Biometrics 63, 758–766
September 2007

DOI: 10.1111/j.1541-0420.2006.00729.x

A Segmentation/Clustering Model for the Analysis of Array CGH Data

F. Picard,* S. Robin, E. Lebarbier, and J.-J. Daudin

UMR INA P-G/ENGREF/INRA MIA 518, Paris, France
**email:* picard@inapg.fr

Summary. Microarray-CGH (comparative genomic hybridization) experiments are used to detect and map chromosomal imbalances. A CGH profile can be viewed as a succession of segments that represent homogeneous regions in the genome whose representative sequences share the same relative copy number on average. Segmentation methods constitute a natural framework for the analysis, but they do not provide a biological status for the detected segments. We propose a new model for this segmentation/clustering problem, combining a segmentation model with a mixture model. We present a new hybrid algorithm called dynamic programming–expectation maximization (DP–EM) to estimate the parameters of the model by maximum likelihood. This algorithm combines DP and the EM algorithm. We also propose a model selection heuristic to select the number of clusters and the number of segments. An example of our procedure is presented, based on publicly available data sets. We compare our method to segmentation methods and to hidden Markov models, and we show that the new segmentation/clustering model is a promising alternative that can be applied in the more general context of signal processing.

Key words: Array CGH; Dynamic programming; EM algorithm; Mixture models; Segmentation.

1. Introduction

Chromosomal aberrations often occur in solid tumors: tumor suppressor genes may be inactivated by physical deletion, and oncogenes activated via duplication in the genome. The purpose of array-based comparative genomic hybridization (array CGH) is precisely to detect and map chromosomal aberrations, on a genomic scale, in a single experiment. In a CGH experiment, genomic DNAs from a test and a reference sample are hybridized on a slide whose probes can be oligonucleotides, cDNAs, or BACs. The size of the probes and the distance between probes then define the resolution of the technique (see Davies, Wilson, and Lam, 2005, for a complete review of array CGH platforms and techniques).

A CGH profile is constituted when log ratios are plotted according to the physical position of their corresponding probe on the genome. Each profile can be viewed as a succession of segments that represent homogeneous regions in the genome whose probes share the same log ratio on average. The true copy number in the test sample is then inferred from the relative ratios. Array CGH data present two major characteristics. First, the signal is spatially ordered along the genome, and shows abrupt changes at unknown coordinates that correspond to breakpoints on the genome. The second characteristic is that the underlying biological process (deletion/amplification of DNA sequences) is discrete whereas the signal is continuous. Consequently, fluorescence log ratios are structured according to an unknown number of clusters at unknown levels, each cluster corresponding to a possible copy number. Regarding these characteristics, two major categories of analysis methods have been proposed.

One strategy is the search for abrupt changes in the signal to detect breakpoints on the genome. This is the purpose of segmentation methods (Hupe et al., 2004; Olshen et al., 2004; Picard et al., 2005). These methods focus on the ordered structure of the data and they aim to provide a partition of the data into segments characterized by a mean and a variance. Consequently they need the definition of an external threshold to cluster segments into biological groups. The clustering result is then dependent on the chosen threshold. Another strategy is to cluster genome-ordered probes into groups with biological relevance. This is the purpose of hidden Markov models (HMM; Fridlyand et al., 2004) and of the modified hierarchical clustering (CLAC) developed by Wang et al. (2005). More than pure clustering, these procedures consider the order of the data, with a Markovian distribution for the hidden variables in HMMs, and with a classification tree whose leaves are ordered along chromosomes for CLAC. Nevertheless, these method do not consider abrupt changes in the signal.

In a recent publication, Lai et al. (2005) reviewed and compared 11 methods for the analysis of array CGH data. They showed that segmentation methods perform consistently well for the identification of chromosomal aberrations even if the level of noise is high, indicating that the strategy of breakpoints identification is more efficient for array CGH data analysis. Considering this result and the limitations of segmentation methods mentioned above, we propose a segmentation/clustering (SegClust) model that combines a segmentation model and a mixture model to assign a biological status to segments. Section 2 is devoted to the precise definition of such

a model. In Section 3, we propose a hybrid algorithm called dynamic programming–expectation maximization (DP–EM) that combines DP and the EM algorithm to alternatively estimate the breakpoint coordinates and the parameters of the mixture. The convergence properties of this algorithm are presented. Once the parameters of the model have been estimated, a key issue is the estimation of the number of segments and of the number of clusters. One originality of our model is that this double model selection issue is new, and no method has yet been proposed. In Section 4, we propose a heuristic for this choice based on a penalized version of the likelihood.

In a last section, we propose to compare the performance of SegClust with HMMs and other segmentation methods on simulated and real data sets. Simulations are also used to validate the performance of our model selection procedure. We show the efficiency of our method in terms of clustering performance and breakpoint identification.

2. A New Model for Segmentation/Clustering

Let y_t be the $\log_2$ ratio of the tth probe on the genome and $y = \{y_1, \ldots, y_n\}$ the entire CGH profile constituted by n data points. We suppose that y is the realization of a Gaussian process Y, whose mean and variance are affected by $K-1$ abrupt changes at unknown coordinates $T = \{t_1, \ldots, t_{K-1}\}$ with the convention $t_0 = 1$ and $t_K = n$. This defines a partition of the data into K segments of length n_k. We write Y as $\{Y^1, \ldots, Y^K\}$, where $Y^k = \{Y_t, t \in I_k\}$, with $I_k = \{t, t \in]t_{k-1}, t_k]\}$. Then we add constraints on the values of the parameters, assuming that the mean and variance of segment Y^k can only take a limited number of values with $\mu_k \in \{m_1, \ldots, m_P\}$, and $\sigma_k^2 \in \{s_1^2, \ldots, s_P^2\}$. In addition to the spatial organization of the data, via partition T, there exists a secondary structure of the process into P clusters, and we adopt a mixture model approach to handle it.

