model (2)] rather than model (1) is that estimation may be performed efficiently via the exact algorithm discussed in Muggeo (2003, 2008b), which does not work for piecewise constant models. We do not detail the procedure here, but given starting values for the change points, $\tilde{\psi}_k$, the problem reduces to fit iteratively the linear model.

$$E[z_i | x_i] = \beta_1 x_i + \sum_k \delta_k \widetilde{U}_{ik} + \sum_k \gamma_k \widetilde{V}_{ik}^-, \tag{3}$$

where $\widetilde{U}_{ik} = (x_i - \tilde{\psi}_k)_+$, $\widetilde{V}_{ik}^- = -I(x_i > \tilde{\psi}_k)$, and where we have used lower case letters to mean observed and random variables as well. The parameters β_1 and δs are from Equations (1) and (2), and the γs are 'working' coefficients useful for the estimation procedure (Muggeo, 2003). At each step the working model (3) is fitted and new estimates of the change points are obtained via

$$\hat{\psi}_k = \tilde{\psi}_k + \hat{\gamma}_k / \hat{\delta}_k. \tag{4}$$

The process is iterated up to convergence when parameter estimates are available and model (1) is readily estimated. Notice we are using standard linear models to fit the change point model, resulting in a very efficient algorithm even with n large and many change points to be estimated.

As previously discussed, estimation of model (3) actually may be carried out via least squares, which ensures unbiased parameter estimates by requiring only zero mean errors; in fact, when this assumption is met in the original data, i.e. $E[\epsilon_i] = 0$, it is also fulfilled for model (2) [and its working version (3)], i.e. $E[\eta_i] = 0$. An important advantage of the ordinary least squares procedure is its robustness against departures from normality, homoscedasticity or uncorrelation/independence; we will assess robustness with regard to change point detection via simulations later.

When the number of change points, K_0, is known and the starting values are not unreasonably 'far' from solutions, the algorithm converges in a few iterations by returning exactly the estimated locations of these K_0 change points.

2.3 Model estimation: overdetecting the change points (output 1)

The aforementioned 'segmented' algorithm allows us to get parameter estimates efficiently when the exact number of change points is known and 'reasonable' starting values are set. However, the true number K_0 of change points is unknown in practice and we propose to overcome these issues as follows.

We start the algorithm using a large number K of candidate change points, possibly quite larger than the supposed one in order to make irrelevant the location of starting values; we set the starting values at equally spaced values or quantiles of the x_is. Then the working segmented model (3) is iteratively fitted, and at each iteration the k-th change point is discarded (and the corresponding covariates U_k and V_k deleted) when it is not *admissible*. We say the estimate $\hat{\psi}_k$ is *not admissible* when

- it does not belong to the allowed range. Whether the generic value $\tilde{\psi}_k$ is not a change point, the relevant $\hat{\delta}_k$ is approximately zero, leading the ratio $\hat{\gamma}_k/\hat{\delta}_k$ to assume a very large value, and causing the corresponding updated estimate (4) to fall outside the range of the x_is.

- the corresponding covariates (U_k,V_k) are redundant with any other couple $(U_{k'},V_{k'})$. Especially when a very large K is used, it could happen that for two estimated change points $\hat{\psi}_k \approx \hat{\psi}_{k'}$, it is $U_k \approx U_{k'}$ and $V_k \approx V_{k'}$. This redundancy may prevent estimation of model (3), and as consequence, only one of the two current change points may be updated.

As the algorithm goes on, only *admissible* values are retained, and at the convergence only the $K^*(<K)$ likelier values are returned. We refer this step of the algorithm as 'output 1' that produces for the original data the fitted model

$$\hat{\mu}_i^* = \hat{\beta}_1 + \hat{\delta}_1 V_{i1} + \cdots + \hat{\delta}_{K^*} V_{iK^*}, \tag{5}$$

where $V_{ik} = I(x_i > \hat{\psi}_k)$ for $k = 1,2,\ldots,K^*$.

