A mixture model to characterize genomic alterations of tumors

Christine Keribin, Yi Liu, Tatiana Popova, Yves Rozenholc

Journal de la Société Française de Statistique
Vol. 160 No. 1 (2019)

A mixture model to characterize genomic alterations of tumors

Title: Un modèle de mélange pour caractériser les altérations génomiques tumorales

Christine Keribin[1,2], Yi Liu[2,3], Tatiana Popova[4] et Yves Rozenholc[2,5]

Résumé : La caractérisation des altérations du nombre de copies dans le génome est d'importance capitale pour développer une médecine personnalisée en cancérologie. Les puces à SNPs (Single Nucleotide Polymorphism), une variante de puce à ADN, sont toujours utilisées pour mesurer les profils d'altération du nombre de copies. Parmi les méthodes d'analyse de profil de SNPs, la méthode GAP (Genome Alteration Print) de Popova et al, basée sur une segmentation préliminaire de profils issus de puces SNPs, utilise une approche déterministe pour déterminer le profil du nombre absolu de copies. Nous développons un modèle probabiliste pour la méthode GAP et définissons un modèle de mélange gaussien dont les centres sont contraints d'appartenir à un réseau dépendant de paramètres inconnus tels que la proportion de tissu tumoral dans le prélèvement. L'estimation est effectuée à l'aide d'un algorithme EM (expectation-maximization) permettant d'accéder non seulement aux paramètres mais aussi au nombre altéré de copies le plus probable sur chaque segment ainsi que la proportion tumorale inconnue. Nous proposons de déduire la ploïdie tumorale en utilisant un critère pénalisé de choix de modèle. Notre modèle est testé sur des données simulées et appliqué à un exemple de données de cancer du côlon.

Abstract: Characterizing the genomic copy number alterations (CNA) in cancer is of major importance in order to develop personalized medicine. Single nucleotide polymorphism (SNP) arrays are still in use to measure CNA profiles. Among the methods for SNP-array analysis, the Genome Alteration Print (GAP) by Popova et al, based on a preliminary segmentation of SNP-array profiles, uses a deterministic approach to infer the absolute copy numbers profile. We develop a probabilistic model for GAP and define a Gaussian mixture model where centers are constrained to belong to a frame depending on unknown parameters such as the proportion of normal tissue. The estimation is performed using an expectation-maximization (EM) algorithm to recover the parameters characterizing the genomic alterations as well as the most probable copy number change of each segment and the unknown proportion of normal tissue. We claim to deduce the tumor ploidy from penalized model selection criterion. Our model is tested on simulated and real data.

Mots-clés : modèle de mélange, algorithme EM, critère BIC, heuristique de pente, cancer, méthode GAP, puce à ADN, SNP
Keywords: mixture model, EM algorithm, BIC criterion, slope heuristics, cancer, GAP method, SNP-array
Classification AMS 2000 : 62H12, 62P10

1. Introduction

Recent research reveals that personalized medicine is arguably one major way to treat cancer because of, for example, the immense diversity of underlying genomic alterations. In order

[1] Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France. E-mail : christine.keribin@math.u-psud.fr
[2] INRIA-Saclay Ile de France - Equipe SELECT.
[3] INSERM UMR-S1147 - Université Paris Descartes.
[4] INSERM U830 – Institut Curie.
[5] Université Paris Descartes - USPC, France - EA 7537 - BioSTM.

 ISSN: 2102-6238

to develop personalized medicine, characterizing the genomic copy number alterations (CNA) is a vital component. One way to characterize this alteration is to use a Single Nucleotide Polymorphism (SNP) array. A SNP is a variation in a single nucleotide (at one given locus in the genome) showing appreciable variability in the population : namely, >1% occurrence for the minor allele for the common SNPs, see Figure 1.

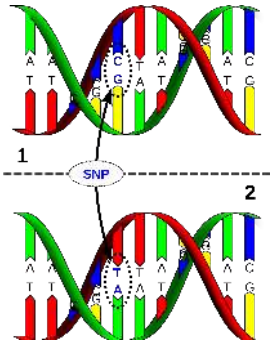

FIGURE 1. *A SNP example. Credit :* https://fr.wikipedia.org/wiki/Polymorphisme_nucléotidique

Arbitrarily, two variants are denoted as A- and B-allele. Since human chromosomes come in pairs, three genotypes AA, BB, AB (=BA) are possible to be called for each SNP in the individual genome. AA or BB genotypes refer to homozygous SNP, and AB (=BA) genotype refers to heterozygous SNP. SNP-arrays are used to measure the polymorphism in the predefined positions on a whole genome scale.

Using SNP-arrays in cancer, one can detect genomic alterations such as copy-number change (when a part of the chromosome is either deleted or duplicated more than once) and allelic imbalance (when part of chromosome has more copies of one allele). Copy number (CN) estimation is based on the signal intensity from the two alleles (A+B), while allelic imbalance (AI) is characterized by the "ratio" of signals (A/(A+B)). SNP-array analysis of a cancer sample is complicated by (1) presence of normal tissue, which diminishes the amplitude of CN and AI variation; (2) technical issues affecting signal-to-noise ratio; and (3) absence of "reference" point and relative nature of copy number alteration (CNA) profile. Measuring paired tumor-normal SNP-arrays allows overcoming the issues of signal-to-noise ratio and improving the detection of unknown proportion p of normal DNA in the tumor sample.

But now, recognition methods have been developed that perform CNA annotation without the normal pair. They proceeds in two steps : first, CN and AI profile segmentations, and second, annotation of each segment to have the actual CN and AI discrete states. GenoCNA (Sun et al., 2009), OncoSNP (Yau et al., 2010), and GPHMM (Li et al., 2011) employ a Hidden Markov Model (HMM) integrating both profile segmentation and CNA characterization in a single step. GAP (Popova et al., 2009) and ASCAT (Van Loo et al., 2010) adopt a such two-steps approach in which the CN and AI profiles are first segmented followed by an optimization step of a deterministic quality criterion with respect to proportion of normal DNA p. Taking into account AI and CN profiles, the criterion used in ASCAT (Van Loo et al., 2010) measures a weighted discrepancy based on several heuristics. Noticing that, for a given p, all possible alterations are precisely localized in a bi-dimensional plane (CN,AI), the GAP method (Popova et al., 2009)

 ISSN: 2102-6238

recovers manually the corresponding positions and allocates the mutated segments to a center given a predefined proximity criterion. Comparison of these methods (Mosén-Ansorena et al., 2012) shows that the two-steps approaches have better performance.

Using the pattern of CN and AI in the bi-dimensional plane introduced by Popova et al. (2009), we develop a family of parametric probabilistic models, each driven by the unknown proportion of normal tissue, and perform the estimation of its parameters, providing not only the most probable alteration types of each segment, but also a probabilistic distribution of these alterations. In each model, the estimation uses the maximization of the log-likelihood function with respect to p together with the estimation of the other parameters such as the variances of the observations. Moreover, our approach does not use any heuristic or any given tuning parameter. We expect our strategy to be not only satisfying from a mathematical point-of-view but also brings to the clinicians an expected probabilistic model for the alterations in a cancer genome.

