Statistical estimation of genomic alterations of tumors

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

[1]UMRS-1147 Université Paris Descartes
[2]INRIA Equipe Select
[3]Laboratoire de Mathématique, Univ. Paris Sud, CNRS, Université Paris Saclay
[4]INSERM U830 – Institut Curie
[5]MAP5 UFR Math-Info Université Paris Descartes

June 2015

Abstract

Recent research reveals that personalized medicine is one major way to treat cancer. In order to develop personalized medicine, characterizing the genomic alterations is a vital component. Several methods have been proposed to this end. One of the first methods is the Genome Alteration Print (GAP) by Popova et al, which uses a deterministic approach. We follow this approach and develop a parametric probabilistic model for GAP, together with its statistical estimation, based on a preliminary segmentation of SNP measurements obtained from microarray experiments. For estimation, we implement the expectation-maximization (EM) algorithm to maximize the likelihood of this model and get the parameter estimation which characterizes the genomic alterations. In our approach, the tumoral ploidy is deduced from penalized model selection. Our model is tested on simulated data and real data.

Keywords EM algorithm, penalized maximum likelihood, tumoral genomic alterations, GAP, SNP

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 to develop personalized medicine, characterizing the genomic alterations is a vital component. One way to characterize this alteration is to use a Single Nucleotide Polymorphism (SNP) microarray. A SNP is a nucleotide showing variability in the population. In theory, there are four possible variations. However in practice, only two variations are observed which are called A-allele and B-allele, one being common in a large proportion of the population. Since the chromosomes in human come in pairs, it is possible for a SNP to have the genotype AA, BB, AB, or BA. The two former cases are called homozygous SNP, and the two latter, which are indistinguishable, are called heterozygous SNP.

Using microarrays, one can detect genomic alterations such as copy-number variation and allele-imbalance. Having at hand two microarrays, one for the tumor, the other for the normal tissue, one can get rid of the unknown proportion p of normal tissues in the tumor sample which acts as a confusing parameter in the tumoral alteration characterization. However, clinicians are expecting to retrieve this information from only a single tumor sample microarray. Several methods have already been developed for this goal. GenoCNA[1], OncoSNP[2], and GPHMM[3] employ a Hidden Markov Model (HMM) integrating both segmentation and mutation characterization in a single step. GAP[4] and ASCAT[5] adopt a two-step approach in which the data are first segmented and then the mutation types are estimated. Both methods are based on an optimization step with respect to p of a deterministic quality criterion. Taking into account allelic imbalance and copy number aberration, the criterion used in ASCAT[5] measures a weighted

discrepancy based on several heuristics. Noticing that, for a given p, the possible mutations are precisely localized in the a bi-dimensional plane to be detailed later, the GAP[4] criterion is defined as the number of segmented observations that are close to these locations within a predefined proximity value. Comparison of these methods [6] shows that, the two-step approaches have better performance.

Using the mutation localization in the bi-dimensional plane introduced in [4], we develop a parametric probabilistic model and realize the estimation of its parameters, providing not only the most probable mutation types of each segment, but also a probabilistic distribution of these mutations. The estimation uses an optimization with respect to p based on the maximization of the log-likelihood function 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 bring to the clinicians the expected probabilistic model for mutations.

Biological model for tumoral mutations with SNP

For a given SNP s, the tumoral mutation type is characterized by the number of replicates of each strand denoted u and v, with $u, v \geq 0$ (the value 0 corresponding to a deletion). Depending on the zygosity (AA, AB, BA, or BB) of the germline cells (*i.e. normal tissue*), one can compute, as shown in Table. 1, the number of each allele n_A^g, n_B^g in the germline cells and n_A^t, n_B^t in the tumor. From these numbers, we have access to the two quantities of interest, namely the *copy number*

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

and the *B-allele frequency*

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

In microarray experiment, measurements provide direct access to the B-allele frequency and to the log-*R-ratio*, which is linked to the copy number by the relation

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

where α is a contraction factor depending on the microarray platform and β is a constant shift due to tumor ploidy.