We assume that the partitionned data $\{Y^1, \ldots, Y^K\}$ are structured into P clusters with weights $\pi_p (\sum_p \pi_p = 1)$. We introduce a sequence of independent hidden random variables, $Z^k = \{Z^k{}_1, \ldots, Z_P^k\}$ such that Z^k is distributed according to a multinomial distribution consisting of one draw on P categories with probabilities $\pi_1, \ldots, \pi_P$. The mixing proportions $\pi_1, \ldots, \pi_P$ thus represent the *prior* probability for segment Y^k to belong to the pth component, while the *posterior* probability of membership to the pth component with Y^k having been observed is: $\tau_p^k = \Pr\{Z_p^k = 1 \,|\, Y^k = y^k\}$. Contrary to classical mixture models, where the indicator variables provide information about the labeling of individual data points (which would be Y_t in our case), our model focuses on the belonging of the segments Y^k to different clusters.

We focus on the case where the data are supposed to be drawn from a mixture of Gaussian densities, with parameters $\theta_p = (m_p, s_p^2)$. Note that the appropriateness of the Gaussian distribution for array CGH data has previously been demonstrated by Hodgson et al. (2001) when the noise is moderate. The robustness of our model to this hypothesis will be studied in Section 5. If we suppose the independence of data points Y_t within a segment, the model can be formulated as follows:

$$Y^k \,\big|\, Z_p^k = 1 \sim \mathcal{N}_{n_k}\left(m_p \mathbf{1}_{n_k}, s_p^2 I_{n_k}\right).$$

We note $\psi = \{\pi_1, \ldots, \pi_{P-1}, \theta_1, \ldots, \theta_P\}$ the vector of unknown independent parameters of the mixture. The log likelihood of the model is:

$$\log \mathcal{L}_{KP}(T, \psi) = \sum_{k=1}^{K} \log \left\{ \sum_{p=1}^{P} \pi_p f\left(y^k; \theta_p\right) \right\}.$$

$f(y^k; \theta_p)$ represents the conditional density of a vector of size n_k. Our purpose is to optimize this likelihood to estimate the parameters of the model using a hybrid algorithm.

3. DP–EM: A Hybrid Algorithm to Optimize the Likelihood

The principle of DP–EM is as follows: when breakpoint coordinates T are known, the EM algorithm is used to estimate mixture parameters ψ, and once ψ has been estimated, breakpoint coordinates are computed using DP. This algorithm is run for fixed P and K.

3.1 *Estimating Breakpoint Coordinates When Mixture Parameters Are Known*

When the number of segments K and the parameters of the mixture are known, the problem is to find the best K-dimensional partition of the data according to the log likelihood $\log \mathcal{L}_{KP}(T, \psi)$. Because the number of partitions of a set with n elements into K segments is $\mathcal{C}_{n-1}^{K-1}$, and because of the additivity in K of the log likelihood, we use a DP approach to reduce the computational load from $\mathcal{O}(n^K)$ to $\mathcal{O}(n^2)$. This approach has been introduced by Auger and Lawrence (1989) and Picard et al. (2005) showed its efficiency in the context of array CGH data analysis.

Let $\hat{C}_{k+1,P}(i, j; \psi)$ be the maximum log likelihood obtained by the best partition of the data $Y^{ij} = \{Y_i, Y_{i+1}, \ldots, Y_j\}$ into $k+1$ segments, when the mixture parameters ψ are known. The algorithm starts as follows: for $k = 0$ and for $(i, j) \in [1, n]^2$, with $i < j$, calculate:

$$\begin{aligned}\hat{C}_{1,P}(i, j; \psi) &= \log \left\{ \sum_{p=1}^{P} \pi_p f\left(y^{ij}; \theta_p\right) \right\} \\ &= \log \left\{ \sum_{p=1}^{P} \pi_p \prod_{t=i+1}^{j} f(y_t; \theta_p) \right\}.\end{aligned}$$

$\hat{C}_1(i, j; \psi)$ represents the local log likelihood for segment Y^{ij}. Then the algorithm is run as follows:

$$\begin{aligned}\forall k \in [1, K] \qquad & \hat{C}_{k+1,P}(1, j; \psi) \\ & = \max_t \{\hat{C}_{k,P}(1, t; \psi) + \hat{C}_{1,P}(t+1, j; \psi)\}.\end{aligned}$$

More than a reduction in the computational load, this approach provides an exact solution for the global optimum of the likelihood that will be central for downstream model selection procedures.

3.2 *Estimating Mixture Model Parameters When Breakpoint Coordinates Are Known*

When breakpoint coordinates T are known, we have a partition of the data into K segments $\{Y^1, \ldots, Y^K\}$. The objective is to maximize the log likelihood of the model $\log \mathcal{L}_{KP}(T, \psi)$ according to ψ. The optimization of the likelihood can be

handled using the EM algorithm in the complete-data framework proposed by Dempster, Laird, and Rubin (1977). We note X^k the complete data vector for segment k, with $X^k = (Y_{t_{k-1}+1}, \ldots, Y_{t_k}, Z^k)$. Let us define the complete-data log likelihood:

$$\log \mathcal{L}^c_{KP}(T, \psi) = \sum_{k=1}^{K} \sum_{p=1}^{P} z_p^k \log \left\{ \pi_p f\left(y^k; \theta_p\right) \right\}.$$

The EM algorithm is as follows:

- *E*-step: compute the conditional expectation of the complete-data log likelihood, given the observed data Y, using the current fit $\psi^{(h)}$ for ψ.

$$Q_{KP}(\psi \,|\, \psi^{(h)}; T) = \sum_{k=1}^{K} \sum_{p=1}^{P} \tau_p^{k(h)} \log \left\{ \pi_p f\left(y^k; \theta_p\right) \right\},$$

$$\text{with} \quad \tau_p^{k(h+1)} = \mathbb{E}_{\psi^{(h)}}[z^k \,|\, Y] = \frac{\pi_p^{(h)} f\left(y^k; \theta_p^{(h)}\right)}{\sum_{\ell=1}^{P} \pi_\ell^{(h)} f\left(y^k; \theta_\ell^{(h)}\right)}.$$

- *M*-step: The M-step on the $(h+1)$ th iteration requires the global maximization of $Q_{KP}(\psi \,|\, \psi^{(h)}; T)$ with respect to ψ to give the updated estimate $\psi^{(h+1)}$:

$$\psi^{(h+1)} = \underset{\psi}{\text{Argmax}} \left\{ Q_{KP}\left(\psi \,|\, \psi^{(h)}; T\right) \right\}.$$