This approach may be considered essentially as a smoothing algorithm based on the 0-degree truncated power functions bases (the V_ks) with 'automatic' selection of the knots, i.e. the change points. The output is similar to that returned by the fused lasso (Tibshirani and Wand, 2008), namely a wiggly fitted 'curve'. While the number of change points will be strongly overestimated, the fitted values from this output may be used to determine, for instance, which segments are aberrant, namely which means may be assumed to be less, equal or greater than 0. In aCGH analysis, typically this is achieved via a post-processing step, namely by applying a given procedure on the fitted values returned by the relevant algorithm. For instance, a simple strategy consists in labeling as aberrant the values outside m times, say, the computed SD. More sophisticated procedures include the so-called 'MergeLevels' (Willenbrock and Fridlyand, 2005), 'CGHcall' (van de Wie *et al.*, 2007) and 'false discovery rate' (FDR)-based approach used in Huang *et al.* (2005) and Tibshirani and Wand (2008).

While the fitted values (5) may be employed by any of the aforementioned approaches to yield 'calls for gains and losses', in this article we are specifically interested in the change points. As previously discussed, this can be relevant in several biological contexts, not necessarily involving aCGH data.

2.4 Model estimation: selecting the number of change points (outputs 2 and 3)

Output 1 returns estimates for the model with K^* change points. However, when the algorithm is started with $K \gg K_0$, it is likely that the K^* returned change points are more than the actual ones, and therefore it is of interest to assess whether K^* change points are really necessary. In other words, we need to select only the 'significant' change points by removing the spurious ones.

Whether the generic $\hat{\psi}_k$ is not a change point, the corresponding covariate V_k should be a *noise* variable, as it would be $\hat{\delta}_k \approx 0$. Thus, selecting the number of change points actually reduces to selecting the significant variables among $V_1,V_2,\ldots,V_{K^*}$. This is a variable selection problem, and to solve it efficiently we use the *lars* algorithm proposed by Efron *et al.* (2004). At the cost of a single least squares computation, the *lars* algorithm returns the solutions for the *entire path*, namely the parameter estimates from the 'null' (only-intercept) model to the full model when every variable V_k is included in the model. The fitted 'optimal' model having $\hat{K} \le K^*$ change points selected by any criterion, such as expression (6), is referred as 'output 2'.

The 'optimal' model may be selected via any goodness-of-fit criterion penalized for the model dimension. However, at this aim we need to make some assumptions on the errors ϵ_i; among the well-known Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) or the Minimum Description Length (MDL), we observe that the 'generalized' BIC based on Gaussian iid errors

$$\mathrm{BIC}_{C_n} = \log(\hat{\sigma}^2) + edf\,\frac{\log(n)}{n}C_n \tag{6}$$

appears to outperform its competitors. Here $\hat{\sigma}^2$ is the residual variance estimate, *edf* is the actual model dimension quantified by the number of estimated parameters given by $edf = 1+2\{\#\text{changepoints}\}$ (including the intercept, the δs and the ψs) and C_n is a known constant. Note the generalized expression (6) reduces to the usual BIC when $C_n = 1$. Wang *et al.* (2009) discuss the use of $C_n > 1$ in identifying the true model when the number of parameters is not fixed but it diverges as $n \to \infty$. In the analysis of genomic sequences, it is reasonable to think that the number of change points depends on the series length, therefore 'diverging' parameters appear to be appropriate here. Simulation studies presented later show that $C_n = \log\log n$ appears to be the most suitable value, and that it works reasonably well even if the stochastic terms are not Gaussian iid: this is an important feature since the error structure is typically unknown in practice.

The change points $\hat{\psi}_1,\ldots,\hat{\psi}_{\hat{K}}$ from output 2 are actually a subset of the estimates $\hat{\psi}_1,\ldots,\hat{\psi}_{K^*}$ from output 1, since one or more change points are not included due to deletion of one or more variables V_k by means of the selection criterion (6). However, it should be recognized that while $\hat{\psi}_1,\ldots,\hat{\psi}_{K^*}$ are 'the best' estimates, in that they minimize the residual sum of squares for K^* change points, there is no guarantee that the subset $\hat{\psi}_1,\ldots,\hat{\psi}_{\hat{K}}$ constitutes also 'the best' estimate for the number of change points, K_0. In general, differences will be negligible; however, it is possible to refine the estimates for the model having $\hat{K}$ change points: at this aim it suffices to run again the segmented algorithm of Section 2.2 using $\hat{\psi}_1,\ldots,\hat{\psi}_{\hat{K}}$ as starting values. Notice that, due to the quite fair starting values, now the segmented algorithm will converge in few iterations; we refer this output as 'output 3'.