Table 1: Allele counts from u and v replicates of each strand

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

Since neighboring SNP's tend to have the same mutation process, the baf and lrr signals obtained in microarray experiment are assumed to follow piecewise constant distributions. Hence, they can be segmented into homogeneous intervals with same tumoral mutation. On one interval characterized by u and v, one can remark that the copy number is constant, however the B-allele frequency takes two pairs of symmetrical values around baf $= 0.5$: the homozygous SNP take value 0 (AA) or 1 (BB) and the heterozygous SNP (AB and BA) take two symmetrical values in $(0, 1)$, as soon as $p > 0$. As labels A and B have been attributed at random for each SNP, their order is non informative, so that we use the symmetry and aggregate the information to have baf ≥ 0.5.

After the symmetry, one interval with a mutation k, characterized by $0 \leq u \leq v$, is associated with two centers $c_k^0 = (\mathrm{baf}_k^0, \mathrm{lrr}_k)$ and $c_k^1 = (\mathrm{baf}_k^1, \mathrm{lrr}_k)$ in the $(\mathrm{baf}, \mathrm{lrr})$ plane. The point c_k^0, which corresponds to heterozygous SNP, satisfies

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

and is shown in Figure 1 by point (AB,uAvB). Similarly, the point c_k^1 corresponds to homozygous SNP with $\mathrm{baf}_k^1 = 1$ and correspond to the point (BB,$(u+v)$B). Two mutations k and k' leading to same copy number share

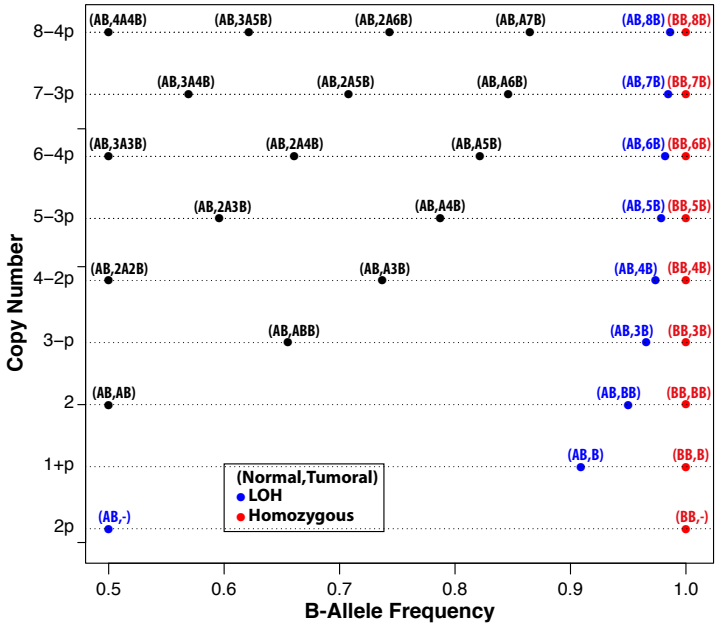

Figure 1: Schematic illustration of the correspondence between the tumoral mutation and the $(\mathrm{baf}, \mathrm{lrr})$ values assuming $0 \leq u \leq v$. Mutations of germ line homozygous are in red. Mutations of germ line heterozygous are in black and blue. The mutation of the latter being characterized by a Loss of Heterozygosity (LOH). Mutations are characterized by $(n_A^h, n_B^h, n_A^t, n_B^t)$.

the same homozygous center on the $(\mathrm{baf}, \mathrm{lrr})$ plane $c_k^1 = c_{k'}^1$. For example, the mutations $(u = 2, v = 2)$ and $(u = 1, v = 3)$ share the same $\mathrm{cn} = 4 - 2p$ and correspond to the heterozygous mutation centers $(AB, 2A2B)$ and $(AB, A3B)$ respectively in Figure 1, however they are associated with the same homozygous mutation center $(BB, 4B)$.

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

Probabilistic model for tumoral mutations with SNP

Having at hand a segmentation into n intervals with homogenous mutation type, for the i-th interval $(i = 1, \ldots, n)$ with N_i SNP's, we denote the number of heterozygous SNP's by N_i^0 and the number of homozygous SNP's by $N_i^1 = N_i - N_i^0$. Three summary variables are extracted from the SNP's 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. According to Central Limit Theorem, these variables follow asymptotically a Gaussian distribution

$$\mathrm{BAF}_i^0 = \mathrm{baf}_{k(i)}^0 + \sigma \frac{\varepsilon_i^0}{\sqrt{N_i^0}},$$
$$\mathrm{BAF}_i^1 = \mathrm{baf}_{k(i)}^1 + \sigma \frac{\varepsilon_i^1}{\sqrt{N_i^1}},$$
$$\mathrm{LRR}_i = \mathrm{lrr}_{k(i)} + \eta \frac{\xi_i}{\sqrt{N_i}},$$

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 mutation type of the i-th interval. The simple underlying assumption for this modelization is that the measurements coming from individual SNP in an homogenous interval follow a distribution having finite first two moments.