After this introduction, section 2 describes the biological model and precises how the CNAs can be localized in a bi-dimensional plane. Section 3 defines the probabilistic model and the methodology to implement the EM algorithm in this case. Section 4 presents the results on synthetic and real data sets. The last section is devoted to a discussion.

2. Describing tumor CNA with SNP-arrays

For a given genomic locus, the tumor copy number state is characterized by the number of replicates of each alleles denoted u and v, with $u, v \geq 0$ (the value 0 corresponding to a deletion). At each SNP, given a germline (i.e. normal) status (AA, AB, or BB) characterized by the number n_A^g of allele A and n_B^g of allele B, the tumor genotype depends on the number of each alleles in tumor denoted n_A^t and n_B^t, which could be non-zero if and only if their normal counter part is also non-zero, as illustrated in Table 1.

germline → tumoral	n_A^g	n_B^g	n_A^t	n_B^t
(AA→ uAvA)	2	0	$u+v$	0
(BA→ uBvA)	1	1	v	u
(AB→ uAvB)	1	1	u	v
(BB→ uBvB)	0	2	0	$u+v$

TABLE 1. *Allele counts from u and v replicates of each strand*

The tumor genotype is characterized by the overall number of alleles, i.e. *copy number* (cn), and the number of B alleles or proportion of B alleles in the genotype, i.e. *B-allele frequency* (baf) for allelic imbalance. Measured tumor sample usually represents a mixture of tumor and normal cells. Hence, denoting the proportion of normal DNA as p, we have :

$$\mathrm{cn} = 2p + (1-p)(n_A^t + n_B^t) = 2p + (1-p)(u+v),$$

$$\mathrm{baf} = \frac{p\, n_B^g + (1-p)\, n_B^t}{2p + (1-p)(n_A^t + n_B^t)}.$$

 ISSN: 2102-6238

SNP-arrays SNP-arrays are a particular case of CGH-arrays. In the late 1990s, Solinas-Toldo et al. (1997) and Pinkel et al. (1998) demonstrated that it is possible to perform a Comparative Genome Analysis (CGH) on DNA fragments called probes fixed on glass slides, called chips. But this method needs to have two distinct samples, one containing normal cells while the other contains tumor cells, each of them being marked with a different fluorochrome (in general, tumor cell DNA is labeled in green with fluorescein and normal cell DNA is red with rhodamine).

With SNP-arrays, it becomes possible to have only one sample (Carr et al., 2008). Probes are oligonucleotides which are complementary fragments of the different alleles for each SNP. Fluorescently-tagged genomic DNA fragments combine preferentially to those oligonucleotides with which they are perfectly complementary. A computer reads the position of the fluorescent tags and identifies heterozygote SNP (two fluorescent tags) and homozygote SNP (one fluorescent tag), as well as the involved nucleotides (A, T, G, C). Hence, SNP-arrays are series of 4x1 vectors, grouped in 4x4 arrays (four SNP locii per 4x4 array), themselves grouped in more larger arrays. The current generation of chips includes more than 250 000 oligonucleotides.

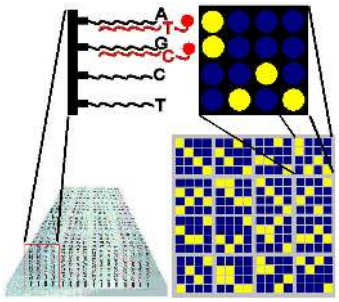

FIGURE 2. *Principle of SNP-array from Carr et al. (2008). Left : the nucleotide probes for a given SNP ; top right : the 4x4 array grouping the results for 4 different SNPs ; lower right : 256-oligo chip for 64 SNPs .*

Figure 2 sketches an example of a 256-oligo chip for 64 SNPs. The probes to combine alleles for a given SNP are represented on the left : an allele with a T binds to the oligonucleotide ending by A, and an allele with a C binds to the nucleotide oligonucleotide ending with a G, revealing a C/T heterozygote locus (first column of the 4x4 array). Similarly, the single spots in the other three columns of the 4x4 array indicate that the individual is homozygous at the three corresponding SNP positions. The 4x4 array fits into one corner of a 256-oligo chip for 64 SNPs.

The advantage of these arrays is to detect situations of loss of hererozygosity.These chips also give information on the number of copies of DNA at the loci studied.

Allelic imbalance and copy number signals from SNP-arrays The SNP-array technology provides measured baf profile related to the direct proportion of the B-allele signal. The value characterizing cn is log-*R-ratio* lrr, which is linked to the copy number by the relation

$$\text{lrr} = \alpha \log_2 \text{cn} + \beta,$$

where α is a contraction factor depending on the SNP-array platform and normalization methods. This parameter also depends on the experimental conditions and this is why it is not possible to have a general calibration. The parameter β is a constant shift due to unknown tumor ploidy (number of complete sets of chromosomes in a cell).

 ISSN: 2102-6238

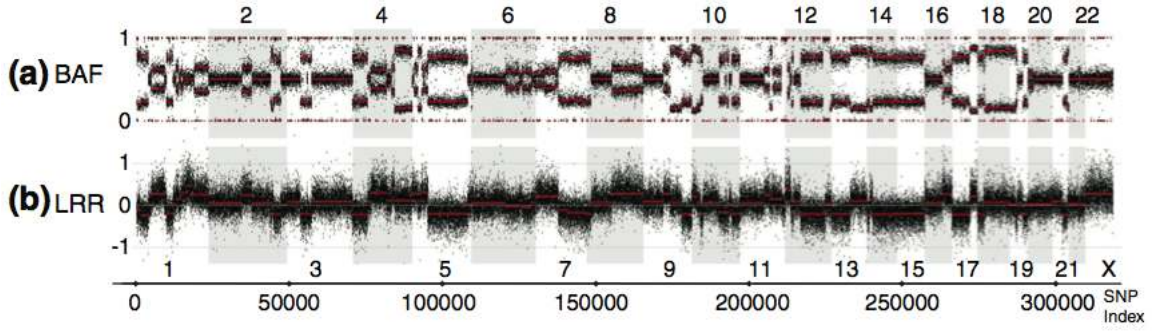

FIGURE 3. *Example of tumoral measurements, from Popova et al. (2009)*

Measured profiles from SNP-array data involve hundreds of thousands of SNPs characterized by values estimating cn and baf (Figure 3). Since neighboring SNPs tend to have the same B-allele and CN status, the baf and lrr signals in SNP-arrays are assumed to follow piece-wise constant distributions corresponding to the "large-scale" CNA profile in the tumor. "Large-scale" relates to the fact that SNPs are on average separated by a hundred to thousand nucleotides, and segments may have several ten thousands nucleotides.

Pattern of genomic alterations Hence, the profiles can be segmented into intervals with the same CNA and characterized by actual u and v. In a given segment with copy number cn, baf is 0 or 1 for a homozygous germline SNP AA or BB ; for a germline heterozygous SNP AB, tumoral alteration uAvB gives baf $= p(1-p)v/\text{cn}$ (see Table 1) and for a germline heterozygous SNP BA, tumoral alteration uBvA gives baf $= p(1-p)u/\text{cn}$: these two baf values are symmetrical with respect to 0.5. Labels A and B being attributed at random (their order is non informative) along a given segment, there are two symmetrical profiles that are observed on the BAF signal, as in Figure 3. The two symmetrical profiles are then aggregated to the $(0.5,1)$ interval, leading to consider only CNAs with $0 \leq u \leq v$.

