Journal of Statistical Computation & Simulation
Vol. 00, No. 00, Month 200x, 1–16

RESEARCH ARTICLE

A Bayesian HMM with random effects and an unknown number of states for DNA copy number analysis

Oscar M. Rueda*†, Cristina Rueda‡ and Ramón Díaz-Uriarte††

(*Received 00 Month 200x; in final form 00 Month 200x*)

Hidden Markov Models (HMMs) have been shown to be a flexible tool for modelling complex biological processes. However, choosing the number of hidden states remains an open question and the inclusion of random effects also deserves more research, as it is a recent addition to the fixed effect HMM in many application fields. We present a Bayesian mixed HMM with an unknown number of hidden states and fixed covariates. The model is fitted using Reversible Jump Markov Chain Monte Carlo (RJMCMC), avoiding the need to select the number of hidden states. We show through simulations that the estimations produced are more precise than those from a fixed effect HMM and illustrate its practical application to the analysis of DNA copy number data, a field where HMMs are widely used.

Keywords: array Comparative Genomic Hybridization; Bayesian Inference; Copy Number Variation; Hidden Markov Models; Reversible Jump Markov Chain Monte Carlo.

AMS Subject Classification: 62F15; 62M05; 62P10.

1. Introduction

DNA copy number alterations (CNAs) are events produced by a failure in the replication machinery of the genome that result in a change in the number of copies of a particular chromosome region. These alterations have been related to a number of diseases, in particular to cancer. Amplifications (copy number gains) of oncogenes produce tumour activation (see [1, 2]), while deletions (copy number losses) can produce the inactivation of tumour suppressor genes.

Several techniques have been developed to measure genomic copy numbers. Array-based Comparative Genomic Hybridization (aCGH) and Single Nucleotide Polymorphism arrays (SNP arrays) are probably the most widely used (see [3], [4] and [5] for a review of the different platforms). The data obtained, after proper normalization (see for example [6, 7]), are log_2 ratios of color intensities for a number of genomic regions, and the spatial location of these points (often called probes) and their length depend on the particular microarray platform. The observations are expected to present some correlation, as alterations occur in contiguous chromosomal regions, and this dependence will be stronger the closer two particular observations are in the genome.

There are a number of methods available for the analysis of these data. Some of them use different segmentation techniques to identify breakpoints, such as DNAcopy from [8], HaarSeg from [9] or GADA from [10]. Most of these methods only

†Cancer Research UK Cambridge Research Institute, Cambridge, UK. e-mail: Oscar.Rueda@cancer.org.uk
‡Department of Statistics and Operations Research, University of Valladolid, Spain. e-mail: crueda@eio.uva.es
††Department of Biochemistry, Universidad Autónoma de Madrid-Instituto de Investigaciones Biomédicas 'Alberto Sols', CSIC-UAM, Madrid, Spain. e-mail: rdiaz02@gmail.com

ISSN: 0094-9655 print/ISSN 1563-5163 online

DOI: 10.1080/0094965YYxxxxxxxx
http://www.informaworld.com

identify segments with the same copy number, but do not assign copy numbers to those regions. Hidden Markov Models (HMMs) have been used extensively for aCGH and SNP analysis. [11] were the first to apply them to this problem and soon others proposed some improvements, such as the non-homogeneous HMM (NH-HMM) of [12], the robust HMM of [13], the continuous-index HMM of [14] or the NH-HMM with an unknown number of states of [15] (this algorithm, named RJaCGH, was shown to perform better than the alternatives and is the basis for the model presented in this paper). Furthermore, there are specific HMMs for some microarray platforms, such as quantiSNP from [16], pennCNV from [17] or PICNIC from [18].

In typical experiments the researcher has to analyze a set of arrays corresponding to different individuals that can show a high level of heterogeneity. Most of the existing methods analyze each of them independently, while a few use multivariate models, like the multivariate version of GADA by [19]. In general terms, fitting the same model to all of the arrays will be very restrictive while fitting a different model will be time consuming and will result in a huge number of parameters and a loss of efficiency, therefore none of these options is fully satisfactory. In this paper we present an extension to RJaCGH, an NH-HMM with an unknown number of states, that incorporates random effects. It allows us to model heterogeneity among individuals while keeping the number of parameters to be estimated at a reasonable amount. HMMs with random effects have received some treatment in the literature and have been applied to a wide range of areas, other than the analysis of CGH data: [20] introduce random effects for a particular type of HMMs called segmental HMMs, while [21] gives a complete treatment for the frequentist approach. Several applications include [22] a latent-state model with one random effect for modelling animal behaviour, [23] a longitudinal model for metastatic brain tumour patients, [24] a semiparametric model or [25] a semi-Markov switching linear mixed model. From a Bayesian point of view, [26] presents a model with a random effect fitted by Metropolis-Hastings step, and [27] fit a model with random effects per individual and hidden state through Gibbs Sampler. [28] fit a hierarchical NH-HMM for longitudinal observations, [29] presents a model for the analysis of infectious disease biomarkers and [30] describes several models for alcoholism.
For the analysis of aCGH data, [31] use a pseudolikelihood approach to fit an HMM with random effects per array and chromosome.
All of these approaches have a fixed number of hidden states, that is, the researcher must specify a priori the number of hidden states. This situation is unrealistic from a biological point of view because, in the case of DNA copy number analysis, the number of different aberrations can be very different in each individual. Alternatively, models with a different number of hidden states can be run, so that the user selects one of them a posteriori. However, the use of model selection criteria such as AIC or BIC is not justified in HMMs, see the next Section.

In this paper we present an NH-HMM Reversible Jump (RJ) model for aCGH data with random effects and an unknown number of states fitted through RJMCMC using Gibbs Sampler or Metropolis-Hastings. The use of Bayesian Model Averaging (BMA) incorporates the uncertainty of model selection in the estimation of the probabilities of alteration for each array/probe. This model allows us to include the natural heterogeneity present in most cancer data while giving clear biological interpretations to the parameters. The combination of an NH-HMM that allows the inclusion of the effect on the distance between probes, the introduction of random effects to reflect individual variability and the fact that the analyst does not need to predefine the number of hidden states (nor choose it a posteriori) makes this model a very valuable tool for the analysis of aCGH data.

The next two Sections introduce the model and describe the algorithm. Then results on simulated data are presented, showing that the model produces more precise estimators, and on a real data example, helping to identify common regions of copy number variation (CNVs). The supplementary material available online contains the mathematical details of the Reversible Jump algorithm.

2. A Bayesian non-homogeneous HMM for aCGH data

RJaCGH [15] is an NH-HMM model with an unknown number of hidden states (related to the different copy numbers) for the analysis of aCGH data. The observed log_2 ratios in a given array/chromosome are modeled using a mixture of normal distributions and the Markov dependence structure reflects that neighbor probes should share the same copy number (unless an abrupt change occurs). The distance between probes (typically very variable depending on the microarray platform) is incorporated into the model using transition functions between hidden states that describe the probability of remaining in the same hidden state (that is, to share the same copy number) as a decreasing function of the distance between two probes (in base pairs and scaled between 0 and 1).