The triplet $(\mathrm{BAF}_i^0, \mathrm{BAF}_i^1, \mathrm{LRR}_i)$ is split into two independent observations:

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

corresponding respectively to the heterozygous and homozygous observations. Here ζ_i^j is a

bi-dimensional centered Gaussian distribution with covariance matrix

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

Let us assume that we face at most K mutations, corresponding to $L \leq 2K$ mutation centers $c_{k(i)}^j$ that we label as $c_\ell := (\mathrm{baf}_\ell, \mathrm{lrr}_\ell)$, $\ell = 1, \ldots, L$. We modelize the split observations with a Gaussian mixture model with probability density function (pdf):

$$f(C_i^j; N_i^j, N_i) = \sum_{\ell=1}^{L} \pi_\ell \phi(C_i^j; c_\ell, \Sigma_i^j)$$

where $\phi(., c, \Sigma)$ is the pdf of a bi-dimensional Gaussian random variable centered on c with covariance matrix Σ and where π_ℓ, $\ell = 1, \ldots, L$, are the mixture proportions.

In this modelization, the parameter to estimate is

$$\theta = (p, \sigma^2, \eta^2, \alpha, \beta, \{\pi_\ell, \ell = 1, \ldots, L\})$$

and the log-likelihood of the observations is

$$LL\left(\theta; \{C_i^j; N_i^j, N_i\}_{i=1,\ldots,n,\, j=0,1}\right) = \sum_{j=0}^{1} \sum_{i=1}^{n} \log\left(\sum_{\ell=1}^{L} \pi_\ell \phi(C_i^j; c_\ell, \Sigma_i^j)\right).$$

Maximum Likelihood Estimation

We use a maximum likelihood approach to estimate the parameter. For our mixture model, we propose an expectation-maximization (EM) algorithm [7] and introduce the latent variables $z_{i\ell}^j$ which equals 1 if $(\mathrm{BAF}_i^j, \mathrm{LRR}_i)$ is from mutation center c_ℓ, 0 otherwise. The EM is known to maximize the expectation, conditionally to the observations, of the complete log-likelihood defined by

$$LL_c(\theta; \{C_i^j; N_i^j, N_i\}, \{z_{i\ell}^j\}) = \sum_{i=1}^{n} \sum_{j=0}^{1} \sum_{\ell=1}^{L} z_{i\ell}^j \left[\log \pi_\ell + \log \phi(C_i^j; c_\ell, \Sigma_i^j)\right].$$

This leads in turn to maximize the log-likelihood.

The parameter p, unlike the other parameters $(\sigma^2, \eta^2, \alpha, \beta)$, cannot be straightforwardly optimized inside the maximization step of the EM, hence the optimization procedure is designed into two nested levels using that

$$\max_{p, \sigma^2, \eta^2, \alpha, \beta} = \max_{p} \max_{\sigma^2, \eta^2, \alpha, \beta}.$$

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

The above estimation is done with a fixed number of mutation centers corresponding to copy numbers confined in the interval $[\mathrm{cn}_{\min}, \mathrm{cn}_{\max}]$. We adopt a penalized log-likelihood approach to select the range $[\mathrm{cn}_{\min}, \mathrm{cn}_{\max}]$ comparing AIC and BIC criteria.

EM for fixed p and fixed range of mutations

Given a fixed number of mutation centers corresponding to copy numbers confined in the interval $[\mathrm{cn}_{\min}, \mathrm{cn}_{\max}]$ and a fixed p value, the EM iterates two steps: one expectation and one maximization.