One genomic segment with CNA k, characterized by $0 \leq u \leq v$, is associated with two centers $c_k^0 = (\text{baf}_k^0, \text{lrr}_k)$ and $c_k^1 = (\text{baf}_k^1, \text{lrr}_k)$ in the (baf, lrr) plane (Figure 4). The point c_k^0 corresponding to the germline heterozygous SNPs from the genomic segment k, satisfies

$$\text{baf}_k^0 = \frac{p+(1-p)v}{2p+(1-p)(u+v)}$$

and is denoted as (AB,uAvB). Similarly, the point c_k^1 corresponds to the germline homozygous SNPs with $\text{baf}_k^1 = 1$ and denoted as (BB,$(u+v)$B). Two CNAs k and k' leading to the same copy number share the same homozygous center on the (baf, lrr) plane, ie $c_k^1 = c_{k'}^1$. For example, the CNAs $(u=2, v=2)$ and $(u=1, v=3)$, having the same $\text{cn} = 4-2p$, occupy different heterozygous CNA centers $(AB, 2A2B)$ and $(AB, A3B)$, respectively, and share homozygous center $(BB, 4B)$ (Figure 4).

Note that the positions of the points (AB,uAvB) and (BB,$(u+v)$B) in the (baf, lrr) plane are uniquely defined by the unknown parameters p, α, and β.

 ISSN: 2102-6238

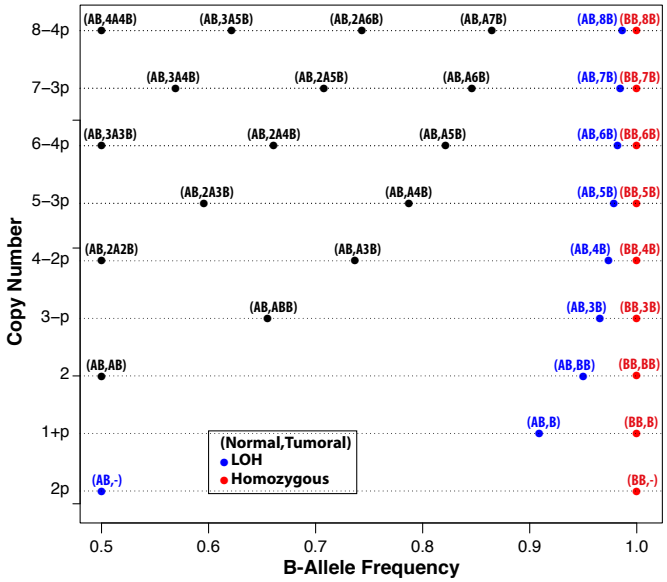

FIGURE 4. *Schematic illustration of the correspondence between the tumoral mutation and the* $(\mathrm{baf}, \mathrm{lrr})$ *values assuming* $0 \leq u \leq v$. *Mutations of germ line homozygous are in red. Mutations of germ line heterozygous are in black and blue. The blue centers characterize a loss of heterozygosity (LOH).*

3. Constrained mixture model for SNP tumoral mutations

Having at hand a segmentation into n intervals with homogeneous CNA, for the i-th interval ($i = 1, \ldots, n$) with N_i SNPs, we denote the number of heterozygous SNPs by N_i^0 and the number of homozygous SNPs by $N_i^1 = N_i - N_i^0$. Three summary variables are extracted from the SNP observations of the i-th interval : LRR_i the average over the N_i lrr observations, BAF_i^0 (resp. BAF_i^1) the average over the N_i^0 heterozygous (resp. N_i^1 homozygous) baf observations. Under the simple underlying assumption that the measurements coming from individual SNPs in an homogenous interval follow a distribution having finite first two moments and are independent, we can propose a Gaussian model for these measurements

$$\begin{aligned}
\mathrm{BAF}_i^0 &= \mathrm{baf}_{k(i)}^0 + \frac{\sigma}{\sqrt{N_i^0}}\varepsilon_i^0, \\
\mathrm{BAF}_i^1 &= \mathrm{baf}_{k(i)}^1 + \frac{\sigma}{\sqrt{N_i^1}}\varepsilon_i^1, \quad \text{with } \mathrm{baf}_{k(i)}^1 = 1, \\
\mathrm{LRR}_i &= \mathrm{lrr}_{k(i)} + \frac{\eta}{\sqrt{N_i}}\xi_i, \quad \text{with } \mathrm{lrr}_{k(i)} = \alpha \log_2(\mathrm{cn}_{k(i)}) + \beta,
\end{aligned}$$

where ε_i^0, ε_i^1, and ξ_i are Gaussian random variables with zero mean and unit standard deviation, and $k(i)$ denotes the CNA type of the i-th interval. Notice that BAF and LRR refer to random variables while baf and lrr refer to the deterministic unknown center coordinates.

The triplet $(\mathrm{BAF}_i^0, \mathrm{BAF}_i^1, \mathrm{LRR}_i)$ is recomposed into two bi-dimensional observations, corres-

 ISSN: 2102-6238

ponding respectively to the heterozygous and homozygous observations in the (baf, lrr) plane :

$$Y_i^j = (\text{BAF}_i^j, \text{LRR}_i) = c_{k(i)}^j + \zeta_i^j, \quad j = 0, 1$$

$c_{k(i)}^j = (\text{baf}_{k(i)}^j, \text{lrr}_{k(i)})$ is the theoretical unknown position of the homozygous ($j = 1$) or heterozygous ($j = 0$) CNA of segment i and ζ_i^j is a bi-dimensional centered Gaussian distribution with covariance matrix

$$\Sigma_i^j := \begin{pmatrix} \sigma^2/N_i^j & 0 \\ 0 & \eta^2/N_i \end{pmatrix}.$$

Define $K(L)$ the count of all different centers occurring with the maximum copy number L. We model the Y_i^j observation with a Gaussian mixture model of centers $c_k := (\text{baf}_k, \text{lrr}_k)$, $k = 1, \ldots, K = K(L)$. The probability density function is

$$f(Y_i^j) = \sum_{k=1}^{K(L)} \pi_k \phi(Y_i^j; c_k(\alpha, \beta, p), \Sigma_i^j(\sigma^2, \eta^2)).$$

where $\phi(., c, \Sigma)$ is the pdf of a bi-dimensional Gaussian random variable centered on c with covariance matrix Σ and $\pi_k \geq 0$, $k = 1, \ldots, K(L)$, $\sum_k \pi_k = 1$, are the mixture proportions. The centers are depending on p, α, β, and the parameter to estimate is

$$\theta = (p, \sigma^2, \eta^2, \alpha, \beta, \{\pi_k, k = 1, \ldots, K(L)\}).$$

Using the independence of the observations $Y = \left\{ Y_{i=1,\ldots,n}^{j=0,1} \right\}$ between segments, the log-likelihood is

$$\mathscr{L}(\theta; Y) = \sum_{j=0}^{1} \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K(L)} \pi_k \phi(Y_i^j; c_k(\alpha, \beta, p), \Sigma_i^j(\sigma^2, \eta^2)) \right).$$

Comments on the assumptions In developing the model, we made the following assumptions : (i) the noises in B-allele frequency and log ratio individual measurements have a finite second moments ; (ii) the contraction factor α and shift constant β of LRR is the same for all measurements ;(iii) the variance of B-allele frequency and of the log ratio individual measurements do not depend on the mutation type.

