Brazilian Journal of Probability and Statistics
2016, Vol. 30, No. 3, 345–365
DOI: 10.1214/15-BJPS283

A semiparametric Bayesian model for comparing DNA copy numbers

Luis Nieto-Barajas[a], **Yuan Ji**[b] **and Veerabhadran Baladandayuthapani**[c]

[a]*ITAM*
[b]*NorthShore University HealthSystem and University of Chicago*
[c]*University of Texas MD Anderson Cancer Center*

Abstract. We propose a two-step method for the analysis of copy number data. We first define the partitions of genome aberrations and conditional on the partitions we introduce a semiparametric Bayesian model for the analysis of multiple samples from patients with different subtypes of a disease. While the biological interest is to identify regions of differential copy numbers across disease subtypes, our model also includes sample-specific random effects that account for copy number alterations between different samples in the same disease subtype. We model the subtype and sample-specific effects using a random effects mixture model. The subtype's main effects are characterized by a mixture distribution whose components are assigned Dirichlet process priors. The performance of the proposed model is examined using simulated data as well as a breast cancer genomic data set.

1 Introduction

There has been increasing interest in constructing the genomic architecture of breast cancer based on alterations in the DNA copy number. The idea is to characterize different subtypes of breast cancer by examining the whole-genome copy number profiles. In this paper, we present a Bayesian semiparametric model to analyze DNA copy number data for multiple samples with multiple conditions, for example, disease subtypes.

An example of such data is a set of 122 breast cancer samples known to fall into three subtypes of breast cancer, namely estrogen receptor-positive (ER+), progesterone receptor-positive (PR+) and triple negative (TN) breast cancer. These three subtypes potentially possess different copy number profiles in various regions of the genome. The scientific aims are to assess: (1) the segmented regions of copy number alterations within each subtype across the genome, and (2) the regions of differential copy numbers between subtypes.

We propose a two-step method that defines the partitions of genome aberrations for each sample in the initial step, and a flexible semiparametric Bayesian framework for jointly modeling all samples known to belong to either of two disease

Key words and phrases. Bayesian nonparametrics, bivariate spike and slab prior, circular binary segmentation, comparative genomic hybridization, Dirichlet process mixture model, random effects.

Received January 2014; accepted February 2015.

subtypes as a second step. We report posterior inferences based on the proposed Bayesian models and make decisions by controlling the false discovery rate (FDR) (e.g., Newton et al., 2004). To start, we introduce the biological background and present a brief literature review.

1.1 DNA copy number and arrayCGH

The normal DNA of human beings has 23 pairs of chromosomes. One pair is the sex chromosomes and the other 22 pairs are autosomal chromosomes, or autosomes. The chromosomes in an autosomal pair are identical; hence, the *copy number* of DNA is two for each autosome.

Copy number alterations (CNAs) refer to the variations (from two) in the copy number of DNA, which are common in cancer and other genetic diseases. CNAs often result from abnormal genetic events, such as a series of mutations that cause discrete gains or losses in contiguous segments of DNA. To detect genome-wide CNAs, array-based hybridization technology, such as the array comparative genomic hybridization (arrayCGH), has been widely applied (Pinkel et al., 1998; Snijders et al., 2001; Pinkel and Albertson, 2005). In short, arrayCGH uses microarrays consisting of thousands or millions of genomic targets or "probes" that are spotted or printed on a glass surface. These probes usually span the whole genome with a resolution of the order of magnitude ranging from one probe per one million base pairs (1 MB) for a bacterial artificial chromosome, to one probe per 50–100 kilo base pairs (kb). A DNA *test* sample of interest is labeled with a dye (say, Cy3) and then mixed with a diploid reference sample that is labeled with a different dye (say, Cy5). The combined sample is then hybridized to the microarrays and the intensities of both colors are measured through an imaging process.

The quantity of interest is the $\log_2$ ratio of the two intensities for each probe. The collection of the intensity ratios then contains useful information about genome-wide changes in copy numbers. In the reference, the copy number of each DNA fragment is always two; thus, the intensity ratio is determined by the copy number of the DNA in the test sample. If that copy number is also two, the $\log_2$ intensity ratio equals zero, that is, no CNA. If there is a single copy loss in the test sample, the theoretical ratio is $\log_2 1/2 = -1$, assuming all the cells in the test sample lost one copy of the DNA fragment. Likewise, if there is a single copy gain, the theoretical ratio is $\log_2 3/2 = 0.58$. Multiple copy gains are called *amplifications*, and the corresponding theoretical intensity ratios are $\log_2 4/2$, $\log_2 5/2$, etc. When both copies are lost (called *deletion*), the theoretical ratio is $-\infty$, and a large negative value is usually observed in experiments. Due to the fact that not all the cells in the test sample have the same copy number and there is a possibility of tumor heterogeneity and other genetic contamination such as cross-hybridization, the absolute values of the observed intensity ratios are often smaller than their theoretical values.