3.3 *Convergence Properties of the Hybrid Algorithm*

THEOREM 1: *The hybrid algorithm generates a sequence* $(T^{(\ell)}, \psi^{(\ell)})_{\ell \geq 0}$ *such that*

$$\log \mathcal{L}_{KP}(T^{(\ell+1)}, \psi^{(\ell+1)}) \geq \log \mathcal{L}_{KP}(T^{(\ell)}, \psi^{(\ell)}).$$

Proof. The proof of the convergence of our algorithm is based on the properties of both DP and EM. Both algorithms are linked through the likelihood they alternatively optimize: the incomplete-data likelihood of the mixture of segments. DP globally optimizes the likelihood with respect to T. So at iteration (ℓ) we have:

$$\log \mathcal{L}_{KP}(T^{(\ell+1)}; \psi^{(\ell)}) \geq \log \mathcal{L}_{KP}(T^{(\ell)}, \psi^{(\ell)}).$$

On the other hand, the key convergence property of the EM algorithm is the increase of the incomplete-data log likelihood at each step (Dempster et al., 1977):

$$\log \mathcal{L}_{KP}(T^{(\ell+1)}, \psi^{(\ell+1)}) \geq \log \mathcal{L}_{KP}(T^{(\ell+1)}, \psi^{(\ell)}).$$

Put together, our algorithm generates a sequence $(T^{(\ell)}, \psi^{(\ell)})_{\ell \geq 0}$ which increases the incomplete-data log likelihood and the result follows.

As for the complexity of DP–EM, it remains in $\mathcal{O}(n^2)$. The algorithm has been implemented using the `R` software (`http://www.R-project.org`) with C functions. We are currently developing a `R` package for distribution.

4. Model Selection

In practice neither the number of segments nor the number of clusters are known, and they should be estimated. Nevertheless, the joint estimation of the number of segments and groups is new and no method has yet been proposed. Classical model selection procedures are largely based on penalized likelihood criteria whose purpose is to establish a trade-off between a good quality of fit of the model to the data and a reasonable number of parameters to estimate. When using such criteria for model selection, it is assumed that the likelihood of the model increases with the number of parameters. In the following, we show that this property is not true for the SegClust model due to models nonnestedness. We show that the likelihood is not necessarily increasing with respect to the number of segments, whereas it increases with the number of clusters. This particular behavior motivates the construction of a two-step heuristic for model selection.

4.1 *Nested and Nonnested Models*

Let us denote $\mathcal{T}_K$ the set of possible breakpoints with K segments and T_K a particular configuration. Then we note Ψ_P the set of mixture parameters for P clusters. We define $\mathcal{M}(K, P)$ the set of all SegClust models with K segments and P clusters such that:

$$\mathcal{M}(K, P) = \{\mathcal{M}(T_K, \psi_P), T_K \in \mathcal{T}_K, \psi_P \in \Psi_P\}.$$

LEMMA 1:

$$\begin{cases} \mathcal{M}(K, P) \not\subset \mathcal{M}(K+1, P), \\ \mathcal{M}(K, P) \subset \mathcal{M}(K, P+1). \end{cases}$$

Proof. The fact that $\mathcal{M}(K, P)$ and $\mathcal{M}(K+1, P)$ are not nested is due to the discrete nature of breakpoints. Because segments of null size are not allowed in the model, it follows that $\mathcal{T}_K \not\subset \mathcal{T}_{K+1}$, which implies the first part of the lemma. For the second part of the proof it is clear that:

$$\forall T_K \in \mathcal{T}_K, \quad \Psi_P \subset \Psi_{P+1},$$

because a proportion π_{P+1} can be set to 0 to ensure the nesting, and the lemma follows.

From Lemma 1 it follows that the log likelihood increases with P whereas it may decrease with K. A classical result is that if models are nested, the likelihoods at their maximum increase. Nevertheless, no ranking can be inferred if models are not nested. Moreover, we provide an example for which the likelihood decreases.

Example. Let us assume that $\mathcal{M}(\hat{T}_K, \hat{\psi}_P)$ is the model that maximizes the likelihood for a given K and P and a given breakpoint configuration. Then we suppose that segment Y^ℓ belongs to cluster $p^\star$ such that $f(Y^\ell; \hat{\psi}) \simeq \hat{\pi}_{p^\star} f(Y^\ell; \hat{\theta}_{p^\star})$. The log likelihood of this configuration is:

$$\log \mathcal{L}_{KP}(\hat{T}_K; \hat{\psi}_P) \simeq \sum_{k \neq \ell}^{K-1} \log \left\{ \sum_{p=1}^{P} \hat{\pi}_p f(Y^k; \hat{\theta}_p) \right\} + \log \left\{ \hat{\pi}_{p^\star} f\left(Y^\ell; \hat{\theta}_{p^\star}\right) \right\}$$

Then consider a new configuration of breakpoints $\hat{T}_{K+1} = \{\hat{T}_K \cup t_{\text{new}}\}$ for which Y^ℓ is split into (Y^{ℓ_1}, Y^{ℓ_2}) without any change in the labeling. It follows that:

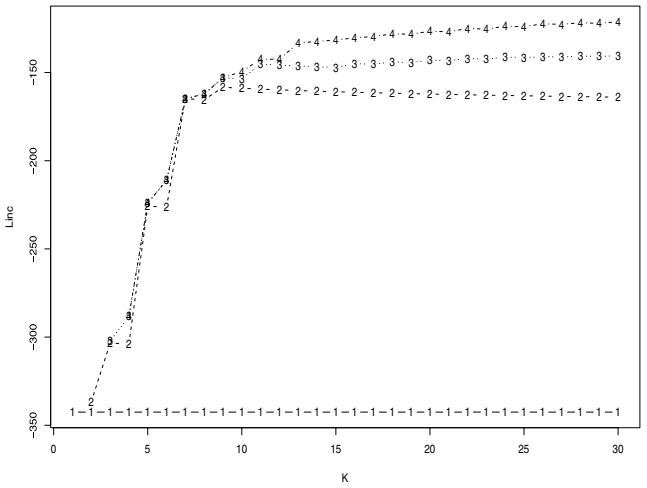

Figure 1. Incomplete-data loglikelihood $\log \mathcal{L}_{KP}$ according to K for different values of P. Data: Glioblastoma Multiform (GBM29, chromosome 7) analyzed in Section 5.