The first assumption is very weak in that no assumption about the form of the underlying distributions is used, and is thus applicable to a wide range of measurement platforms. The Gaussian mixture model follows from the central limit theorem when obtaining the segmented data by averaging over the homogeneous intervals. In the second assumption, the correction of GC content is neglected since this can be treated in the segmentation step (Staaf et al., 2008).

3.1. Estimation

We first assume that the maximum copy number is known. Maximum likelihood estimation for mixture models is traditionally performed with an expectation-maximization (EM) algorithm (Dempster et al., 1977), and we introduce the latent variables z_{ik}^j such that $z_{ik}^j = 1$ if $(\text{BAF}_i^j, \text{LRR}_i)$ is from mutation center c_k, 0 otherwise.

 ISSN: 2102-6238

In our case, the E step that computes the expectation, conditionally to the observations and under a given current parameter, of the complete log-likelihood

$$\mathscr{L}_c(\theta;Y,Z=\{z_{ik}^j\}) = \sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{k=1}^{K} z_{ik}^j \left[\log \pi_k + \log \phi(Y_i^j; c_k(\alpha,\beta,p), \Sigma_i^j(\sigma^2,\eta^2))\right]$$

is standard and reduces to compute the conditional expectation of the labels z_{ik}^j. However, if the M step is also straightforward and closed form for parameters $(\sigma^2,\eta^2,\alpha,\beta)$, it is not the case for p, that introduces non linearity. As using a nonlinear optimization procedure inside the M step did not give satisfactory results, we designed a two nested levels procedure :

$$\max_{\theta} \mathscr{L}(\theta;Y) = \max_{p} \max_{(\alpha,\beta,\pi,\sigma^2,\eta^2)} \mathscr{L}\left(p,\sigma^2,\eta^2,\alpha,\beta,\pi;Y\right).$$

Hence, we use an EM algorithm to deal with the maximization over $(\alpha,\beta,\pi,\sigma^2,\eta^2)$ nested in a gradient descent over p.

EM for fixed p and fixed range of mutations Given a fixed number of CNA centers corresponding to copy numbers belonging to the interval $[0,L]$ and a fixed p value, the EM iterates two steps.

The expectation step computes the conditional expected value of z_{ik}^j given the parameter obtained in the previous iteration denoted $\breve{\theta} = (\breve{\alpha},\breve{\beta},\breve{\sigma}^2,\breve{\eta}^2,p)$,

$$\tau_{ik}^j \leftarrow E(z_{ik}^j|Y;\breve{\theta}) = \frac{\breve{\pi}_k \phi(Y_i^j;\breve{c}_k,\breve{\Sigma}_i^j)}{\sum_k \breve{\pi}_k \phi(Y_i^j;\breve{c}_k,\breve{\Sigma}_i^j)}.$$

Using this updated value of τ_{ik}^j, the maximization leads to update the parameters according to

$$\begin{aligned}
\pi_k &\leftarrow \frac{\sum_{i,j}\tau_{ik}^j}{\sum_{i,j,k}\tau_{ik}^j},\\
\sigma^2 &\leftarrow \frac{\sum_{i,j,k}\tau_{ik}^j N_i^j (\mathrm{BAF}_i^j - \mathrm{baf}_k)^2}{\sum_{i,j,k}\tau_{ik}^j},\\
\alpha &\leftarrow \frac{CD-BE}{AC-B^2},\\
\beta &\leftarrow \frac{BD-AE}{B^2-AC},\\
\eta^2 &\leftarrow \frac{\sum_{i,j,k}\tau_{ik}^j N_i (\mathrm{LRR}_i - \alpha\log_2 \mathrm{cn}_k - \beta)^2}{\sum_{i,j,k}\tau_{ik}^j},
\end{aligned}$$

 ISSN: 2102-6238

where

$$A = \sum_{i,j,k} \tau_{ik}^j N_i (\log_2 \mathrm{cn}_k)^2,$$
$$B = \sum_{i,j,k} \tau_{ik}^j N_i \log_2 \mathrm{cn}_k,$$
$$C = \sum_{i,j,k} \tau_{ik}^j N_i,$$
$$D = \sum_{i,j,k} \tau_{ik}^j N_i \mathrm{LRR}_i \log_2 \mathrm{cn}_k,$$
$$E = \sum_{i,j,k} \tau_{ik}^j N_i \mathrm{LRR}_i.$$

The above two steps are repeated until convergence criterion is met.

Initialization of the parameters Because the log-likelihood function in the mixture model is not globally convex, the performance of the EM algorithm is sensitive to the choice of initial values of parameters. In our implementation, σ^2 and η^2 are initialized with the observed variance of the BAF and LRR signals :

$$(\sigma^2)^0 = \mathrm{var}(\mathrm{BAF})$$
$$(\eta^2)^0 = \mathrm{var}(\mathrm{LRR})$$

As the minimum and maximum values of the BAF signal are roughly delineated by the centers position of the minimum and maximum copy numbers, one initial value for α can be set as

$$\alpha^0 = \frac{\mathrm{LRR}_{\max} - \mathrm{LRR}_{\min}}{\log_2 \mathrm{cn}_{\max} - \log_2 \mathrm{cn}_{\min}}$$

and provides good results when minimum and maximum copy numbers are known. If it is not the case, it is interesting to use all combinations of cn_l and cn_h such that $0 \leq \mathrm{cn}_l < \mathrm{cn}_h \leq L$ to provide different initial values for α

$$\alpha^0 = \frac{\mathrm{LRR}_{\max} - \mathrm{LRR}_{\min}}{\log_2 \mathrm{cn}_h - \log_2 \mathrm{cn}_l}.$$

The initial values for β are then derived according to

$$\beta^0 = \mathrm{LRR}_{\max} - \alpha^0 \log_2 \mathrm{cn}_h.$$

Now, using the knowledge of p, α^0 and β^0, we compute the centers c_k and affect each Y_i^j to the closest center c_k, giving access to an estimation of the mixing weights π_k, $k = 1, \ldots, K$.

Maximization with respect to p The behavior of the expectation, conditionally to the observations, of the complete log-likelihood viewed as a function of p is smooth though not necessarily convex globally. Hence we use a grid to start the search of the optimal value of p.

 ISSN: 2102-6238

Probabilistic mutation characterization per segment The Maximum A Posteriori rule from the conditional expectations τ_{ik}^j is a natural way to characterize the segment mutation, by attributing each segment to its most probable mutation center.

This is a real added value compared to Popova et al. (2009) : their hand made process is not only automated by using our mixture model, but also improved, as the segments are not attributed to the nearest center, but to the most probable CNA.

3.2. Model selection with penalized log-likelihood

As so far, the estimation is performed with all CNA centers corresponding to copy numbers belonging to the interval $[0, L]$. Two model selection cases may be considered : (1) for a given copy number, are there CNA centers that do not appear ? (2) Which are the minimum and maximum copy numbers appearing in the sample ? As the positions of the centers depend on the proportion of normal DNA p, targeting these two goals simultaneously can lead non identifiable situations that cannot be distinguished. This situation has already be discussed and Popova et al. (2009) proposes to choose the configuration with the lower maximum copy number in such case.