In summary, the proposed algorithm may be expressed as follows

(1) Fix a high initial number K of change points and fix the corresponding starting values (their values are not important, provided that K is large);

(2) Compute the variables U_k,V_k for $k=1,\ldots,K$ and fit the segmented linear model for the transformed response via ordinary least squares;

(3) Extract the coefficient estimates $\hat{\gamma}_k$ and $\hat{\delta}_k$ to update the change point estimates: delete the variables (U_k,V_k) if the corresponding estimate $\hat{\psi}_k$ is not admissible, and update the admissible change points;

(4) Repeat steps 2 and 3 up to some convergence criterion is met;

(5) Apply the *lars* algorithm to discard the noise variables V_k and to select significant change points using the generalized BIC with $C_n = \log\log n$;

(6) Refine, if requested, the parameter estimates by running a new segmented algorithm.

Notice that if one is only interested in obtaining fitted values to apply some classification procedure (as in aCGH analysis), the algorithm may stop at step 4; otherwise the algorithm finishes at step 5, or possibly 6: in this case it suffices to run the segmented algorithm in step 4 with three to five iterations, without obtaining complete convergence.

3 SIMULATION STUDY

We have carried out some simulation experiments to assess the performance of the proposed approach in identifying the correct number of change points. At this aim, we have used the language R (R Development Core Team, 2006) with some additional packages as reported below. The first issue to be discussed concerns the criterion to be used to select the change points. We consider AIC, MDL and 'generalized' BIC (6) with different values of C_n (for instance $1,\sqrt{n},\log n,\log\log n$). Preliminary simulations (not reported here) have revealed that the conventional BIC and especially the AIC strongly overestimate the number of change points, as also reported in Picard *et al.* (2005). Among the different values $C_n > 1$ examined, $C_n = \log\log n$ performs the best and we have used this value throughout the rest of simulations. We have compared our approach (‘cumSeg’) against three procedures: (i) circular binary segmentation (CBS; Olshen *et al.*, 2004; Venkatraman and Olshen, 2007) as implemented in the R package DNAcopy (Seshan and Olshen, 2008); (ii) CGHseg (Picard *et al.*, 2005) based on dynamic programming and implemented in the package tilingArray (Huber *et al.*, 2006b) via the function findSegments() [or equivalently segment()]; (iii) LB, the lasso-based discussed by Huang *et al.* (2005, at time of writing code at http://bioinformatics.med.yale.edu/DNACopyNumber/). We have focussed on robustness to some violations of the common model assumptions, namely performance of the procedures in identifying the true number of change points when the standard assumption of Gaussian iid is employed for estimation, but the true model has different stochastic terms: Gaussian iid (zero mean and variance equaling 0.2^2), Gaussian autoregressive (first-order autoregressive errors with zero mean are coefficient equal to 0.6, and variance 0.123^2), Gaussian heteroscedastic [zero mean and variance equal to $0.2u$ being $u\sim\mathrm{Unif}(0.5,3)$] and Laplace iid (scale parameter equal to $0.5/\sqrt{2}$). We have considered three sample sizes ($n=100,500,1000$) and two values for the true number of change points ($K_0=0$ and 4), being the true signal $\mu_i=0$ or $\mu_i=I(i>.3)-2\,I(i>.4)+1.5I(i>.8)-.5I(i>.85)$; the ‘covariate’ i includes equispaced values ranging from 0 to 1.

It should be stressed that the performance in the ‘null’ case ($K_0=0$) appears to be crucial, as it represents the well-known type 1 error that has to be controlled; comparisons concerning the capability in identifying change points cannot be apart from their performance when no change point exist. Perhaps surprisingly, it appears that most of the papers in literature have contrasted the different methods only with respect to their ability in detecting aberrant regions when these exist (Picard *et al.*, 2005); to the best of our knowledge, only few papers (e.g. Barry and Hartigan, 1993) have carried out comparisons in the null case.

Table 1 reports the results in terms of means and SDs of the number of selected change points over the 1000 replicates in each of the 24 simulation scenarios considered.

