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 L1-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,
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 \((y_i, x_i)\) be the observed data; \(y_i\) is the outcome, e.g. the log fluorescence ratio, and \(x_i\) represents the genome location for \(i = 1, 2, \ldots, n\) genes/clones. As usual in statistical modeling, we assume the datum is the sum of the \(K\) differences in the mean levels at the change points; for instance \(K\) signal and noise, i.e.

\[
y_i = \beta_1 x_i + \delta (x_i > \bar{\psi}_1) + \cdots + \delta (x_i > \bar{\psi}_K) + \epsilon_i
\]

where \(\epsilon_i\) is noise which, depending on its ‘strength’, distorts the signal \(\beta_i\) and complicates change point detection. In Equation (1), \(\epsilon_i\) is the indicator function being equal to one when its argument is true, and the \(\psi\)'s represent the \(K\) change points on the genome, (e.g. the locations of the changes), \(\beta_1\) is the mean level (e.g. the log2 ratio) for \(x_i < \bar{\psi}_1\) and the \(\lambda\)'s are the differences in the mean levels at the change points; for instance \(\beta_1 - \beta_2\) is the mean level when \(\psi_1 < x_i < \psi_2\). Typically it is assumed \(\epsilon_i \sim X(0, \sigma^2)\), 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 \(i = 1, 2, \ldots, n\) 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 \(\psi_k\)'s) and the mean levels assumed in between (the regression coefficients \(\beta_1\) and \(\beta_2\)) 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

\[
z_i = \beta_1 y_i + \delta (y_i > \bar{\psi}_1) + \cdots + \delta (y_i > \bar{\psi}_K) + \nu_i
\]

where for each \(k\) and \(i \geq 1\), \(z_i = \sum_{j=1}^i y_j, z_i = \sum_{j=1}^i 1, \eta_i = \sum_{j=1}^i \epsilon_j, \text{and } (y_i - \psi_1) = \sum_{j=1}^i (y_j - \psi_1) = \hat{y}_k = \bar{y}_k - \bar{y}_k\). 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 \(\psi_k\). 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).

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

![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. (2009).](image-url)

starting values for the change points, \(\hat{\psi}_K\), the problem reduces to fit iteratively the linear model

\[
E[y_i | z_i] = \beta_1 y_i + \sum_{k=1}^K \hat{\beta}_k \hat{y}_k + \sum_{k=1}^K \hat{\gamma}_k \hat{y}_k^2.
\]

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 a 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. \(E[\epsilon_i] = 0\). 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 \(x_i\). 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 \(V_k\) is not admissible when

- it does not belong to the allowed range.
- Whether the generic value \(\hat{\psi}_k\) is not a change point, the relevant \(\hat{\beta}_i\) is approximately zero, leading the
specifically interested in the change points. As previously discussed, this can happen when the algorithm is started with penalized for the model dimension. However, at this aim we need to select only the 'significant' change points by removing the spurious ones. This approach may be considered essentially as a smoothing algorithm based on the 0-degree truncated power functions bases the (νk) with 'automatic' selection of the knots, i.e. the change points. The output is similar to that returned by the fused lasso (Tibshirani and Wang, 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 mean 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' (Willebrord and Friddlyand, 2005), 'CGHcall' (van de Wee et al, 2007) and 'false discovery rate' (FDR)-based approach used in Huang et al. (2005) and Tibshirani and Wang (2008).

While the fitted values (5) may be employed by any of the aforementioned approaches, it may be of interest to have 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≥K0, 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 Vk is not a change point, the corresponding covariate Vk should be a noise variable, as it would be Vk=0. Thus, selecting the number of change points actually reduces to selecting the significant variables among V1, V2,..., VK. 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 Vk is included in the model. The fitted 'optimal' model having K≤K* change points selected by any criterion, such as expression (6), is referred to 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 ei; 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

\[ BIC_{\text{G}} = \log(n) + \frac{1}{n} \sum_{i=1}^{n} \log(e_i) \]  