We adopt a penalized log-likelihood approach to select the maximum copy number, and first use BIC as it is known to be a suited model section criterion for simple mixture models (Keribin, 2000). The only parameter depending on the copy number being the mixing weights π, the model complexity is $K(L)$. We deduce the following expression of BIC :

$$BIC(L) = -\mathscr{L}(\widehat{\theta}; Y) + \frac{\log(2n)}{2} K(L)$$

where $2n$ stands for the number of observations, two times the number n of segments, to take into account the homozygous and heterozygous sites. The model with the smallest BIC is chosen.

BIC is an asymptotic approximation of the logarithm of the integrated likelihood in a Bayesian context. As an alternative, Birgé and Massart (2001), Birgé and Massart (2007) proposed a non-asymptotic method called the slope heuristics, a data-driven method which allows to calibrate a multiplicative constant $\kappa > 0$ in the penalized criterion defined in our context as follows :

$$crit(L; \kappa) = -\mathscr{L}(\widehat{\theta}; Y) + \kappa K(L)$$

Indeed, the log-likelihood in expectation should behave linearly with respect to the complexity for large models, as more complex models do not provide any further improvement of the bias. An effective choice for the constant κ is shown to be twice the slope provided by this linear behavior. An equivalent methodology is to use twice the penalty which brings the most abrupt change in the complexity of the selected model. See Baudry (2009) for application to mixtures and Arlot (2019) for a comprehensive survey.

Finally, we propose to deduce the tumoral ploidy as a weighted average of the copy numbers :

$$\frac{\sum_{i,j,k} \tau_{ik}^j N_i^j \, \mathrm{cn}_k}{\sum_{i,j,k} \tau_{ik}^j N_i^j}.$$

Note that we do not intend, for a given copy number, to refine the model by suppressing CNA centers, but consider that the identity card of the tumor is described by the estimated mixture weights $\widehat{\pi}$ once the maximum copy number is selected by the model criteria.

 ISSN: 2102-6238

4. Results

We first test our methodology on simulated data, then apply it to a real cancer data set.

4.1. Simulated data

To evaluate the performance of the algorithm, we devised a strategy to generate simulated data with a maximum copy number up to $\text{cn}_{\max} = 8$ for the tumor tissues and compared the estimation results with the real parameter used to generate the data sets. Once given values for θ, the data generation strategy is as follows :

1. Generate randomly n segments on a profile of N SNPs. We choose $n = 200$ and $N = 261976$ as values of the real SNP-array experiment.
2. For each segment, generate independently the number (u or v) of each allele following a multinomial distribution $\mathscr{M}(1,\mu)$ on $\{0,1,2,3,4\}$ with $\mu = (0.15, 0.5, 0.2, 0.1, 0.05)$.
3. Compute baf and lrr correspondingly and symmetrize baf into the interval $[0.5, 1]$.
4. On each segment, generate the number of homozygous SNPs following a binomial distribution of $P(homozygous) = 0.8$.
5. Add noise to the baf and lrr values and form the final observations with given α and β.

A data set was generated with parameter $\theta = (p = 0.1, \alpha = 1, \beta = 0, \sigma^2 = 0.04, \eta^2 = 0.25)$ and maximum copy number 8. Some centers are not occupied (Figure 5 top left).

Models with $\text{cn}_{max} \in [3, 10]$ were used for the estimation. Using the BIC criterion, the model with maximum copy number 8 is selected, corresponding to the underlying parameter. The same holds with the adaptive penalized criterion based on the slope heuristics. With this model, the parameter estimation gives $\hat{\theta} = (\hat{p} = 0.1, \hat{\alpha} = 1, \hat{\beta} = -0.00316, \hat{\sigma}^2 = 0.0404, \hat{\eta}^2 = 0.255)$, which is in good agreement with the underlying data (Figure 5 top right). Also, the classification error rate based on the MAP rule is 0.025 and the selected CNAs are displayed on Figure 5 bottom right.

4.2. Influence of the normal DNA proportion

To study the influence of the proportion of normal DNA p on the previous example, we generated a series of tumor samples with the same CNA profile but with different proportions of normal cells, similar to a diluted cell line samples. Nine samples were generated with $p = (0.1, 0.2, \ldots, 0.9)$ respectively. The other parameters were set to be $\sigma^2 = 0.04$, $\eta^2 = 0.25$, $\alpha = 1$, and $\beta = 0$ in all the data sets. Among copy number from 3 to 14, both BIC and slope heuristic criterion choose the model with maximum copy number 8 in all nine data sets, and the parameter estimation is good even with a high normal tissues contamination ($p = 0.9$) for which the inference is more difficult, see Table 2. In this case, our method still shows a relatively consistent behavior and the MAP classification error is 0.13.

We performed Monte Carlo sampling to assess the variability of the estimation for three values of normal DNA proportion : $p = 0.1125, 0.5125, 0.8125$. Results in Table 3 show that the algorithm is rather robust against the normal tissue proportion. A moderate proportion of normal