In our simulation, no method outperforms the others over all the scenarios. For the null (i.e. ‘only-intercept’) case, CGHseg breaks down in each scenario, regardless of the sample size and the error structure, as it selects too many change points. CBS heavily fails in the autoregressive scenario, especially at larger sample sizes; sensitivity of CBS to‘local trends’ induced by autocorrelation has also been reported by Olshen *et al.* (2004) and Huang *et al.* (2005). LB works adequately in all but the heteroscedastic case, where more change points are selected at larger samples. The proposed approach (cumSeg) selects the correct number of change points in all but the autoregressive scenario; however in this context, the average number of detected ψs is small and, more importantly, it decreases along with the SD as the sample size increases. When there exist $K_0=4$ change points, cumSeg performs reasonably well in all the scenarios, being slightly below CBS in the heteroscedastic case. However, CBS selects too many segments with autocorrelated errors as in the null case, while heteroscedasticity appears to affect the performance of CGHseg, which does not improve at larger sample sizes. Results from LB are somewhat unsatisfactory depending strongly on the sample size.

Simulation results concerning the the estimated locations of the change points and the fitted conditional means can be found in the Supplementary Material.

Table 1. Mean (m) and standard deviation (s) of the detected number of change points in two models ($K_0=0$ and $K_0=4$) with four different error structures and three sample sizes according to the four different approaches

n		$K_0=0$				$K_0=4$			
		(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
					Gaussian iid				
100	m	0.03	2.60	0.07	0.01	3.61	3.54	2.81	3.75
	s	0.23	2.46	0.36	0.10	0.58	0.50	0.55	0.57
500	m	0.03	2.40	0.19	0.00	4.08	3.95	4.94	4.05
	s	0.23	2.23	0.67	0.04	0.37	0.23	1.15	0.23
1000	m	0.02	2.38	0.33	0.00	4.15	4.00	5.80	4.05
	s	0.18	2.20	0.84	0.03	0.53	0.04	1.31	0.25
					Gaussian autoregressive				
100	m	2.64	3.51	0.02	1.05	5.92	3.65	2.20	3.86
	s	2.09	2.28	0.20	1.44	1.55	0.62	0.44	0.97
500	m	13.23	2.54	0.04	0.33	17.32	3.95	3.88	4.07
	s	5.53	2.25	0.19	0.74	4.76	0.22	0.78	0.49
1000	m	25.18	2.62	0.03	0.16	30.63	3.99	5.32	4.11
	s	8.46	2.31	0.17	0.51	7.23	0.08	1.43	0.47
					Gaussian heteroscedastic				
100	m	0.02	3.23	0.96	0.01	3.11	3.13	2.66	3.06
	s	0.19	2.48	0.92	0.14	0.46	0.50	0.67	0.40
500	m	0.00	3.64	4.19	0.00	4.00	3.12	6.90	3.74
	s	0.09	2.52	2.22	0.00	0.65	0.33	1.77	0.49
1000	m	0.02	4.04	6.76	0.00	4.20	3.04	10.5	4.03
	s	0.22	2.65	2.79	0.00	0.52	0.20	2.56	0.24
					Laplace iid				
100	m	0.01	3.79	0.31	0.01	2.97	3.41	2.88	2.92
	s	0.15	2.51	0.72	0.14	0.54	1.04	0.42	0.47
500	m	0.02	4.72	0.82	0.00	3.62	3.04	5.42	3.26
	s	0.21	2.46	1.41	0.00	0.75	0.22	1.40	0.46
1000	m	0.03	5.18	0.98	0.00	4.00	3.02	7.24	3.78
	s	0.24	2.39	1.33	0.00	0.63	0.13	2.29	0.48

(a), CBS; (b), CGHseg; (c), LB; (d), cumSeg (proposed approach).

4 EXAMPLES

We present results of the proposed approach in practice. We use two well-known datasets in the aCGH literature: in the former dataset the aberrations and then K_0 are known, so this dataset may be useful to validate the procedure; the latter dataset is interesting as it includes several sequences with different patterns. We compare cumSeg with CBS, CGHseg, LB introduced in the previous section and also with the fused lasso (FL) discussed by Tibshirani and Wand (2008) and implemented in the R package cghFLasso (Johnson *et al.*, 2009). Although the FL is not designed to estimate the number of change points, (typically it returns many change points), we include it for completeness. In both examples, we report results from ‘output 3’ by using $K=30$ equally spaced starting values.

4.1 Fibroblast cell lines data

This dataset is discussed by several authors (e.g. Hsu *et al.*, 2005; Huang *et al.*, 2005; Olshen *et al.*, 2004). The data consist of single experiments on 15 fibroblast cell lines with each array containing over 2000 (mapped) BACs spotted in triplicate. The variable used for analysis is the normalized average of the log base 2 test over reference ratio.