Different arrays and chromosomes are likely to contain a variable number of aberrations, so fixing in advance the number of hidden states is not a realistic assumption. Moreover, scenarios like stromal contamination or intratumoral heterogeneity can produce averaged non-integer copy numbers for some arrays; for example in the case of a sample with only 75% of tumoral cells (a gain of 3 copies would be seen as 0.75×3).

Selecting the number of hidden states based on such popular measures as AIC or BIC is not appropriate for HMMs, as their consistency has not been proved (see [32, 33]; although a consistent method for many HMM models based on penalized minimum distances has been developed by [32]). RJaCGH uses Reversible Jump [34], a generalization of MCMC that can explore simultaneously the sample spaces of several models and estimates the posterior distribution of models. RJ eliminates the need for model selection through the use of BMA and incorporates the uncertainty in model selection. The algorithm can use a Gibbs or Metropolis-Hastings sampler, and incorporates delayed rejection ([35, 36]) and coupled parallel chains ([37]) to improve the estimation process.

If the number of hidden states cannot be fixed a priori for biological reasons, there is no injective mapping between hidden states and copy numbers either. The observed log-ratios for a set of probes with a given copy number may not be approximated accurately enough with a Gaussian distribution, but with a mixture of Gaussian distributions. This means that more than one hidden state may correspond to the same copy number if the distribution of its log ratios is complicated enough. Instead of assigning each hidden state to a copy number, RJaCGH assigns probabilities of alteration to each hidden state based on a normal reference μ_N, defined as the expected log-ratio for a normal copy number; a band is then formed around it with W_L, W_G width and the probability of alteration is computed as the probability for every state of drawing observations outside this region. This window can be asymmetric (the expected loss of one copy is $log_2(1/2)$ and the expected gain of one copy is $log_2(3/2)$), but it might depend on the particular dataset. Algorithm 1 summarizes the method.

Note that different values of W_L, W_G can be chosen for each array depending on the purity of the DNA sample. Scaling the values of these windows by an estimation of the percentage of tumoral cells can fix the problem of comparing copy numbers of samples with different proportions of normal contamination.

Algorithm 1: Fully probabilistic approach for state labelling

Input: normal reference: μ_N.
Input: window: W_L, W_G.
foreach *state* i **do**
  Compute $NI = (\mu_N - W_L, \mu_N + W_G)$;
  $P(i = Loss) = P(N_{(\mu_i,\sigma_i^2)} \leq NI_1)$;
  $P(i = Gain) = P(N_{(\mu_i,\sigma_i^2)} \geq NI_2)$;
  $P(i = Normal) = 1 - P(i = Loss) - P(i = Gain)$;
end
Output: Matrix of probabilities of alteration for every hidden state.

3. A Bayesian Random Effects Model for HMMs

Let us consider a collection of A arrays with the same set of P probes in each of them. We measure $y_{a,p}$, the log-ratio intensity for the array a and the probe p. An HMM for each array with k states is a bivariate stochastic process on (Y, S), which is the vector of observed log ratios and the vector of unobserved hidden states. The same HMM can be defined for all arrays and a random effect can be incorporated on the means of the hidden states per array:

$$(Y_{a,p}|S_{a,p} = j) \sim N(\mu_j + b_a, \sigma_j^2) \tag{1}$$

where

$$b_a \sim N(0, \sigma_b^2)$$

If the probe p is in the hidden state j for the array a, the observed $y_{a,p}$ follows a normal distribution with a mean that depends on the hidden state plus another mean that depends only on the array, and is the same for all the probes, regardless of the state. The effect is to shift the log-ratios per array by the same amount, a situation that is very typical in microarray experiments. σ_b^2 is the variance of the random effects and σ_j^2 can be the same for all the hidden states, reducing to a homoscedastic model.

The transition functions $q_{i,j}$ between two states i, j are:

$$q_{i,j}(x) = \frac{exp\{-\beta_{i,j} + \beta_{i,j}x\}}{\sum_{m=1}^{k} exp\{-\beta_{i,m} + \beta_{i,m}x\}} \tag{2}$$

where $\beta_{i,j} \geq 0 \quad \forall i, j$ and $\beta_{i,i} = 0 \quad \forall i$ to ensure that the parameters are uniquely defined. x is the distance between the two probes.

3.1. *Distribution of the parameters*

Let us consider the following priors, typical in mixture models [38]. The variance of the random effects and the variance of the hidden states follow Inverse Gamma distributions, the random effects and the means of the hidden states follow Gaussian distributions and the parameters of the transition functions follow Gamma distributions:

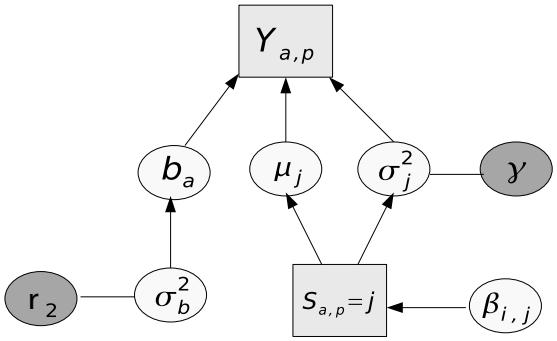

Figure 1. Model with a random effect per array. Squares: random sequences; log-ratios (visible) and hidden states (invisible). Light circles: random variables of main interest. Dark circles: hyperparameters.

$$\sigma_b^2 \sim \Gamma^{-1}(r_1, r_2)$$

$$b_a \sim N(0, \sigma_b^2)$$

$$\beta_{i,j} \sim \Gamma(1, 1)$$

$$\sigma_j^2 \sim \Gamma^{-1}(\kappa, \gamma)$$

$$\mu_j \sim N(\alpha_\mu, \beta_\mu)$$

We can make the posterior distributions less dependent on the priors, defining hyperprior Gamma distributions for the hyperparameters r_2 and γ:

$$r_2 \sim \Gamma(gb_1, gb_2)$$

$$\gamma \sim \Gamma(g_1, g_2)$$

Figure 1 shows a graphical representation of the model.

We can set sensible values for the hyperparameters such as $\alpha_\mu = median(y)$, $\beta_\mu = range(y)$, $k = 2$, $g_1 = 0.2$ and $g_2 = 1/range^2(y)$ where y is the vector obtained concatenating the log ratios of all the arrays. For gb_1 a small value such as 0.2 can be chosen, and for gb_2 $1/range^2(y)$, as in [38]. Finally, for r_1, a small value such as 1 can also be selected. When we have a small number of arrays, these values may have more influence on the posterior distribution of σ_b^2.