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

$$\tau_{i\ell}^j \leftarrow E(z_{i\ell}^j | \breve{\theta}) = \frac{\breve{\pi}_\ell \phi(C_i^j; \breve{c}_\ell, \breve{\Sigma}_i^j)}{\sum_\ell \breve{\pi}_\ell \phi(C_i^j; \breve{c}_\ell, \breve{\Sigma}_i^j)}.$$

Using this updated value of $\tau_{i\ell}^j$, the maximization leads to update the parameters according to

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

where

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

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

Initialization of 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, the parameters $(\sigma^2, \eta^2, \alpha, \beta)$ are initialized with different values according to

$$(\sigma^2)^0 = \mathrm{var}(\mathrm{BAF}),$$
$$(\eta^2)^0 = \mathrm{var}(\mathrm{LRR}).$$
$$\alpha^0 = \frac{\mathrm{LRR}_{\max} - \mathrm{LRR}_{\min}}{\log_2 \mathrm{cn}_h - \log_2 \mathrm{cn}_l},$$
$$\beta^0 = \mathrm{LRR}_{\max} - \alpha^0 \log_2 \mathrm{cn}_h,$$

with $\mathrm{cn}_{\min} \leq \mathrm{cn}_l < \mathrm{cn}_h \leq \mathrm{cn}_{\max}$, being all possible combinations. Using the knowledge of p, α^0 and β^0, we compute the centers c_ℓ and affect each C_i^j to the closest c_ℓ in order to compute the parameters π_ℓ^0 for $\ell = 1, \ldots, L$.

Maximization with respect to p

The behaviour 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 search algorithm to find the optimal value of p on the grid $p \in \{0.025, 0.05, 0.075, \ldots, 0.975\}$.

Model selection with penalized log-likelihood

Our model depends on the K considered mutations constraint to the interval $[\mathrm{cn}_{\min} = 0, \mathrm{cn}_{\max}]$. We select the values $\mathrm{cn}_{\max}$ using a penalized maximum likelihood approach [8, 9] by miniimizing

$$-2LL(\widehat{\theta}; \{C_i^j; N_i^j, N_i\}) + \mathrm{pen}(L)$$

where L is the number of parameters used in the model and pen is the BIC penalty with the form $pen(L) = L \log 2n$.

Finally, the tumoral ploidy is computed as the weighted average of the copy numbers deduced from our estimation:

$$\frac{\sum_{i,j,\ell} \tau_{i\ell}^j N_i^j \mathrm{cn}_\ell}{\sum_{i,j,\ell} \tau_{i\ell}^j N_i^j}.$$

Results

Simulated data

To evaluate the performance of the algorithm, we devised a strategy to generate simulated data and compared the estimation results with the real parameter used to generate the data sets. The data generation strategy is as follows: 1) generate randomly 200 segments on a microarray measurements of 261976 SNP's (the number of SNP's used in a real microarray experiment), 2) for each segment, generate independently the number of two alleles following a multiple Bernoulli distribution with $P(n = 0) = 0.15$, $P(n = 1) = 0.5$, $P(n = 2) = 0.2$, $P(n = 3) = 0.1$, and $P(n = 4) = 0.05$, 3) calculate the baf and lrr correspondingly and symmetrize baf into the interval $[0.5, 1]$, 4) on each segment generate the number of homozygous SNP's following a binomial distribution of $P(homozygous) = 0.8$, and 5) add noise to the baf and lrr values and form the final observations with given α and β. With this strategy, we can generate simulated data with a maximum copy number of 8 for the tumor tissues.

First the implementation was tested on a single simulated data set. The data set was generated with the parameter $\theta = (p = 0.1, \alpha = 1, \beta = 0, \sigma^2 = 0.04, \eta^2 = 0.25)$ with the maximum copy number 8. Models with $\mathrm{cn}_{max} \in [3, 10]$ were used for the estimation. Using the BIC criterion, the model with maximum copy number 8 is selected, corresponding to the underlying parameter. With this model, the parameter estimation gives $\hat{\theta} = (\hat{p} = 0.1, \hat{\alpha} =$

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	0.00342	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: Estimation result on simulated diluted data sets.

$1, \hat{\beta} = -0.00316, \hat{\sigma}^2 = 0.0404, \hat{\eta}^2 = 0.255$), which is in good agreement with the underlying data. Also, the classification error rate based on maximum a posterior (MAP) probability is 0.025. The estimation result is show in Figure 2.

Next, to test the coherence of the algorithm, we generated a series of tumor sample with the same mutation type but with different proportions of normal tissues, 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. The algorithm chooses the model with maximum copy number 8 in all nine data sets with the BIC criterion. And the parameter estimation is good even in the data sets with normal tissues proportion 0.9 (see TABLE 2). Even in the strong contamination case with $p = 0.9$, the algorithm still shows relatively coherent behavior and the MAP classification error is 0.13.