(where e_i is the residual variance estimate, edf is the actual model dimension quantified by the number of estimated parameters given by edf = 1 + 2(#change points) (including the intercept, the \( \delta \) and the \( \psi \)) and \( C_0 \) is a known constant. Note the generalized expression (6) reduces to the usual BIC when \( C_0 = 1 \). Wang et al. (2009) discuss the use of \( C_0 \geq 1 \) in identifying the true model when the number of parameters is not fixed but it diverges as n→∞. In the analysis of genomic sequences, it is reasonable to think that the number of change points depends on the sample length, therefore 'diverging' parameters appear to be appropriate here. Simulation studies presented later show that \( C_0 = \log(n) \) 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 \( \hat{\psi}_1, ..., \hat{\psi}_K \) from output 2 are actually a subset of the estimates \( \hat{\psi}_1, ..., \hat{\psi}_K \) from output 1, since one or more change points are not included due to deletion of one or more variables Vj by means of the selection criterion (6). However, it should be recognized that while \( \hat{\psi}_1, ..., \hat{\psi}_K \) are 'the best' estimates, in that they minimize the residual sum of squares for \( K^* \) change points, there is no guarantee that the subset \( \hat{\psi}_1, ..., \hat{\psi}_K \) constitutes also 'the best' estimate for the number of change points, \( K^* \).

In general, differences will be negligible; however, it is possible to refine the estimates for the model having \( K \) change points: at this aim it suffices to run again the segmented algorithm of Section 2.2 using \( \hat{\psi}_1, ..., \hat{\psi}_K \) as starting values. Notice that, due to the quite flat 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 fit the corresponding starting values (their values are not important, provided that K is large);
(2) Compute the variables U0, V0 for k = 1,...,K and fit the segmented linear model for the transformed response via ordinary least squares;
(3) Extract the coefficient estimates \( \hat{\beta}_k \) and \( \hat{\delta}_k \) to update the change point estimates; delete the variables (U0, V0) if the corresponding estimate \( \hat{\delta}_k \) 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 V0 and to select significant change points using the generalized BIC with \( C_0 = \log(n) \);
(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 \( C_0 \) (for instance 1, \( \sqrt{n} \log(n) \), \( \log(n) \)). 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 \( C_0 > 1 \) examined, \( C_0 = \log(n) \) 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; Oshen 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 focused 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 0.6, and variance 0.123); Gaussian heteroscedastic [zero mean and variance equal to 0.2π being as Unif(0.5, 3)] and Laplace iid (scale parameter equal to 0.5/√2). We have considered three sample sizes (n = 100, 500, 1000) and two values for the true number of change points (K0 = 0 and 4), being the true signal μ = 0 or μ = I(i > 3)−2I(i > 4)+1.5I(i > 8)−5I(i > 85); the ‘covariate’ i includes equi-spaced values ranging from 0 to 1.

It should be stressed that the performance in the ‘null’ case (K0 = 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 case (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 βs is small and, more importantly, it decreases along with the SD as the sample size increases. When there exist K0=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>
<thead>
<tr>
<th>n</th>
<th>K0=0</th>
<th>K0=4</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>m</td>
<td>0.03</td>
</tr>
<tr>
<td>500</td>
<td>m</td>
<td>0.03</td>
</tr>
<tr>
<td>1000</td>
<td>m</td>
<td>0.02</td>
</tr>
<tr>
<td>1000</td>
<td>m</td>
<td>2.64</td>
</tr>
<tr>
<td>500</td>
<td>m</td>
<td>13.23</td>
</tr>
<tr>
<td>1000</td>
<td>m</td>
<td>25.18</td>
</tr>
</tbody>
</table>

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 K0 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. Hu et al., 2005; Huang et al., 2005; Oshen 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.
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).

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 \( \text{lars} \) 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 cunseg (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’ not simply the position index and the model should include this the measurement. For instance, $y_i = \sum_j \gamma_j y_j$ (rather than simply $y_i = \sum_j y_j$), where $\Delta_1 = x_{i-1} - x_i$ and $\Delta_2 = x_i - x_{i-1}$ for $i = 2, \ldots, n$, and $x_{i+1} > x_i$. 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 example, $w_i$ is the inverse of the empirical SD of the pixels corresponding to $x_i$, 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 ‘tidelity’ 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


Huber,W. et al. (2006a) Transcript mapping with high-density oligonucleotide tiling arrays. Bioinformatics, 22, 1063–1070.


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