If the sequence of hidden states were known, the likelihood would be the following:

$$L(y|k,s,\mu,\sigma^2,\beta,b_a,\sigma_b^2) = \prod_{a=1}^{N} v(s_{a,1}) N(y_{a,1}|\mu_{s_{a,1}} + b_a, \sigma^2_{s_{a,1}}) \times$$
$$\prod_{p=2}^{P} q_{s_{a,p-1},s_{a,p}}(x_{p-1}|\beta) N(y_{a,p}|\mu_{s_{a,p}} + b_a, \sigma^2_{s_{a,p}}) \quad (3)$$

where v is the distribution of the initial states. This likelihood is often referred to as the complete likelihood in the literature. As we only observe the log-ratios, we have to average over all possible hidden paths for the likelihood, obtaining what is usually called the incomplete likelihood:

$$L(y|k,\mu,\sigma^2,\beta,b_a,\sigma_b^2) = \prod_{a=1}^{N} \sum_{s\in k^P} (v(s_{a,1}) N(y_{a,1}|\mu_{s_{a,1}} + b_a, \sigma^2_{s_{a,1}}) \times$$
$$\prod_{p=2}^{P} q_{s_{p-1},s_p}(x_{p-1}|\beta) N(y_{a,p}|\mu_{s_{a,p}} + b_a, \sigma^2_{s_{a,p}})) \quad (4)$$

Using conditional independence, the full conditional distributions (see [39] for an introduction) for all the parameters can be computed; technically, we obtain these distributions by multiplying the priors and the likelihood, keeping the relevant parts involved for that particular term and identifying the resulting distributions by their functional forms. Our particular choice of prior distributions results in closed form expressions for the full conditional distributions, namely:

$$r_2|. \sim \Gamma(gb_1 + r_1, gb_2 + \frac{1}{\sigma_b^2})$$

$$\gamma|. \sim \Gamma(g_1 + k\kappa, r_2 + \sum_{j=1}^{K} \sigma_j^{-2}$$

$$\sigma_b^2|. \sim \Gamma^{-1}(r_1 + N/2, r_2 + \sum_{a=1}^{N} b_a^2/2)$$

$$\mu_j|. \sim N\left(\frac{\frac{\alpha_\mu}{\beta_\mu} + \frac{\sum_{a=1}^{N} B_{a,j}}{\sigma_j^2} - \frac{\sum_{a=1}^{N} n_{a,j} b_a}{\sigma_j^2}}{\frac{1}{\beta_\mu} + \frac{\sum_{a=1}^{N} n_{a,j}}{\sigma_j^2}}, \frac{1}{\frac{1}{\beta_\mu} + \frac{\sum_{a=1}^{N} n_{a,j}}{\sigma_j^2}} \right)$$

$$b_a|. \sim N\left(\frac{\sum_{j=1}^{k} \frac{B_{a,j} - n_{a,j}\mu_j}{\sigma_j^2}}{\frac{1}{\sigma_b^2} + \sum_{j=1}^{k} \frac{n_{a,j}}{\sigma_j^2}}, \frac{1}{\frac{1}{\sigma_b^2} + \sum_{j=1}^{k} \frac{n_{a,j}}{\sigma_j^2}} \right)$$

The conditional distribution of variances, in the case of the heteroscedastic and homoscedastic models are:

$$\sigma_j^2|. \sim \Gamma^{-1}(C_1, C_2)$$

with

$$C_1 = \kappa + \frac{n.j}{2}$$

$$C_2 = \gamma + \frac{1}{2}\sum_{a=1}^{N}\left(B_{a,j}^2 + n_{a,j}b_a^2 - 2b_a \sum_{p/s_p=j}(y_{a,p} - \mu_j)\right)$$

and σ^2 follows an Inverse Gamma distribution:

$$\sigma^2|. \sim \Gamma^{-1}(C_1, C_2)$$

with

$$C_1 = \kappa + \frac{n..}{2}$$

$$C_2 = \gamma + \frac{1}{2}\sum_{a=1}^{N}\left(B_{a.}^2 + n_{a,.}b_a^2 - 2b_a \sum_{p}(y_{a,p} - \mu_{S_p})\right)$$

where $n_{a,j}$ is the sum of the probes in the array a with the hidden state j, $B_{a,j} = \sum_{p/s_p=j} y_{a,p}$, $B_{a,j}^2 = \sum_{p/s_p=j}(y_{a,p} - \mu_j)^2$ and $n_{.j} = \sum_{a=1}^{N} n_{a,j}$.

It is more convenient to update $\beta_{i,j}$, the parameters of the transition functions, using a Metropolis move, because it is very complicated to sample from the conjugate distribution:

Given the values $\beta_{i,j} \quad (i \neq j)$, we propose a candidate vector:

$$\log \beta_{i,j}^C = \log \beta_{i,j} + \epsilon_{i,j}^{\beta} \tag{5}$$

for all $i, j = 1, \ldots, k, i \neq j$, where $\epsilon_{i,j}^{\beta} \sim N(0, \tau_\beta^2)$.
We accept the candidate with probability $min(1, r)$, where:

$$r = \frac{\prod_{i,j=1,\ldots,k;i\neq j}\pi(\beta_{i,j}^C)}{\prod_{i,j=1,\ldots,k;i\neq j}\pi(\beta_{i,j})}\frac{L(y|k,s,\mu,\sigma^2,b_a,\sigma_b^2,\beta^C)}{L(y|k,s,\mu,\sigma^2,b_a,\sigma_b^2,\beta)}\prod_{i,j=1,\ldots,k;i\neq j}\frac{\beta_{i,j}^C}{\beta_{i,j}} \tag{6}$$

τ_β controls the magnitude of the jump to the next candidate and can be tuned to obtain a probability of acceptance around 0.23, as suggested, for example, by [40].

There are several schemes for sampling from the distribution of hidden states. The simplest one is probably local updating:

$$P(S_{a,p} = j|.) \propto q_{s_{a,p-1},j}(x_{p-1})\frac{1}{\sigma_j}\exp\left\{\frac{-(y_{a,p} - \mu_j - b_a)^2}{2\sigma_j^2}\right\} q_{j,s_{a,p+1}}(x_p) \tag{7}$$

For $S_{a,1}$, we replace $q_{S_{a,0},j}(x_0)$ with $v(j)$ and for $S_{a,P}$, we replace $q_{j,S_{a,P+1}}(x_j)$ with 1.

3.2. *Reversible Jump algorithm*

Our model does not assume a fixed number of hidden states, although for computational convenience we fix a maximum number of hidden states K. For each of the models M_k considered, we have a parameter vector $\theta(M_k) = \{\mu_{M_k}, \sigma^2_{M_k}, \beta_{M_k}, b_{M_k}, \sigma^2_{b,M_k}, s_{M_k}\}$

Reversible Jump allows us to jump between models using special moves. We use birth/death and split/combine, which are commonly used in the mixture modelling and HMM literature ([33, 38, 41]). A birth move consists in creating a new hidden state, and a death move in deleting an existing hidden state. A split move consists in taking an existing hidden state and dividing it into two, and the combine move in taking two adjacent states and joining them. Full details can be found in the supplementary material. Algorithm 2 shows the whole procedure using a Gibbs sampler.