To evaluate the robustness of the algorithm, we introduced more variability in the data generation strategy and tested on different parameter values. The modified strategy is as follows: 1) generate the number of segments on the SNP's n following a Poisson distribution $Poisson(\lambda = 200)$, 2) generate $n - 1$ break points on the 261976 SNP's (the number of SNP's used in a real microarray experiment), 3)

Influence of p

We used three values of normal tissues proportion in the Monte Carlo sampling $p = 0.1125, 0.5125, 0.8125$. The result is shown in Table 3. It is evident that the algorithm is rather robust against the normal tissue contamination. And a moderate degree of normal tissues contamination might be conducive to the estimation of genomic mutations, which is in accordance with Popova et al.[4].

Influence of σ

We used three values $\sigma = 0.2, 1.5, 3$ in the Monte Carlo sampling to determine its influence on the estimation. The result is 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 is fixed not only by the underlying mutation type and p, but also by α and β, BAF measurements is uniquely determined by p and the mutation type. Thus a large variance of BAF will deteriorate greatly the estimation result. 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.

Influence of η

The influence of η was tested with three different values $\eta = (0.5, 5, 10)$. The result for the Monte Carlo simulation is listed in Table 5.

Cancer sample data

The implementation is also tested on real cancer sample data. The BIC criterion chooses the model with maximum copy number 12, among the models with maximum copy number $\{3, 4, \ldots, 14\}$. However, using the slope heuristic criterion[10, 11, 12], the model with maximum copy number 4 is preferred, which agrees with the result obtained by GAP. The estimation results of the two models are shown in Figure 3.

Discussion

We developed a parametric probabilistic model for the characterization of genomic alterations in tumors for segmented SNP microarray data, using a Gaussian mixture model. The model is