$$\log \mathcal{L}_{K+1,P}(\hat{T}_{K+1};\hat{\psi}_P) \simeq \sum_{k\neq \ell_1,\ell_2}^{K-1} \log\left\{\sum_{p=1}^{P} \hat{\pi}_p f(Y^k;\hat{\theta}_p)\right\} + \log\left\{\hat{\pi}_{p^\star} f\left(Y^{\ell_1};\hat{\theta}_{p^\star}\right)\right\} + \log\left\{\hat{\pi}_{p^\star} f\left(Y^{\ell_2};\hat{\theta}_{p^\star}\right)\right\}.$$

It follows that the log likelihood can decrease because:

$$\log \mathcal{L}_{K+1,P}(\hat{T}_{K+1};\hat{\psi}_P) - \log \mathcal{L}_{KP}(\hat{T}_K;\hat{\psi}_P) \simeq \log(\hat{\pi}_{p^\star}) \leq 0.$$

An example of such log likelihood is provided in Figure 1, where $\log \mathcal{L}_{KP}$ is decreasing when $P = 2$. Because SegClust models show a disymetrical behavior with respect to P and K we propose to adopt a two-step strategy for penalization, choosing the number of clusters first.

4.2 *Choosing the Number of Clusters*

The first step of the heuristic is to select the number of clusters whatever the number of segments. To do so, we propose to construct a sequence of increasing likelihoods as follows.

$$\text{Hypothesis (H): } \forall P \in \{1,\ldots,P_{\max}\}, \exists \tilde{K}_P, \quad \tilde{K}_P = \underset{K}{\text{Argmax}}\{\log \mathcal{L}_{KP}(\hat{T};\hat{\psi})\}.$$

THEOREM 2: *For a set of SegClust models with P clusters, $P \in \{1,\ldots,P_{\max}\}$ and K segments, under hypothesis (H) there exists a sequence of increasing log likelihoods noted* $\log \tilde{\mathcal{L}}_P$ such that $\log \tilde{\mathcal{L}}_1 \cdots \leq \log \tilde{\mathcal{L}}_P \leq \cdots \log \tilde{\mathcal{L}}_{P_{\max}}$ *with*

$$\log \tilde{\mathcal{L}}_P = \max_K \{\log \mathcal{L}_{KP}(\hat{T}_K;\hat{\psi}_P)\}.$$

The proof of this theorem is provided in the Supplementary Material section. The sequence $\{\log \tilde{\mathcal{L}}_P\}$ can be interpreted as the maximal fit that can be reached by a SegClust model with P clusters. This is why we use these likelihoods as targets to penalize.

In order to select P, we propose to use the adaptive method proposed by Lavielle (2005). Applied to our problem, this method aims at finding the number of clusters for which the log likelihood ceases to increase significantly. It is based on the calculus of the empirical second derivative of the log likelihood. Denoting $J_P = -\log \tilde{\mathcal{L}}_P$, the first step consists the calculus of $\tilde{J}_P$ such that:

$$\tilde{J}_P = \frac{J_{P_{\max}} - J_P}{J_{P_{\max}} - J_1} \times (P_{\max} - 1) + 1.$$

This normalization step ensures that $\tilde{J}_1 = P_{\max}$ and that $\tilde{J}_{P_{\max}} = 1$. In a second step, calculate:

$$\forall P \in \{2,\ldots,P_{\max}-1\}, \quad D_P = \tilde{J}_{P-1} - 2\tilde{J}_P + \tilde{J}_{P+1}.$$

Then select the number of clusters, such that:

$$\hat{P} = \begin{cases} \max_P\{P \in \{2,\ldots,P_{\max}-1\} \,|\, D_P \geq s\}, \\ 1 \text{ if } \forall P, D_P < s. \end{cases}$$

with s a threshold whose choice is discussed in Section 5. The default value of this tuning parameter is set to 0.75.

Lavielle (2005) showed that this selection procedure is equivalent to the use of the penalized criterion $\log \tilde{\mathcal{L}}_P - \beta_P \times pen(P)$, where $pen(P)$ is a increasing function with P (number of parameters of the mixture), and where β_P is estimated by $(\log \tilde{\mathcal{L}}_{P+1} - \log \tilde{\mathcal{L}}_P)/(pen(P+1) - pen(P))$.

4.3 *Choosing the Number of Segments*

Once the number of clusters has been chosen, the second step consists in the estimation of the number of segments $\hat{K}_{\hat{P}}$. Because breakpoint parameters are discrete the likelihood is not continuous with respect to these parameters. Therefore, classical model selection techniques cannot be applied to this case. If a Bayesian information criterion (BIC) criterion is derived to select $K_{\hat{P}}$, the breakpoints need to be fixed. In this case, the penalty is proportional to the number of *continuous* independent parameters of the model, which are the $3P - 1$ parameters of the mixture. This means that when the number of discrete parameters increases, the number of continuous parameters is constant for a given P. Consequently, deriving a penalty term in the Bayesian setting does not penalize the addition of new segments in the SegClust model. In order to circumvent this difficulty, we propose to use a *pseudo*-BIC criterion, which penalizes the addition of new segments as if they were continuous parameters. Therefore, we consider the following criterion to select the number of segments:

$$\hat{K}_{\hat{P}} = \underset{K}{\text{Argmax}} \left\{\log \mathcal{L}_{K\hat{P}}(\hat{T},\hat{\psi}) - \frac{1}{2}\log(n) \times K\right\}.$$

Because the model selection procedure we propose is heuristically based, it is evaluated based on its performance on simulated data.