1.2 Previous work on arrayCGH analysis

There have been a number of approaches proposed for analyzing arrayCGH data depending on the scientific question of interest. A common starting point of most investigations is in locating genomic regions that have abnormal copy numbers and determining the number of DNA copies in that region. In the frequentist domain, these include hidden Markov models (Fridlyand et al., 2004), finite mixture models (Hodgson et al., 2001) and change-point models (Olshen et al., 2004; Pollack et al., 2002) and penalization approaches such as least squares criterion (Huang et al., 2005), penalized quantile smoothing (Eilers and de Menezes, 2005), and fused lasso penalty (Tibshirani and Wang, 2008). While these approaches (at times) enable fast fitting due to their model construction and provide point estimations, however, they do not explicitly provide quantification of uncertainties associated with the genomic copy number aberrations. To overcome these challenges, Bayesian probabilistic approaches have been proposed by Guha, Li and Neuberg (2008) who use a parametric hidden Markov model to assess copy number aberrations at the probe-level. In a Bayesian nonparametric setting, Yau et al. (2011) proposed a mixture model that combines a hidden Markov model for the locations (states), with a Dirichlet process prior for the scales. However, all the above approaches are only applicable for single sample analyses and do not provide a mechanism to borrow strength between samples to detect population level copy number changes.

Recently, several approaches have been developed to allow joint modeling of multiple arrayCGH samples. These include segmentation methods based on generalized fused lasso (Zhang, Lange and Sbatti, 2012) and correlated random-effect models of Teo et al. (2011). Bayesian methods for single samples analysis have been provided by Baladandayuthapani et al. (2010) who use a Bayesian segmentation model to detect shared aberrations between multiple samples, and Shah et al. (2007) who propose a class of novel hierarchical hidden Markov models for recurrent copy number aberrations. However, these class of methods suffer from two drawbacks. First, they rely on parametric models that do not allow more flexible structures to be determined from the data, and second, they do not explicitly test (or model) differential aberrations between multiple populations of samples—a gap in literature this work aims to fill.

In this paper, we generalize the previous methods in two key directions:

(i) First, from a statistical modeling point of view, we propose a semiparametric Bayesian model for arrayCGH data, where we build hierarchical models with mixture specifications and Dirichlet process ($\mathcal{DP}$) priors (Ferguson, 1973) for the copy number states of specific genomic regions. The semiparametric formulation allows for a more flexible and adaptive data-driven mechanism for identification of copy number aberrations.

(ii) More importantly, the proposed models account for variability in the samples, both within and between different conditions, such as cancer subtypes. This enables researchers to borrow strength across samples within the same condition, as well as to infer the identification of differential copy numbers between different conditions.

Our approach allows for borrowing strength across repeated experiments and does not rely on specific parametric distributions. A nonparametric specification for the copy number states prevents us from considering a finite number of states (typically loss, neutral and gain) and allows us to cope with more states present in the data. Additionally, we compare different disease subtypes by considering a kind of bivariate spike and slap prior.

The paper is structured thusly: In Section 2, we present the semiparametric model and the corresponding posterior distributions required to make inference; in Section 3, we describe the analysis of simulated and real data; and in Section 4, we conclude with a discussion.

2 Semiparametric modeling

2.1 Notation

For ease of exposition, we illustrate our proposed model for one chromosome. The same model is used for other chromosomes in the analysis. Let n denote the number of probes printed on the microarray for the chromosome. Let $\mathcal{A} = \{t_1, t_2, \ldots, t_n\}$ be the index of probes. These indexes are ordered based on the physical genomic locations of the probes on the chromosome. For example, probe t_1 is located at the very beginning of the chromosome (e.g., at the p-arm) while probe t_n is at the end (e.g., the q-arm). Typically, the number of probes is the same for all samples from the same platform. When different types of microarrays are used, we assume that $\mathcal{A}$ is a union of all the probes in the samples. For each sample j, we assume that there are n_j probes, which are a subset of $\mathcal{A}$.

To build the models, we require $J + 1$ different partitions of $\mathcal{A}$. One partition for each sample j, $\{\Delta_{lj}\}_{l=1}^{L_j}$, $j = 1, \ldots, J$, plus a common partition $\{\Omega_k\}_{k=1}^{K}$ for all samples. For each sample j, the partition $\{\Delta_{lj}\}_{l=1}^{L_j}$ is defined such that $\Delta_{lj} \cap \Delta_{l',j} = \varnothing$ and $\bigcup_{l=1}^{L_j} \Delta_{lj} = \mathcal{A}$. Each element Δ_{lj} contains a consecutive set of indexes in $\mathcal{A}$. That is, denoting $\{t_1 = c_{1j} < c_{2j} \cdots < c_{L_j+1,j} = t_n\}$ a subset of $\mathcal{A}$, we define

$$\{\Delta_{lj} = [c_{lj}, c_{l+1,j}), l = 1, \ldots, L_j\}$$

as a partition of $\mathcal{A}$. Some probes, t_i's, may not be present in sample j, in which case we simply remove those probes t_i's and the partition remains unchanged. These partitions can be obtained, for example, by applying circular binary segmentation (CBS) (Olshen et al., 2004) to each sample, j.