 ISSN: 2102-6238

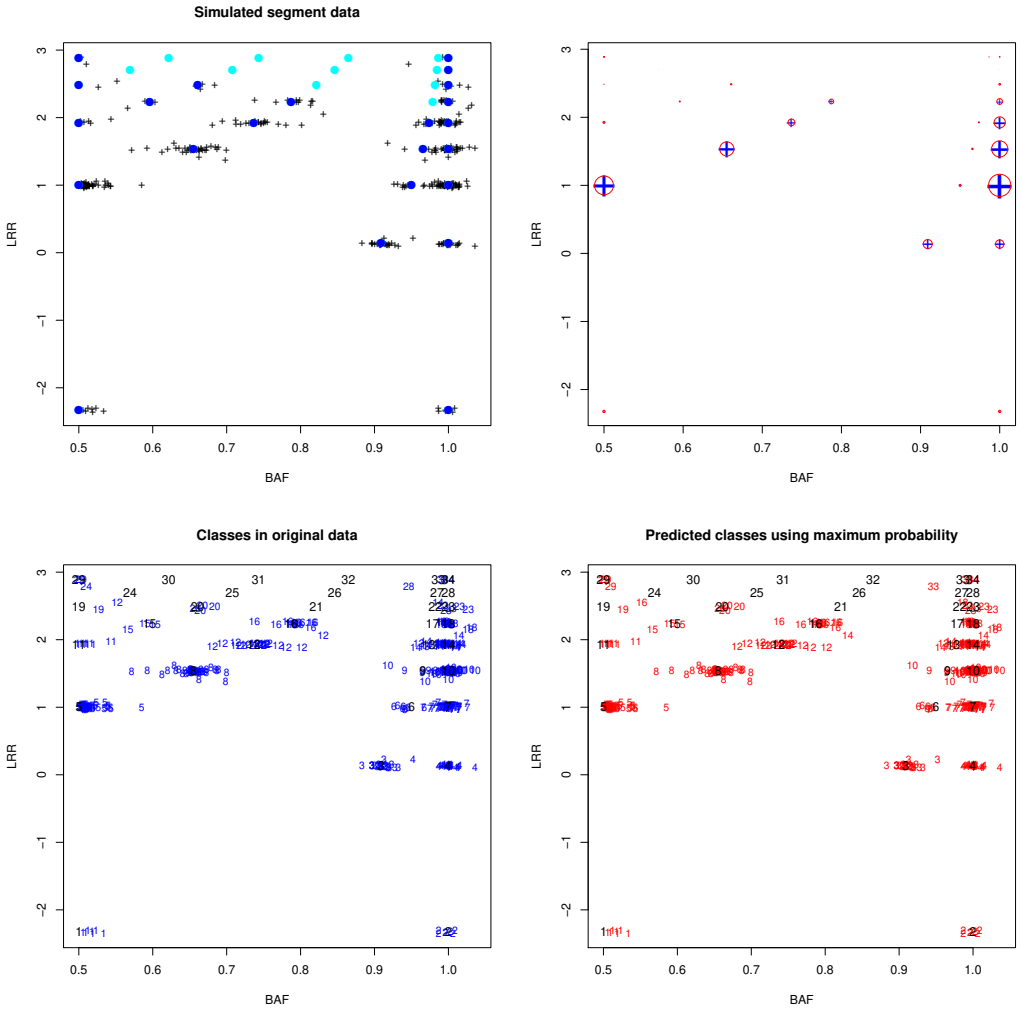

FIGURE 5. *Top left : representation of the simulated data with* $\theta = (p = 0.1, \sigma = 0.2, \eta = 0.5, \alpha = 1, \beta = 0)$ *in the* (baf, lrr) *plane. Black crosses represent the independent observations, blue dots the CNA centers used to generate the observations, cyan dots non-occupied centers in the complete model. Top right : parameter estimation on the simulated data. Blue crosses represent the position of the CNA centers and the relative proportion with its size. Red circles represent the estimated position of the CNA centers and their relative proportion. Bottom left : the class label of the simulated data. Bottom right : the class label based on maximum a posterior probability.*

 ISSN: 2102-6238

p	$\hat{p}$	$\hat{\sigma}^2$	$\hat{\eta}^2$	$\hat{\alpha}$	$\hat{\beta}$
0.1	0.1	0.0350	0.215	0.999	0.00180
0.2	0.2	0.0342	0.216	0.999	0.00191
0.3	0.3	0.0342	0.216	0.999	0.00202
0.4	0.4	0.0347	0.216	0.999	0.00216
0.5	0.5	0.0349	0.216	0.999	0.00235
0.6	0.6	0.0348	0.218	0.998	0.00271
0.7	0.7	0.0340	0.215	0.997	0.00353
0.8	0.8	0.0334	0.208	0.996	0.00466
0.9	0.9	0.0306	0.207	0.998	0.00231

TABLE 2. *Estimates parameters on simulated diluted data sets with different proportions p of normal DNA.*

DNA being even conducived for the estimation of genomic CNAs, which is in accordance with Popova et al. (2009).

p	$\lvert\hat{p}-p\rvert$	Error classification rate
0.1125	$0.0125(1.80e-18)$	0.0278(0.009)
0.5125	$0.0125(5.10e-17)$	0.013(0.0032)
0.8125	$0.0125(5.10e-17)$	0.056(0.014)

TABLE 3. *Monte Carlo simulation for the estimation of p in three cases. The values are shown in the format mean (standard-deviation).*

4.3. Influence of nuisance parameters

Three values of the BAF standard error $\sigma = 0.2, 1.5, 3$ are used in the Monte Carlo sampling to determine the influence of this parameter on the estimation. Results are shown in Table 4. This parameter has a large influence on the quality of estimation result. This is because, contrary to the LRR measurements which are fixed not only by the underlying mutation type and p, but also by α and β, BAF measurements are uniquely determined by p and the mutation type. Thus a large variance of BAF will deteriorate greatly the estimation result.

The LRR standard deviation η does not influence greatly the estimation result because the BAF measurements still provide enough information for the mutation types and normal tissues proportion in the data sets. This appears on results in Table 5, where Monte Carlo simulation are done for three different values $\eta = (0.5, 5, 10)$.

σ	$\lvert\hat{p}-p\rvert$	Error classification rate
0.2	$0.0125(1.80e-18)$	0.0278(0.009)
1.5	0.281(0.353)	0.520(0.383)
3	0.21(0.34)	0.64(0.26)

TABLE 4. *The estimation results for Monte Carlo simulation of different σ values. The values are shown in the format mean (standard-deviation).*

 ISSN: 2102-6238

η	$\lvert\hat{p}-p\rvert$	Error classification rate
0.5	$0.0125(1.80e-18)$	0.0278(0.009)
5	0.0225(0.0387)	0.222(0.202)
10	$0.0125(1.796e-18)$	0.291(0.02)

TABLE 5. *The estimation results for Monte Carlo simulation of different η values. The values are shown in the format mean (standard-deviation).*

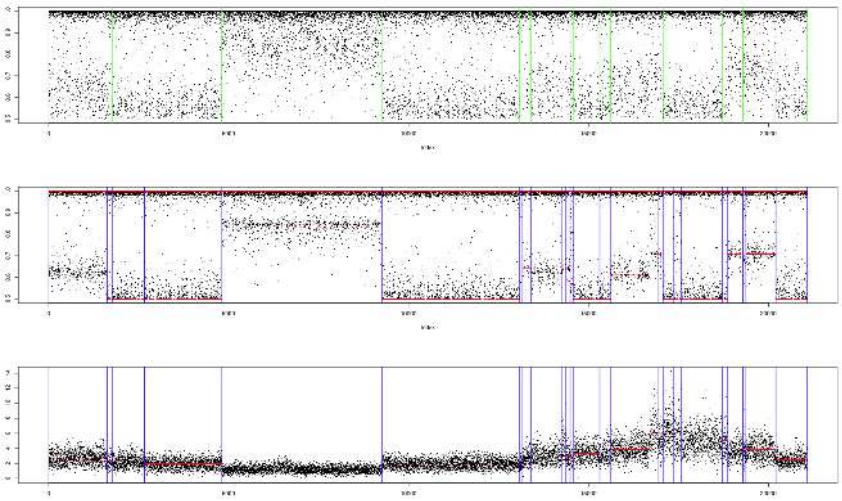

FIGURE 6. *Example of BAF and LRR segmentations of one chromosome in a colon cancer sample. Top : BAF segmentation after mirroring around* 0.5. *Bottom : LRR segmentation. Middle : Resulting LRR segmentation.*

4.4. Cancer sample data

The implementation is also tested on a real tumoral data sample. Colon cancer samples were acquired using 250K Affymetrix SNP-arrays and made available after anonymization by Pr Pierre Laurent-Puig, SIRIC CARPEM Director. BAF and LRR signals are first segmented using the method developed by Comte and Rozenholc (2004). After mirroring around the 0.5 value, the BAF signal is segmented as a bi-modal density where one mode is fixed to be 1, using a dynamic programming algorithm on the penalized log-likelihood. The LRR signal is in turn segmented with a dynamic programming algorithm on the penalized least square function. Combining both segmentations leads to the input data of our method. An illustration is displayed on Figure 6.