5. Performance

Because many methods have been proposed for the analysis of array CGH data, it is crucial to compare their relative performance. Lai et al. (2005) proposed to compare 11 methods using simulated and real data sets. They point out that the comparison is not straightforward because the goals of the methods are different (i.e., clustering or breakpoint identification). The power of the SegClust model is that it combines the advantages of both strategies: it is based on a segmentation model that aims at finding abrupt changes, and the

mixture model directly clusters segments into a finite number of groups. In this section, we show that our model selection heuristic is performant, based on simulated data. We also show that the SegClust model is more efficient than HMMs from the clustering point of view, and as efficient as pure segmentation for the detection of the breakpoints.

5.1 *Simulated Data*

We use the simulated data set proposed by Lai et al. (2005). Simulations have been generated with a factorial design, factors of variations being the width of the amplification (0, 5, 10, 20, and 40 points) and the signal-to-noise ratio (SNR) of 0, 1, 2, 3, and 4. SNR is equal to μ/σ, where μ is the mean of the amplified segment, and σ is the standard deviation of a Gaussian noise (fixed at 0.25). For each combination, 100 artificial chromosomes with 100 data points were generated.

In a first step, we compare the performance of HMMs from Fridlyand et al. (2004) and SegClust in terms of clustering. To do so, we study the selection of the number of clusters P, the sensitivity (proportion of points truly assigned to the amplification), and the specificity (the proportion of points truly assigned to the unaltered group) in Figure 2. A first difference lies in the selection of P. While SegClust selects a constant number around 2 (with a slight overestimation), HMMs tend to overestimate P when SNR is high. This behavior may be linked to the use of Akaike information criterion (AIC) and BIC to select the number of hidden states, while the adaptive method we propose is more stable. When P is overestimated, HMMs should be more sensitive than SegClust with lower specificity. Nevertheless, we observed that the overestimation of P often leads to the creation of empty clusters. Moreover, we report estimation problems due to singular matrices as mentioned by Lai et al. (2005). The clustering performance of both methods could be summarized as follows: they are both highly specific, and SegClust shows a higher sensitivity which increases with the width of the aberration and the SNR.

In a second step we compare the performance of segmentation clustering in terms of breakpoints positioning, with other segmentation methods whose objectives are similar (CGHSeg from Picard et al., 2005 and CBS from Olshen et al., 2004). This study was not performed by Lai et al. (2005). To compare segmentation procedures, we study the estimated number of segments K and we check that the breakpoints are correctly located, using specificity and sensitivity (Figure 3). For the choice of the number of segments we observe three different behaviors. While CGHSeg tends to overestimate K when the size of the aberration and the SNR are small, CBS tends to underestimate this number and SegClust shows a stable estimation. This result shows the efficiency and the robustness of our model selection procedure. As a result, CBS is more conservative, leading to a lower sensitivity and a higher specificity, and the opposite is observed for CGHSeg. Despite slight differences regarding the methods, SegClust establishes

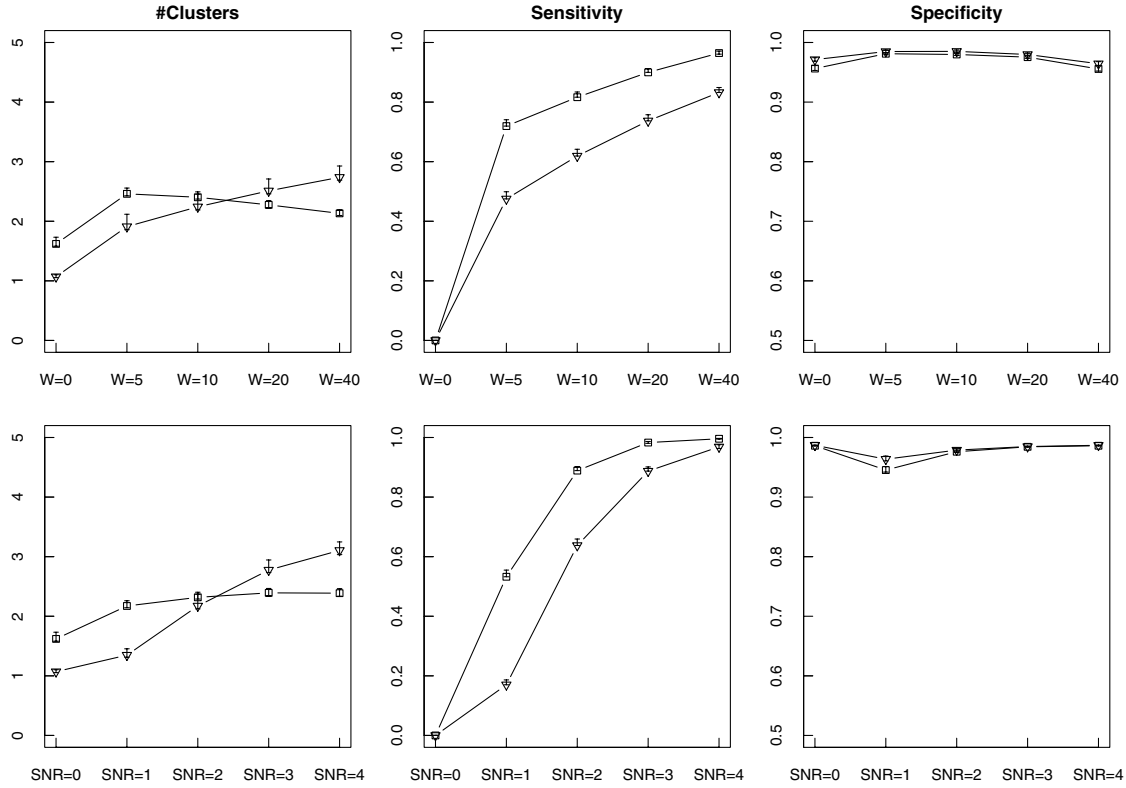

Figure 2. Estimated number of clusters, sensitivity, and specificity for clustering with different aberration widths and SNR. Top: average over width aberration, bottom: average over SNR. Bars correspond to 95% empirical confidence intervals. $-\square-$ SegClust, $-\triangledown-$ HMMs.