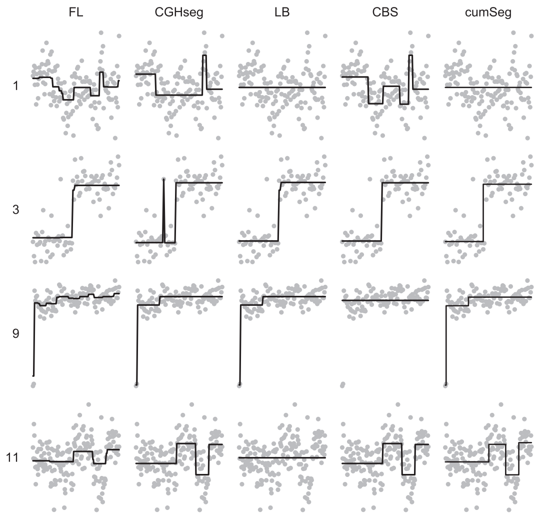

Fig. 2. Analysis of the fibroblast cell line GM03563 in four chromosomes (the chromosome index is on the left side): data and fitted lines from FL, CGHseg, LB, CBS and cumSeg (the proposed approach).

We analyze the sequences of the fibroblast cell line GM03563 in four chromosomes, such as 1, 3, 9, 11. Data and results are reported in Figure 2.

By spectral karyotyping, it is known that ‘real’ alterations are present only in chromosomes #3 and #9 (in particular $K_0 = 1$ in both the sequences). In chromosome #3, CBS and cumSeg provide the same right answer, while the others return two change points (the two estimates by LB are quite close); in the chromosome #9 no method returns one single change point estimate: CBS fails to find it, and the others select more than one. For the other chromosomes, only LB for chromosome #11 and both LB and cumSeg for chromosome #1 give the correct result (i.e. no change point). By considering chromosomes #1 and #11, results appear coherent with those of the simulation study in previous section. CGHseg always returns too many change points and if there is autocorrelation, CBS and, to some extent, cumSeg tend to overestimate the number of changes, while LB performs the best. However in chromosome #3, LB estimates two, rather than one, change points.

4.2 Breast tumor data

Here we consider the breast cancer cell line (MDA157) from cDNA microarray CGH profiled across 6691 mapped human genes in 44 breast tumor samples and 10 breast cancer cell lines.

This dataset is discussed in Tibshirani and Wand (2008) and available from the Stanford Microarray Database at http://smd.stanford.edu. Figure 3 shows the data along the fitted values from the five different approaches. We observe that cumSeg appears to provide reasonable results quite similar to those of CBS with some differences in chromosomes #5, #6 and #19. Unlike the previous dataset, here LB gives fitted curves too wiggly with an excessive number of change points (chromosomes #8, #13 and #19); simulations have also shown that LB tends to select too many change points in certain scenarios (Gaussian heteroscedastic or Laplace iid errors).

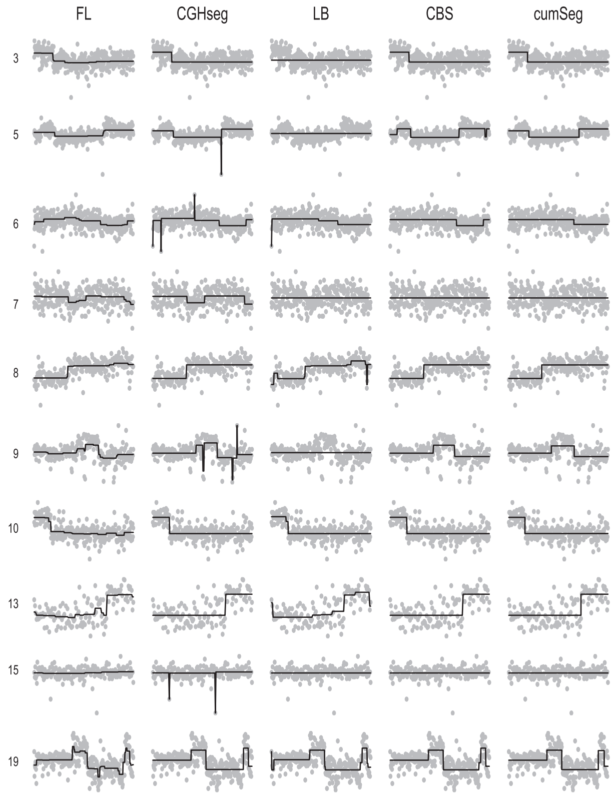

Fig. 3. aCGH profiles for 10 chromosomes of breast cancer cell line MDA157. Each row represents a chromosome (relevant index on the left side) and each column refers to the different estimating method.