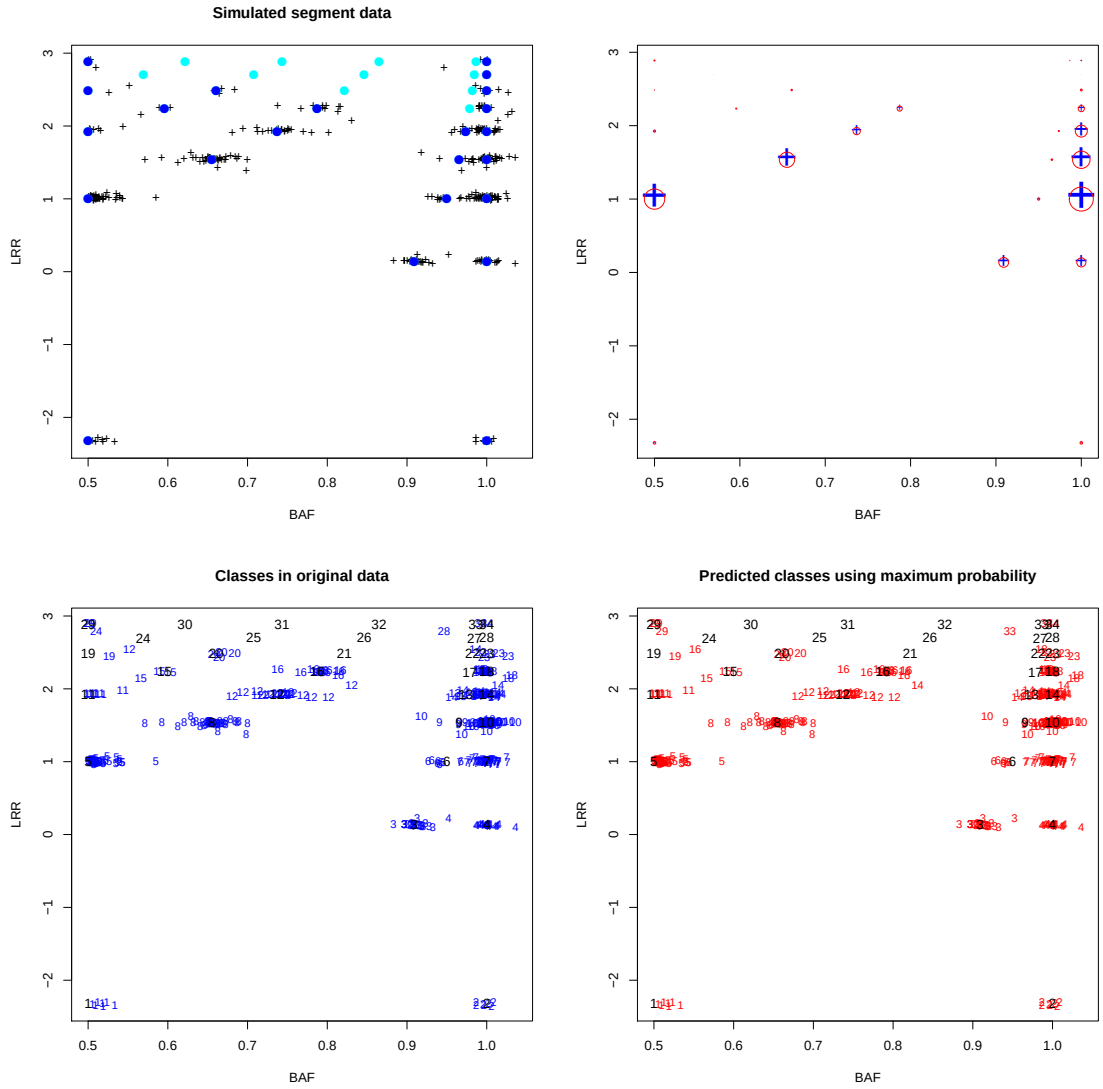

Figure 2: Top left: representation of the simulated data with $\theta = (p = 0.1, \sigma = 0.2, \eta = 0.5, \alpha = 1, \beta = 0)$ in the (baf, lrr) plane. Black crosses represent the independent observations, blue dots the mutation centers used to generate the observations, cyan dots non-occupied centers in the complete model. Top right: parameter estimation on the simulated data. Blue crosses represent the position of the mutation centers and the relative proportion with its size. Red circles represent the estimated position of the mutation 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.

p	$\lvert\hat{p} - p\rvert$	Number of error classes	Error classification rate
0.1125	$0.0125(1.80e-18)$	8.67(2.87)	$0.0278(9.20e-3)$
0.5125	$0.0125(5.10e-17)$	6.67(2.77)	0.013(0.0032)
0.8125	$0.0125(5.10e-17)$	7.27(2.15)	0.056(0.014)

Table 3: The estimation results for Monte Carlo simulation of different p values. The values are shown in the format mean(standard-deviation).