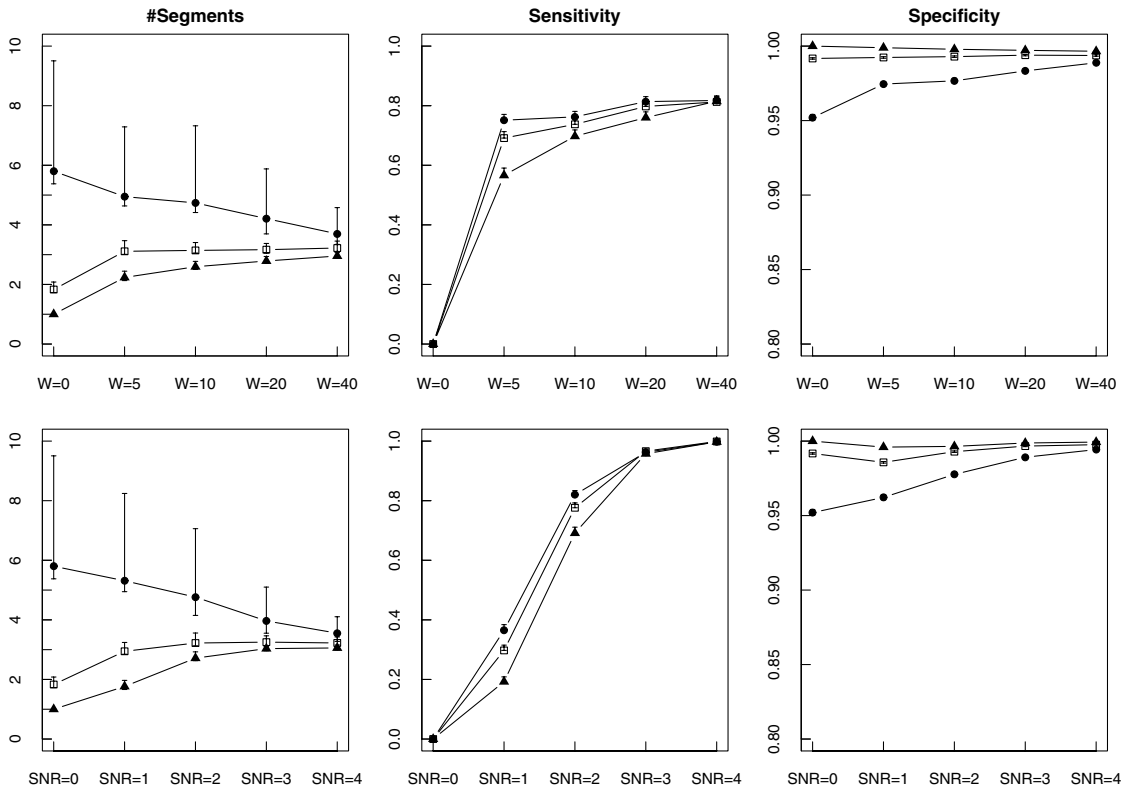

Figure 3. Estimated number of segments, sensitivity, and specificity for breakpoints positioning with different aberration widths and SNR. Top: average over width aberration, bottom: average over SNR. Bars correspond to 95% empirical confidence intervals. −●− CGHSeg, −□− SegClust, and −▲− CBS.

a good trade-off between a correct estimated number of segments and a good specificity/sensitivity.

In a last step, we study the robustness of SegClust to the normality assumption. We consider simulations with the same experimental design, but with a noise which follows a mixture of Gaussian distributions $(1-\alpha)\mathcal{N}(0,\sigma^2)+\alpha\mathcal{N}(0,4\sigma^2)$, where $\alpha = (0, 0.25, 0.5)$. Parameter α is used to increase the weight of the distribution tails. As shown in Figure 4, increasing the weight of the distribution tails leads to an increase in the estimated number of segments, as heavier tails of distribution lead to a higher dispersion of the data. Nevertheless, because the overestimation of P remains moderate (Figure 5), the addition of new segments does not lead to a decrease in terms of specificity/sensitivity for clustering, meaning that additional segments are correctly clustered. However, we observe a slight decrease in the specificity for segmentation, which remains higher than 0.95. Nevertheless, when parameter α equals 0.5, the distribution of the data is far from Gaussian. Consequently this result shows that the performance of SegClust are not sensitive to the normality assumption.

Finally, from the practical point of view, we notice that HMM and SegClust require more computational time than CBS and CGHSeg, because models with hidden variables require the use of iterative algorithms such as the EM algorithm. Moreover, HMMs and SegClust both need to estimate the parameters for different model dimensions in order to perform a downstream model selection procedure, which increases the computational time of both methods. However, the computational complexity of DP in $O(n^2)$ is not problematic, because it has recently been applied to tiling arrays (Huber, Toedling, and Steinmetz, 2006).

5.2 *Glioblastoma Multiform Data*

In order to compare the performance of different methods on real CGH data sets, we use the Glioblastoma Multiform (GBM) data that are described in Bredel et al. (2005) and were used by Lai et al. (2005). They consist of 26 samples cohybridized with pooled human controls onto cDNA microarrays. We focus on the GBM29 sample, chromosome 7 that is known to show at least three high amplifications around EGFR. This means that an appropriate method should detect six breakpoints and two clusters. Lai et al. (2005) compared the performance of CGHSeg, CBS, and HMMs on this example. They show that HMMs do not detect any amplification, whereas SegClust detects two clusters as shown in Figure 6. This result is in accordance with the simulation results, and shows that segmentation/clustering is more powerful than HMMs for the detection of aberrations. From the segmentation point of view, CBS does not detect the third segment that is only constituted of four clones. On the contrary, CGHSeg detects six breakpoints which correspond to biologically relevant events. Figure 6 (top) shows that SegClust detects an additional breakpoint because this probe shows a signal that is close to the mean signal of the unaltered group. We use this example to give some guidance for the choice of parameter s used to select the number of clusters. When this tuning