Algorithm 2: RJaCGH with random effects per array with Gibbs Sampler

Input: T: number of iterations.
Input: K: maximum number of hidden states.
Input: τ_β: standard deviations of the random walk for updates within models.
Input: $\tau_{sp,\mu}$: parameter for the split move.
Input: $\mu_{\cdot}^{(0)}\sigma_{\cdot}^{2(0)}, \beta_{\cdot}^{(0)}, b_{\cdot}^{(0)}, \sigma_b^{2(0)}, s_{\cdot}^{(0)}, k^{(0)}$: initial values for each model.
for $t = 1$ *to* T **do**
  Gibbs Update: $\mu_k^{(t)}, \sigma_k^{2(t)}, s_k^{(t)}, b_k^{(t)}, \sigma_{b,k}^{2(t)}$;
  Metropolis Update: $\beta_k^{(t)}$;
  Select and try Birth or Death with delayed rejection ;
  Select and try Split or Combine ;
end
Output: Chain with samples from the joint distribution
$\bigcup_k \{k\} \times (\mu_k, \sigma_k^2, \beta_k, b_k, \sigma_{b,k}^2, s_k)$

If we want to use a Metropolis-Hastings (MH) sampler, conjugate priors do not have to be used, so the priors for the variances can be simplified:

$$\sigma_j \sim U(0, R)$$

$$\sigma_b \sim U(0, R_{\sigma_b^2})$$

where $R = range(y)$ and $R_{\sigma_b^2}$ can be a smaller value to prevent the variability among arrays being higher than the variability among probes.

In addition, we do not need to sample from the distribution of hidden states because the incomplete likelihood is used (see equation 4). We update the rest of the parameters generating candidates with random walks:

$$\mu_i^C = \mu_i + \epsilon_i^\mu \tag{8}$$

for all $i = 1, \ldots, k$, where $\epsilon_i^{\mu} \sim N(0, \tau_{\mu}^2)$.

$$\log \sigma_i^{2C} = \log \sigma_i^2 + \epsilon_i^{\sigma^2} \tag{9}$$

$$\log \sigma_b^{2C} = \log \sigma_b^2 + \epsilon^{\sigma_b^2}$$
$$b_a^C = b_a + \epsilon_a^b$$

with auxiliary variables generated with normals:

$$\epsilon^{\sigma_b^2} \sim N(0, \tau_{\sigma^2}^2)$$
$$\epsilon_a^b \sim N(0, \tau_{\mu}^2)$$

The step sizes can be the adjusted in the same way as for τ_{β} and the acceptance ratio for each parameter can be easily computed. The whole procedure is summarized in algorithm 3.

Algorithm 3: RJaCGH with random effects per array with Metropolis-Hastings and coupled chains

Input: NC: number of coupled chains.
Input: T: number of iterations.
Input: K: maximum number of hidden states.
Input: $\tau_{\mu}, \tau_{\sigma^2}, \tau_{\beta}$: standard deviations of the random walk for updates within models.
Input: $\tau_{sp,\mu}$: parameter for the split move.
Input: $\mu_{\cdot}^{(.,0)}, \sigma_{\cdot}^{2(.,0)}, \beta_{\cdot}^{(.,0)}, b_{\cdot}^{(.,0)}, \sigma_{b,\cdot}^{2(.,0)}, k^{(.,0)}$: initial values for each chain and each model.
for $t = 1$ *to* T **do**
  for $h = 1$ *to* NC **do**
    Metropolis Update: $\mu_k^{(h,t)}, \sigma_k^{2(h,t)}, \beta_k^{(h,t)}, b_k^{(h,t)}, \sigma_{b,k}^{2(h,t)}$;
    Select and try Birth or Death with delayed rejection ;
    Select and try Split or Combine ;
  end
  Select two chains and try to swap them ;
end
Output: Cool chain with samples from the joint distribution $\bigcup_k \{k\} \times (\mu_k, \sigma_k^2, \beta_k, b_k, \sigma_{b,k}^2)$

3.3. *Probabilities of alteration*

The algorithm returns samples from the posterior distributions of the parameters of each HMM (μ, σ^2, β, b, σ_b^2 and s in the case of the Gibbs Sampler) for all the models visited by the algorithm. Moreover, it returns samples from the posterior distribution of the number of hidden states. We are particularly interested in the probabilities for each array and each probe of belonging to any hidden state; probabilities that can be obtained for each model using the Viterbi algorithm (see [42]).

Furthermore, the probabilities for each hidden state of being a state of gain or loss can be computed using algorithm 1 and all this information can be combined using BMA, obtaining the following estimators of the probabilities of gain for a probe p in an array a (informally expressed as $P(Y_{a,p} = G)$):

$$P(Y_{a,p} = G) = \sum_{m=1}^{K} P(M = m) \sum_{j=1}^{m} P(Y_{a,p} \in S_j|m)P(S_j = G|m) \qquad (10)$$

That is, the uncertainty in the classification of probes into hidden states is incorporated, the uncertainty in the assignments of states into states of copy number is included too and also the uncertainty in model selection. This methodology leads to better estimators; there are theoretical results that state that, under certain conditions, BMA estimators minimize the mean squared error (MSE) among point estimators and that their predictive distribution is optimal (see [43] for details). The probability of loss can be computed in a similar way.

3.4. *Convergence*

There seems to be no perfect method for checking convergence, specially for RJMCMC. Besides, as [37] states, these methods only can tell us if the chain has not converged. In our experience, a careful design of the moves and good starting values can help reach convergence. In particular, techniques like delayed reaction and coupled parallel chains can help the sampler to explore the state parameter easily, helping to reach convergence. We have also observed that unrealistic initial values for σ_b^2 can increase the number of iterations needed to attain convergence.

4. Results

4.1. *Simulations*

We generated 150 observations (probes) from 25 individuals (arrays) from a mixture of 3 normal distributions (50 observations belonging to each of the groups) with means $\mu = \{-1.5, 0, 1.5\}$ and variances $\sigma^2 = \{1, 1, 1\}$. A random effect for each individual was added to these means generating 25 observations from a normal distribution centered in zero and with a variance chosen from 5 different values $\sigma_b^2 = \{0.1, 0.25, 0.5, 0.75, 1\}$. 100 replicates were generated for each of these five datasets with $150 * 25$ observations. We then ran four versions of RJaCGH: fitting the same HMM to the 25 arrays using a MH sampler, fitting a different HMM to each array using an MH sampler, fitting RJaCGH with a random effect per array using MH sampler and also fitting RJaCGH using a Gibbs sampler. We used a maximum of 6 states and ran the algorithm for $10,000$ samples as burn-in and another $10,000$ for the inferences. For every replication we computed the Mean Squared Error (MSE) of the $10,000$ samples as the average of the MSEs in the estimation of each parameter (using the median) in each hidden state. We also relabeled each hidden state as loss, normal or gain using a window of 0.75 (algorithm 1) and computed the MSE of estimating the probability for each probe to be correctly classified. According to the model, these probabilities should be 0.77, 0.55 and 0.77 for loss, normal and gain.