5 DISCUSSION AND FUTURE WORK

We have presented a simple and computationally efficient framework, ‘cumSeg’, for detection of the number of change points in genomic sequences; we have not discussed in detail estimation of the location of the change points or the mean levels, although some simulation results are provided in the Supplementary Material. ‘cumSeg’ is based on a quite efficient algorithm to estimate the change points and the *lars* algorithm to discard the spurious ones via a generalized version of the BIC. The number of change points (K) to initialize the algorithm does not appear to affect the final result (i.e. ‘output 2’ and ‘3’), provided that it is large relatively to the supposed number of change points; for instance 30 to 50 or even equal to some fraction of the series length, $n/4$ say. Simulations and analyses of some datasets corroborate this point. The method turns out to be quite efficient even for large datasets, and for huge datasets it may benefit from strategies such as the R packages biglm (Lumley, 2009) and biglars (Seligman *et al.*, 2010), although we have not tried them in practice. The main appealing of the proposed approach is its robustness to the deviations from the common assumptions which are usually employed in real data analyses: Gaussian, independent and homoscedastic errors. Some simulations have shown that the our approach works reasonably well with non-normal errors, spatial dependence due to the physical dependence of nearby markers and heteroscedasticity coming from different quality of measurements. Simulation studies have shown that the other approaches perform poorly in other specific contexts, especially when there is no change point. Like other approaches proposed in literature (Eilers and Menezes, 2004; Huang *et al.*, 2005; Olshen *et al.*, 2004; Tibshirani and Wand, 2008), our framework also does not deal with 'single shocks', i.e. unit-width aberrations concerning single observations, genes/clones. As discussed by Tibshirani and Wand (2008), it is sometimes difficult to distinguish between single-gene copy number mutations and outliers, since an isolate high/low outcome may be caused by measurement errors/inaccuracies.

The cumSeg algorithm, currently implemented in R package cumSeg (Muggeo, 2010), has been illustrated for the basic, simplest case. However, some extensions may be easily supported by our framework. We give a brief outline of the possible topics which may represent hints for development and future research. Firstly, when the real genomic location for each probe (gene/clone) is available (e.g. Huang *et al.*, 2005), the 'segmented' variable x_i is not simply the position index and the model should include this information. In this context, it is easy to verify that to 'convert' the jumpoint model (1) into the (continuous) segmented model (2), it suffices to consider $z_i=\sum_j^i y_j\Delta_j$ (rather than simply $z_i=\sum_j^i y_j$), where $\Delta_1=x_{(1)}$ and $\Delta_i=x_{(i)}-x_{(i-1)}$ for $i=2,\ldots,n$, and $x_{(i)}>x_{(i-1)}$. The rest of the algorithm remains unchanged. Quality of measurement data may exist in some platforms, see Lipson *et al.* (2005) for a discussion. The reliability for the outcome variable y_i is quantified by the value w_i, such that the larger w_i, the better the measurement. For instance, w_i is the inverse of the empirical SD of the pixels corresponding to y_i, i.e. associated with the same probe i. In wider statistical context, w_i represents the weight which may be included in the objective function, therefore to account for quality of measurements it suffices to estimate model (3) via weighted (rather than ordinary) least squares. Yet another development is the simultaneous analysis of multiple sequences: to deal with such extension the 'segmented' algorithm could include some modifications discussed in Muggeo (2008a) to deal with segmented relationships in each levels of some grouping variable. There is another possible generalization of our framework. Halpern (2000) discusses detecting multiple change point in binary 0/1 genetic sequence for the analysis of recombination of viruses in HIV studies. Here the response variable is Bernoulli and the underlying piecewise constant signal may be expressed on the logit scale, in the spirit of classical generalized linear models (GLMs). The standard iterative weighted least squares algorithm employed in estimation of usual GLMs may be modified by considering the cumulative sums of the working variate. Moreover, the penalized selection criterion (6) should be also modified, but only in the 'fidelity' term: the log variance should be replaced by the bernoulli deviance, remaining unchanged by the penalty term.

ACKNOWLEDGEMENT

The authors thank the reviewers for their useful comments and suggestions, which improved the quality of the article. Thanks also to Dr Marcello Tagliavia (Dip. di Biologia Cellulare e dello Sviluppo, Università di Palermo) for early reading of the draft of the manuscript.