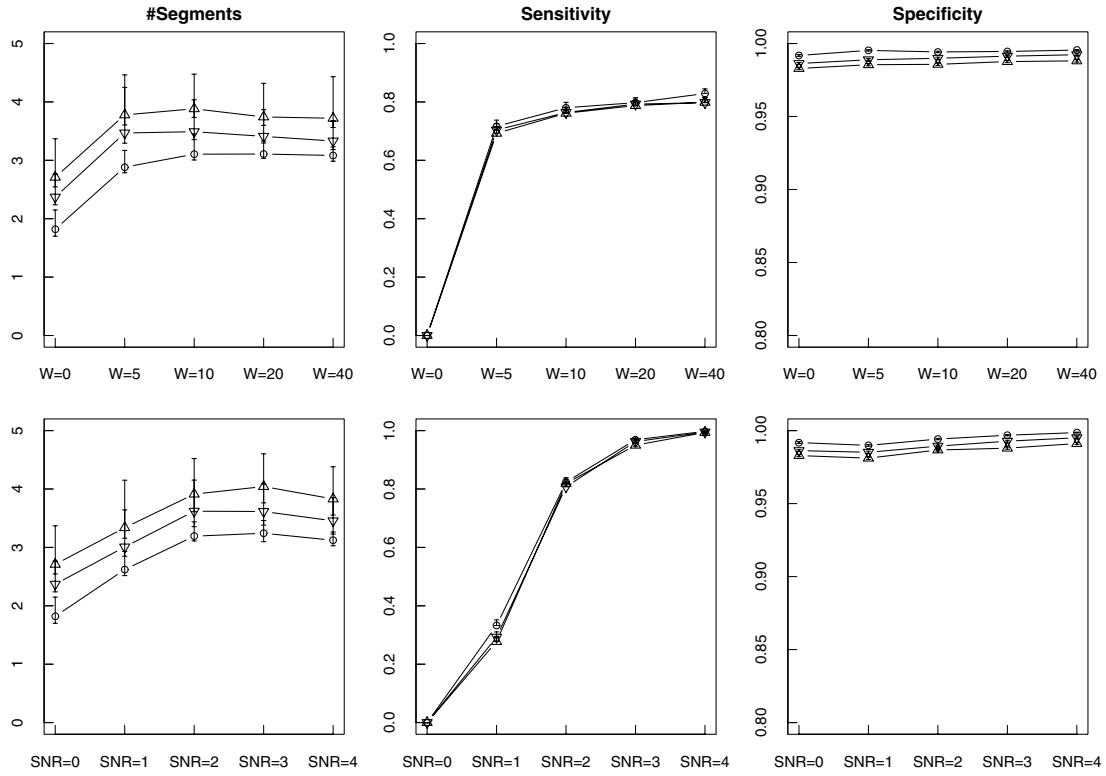

Figure 4. Robustness to the normality assumption. Estimated number of segments, sensitivity, and specificity for breakpoints positioning with different aberration widths and SNR. Top: average over width aberration, bottom: average over SNR. Bars correspond to 95% empirical confidence intervals. $-\circ-\ \alpha=0$, $-\triangledown-\ \alpha=0.25$, and $-\triangle-\ \alpha=0.5$.

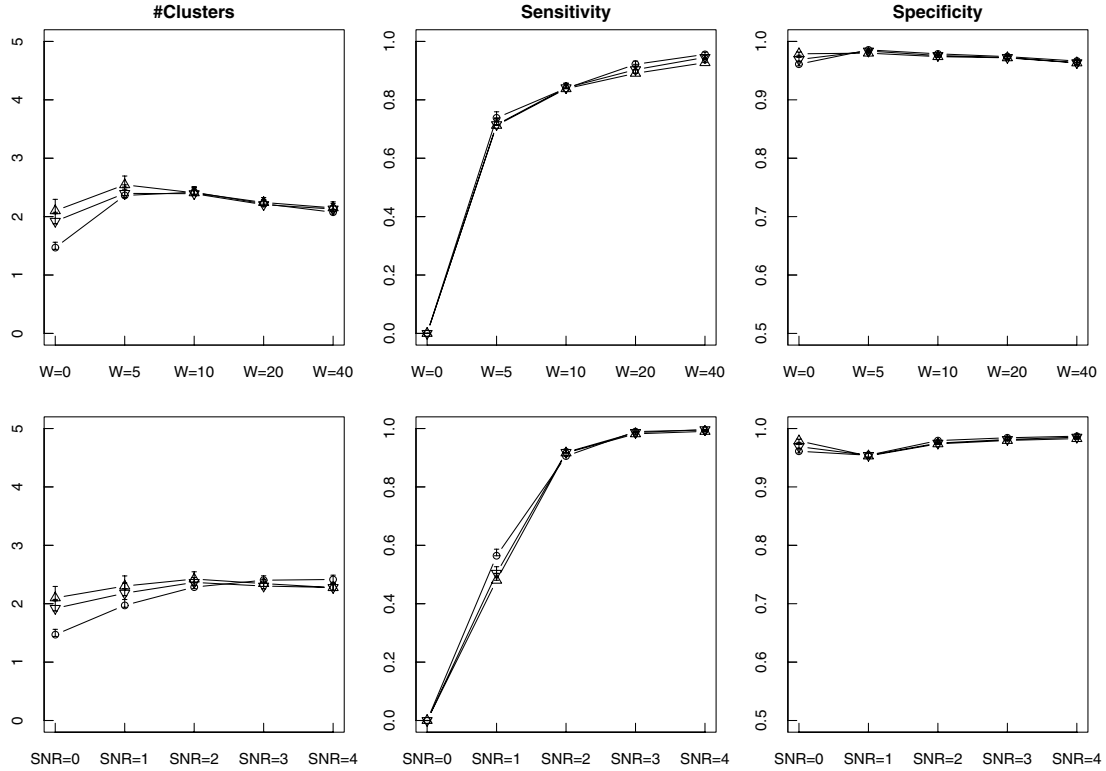

Figure 5. Robustness to the normality assumption. Estimated number of clusters, sensitivity, and specificity for clustering with different aberration widths and SNR. Top: average over width aberration, bottom: average over SNR. Bars correspond to 95% empirical confidence intervals. $-\circ-\ \alpha=0$, $-\triangledown-\ \alpha=0.25$, and $-\triangle-\ \alpha=0.5$.