Table 1 shows the results for the simulations (medians and IQR of the 100 replicates). For the computation of the MSEs of the parameters of a 3-state HMM, we only use the visits of the sampler to the model with 3 hidden states. The es-

Table 1. Results of the simulations.The values are medians of 100 replicates and the IQR is shown in parenthesis. θ is the mean per hidden state and array ($\mu + b$). The MSE are Mean Squared Errors of the estimation of the parameters using the model with 3 states for the parameters and Bayesian Model Averaging over all models for the probabilities of correct classification. For each replication, they are the average of the MSE of each parameter and hidden state.

			Same Model		
σ_b^2	0.1	0.25	0.5	0.75	1
MSE μ	0.004 (0.01)	0.017 (0.02)	0.066 (0.07)	0.088 (0.10)	0.174 (0.16)
MSE θ	0.092 (0.03)	0.267 (0.10)	0.509 (0.22)	0.783 (0.35)	1.108 (0.51)
MSE σ^2	0.004 (0.00)	0.017 (0.02)	0.036 (0.03)	0.063 (0.04)	0.094 (0.06)
MSE σ_b^2	-	-	-	-	-
MSE P_{Loss}	0.039 (0.01)	0.072 (0.03)	0.103 (0.04)	0.125 (0.04)	0.142 (0.05)
MSE P_{Normal}	0.010 (0.00)	0.026 (0.01)	0.042 (0.02)	0.056 (0.02)	0.071 (0.02)
MSE P_{Gain}	0.040 (0.01)	0.066 (0.03)	0.099 (0.04)	0.112 (0.04)	0.146 (0.04)
			Different Model		
σ_b^2	0.1	0.25	0.5	0.75	1
MSE μ	0.118 (0.05)	0.278 (0.11)	0.507 (0.24)	0.766 (0.36)	0.999 (0.45)
MSE θ	0.024 (0.01)	0.024 (0.01)	0.026 (0.01)	0.025 (0.01)	0.024 (0.01)
MSE σ^2	0.018 (0.01)	0.018 (0.01)	0.018 (0.01)	0.018 (0.01)	0.018 (0.01)
MSE σ_b^2	-	-	-	-	-
MSE P_{Loss}	0.061 (0.02)	0.078 (0.03)	0.102 (0.03)	0.121 (0.03)	0.129 (0.04)
MSE P_{Normal}	0.026 (0.01)	0.036 (0.01)	0.051 (0.02)	0.069 (0.02)	0.082 (0.03)
MSE P_{Gain}	0.064 (0.02)	0.082 (0.03)	0.100 (0.03)	0.111 (0.03)	0.134 (0.04)
			Random Effects Model (MH)		
σ_b^2	0.1	0.25	0.5	0.75	1
MSE μ	0.003 (0.01)	0.009 (0.02)	0.018 (0.03)	0.030 (0.07)	0.033 (0.07)
MSE θ	0.007 (0.00)	0.008 (0.00)	0.009 (0.00)	0.008 (0.00)	0.008 (0.00)
MSE σ^2	0.000 (0.00)	0.000 (0.00)	0.000 (0.00)	0.000 (0.00)	0.000 (0.00)
MSE σ_b^2	0.001 (0.00)	0.004 (0.01)	0.010 (0.04)	0.021 (0.05)	0.025 (0.08)
MSE P_{Loss}	0.009 (0.01)	0.026 (0.02)	0.043 (0.02)	0.060 (0.03)	0.071 (0.03)
MSE P_{Normal}	0.068 (0.01)	0.100 (0.02)	0.133 (0.04)	0.152 (0.03)	0.167 (0.03)
MSE P_{Gain}	0.008 (0.01)	0.025 (0.01)	0.041 (0.02)	0.055 (0.03)	0.070 (0.03)
			Random Effects Model (Gibbs)		
σ_b^2	0.1	0.25	0.5	0.75	1
MSE μ	0.003 (0.01)	0.007 (0.02)	0.014 (0.03)	0.015 (0.04)	0.033 (0.07)
MSE θ	0.007 (0.00)	0.008 (0.00)	0.009 (0.00)	0.008 (0.00)	0.009 (0.00)
MSE σ^2	0.000 (0.00)	0.000 (0.00)	0.000 (0.00)	0.000 (0.00)	0.000 (0.00)
MSE σ_b^2	0.000 (0.00)	0.004 (0.01)	0.016 (0.05)	0.029 (0.06)	0.057 (0.17)
MSE P_{Loss}	0.003 (0.00)	0.004 (0.00)	0.004 (0.00)	0.005 (0.00)	0.005 (0.01)
MSE P_{Normal}	0.060 (0.01)	0.061 (0.01)	0.063 (0.01)	0.062 (0.01)	0.064 (0.02)
MSE P_{Gain}	0.003 (0.00)	0.003 (0.00)	0.004 (0.00)	0.004 (0.00)	0.005 (0.01)

timation of μ and θ, the mean + the random effect, fitting the same HMM and fitting a different one are not optimal, showing that when there is variability in the mean levels of the arrays they cannot be estimated accurately. The MSE for the variance increases in the model with the same HMM when the random variance increases. The random effects model produces better estimations with both samplers, although it shows some variability in the estimation of the variance of the random effects when it is large. For the estimation of the classification error, we have used BMA over all the HMMs fitted. Again, we can see that the random effects model outperforms the other two, especially using the Gibbs Sampler.

4.2. *Application to real tumour data*

[44] present data from 44 breast tumours and compare frequencies of alteration to several clinical variables. [42] analyze the data and compute common regions of alteration, showing that a certain amount of heterogeneity among individuals is present in this dataset, as expected in breast cancer. We have compared our three approaches to these data to see if this heterogeneity implies that a different HMM is needed for each array. So we ran $20,000$ MCMC runs ($10,000$ of them as burn-in)

for the model with the same HMM (s-HMM) and the model with an independent HMM for each array (i-HMM) and 30,000 MCMC runs (10,000 of them as burn-in) for the HMM with random effects (re-HMM) with a Gibbs sampler (as this model is more complex and requires more iterations). The maximum number of hidden states was 10.

Figure 2 summarizes the fits of the three models. The left panel shows the medians of the posterior distribution for the means of each hidden state. Both models, s-HMM and re-HMM have selected 10 hidden states in all final iterations, and the estimations are similar. Note that neither the re-HMM model nor the s-HMM model seem to have a hidden state for homozygous deletions (zero copies of a gene); while looking at the 44 fits of the i-HMM, some of them include two clear levels of losses (1 and 0 copies). A closer inspection shows that these correspond to cases with only one probe, thus the models that use information from all samples (s-HMM and re-HMM) consider these points as outliers. The top right panel shows the probabilities of the number of hidden states for the i-HMM averaged on the 44 arrays. On average, a typical array needs 3-5 states. Both s-HMM and re-HMM use the maximum number allowed (10), but to widen this limit would lead to a huge increase in computing time and no improvement in the estimations (data not shown). The rest of the right panels show distributions for the variances and the random effects.