Figure 7 shows the estimation results for two models, one with maximum copy number 4 (and estimated $\widehat{p}=0.35$), the other with maximum copy number 8 (and estimated $\widehat{p}=0.525$). The red crosses represents the theoretical centers for the given model with the corresponding estimated contamination rate $\widehat{p}$. One can see that some DNA centers are not present. Blue ellipses are linked to the importance of each CNA center. The two representations are scale such that the copy number lines are aligned.

Bottom graphs on Figure 7 draw the maximum log-likelihood as a function of the normal tissues proportion p for the two models. Both curves are not globally convex. For a maximum copy number $L=4$, the representation shows a good regularity and a marked peak around the

 ISSN: 2102-6238

optimal value $\widehat{p}$, unlike the case $L = 8$ which is less regular.

Model selection performed with BIC chooses the model with maximum copy number $L = 14$, with corresponding optimal value $\widehat{p} = 0.75$. This result is questionable. Indeed, BIC is expected to be a good model selector when the data are generated from one model of the collection, which was the case in our simulations. However, BIC is also known to be not consistent when this is not the case. Due to some assumptions of the statistical model, or due to heterogeneity context of tumor DNA, it is likely that observations are not coming from one model of the collection. To circumvent this issue, we use the slope heuristic criterion which provides a good control of the quadratic risk, and performs a trade-off between bias and variance. This criterion does not need for the observations to come from an existing model in the collection, and should be more suitable in the tumor DNA context, which is confirmed by our results.

Indeed, top left graph on Figure 8 shows a linear trend of the log-likelihood against the complexity $K(L)$ for high complexity. The minimal penalty κ_{min} (the slope) is usually chosen at sight, providing in our case $\kappa_{min} = (930 - 700)/(83 - 30) = 4.34$. It can also be determined with the R package `capushe` (Baudry et al., 2012), leading here to $\kappa_{min} = 3.76$.

The constant κ_{opt} is twice κ_{min}, providing $\kappa_{opt} = 8.68$ at sight, and 7.5 with `capushe`. Both values lead to select the same model $L = 4$, which is in agreement with Popova et al. (2009). The dimension jump, illustrated on the bottom left sub-figure, confirms this model selection. In comparison, the BIC penalty constant is $\log(2 \times 200)/2 = 3 < \kappa_{min}$, which enlightens that BIC is not adapted here.

5. Discussion

We developed a parametric probabilistic model for the characterization of genomic alterations in tumors for segmented SNP-array data. We used a Gaussian mixture model where the centers are constrained to belong to a grid of the (baf, lrr) plane. This automates the pattern recognition method of the GAP method of Popova et al. (2009). The centers only depend on the normal DNA contamination, the contraction factor of LRR measurements, and the shift in LRR due to tumor ploidy. The parameter estimation is achieved by maximum likelihood estimation and no tuning parameter is needed. There is no limit on the number of CNAs in the model, and theoretically we can consider as many CN as necessary. This is particularly useful when the tumor sample has genomic alterations with a very large copy number. A model selection procedure is applied to choose an appropriate model complexity and the algorithms are robust against high normal DNA contamination and measurement noises.

This work could be extended in several directions. If some assumptions on the model are quite weak, one can think to relax the stronger ones as the homocedasticity of the B-allele frequency and log-R ratio signals along the genome : these variances could depend on the mutation type for example. A non constant homozygous proportion could be also considered. Our method, as the GAP method, breaks the link between homozygous and heterozygous sites of a same segment. An extension could be to keep this link tight.

Although an important and common phenomenon in tumor development, tumor heterogeneity is not considered here. To be able to recognize such sub-clones around the main centers could be a nice enhancement.

 ISSN: 2102-6238

Acknowledgments The authors thank Pr Pierre Laurent-Puig, SIRIC CARPEM Director, Paris Descartes University, to give them the ability to access colon cancer data used in this article. The authors also thank anonymous referees for their careful read and relevant comments that have helped to improve the manuscript.

 ISSN: 2102-6238

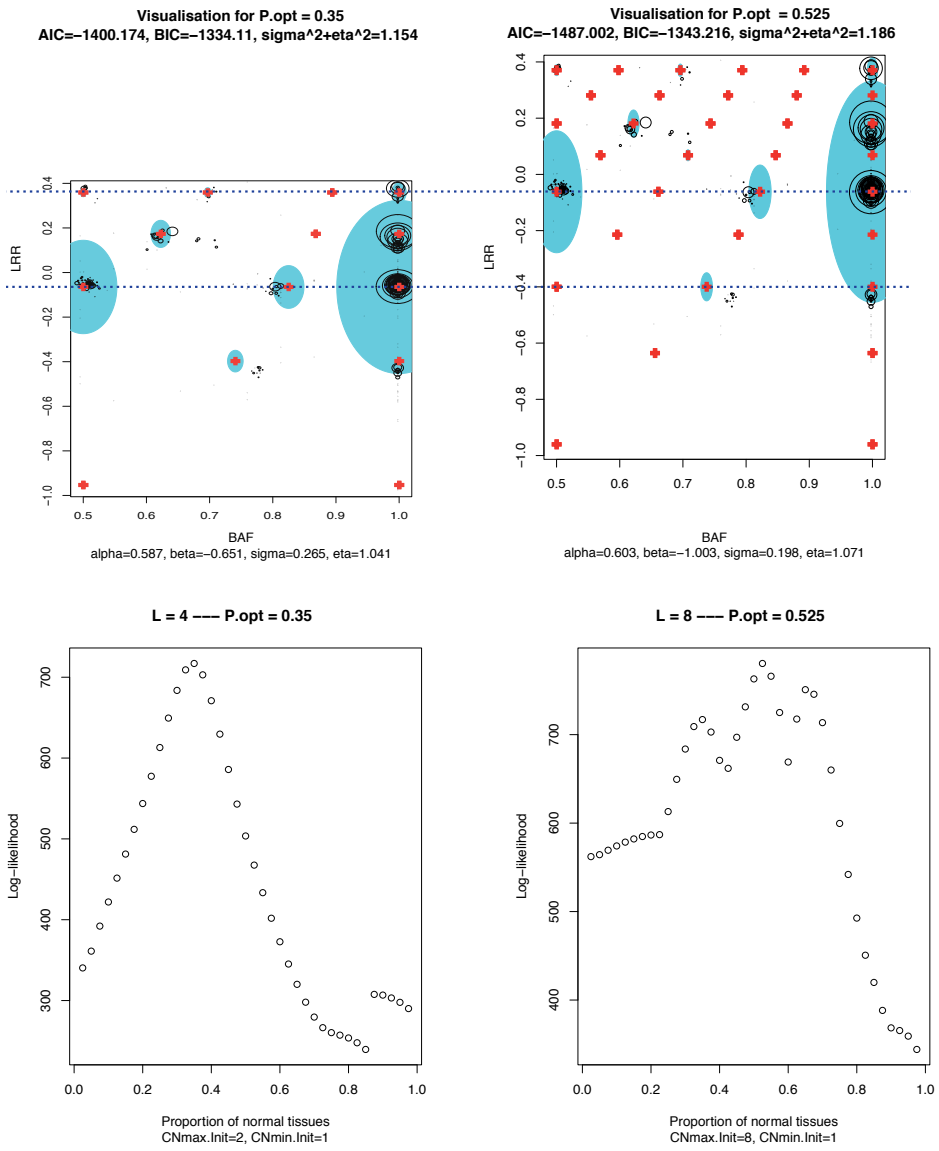

FIGURE 7. *Application to a real cancer sample. Black circles radius are linked to the size of the observed segment. Top left : visualization of the model with maximum copy number* 4. *Top right : visualization of the model with maximum copy number* 8. *Bottom left : optimal log-likelihood of proportion of healthy tissue obtained by models with a maximum copy number of* 4. *Bottom right : optimal log-likelihood of proportion of healthy tissue obtained by models with a maximum copy number of* 8.