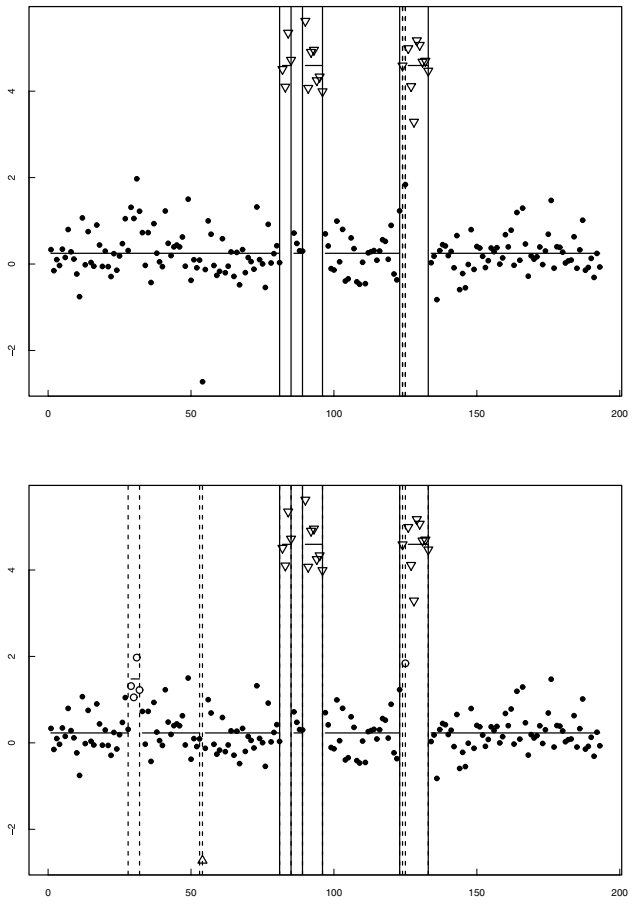

Figure 6. Array CGH profile of chromosome 7 in a Blioblastoma Multiform sample (GBM29). Comparison between CGHSeg and SegClust results. Solid vertical lines: common breakpoints, dashed vertical lines: additional breakpoints identified by SegClust. Top: $P = 2$ clusters ($s = 0.75$), bottom: $P = 4$ clusters ($s = 0.5$). • unaltered regions, ▽ high amplification, ○ low amplification, and △ deletion.

parameter is set to a lower value ($s = 0.5$, Figure 6, bottom), the selection procedure is less conservative, which leads to the selection of four clusters instead of two. In this case, a distinction is made between high- and low-level amplifications and a new cluster of one deleted clone is detected. This example shows the interest of a data-driven selection procedure which is flexible. In practice, this parameter is set to $s = 0.75$, which establishes a good trade-off between sensitivity and specificity. The simulation study was conducted with this value.

6. Conclusion

Microarray CGH currently constitutes the most powerful method to detect gain or loss of genetic material on a genomic scale. We introduced a statistical methodology for the analysis of CGH microarray data, that combines segmentation methods and clustering techniques. Because Lai et al. (2005) showed the efficiency of segmentation methods based on maximum likelihood and DP, we propose to refine this model with the addition of a hidden structure that corresponds to the biological status of genomic regions. In this article we show that the consideration of clusters could lead to segmentation results that are more precise compared with pure segmentation methods. We also show that our model handles the spatial structure of the data whereas HMMs can lead to unstable results when the level of noise is high. Consequently the new SegClust model we propose appears to be a promising alternative to HMMs. Moreover, this model can be applied to a wide variety of signals that are affected by abrupt changes and which show similar characteristics on different segments.

The definition of this new model leads to unusual statistical considerations: it appears that the statistical units of the mixture model (when the segmentation is known) are segments of different size. Because the partition of the data is random, the statistical units of the mixture model themselves are random. This explains the difficulty of the joint estimation of K the number of segments, and P the number of clusters, because classical model selection procedures are based on a trade-off between a reasonable number of parameters to estimate given a fixed number of statistical units. In this article we propose a model selection heuristic for this choice, whose performance has been validated on simulated and real data sets. Nevertheless, further theoretical developments would be valuable to handle this new model selection problem.

7. Supplementary Material

Appendix referenced in Section 4.2 is available under the Paper Information link at the *Biometrics* web site `http://www.tibs.org/biometrics`.

References

Auger, I. and Lawrence, C. (1989). Algorithms for the optimal identification of segments neighborhoods. *Bulletin of Mathematical Biology* **51,** 39–54.

Bredel, M., Bredel, C., Juric, D., Harsh, G., Vogel, H., Recht, L., and Sikic, B. (2005). High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. *Cancer Research* **65,** 4088–4096.

Davies, J., Wilson, I. M., and Lam, W. (2005). Array CGH technologies and their applications to cancer genomics. *Chromosome Research* **13,** 237–248.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* **39,** 1–38.

Fridlyand, J., Snijders, A., Pinkel, D., Albertson, D., and Jain, A. (2004). Hidden markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis* **90,** 132–153.

Hodgson, G., Hager, J., Volik, S., Hariono, S., Wermick, M., Moore, D., Nowak, N., Albertson, D., Pinkel, D., Collins, C., Hanahan, D., and Gray, J. (2001). Genome scanning with array cgh delineates regional alterations in mouse islet carcinomas. *Nature Genetics* **29,** 459–464.

Huber, W., Toedling, J., and Steinmetz, L. (2006). Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics* **22,** 1963–1973.

Hupe, P., Stransky, N., Thiery, J., Radvanyi, F., and Barillot, E. (2004). Analysis of array CGH data: From signal ratio to gain and loss of DNA regions. *Bioinformatics* **20,** 3413–3422.

Lai, W., Johnson, M., Kucherlapati, R., and Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21,** 3763–3770.

Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal Processing* **85,** 1501–1510.

Olshen, A., Venkatraman, E., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5,** 557–572.

Picard, F., Robin, S., Lavielle, M., Vaisse, C., and Daudin, J.-J. (2005). A statistical approach for CGH microarray data analysis. *BMC Bioinformatics* **6,** 27.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B., and Tibshirani, R. (2005). A method for calling gains and losses in array CGH data. *Biostatistics* **6,** 45–58.

Received December 2005. *Revised September* 2006.
Accepted September 2006.