The main differences in the s-HMM and re-HMM appear in the way that the probabilities of alteration are computed for each probe. First, each hidden state was classified as loss, neutral or gain using algorithm 1. The left panel of Figure 2 shows with black squares the hidden states with a maximum probability of being classified as 'gains', and with black triangles those with a maximum probability of being classified as 'losses'. It can be seen that small differences in estimation may lead to differences in probabilities of classification. The re-HMM classifies the fifth hidden state as gain with a probability of 0.59, while the s-HMM gives it only 0.45, although there was just a difference of 0.05 between the estimated means.

We next used BMA to obtain probabilities of alteration for each probe and averaged them over all tumours. These probabilities can be seen in figure 3. It shows the probability of aberration for each genomic region averaged over all 44 arrays. The three fits appear similar, but there are some important differences. For example, there is a region at the beginning of chromosome 1 with a probability of alteration of about 0.25 in re-HMM, but with a probability of only 0.12 in the s-HMM. This is an area with well known CNVs, as listed in [45]. There is another region in chromosome 5 with a probability of alteration in re-HMM of 0.3 and only 0.14 in i-HMM that corresponds to the genes BIRC1 and SMA5, both flanked by a CNV (see [45], Chrom5: 69,865,970,-70,533,749). NCF1 is another CNV (see [46]) with a probability of alteration of 0.29 in re-HMM, 0.18 in s-HMM and 0.20 in i-HMM. There are further examples of differences in these probabilities that show that the random effects model produces better estimations than the i-HMM, because it borrows information from all the samples, and than the s-HMM, because it allows for individual variability.

5. Conclusions and future directions

Random effects HMMs provide an elegant and concise way to model a wide range of biological problems. In particular, they provide a natural framework for the analysis of copy number data. We have shown through simulations that, in situations where individual variability is present, this model leads to more accurate estimates. Copy number data from tumour samples contain multiple sources of

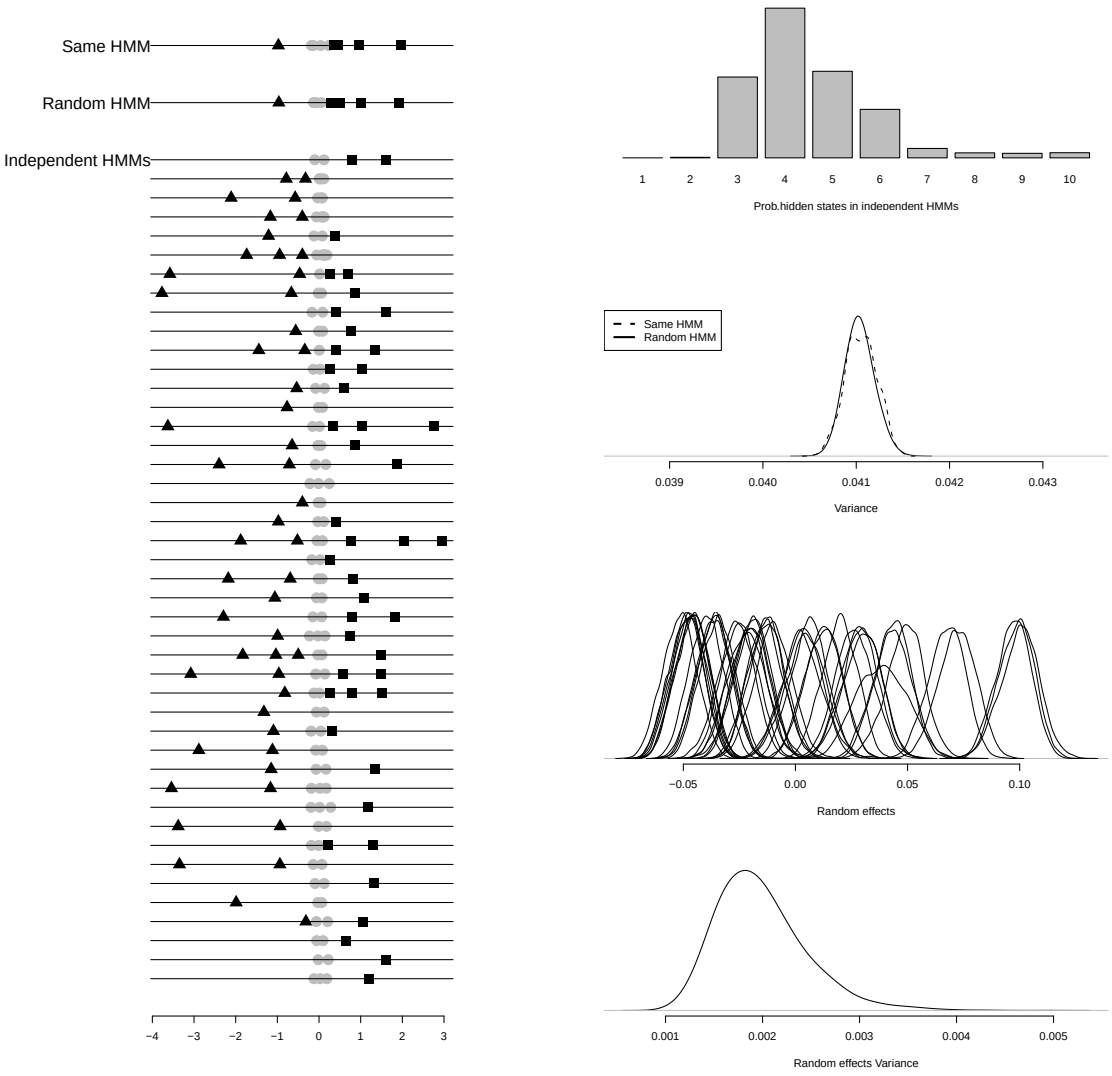

Figure 2. Results of the fits for the three models for the Pollack et al. dataset. Left panel: Medians of the posterior probability of the means of the hidden states. They have been labeled depending on the copy number state with the maximum probability assigned by algorithm 1. Black triangles are states of loss, grey circles are states of normal copy number and black squares states of gain. Right panel: starting from the top, posterior probabilities for the number of hidden states in the model with independent HMMs for each array, posterior distribution of the variances in the model with the same HMM for all arrays and random effects HMM, posterior distributions for the random effects in the model with random effects HMM and posterior distribution of the variance of the random effects in the model with random effects HMM.

heterogeneity, such as contamination of normal cells in the sample, intratumoral heterogeneity, or aneuploidy. The introduction of random effects in the HMM can help in the correct modelling of these effects, leading to better probability estimates of alteration and therefore helping in the identification of potential tumour driver genes. In a real data example we have seen how the introduction of one random effect refines these probabilities and helps to identify regions of CNV. Although the interpretation of random effects is in this case straightforward, it is not the case for the interpretation of the hidden states. They cannot simply be related to individual copy numbers, and as we have seen in the example, values supported