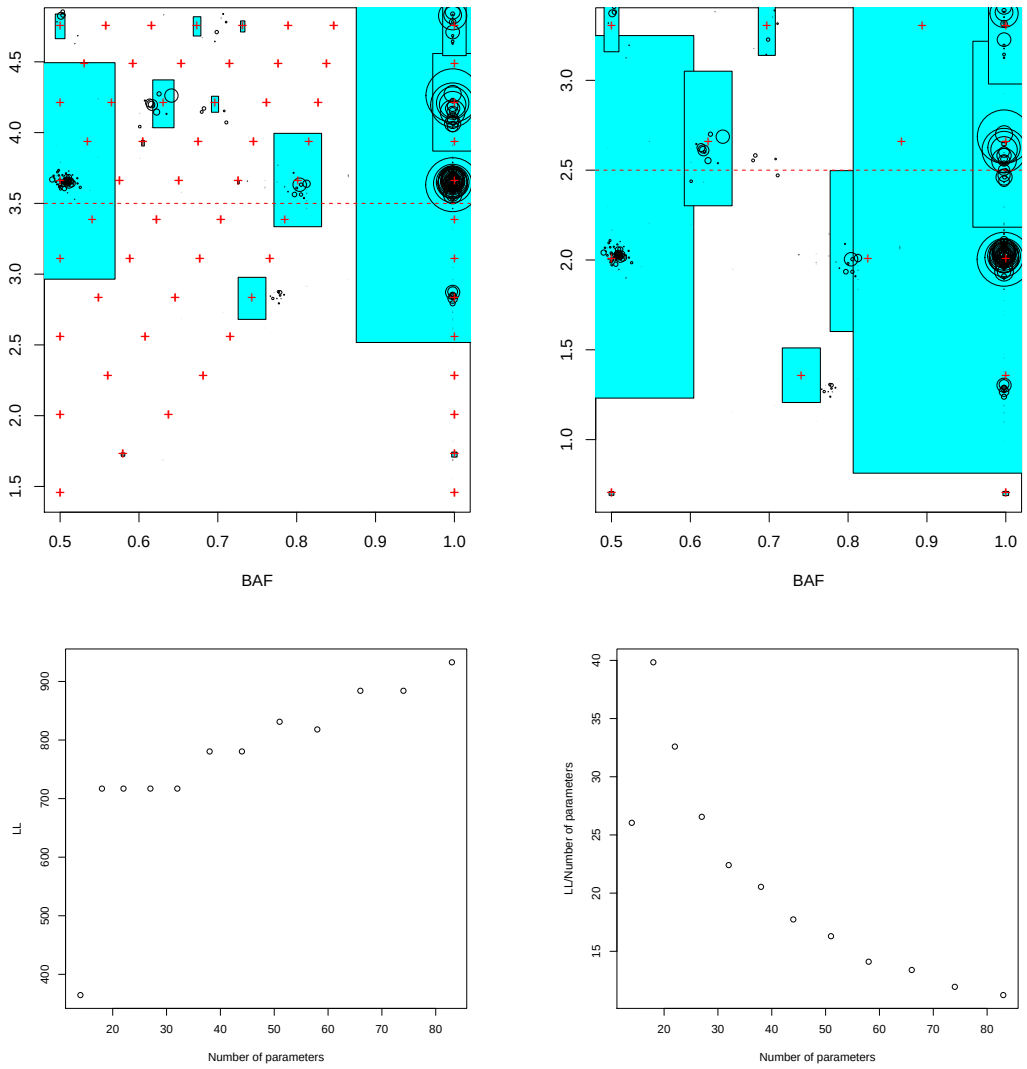

Figure 3: Application of the implementation to a real cancer sample. Top left: visualization of the model with maximum copy number 12. Top right: visualization of the model with maximum copy number 4. Bottom left: optimal log-likelihood obtained by models with different maximum copy numbers. Bottom right: optimal log-likelihood divided by the number of parameters for models with different maximum copy numbers.

σ	$\lvert\hat{p}-p\rvert$	Number of error classes	Error classification rate
0.2	$0.0125(1.80e-18)$	8.67(2.87)	$0.0278(9.20e-3)$
1.5	0.281(0.353)	14.53(9.65)	0.520(0.383)
3	0.21(0.34)	12.27(8.94)	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).

η	$\lvert\hat{p}-p\rvert$	Number of error classes	Error classification rate
0.5	$0.0125(1.80e-18)$	8.67(2.87)	$0.0278(9.20e-3)$
5	0.0225(0.0387)	7.667(2.77)	0.222(0.202)
10	$0.0125(1.796e-18)$	6.067(1.223)	$0.291(2.055e-2)$

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

based on the GAP[4] method, which employs pattern recognition method on the (baf, lrr) plane. It takes into account the normal tissue 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 numbers of mutation type in the model, and theoretically we can consider as many mutation types as necessary. This is particularly useful in the case where the tumor sample has genomic alterations with very large copy number. And a model selection procedure is applied to choose the right model complexity. The algorithms is robust against severe normal tissue contamination and measurement noises.

In developing the model, we made the following assumptions: (i) the noises in BAF and LRR measurements have a finite second moments, (ii) the contraction factor and shift constant of LRR is the same for all measurements. The first assumption is very weak in that no assumptions about the form of the underlying distributions are 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 segmentation step[13].

Although an important and common phenomenon in tumor development, tumor heterogeneity is not considered in this model.

References

[1] Wei Sun, Fred A Wright, Zhengzheng Tang, Silje H Nordgard, Peter Van Loo, Tianwei Yu, Vessela N Kristensen, and Charles M Perou. Integrated study of copy number states and genotype calls using high-density snp arrays. *Nucleic acids research*, 37(16):5365–5377, 2009.

[2] Christopher Yau, Dmitri Mouradov, Robert N Jorissen, Stefano Colella, Ghazala Mirza, Graham Steers, Adrian Harris, Jiannis Ragoussis, Oliver Sieber, Christopher C Holmes, et al. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol*, 11(9):R92–R92, 2010.