 ISSN: 2102-6238

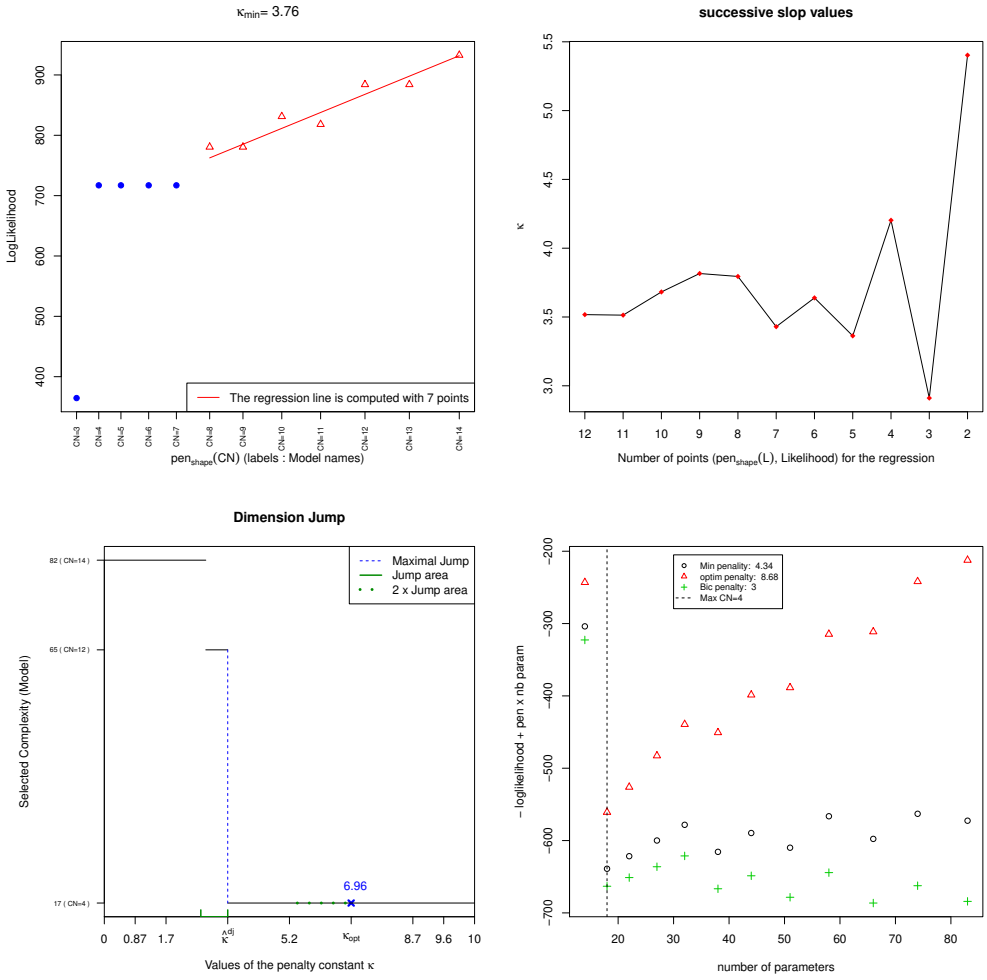

FIGURE 8. *Model selection on a real cancer sample. Top Left : Log-likelihood against the complexity $K(L)$, from maximum copy number $L = 3$ (first point) to $L = 14$ (last point). Top right : graph where the point with the smallest penalty value is removed at each step. Bottom left : dimension jump plot. Bottom right : Comparison of different criteria crit(L, κ), for κ calibrated with slope heuristics (κ_{min} and $\kappa_{opt} = 2\kappa_{min}$) or with BIC ($\kappa = \log(n)/2$).*

 ISSN: 2102-6238

Références

Arlot, S. (2019). Minimal penalties and the slope heuristics: a survey. *arXiv preprint arXiv:1901.07277*.

Baudry, J.-P. (2009). *Sélection de modèle pour la classification non supervisée. Choix du nombre de classes.* PhD thesis, Université Paris Sud-Paris XI.

Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268.

Birgé, L. and Massart, P. (2007). Minimal penalties for gaussian model selection. *Probability theory and related fields*, 138(1-2):33–73.

Carr, S. M., Marshall, H. D., Duggan, A. T., Flynn, S. M., Johnstone, K. A., Pope, A. M., and Wilkerson, C. D. (2008). Phylogeographic genomics of mitochondrial DNA: highly-resolved patterns of intraspecific evolution and a multi-species, microarray-based DNA sequencing strategy for biodiversity studies. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*, 3(1):1–11.

Comte, F. and Rozenholc, Y. (2004). A new algorithm for fixed design regression and denoising. *Annals of the Institute of Statistical Mathematics*, 56(3):449–473.

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 (methodological)*, pages 1–38.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 49–66.

Li, A., Liu, Z., Lezon-Geyda, K., Sarkar, S., Lannin, D., Schulz, V., Krop, I., Winer, E., Harris, L., and Tuck, D. (2011). GPHMM: an integrated hidden markov model for identification of copy number alteration and loss of heterozygosity in complex tumor samples using whole genome SNP arrays. *Nucleic Acids Research*, 39(12):4928–4941.

Mosén-Ansorena, D., Aransay, A. M., and Rodríguez-Ezpeleta, N. (2012). Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data. *BMC bioinformatics*, 13(1):192.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.-L., Chen, C., Zhai, Y., et al. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature genetics*, 20(2):207.

Popova, T., Manié, E., Stoppa-Lyonnet, D., Rigaill, G., Barillot, E., Stern, M. H., et al. (2009). Genome alteration print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome Biol*, 10(11):R128–R128.

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Döhner, H., Cremer, T., and Lichter, P. (1997). Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes, chromosomes and cancer*, 20(4):399–407.

Staaf, J., Lindgren, D., Vallon-Christersson, J., Isaksson, A., Goransson, H., Juliusson, G., Rosenquist, R., Hoglund, M., Borg, A., and Ringner, M. (2008). Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biology*, 9(9):R136.

Sun, W., Wright, F. A., Tang, Z., Nordgard, S. H., Van Loo, P., Yu, T., Kristensen, V. N., and Perou, C. M. (2009). Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic acids research*, 37(16):5365–5377.

Van Loo, P., Nordgard, S. H., Lingjærde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B., et al. (2010). Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915.

Yau, C., Mouradov, D., Jorissen, R. N., Colella, S., Mirza, G., Steers, G., Harris, A., Ragoussis, J., Sieber, O., Holmes, C. C., et al. (2010). A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol*, 11(9):R92–R92.

 ISSN: 2102-6238