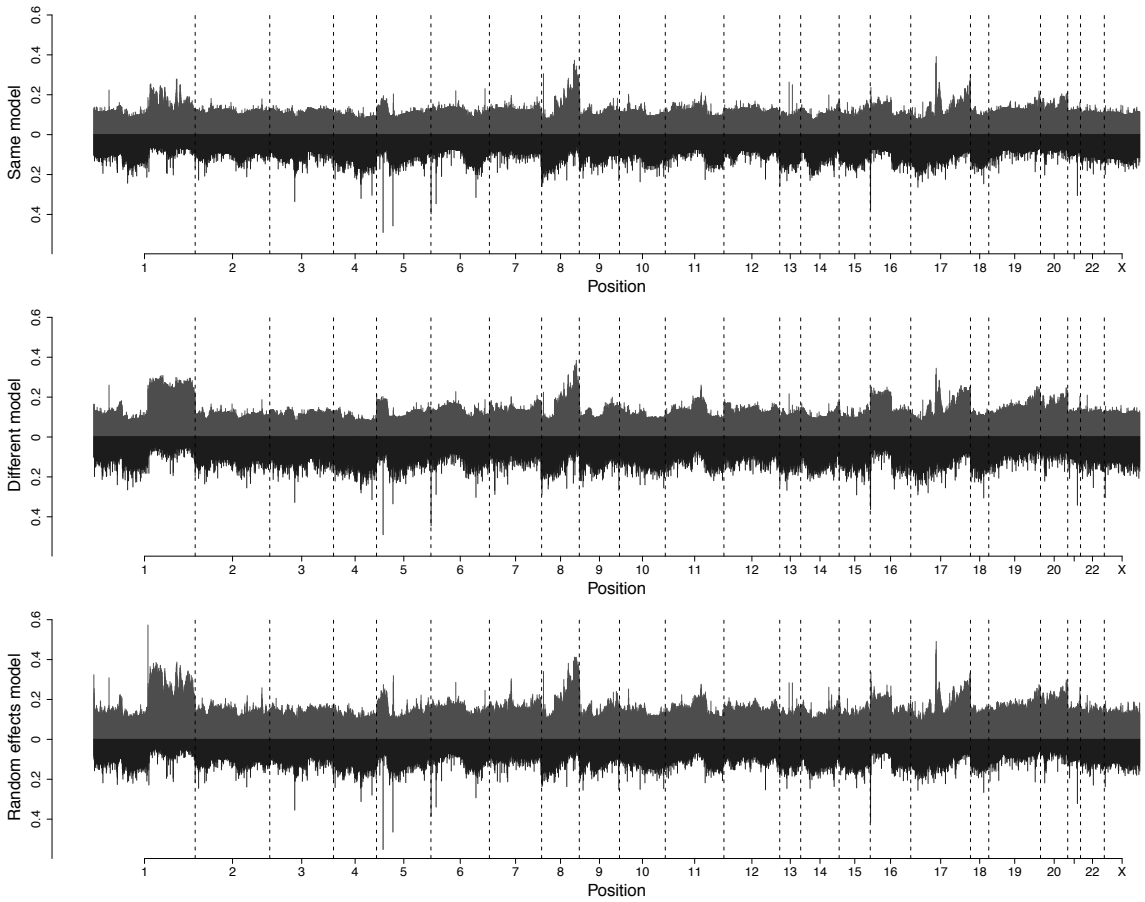

Figure 3. Probabilities of alteration for Pollack et al. Each panel shows the probability of alteration (gains and losses) for a given model (same HMM model for all arrays, different HMM model and random effects HMM) over the genome. These probabilities are averaged over all HMM models with a different number of hidden states and over all arrays.

by just a few observations in a few samples are not guaranteed to form their own hidden state. This, however, is not a major problem if we want to estimate probabilities of gain/loss instead of estimating absolute copy numbers (something that simply cannot be done with aCGH data). More importantly, this situation stresses the danger of using HMMs with a fixed number of states, and it also emphasizes the big influence that the priors can have in the posterior distributions of these models (see [38]). RJMCMC provides a way to deal with these issues and improve estimation averaging predictions from HMMS with a different number of hidden states.

This model can be expanded in several interesting ways to incorporate different sources of variability. A random effect for the means in each hidden state can be added for situations where each individual has different levels of alterations (for example, due to intratumoral heterogeneity). A random effect can also be included in the transition probability functions, modelling scenarios where there is a large difference in proportions of alterations amongst individuals. These situations can easily be adapted to our model and an appropriate RJMCMC algorithm can be designed. However, the method presented here will need strong requirements for the analysis of high-density microarray platforms with millions of probes in terms of memory and computing time. For the analysis of these big data sets we suggest using our method for a detailed analysis of preselected interesting regions or segmenting each array with a fast smoothing algorithm (like [8, 9]), then selecting

regions with copy number changes and finally applying our model to them.

Supplementary Material

R code for the models is available upon request. The material referenced in Section 3.2 is available in the on-line Supplementary Material.

Acknowledgments

This research was supported by the Spanish Ministerio de Ciencia e Innovación (grants MTM 2009-11161 and BIO2009-12458) and Fundación de Investigación Médica Mutua Madrileña. We want to thank an anonymous reviewer for his comments that improved the quality of the paper.

References

[1] M.A. Heiskanen, M.L. Bittner, Y. Chen, J. Khan, K.E. Adler, J.M. Trent, and P.S. Meltzer, *Detection of gene amplification by genomic hybridization to cDNA microarrays.*, Cancer Research 60 (2000), pp. 799–802.

[2] K. Holzmann et al., *Genomic DNA-chip hybridization reveals a higher incidence of genomic amplifications in pancreatic cancer than conventional comparative genomic hybridization and leads to the identification of novel candidate genes.*, Cancer Research 64 (2004), pp. 4428–4433.

[3] B. Ylstra, P. Ijsselvan den , B. Carvalho, R.H. Brakenhoff, and G.A. Meijer, *BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH).*, Nucleic Acids Research 34 (2006), pp. 445–450.

[4] D.S. Tan, M.B. Lambros, R. Natrajan, and R.F.J. S., *Getting it right: designing microarray (and not 'microawry') comparative genomic hybridization studies for cancer research*, Laboratory Investigation 87 (2007), pp. 737–754.

[5] C. Curtis, A.G. Lynch, M.J. Dunning, I. Spiteri, J.C. Marioni, J. Hadfield, S.F. Ching, J.D. Brenton, S. Tavaré, and C. Caldas, *The pitfalls of platform comparison: DNA copy number array technologies assessed*, BMC Genomics 10 (2009).

[6] H. Bengtsson, A. Ray, P. Spellman, and T. Speed, *A single-sample method for normalizing and combining full-resolution copy numbers from multiple platforms, labs and analysis methods*, Bioinformatics 25 (2009), pp. 861–867.

[7] S. Pounds, C. Cheng, C. Mullighan, S. Raimondi, S. Shurtleff, and J. Downing, *Reference alignment of SNP microarray signals for copy number analysis of tumors*, Bioinformatics 25 (2009), pp. 315–321.

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

[9] E. Ben-Yaacov and Y.C. Eldar, *A fast and flexible method for the segmentation of aCGH data*, Bioinformatics 24 (2008), pp. 139–145.

[10] R. Pique-Regi, J. Monso-Varona, A. Ortega, R. Seeger, T. Triche, and S. Asgharzadeh, *Sparse representation and Bayesian detection of genome copy number alterations from microarray data*, Bioinformatics 24 (2008), pp. 309–318.