Conflict of Interest: none declared.

REFERENCES

Albertson,D.G. and Pinkel,D. (2003) Genomic microarrays in human genetic disease and cancer. *Hum. Mol. Genet.*, **12**, R145–R152.

Barry,D. and Hartigan,J.A. (1993) A Bayesian analysis for change point problems. *J. Am. Stat. Assoc.*, **88**, 309–319.

Efron,B. *et al.* (2004) Least angle regression. *Ann. Stat.*, **32**, 407–489.

Eilers,P.H.C. and Menezes,R.X. (2004) Quantile smoothing of array CGH data. *Bioinformatics*, **21**, 1146–1153.

Erdman,C. and Emerson,J. (2008) A fast Bayesian change point analysis for the segmentation of microarray data. *Bioinformatics*, **24**, 2143–2148.

Fridlyand,J. *et al.* (2004) Hidden Markov models approach to the analysis of array CGH data. *J. Multivar. Anal.*, **90**, 132–153.

Halpern,A.L. (2000) Multiple-changepoint testing for an alternating segments model of binary sequence. *Biometrics*, **56**, 903–908.

Hsu,L.I. *et al.* (2005) Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, **6**, 211–226.

Huang,T. *et al.* (2005) Detection of DNA copy number alterations using penalized least squares regression. *Bioinformatics*, **21**, 3811–3817.

Huber,W. *et al.* (2006a) Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, **22**, 1963–1970.

Huber,W. *et al.* (2006b) Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, **22**, 1963–1970.

Hupé,P. *et al.* (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.

Johnson,N.A. *et al.* (2009) *cghFLasso: Detecting Hot Spot on CGH Array Data with Fused Lasso Regression.* R package version 0.2-1.

Jong,K. *et al.* (2003) Chromosomal breakpoint detection in human cancer. *Lect. Notes Comput. Sci.*, **2611**, 54–65.

Karkin,S. and Brendel,V. (1993) Patchiness and correlations in DNA sequences. *Science*, **259**, 677–680.

Lai,W.R. *et al.* (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**, 3763–3770.

Lipson,D. *et al.* (2005) Efficient calculation of interval scores for DNA copy number data analysis. *J. Comput. Biol.*, **13**, 215–228.

Lumley,T. (2009) *biglm: Bounded Memory Linear and Generalized Linear Models*. R package version 0.7.

Muggeo,V.M.R. (2003) Estimating regression models with unknown break-points. *Stat. Med.*, **22**, 3055–3071.

Muggeo,V.M.R. (2008a) Modeling temperature effects on mortality: multiple segmented relationships with common break points. *Biostatistics*, **9**, 613–620.

Muggeo,V.M.R. (2008b) Segmented: an R package to fit regression models with broken-line relationships. *R News*, **8**, 20–25.

Muggeo,V.M.R. (2010) *cumSeg: Change Point Detection in Genomic Sequences*. R package version 1.0.

Myers,C.L. *et al.* (2004) Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics*, **20**, 3533–3543.

Olshen,A.B. *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Picard,F. *et al.* (2005) A statistical approach for array CGH data analysis. *BMC Bioinformatics*, **6**.

Redon,R. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

R Development Core Team (2006) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Seligman,M. *et al.* (2010) *biglars: Scalable Least-Angle Regression and Lasso*. R package version 1.0.1.

Seshan,V.E. and Olshen,A. (2008) *DNAcopy: DNA Copy Number Data Analysis*. R package version 1.16.0.

Tibshirani,R. and Wand,P. (2008) Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, **9**, 18–29.

van de Wie,M.A. *et al.* (2007) CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics*, **23**, 892–894.

Veltman,J.A. *et al.* (2003) Array-based comparative genomic hybridization for genome-wide screening of dna copy number in bladder tumors. *Cancer Res.*, **63**, 2872–2880.

Venkatraman,E.S. and Olshen,A.B. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **23**, 657–663.

Wang,H. *et al.* (2009) Shrinkage tuning parameter selection with a diverging number of parameters. *J. R Stat. Soc. B*, **71**, 671–683.

Wang,P. *et al.* (2005) A method for calling gains and losses in array CGH data. *Biostatistics*, **6**, 45–58.

Willenbrock,H. and Fridlyand,J. (2005) A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, **21**, 4084–4091.