[3] Ao Li, Zongzhi Liu, Kimberly Lezon-Geyda, Sudipa Sarkar, Donald Lannin, Vincent Schulz, Ian Krop, Eric Winer, Lyndsay Harris, and David Tuck. 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, 2011.

[4] Tatiana Popova, Elodie Manié, Dominique Stoppa-Lyonnet, Guillem Rigaill, Emmanuel Barillot, Marc Henri Stern, et al. Genome alteration print (gap): a tool to visualize and mine complex cancer genomic profiles obtained by snp arrays. *Genome Biol*, 10(11):R128–R128, 2009.

[5] Peter Van Loo, Silje H Nordgard, Ole Christian Lingjærde, Hege G Russnes, Inga H Rye, Wei Sun, Victor J Weigman, Peter Marynen, Anders Zetterberg, Bjørn Naume, et al. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915, 2010.

[6] David Mosén-Ansorena, Ana M Aransay, and Naiara Rodríguez-Ezpeleta. Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data. *BMC bioinformatics*, 13(1):192, 2012.

[7] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[8] Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7):719–725, 2000.

[9] Jean-Patrick Baudry, Adrian E Raftery, Gilles Celeux, Kenneth Lo, and Raphael Gottardo. Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*, 19(2), 2010.

[10] Lucien Birgé and Pascal Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268, 2001.

[11] Lucien Birgé and Pascal Massart. Minimal penalties for gaussian model selection. *Probability theory and related fields*, 138(1-2):33–73, 2007.

[12] Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012.

[13] Johan Staaf, David Lindgren, Johan Vallon-Christersson, Anders Isaksson, Hanna Goransson, Gunnar Juliusson, Richard Rosenquist, Mattias Hoglund, Ake Borg, and Markus Ringner. Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome snp arrays. *Genome Biology*, 9(9):R136, 2008.

A Algorithm in summary

Based on the preceding discussion, the algorithm with a grid search of optimal p value for a given model is shown as follows. The gradient descent method is similar, the only difference being that the loop on p values (the red part in the algorithm) is replaced by a gradient descent on p.

input : Micro-array data
output: Finding the best parameters: $(p, \alpha, \beta, (\sigma_k^2), (\eta_k^2), (\pi_k))$

begin
  for CN_{max} *in* CN_{range} **do**
    for CN_{min} *in* $0 : CN_{max} - 1$ **do**
      for p *in pvals* **do**
        Calculate the theoretical centers on the grid;
        (BAF_k, LRR_k);
        `/* Use EM algorithm to find the optimal values of` $(\alpha, \beta, (\sigma_k^2), (\eta_k^2), (\pi_k))$ `*/`
        Initialization of $\theta^c = (\alpha, \beta, (\sigma^2), (\eta^2), (\pi_k))$;
        $\sigma^2 = var(BAF)$;
        $\alpha = (\max LRR - \min LRR)/(\max \log_2 CN - \min \log_2 CN)$;
        $\beta = \max LRR - \alpha \max \log_2 CN$;
        $\eta^2 = \alpha var(LRR)$;
        $\pi_\ell = 1/L$;
        while *Log-likelihood not converged* **do**
          E-step:
          $\tau_{i\ell}^j = \frac{\pi_\ell \phi^c(C_i^j; c_\ell, N_i^j)}{\sum_l \phi^c(C_i^j; c_\ell, N_i^j)}$
          M-step:
          $\pi_\ell = \frac{\sum_i \sum_j \tau_{i\ell}^j}{\sum_i \sum_j \sum_\ell \tau_{i\ell}^j}$
          $\sigma^2 = \frac{\sum_i \sum_\ell \sum_j \tau_{i\ell}^j N_i^j (BAF_i - baf_\ell)^2}{\sum_i \sum_\ell \sum_j \tau_{i\ell}^j}$
          $(\alpha, \beta) =$ solve linear equations
          $\bar{\eta}^2 = \frac{\sum_i \sum_\ell \sum_j \tau_{i\ell}^j N_i (LRR_i - \alpha lrr_\ell - \beta)^2}{\sum_i \sum_\ell \sum_j \tau_{i\ell}^j}$
          $\theta^c = \theta$
        end
      end
      Choose the model with optimal log-likelihood;
      Calculate the BIC and AIC of the corresponding model;
    end
  end
  Choose the best model based on BIC and AIC;
end

Algorithm 1: Algorithm of the estimation with a grid search on p