[11] J. Fridlyand, A.M. Snijders, D. Pinkel, D.G. Albertson, and A.N. Jain, *Hidden Markov models approach to the analysis of array CGH data*, Journal of Multivariate Analysis 90 (2004), pp. 132–153.

[12] J.C. Marioni, N.P. Thorne, and S. Tavaré, *BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data.*, Bioinformatics 22 (2006), pp. 1144–1146.

[13] S.P. Shah, X. Xuan, R.J. Deleeuw, M. Khojasteh, W.L. Lam, R. Ng, and K.P. Murphy, *Integrating copy number polymorphisms into array CGH analysis using a robust HMM.*, Bioinformatics 22 (2006), pp. e431–e439.

[14] S. Stjernqvist, T. Rydén, M. Sköld, and J. Staaf, *Continous-index hidden Markov modeling of array CGH copy number data*, Bioinformatics 23 (2007), pp. 1006–1014.

[15] O.M. Rueda and R. Diaz-Uriarte, *Flexible and Accurate Detection of Genomic Copy-Number Changes from aCGH*, PLoS Computational Biology 3 (2007), p. e122.

[16] S. Colella, C. Yau, J. Taylor, G. Mirza, H. Butler, P. Clouston, A. Bassett, A. Seller, C. Holmes, and J. Ragoussis, *QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data.*, Nucleic Acids Research 35 (2007), pp. 2013–2025.

[17] K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S. Grant, H. Hakonarson, and M. Bucan, *PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data.*, Genome Research 17 (2007), pp. 1665–1674.

[18] C. Greenman et al., *PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data.*, Biostatistics 11 (2010), pp. 164–175.

[19] R. Pique-Regi, A. Ortega, and S. Asgharzadeh, *Joint estimation of copy number variation and reference intensities on multiple DNA arrays using GADA.*, Bioinformatics 25 (2009), pp. 1223–1230.

[20] S. Kim and P. Smyth, *Segmental Hidden Markov Models with random effects for waveform modeling*, Journal of Machine Learning Research 7 (2006), pp. 945–969.
[21] R.M. Altman, *Mixed Hidden Markov Models: An Extension of the Hidden Markov Model to the Longitudinal Data Setting*, Journal of the American Statistical Association 102 (2007), pp. 201–210.
[22] W. Zucchini, D. Raubenheimer, and I.L. MacDonald, *Modeling Time Series of Animal Behavior by Means of a latent-state model with feedback.*, Biometrics 64 (2008), pp. 807–815.
[23] F. Rijmen, E. Ip, S. Rapp, and E. Shaw, *Qualitative longitudinal analysis of symptoms in patients with primary and metastatic brain tumors*, Journal of the Royal Statistic Society, Series A 171 (2008), pp. 739–753.
[24] A. Maruotti and T. Rydén, *A semiparametric approach to hidden Markov models under longitudinal observations.*, Statistics and Computing 19 (2009), pp. 381–393.
[25] F. Chaubert-Pereira, Y. Guédon, C. Lavergne, and C. Trottier, *Markov and Semi-Markov Switching Linear Mixed Models Used to Identify Forest Tree Growth Components*, Biometrics 66 (2010), pp. 753–762.
[26] H.J. Seltman, *Hidden Markov Models for Analysis of Biological Rhythm Data*, , in *Case Studies in Bayesian Statistics* Springer-Verlag, 2002, pp. 397–405.
[27] P.G. Ridall and A.N. Pettitt, *Bayesian Hidden Markov Models for longitudinal counts*, Australian & New Zealand Journal of Statistics 47 (2005), pp. 129–145.
[28] S.L. Scott, G.M. James, and C.A. Sugar, *Hidden Markov Models for Longitudinal Comparisons*, Journal of the American Statistical Association 100 (2005), pp. 359–369.
[29] J.C. Detilleux, *The analysis of disease biomarker data usign a mixed hidden Markov model*, Genetics, selection, evolution 40 (2008), pp. 491–509.
[30] K.E. Shirley, D.S. Small, K.G. Lynch, S.A. Maisto, and D.W. Oslin, *Hidden Markov models for alcoholism treatment trial data*, The Annals of Applied Statistics 4 (2010), pp. 366–395.
[31] D. Engler, G. Mohaptra, D. Louis, and R. Betensky, *A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations*, Biostatistics 7 (2006), pp. 399–421.
[32] R.J. MacKay, *Estimating the order of a hidden Markov model*, The Canadian Journal of Statistics 30 (2002), pp. 573–589.
[33] O. Cappé, E. Moulines, and T. Rydén *Inference in Hidden Markov Models (Springer Series in Statistics)*, Springer, 2005.
[34] P. Green, *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*, Biometrika (1995), pp. 711–732.
[35] L. Tierney and A. Mira, *Some adaptive Monte Carlo methods for Bayesian inference*, Statistics in Medicine 18 (1999), pp. 2507–2515.
[36] P.J. Green and A. Mira, *Delayed Rejection in Reversible Jump Metropolis Hastings*, Biometrika 88 (2001), pp. 1035–1053.
[37] C. Geyer, *Markov Chain Monte Carlo Maximum Likelihood*, in *Computing Science and Statistics*Keramidas ed., , Proceedings of the 23th Symposium on the Interface, 1991, pp. 156–163.
[38] S. Richardson and P.J. Green, *On Bayesian Analysis of Mixtures with an unknown number of components*, Journal of the Royal Statistical Society, Series B 59 (1997), pp. 731–792.
[39] W. Gilks, *Full Conditional Distributions*, in *Markov Chain Monte Carlo in Practice*, W. Gilks, S. Richardson and D. Spiegelhalter, eds., Chapman & Hall/CRC Interdisciplinary Statistics, 1996.
[40] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin *Bayesian Data Analysis, Second Edition*, Chapman & Hall/CRC, 2003.
[41] C. Robert, T. Rydén, and D.M. Titterington, *Bayesian Inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method*, J. R. Statist. Soc. B 62 (2000), pp. 57–75.
[42] O. Rueda and R. Diaz-Uriarte, *Detection of recurrent copy number alterations in the genome: taking among-subject heterogeneity seriously.*, BMC Bioinformatics 10 (2009).
[43] A. Raftery and Y. Zheng, *Discussion: Performance of bayesian model averaging*, Journal of the American Statistical Association 98 (2003), pp. 931–938.
[44] J.R. Pollack, T. Srlie, C.M. Perou, C.A. Rees, S.S. Jeffrey, P.E. Lonning, R. Tibshirani, D. Botstein, A.L. Brresen-Dale, and P.O. Brown, *Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors.*, Proceedengs of the National Academy of Science of USA 99 (2002), pp. 12963–12968.
[45] D. Conrad et al., *Origins and functional impact of copy number variation in the human genome*, Nature 464 (2010), pp. 704–712.
[46] T. Brunson, Q. Wang, I. Chambers, and Q. Song, *A copy number variation in human NCF1 and its pseudogenes*, BMC Genetics 11 (2010).