The common partition $\{\Omega_k\}_{k=1}^{K}$ of size K is defined as the union of all partition segments over $j = 1, \ldots, J$. That is, $\Omega_k = [c_k, c_{k+1})$ with $\{t_1 = c_1 < c_2 \cdots <$

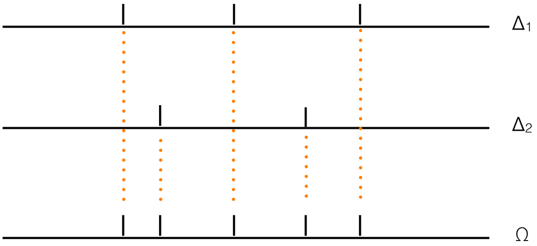

Figure 1 *Toy example with two individual partitions* $\{\Delta_1\}$ *and* $\{\Delta_2\}$ *and the common partition* $\{\Omega\}$.

$c_{K+1} = t_n\} = \bigcup_j \{t_1 = c_{1j} < c_{2j} \cdots < c_{L_j+1,j} = t_n\}$. Therefore, for a given probe t_i, it must be in one and only one Ω_k, for $k = 1, \ldots, K$. Note that this common partition is finer than each of the individual sample partitions. To better understand the relation between individual and common partition, we show in Figure 1 a toy example with two individual partitions and how they relate to form the common partition.

Let $g_j \in \{1, 2\}$ indicate the disease subtype for sample j. In our motivating example, $g_j = 1$ denotes the ER+ subtype and $g_j = 2$ denotes the TN subtype of breast cancer. We also define some distribution notations: $\mathrm{N}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2; $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a p-variate multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance–covariance matrix $\boldsymbol{\Sigma}$; $\mathrm{Ga}(\alpha, \beta)$ is a gamma distribution with mean α/β; $\mathrm{IGa}(\alpha, \beta)$ represents an inverse gamma distribution with mean $\beta/(\alpha - 1)$; and $\mathrm{Ber}(\pi)$ denotes a Bernoulli distribution with success parameter π. $\mathcal{DP}(a, F)$ is a Dirichlet process with precision parameter a and centering measure F. We proceed with the introduction of a sampling model, followed by the priors.

2.2 Probability model

With the two aforementioned objectives in mind, we construct the following hierarchical models. Let Y_{ij} be the $\log_2$ ratio of probe t_i at sample j. We assume that Y_{ij} arises from the sum of a population mean, a sample-specific mean, plus a measurement error. Specifically, let the population mean be μ_{k,g_j}, if probe t_i is in the population segment Ω_k and sample j is from disease subtype g_j; let m_{lj} be a random effect specific to sample j, assuming that probe t_i is in sample-specific segment Δ_{lj}; and denote the measurement error by ϵ_{ij}. For simplicity, let us assume that $t_i = i$. In summary, the model has a linear expression of the form

$$Y_{ij} = \sum_{k=1}^{K} \mu_{k,g_j} I(i \in \Omega_k) + \sum_{l=1}^{L_j} m_{lj} I(i \in \Delta_{lj}) + \epsilon_{ij}, \tag{2.1}$$

for $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$. We assume that the measurement errors ϵ_{ij} are independent and identically distributed $\mathrm{N}(0, \sigma_\epsilon^2)$.

We set up the following prior distributions to fulfill our objectives. Denote by $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2})$ the vector of population copy number levels for subtypes 1 and 2, respectively, with distribution G. We construct G as a mixture of two distributions G_0 and G_1, which in turn are assigned nonparametric Dirichlet process priors. In notation, we have

$$\begin{aligned} \boldsymbol{\mu}_k \mid G &\overset{\text{ind.}}{\sim} G, \qquad \text{for } k = 1, \dots, K \\ G &= (1-\pi)G_0 + \pi G_1 \\ G_r \mid a_r &\overset{\text{ind.}}{\sim} \mathcal{DP}(a_r, F_r), \qquad r = 0, 1, \end{aligned}$$

where a_r and F_r are the precision and centering measure parameters, respectively. Thus, the $\boldsymbol{\mu}_k$'s turn out to be partially exchangeable. For the centering measures of the nonparametric Dirichlet process priors, we take a degenerate bivariate normal on the identity line $F_0(\boldsymbol{\mu}_k) = \mathrm{N}(\mu_{k1} \mid 0, \lambda_0^2) I(\mu_{k1} = \mu_{k2})$ and a proper bivariate normal $F_1(\boldsymbol{\mu}_k) = \mathrm{N}_2(\boldsymbol{\mu}_k \mid \mathbf{0}, \boldsymbol{\Lambda}_1)$, with λ_0^2 as a nonnegative scalar and $\boldsymbol{\Lambda}_1$ a positive defined variance–covariance matrix, where we take the covariance to be zero, that is, $\boldsymbol{\Lambda}_1 = \mathrm{diag}(\lambda_1^2, \lambda_2^2)$, to ensure identification in the mixture. Note that these choices for the centering measures are equivalent to the well-known spike and slab priors (e.g., Mitchell and Beauchamp, 1988) but in two dimensions.

The mixture construction $G = (1-\pi)G_0 + \pi G_1$ and the centering measures for $\mathcal{DP}$, F_0 and F_1, allow us to determine regions along the chromosome for which the two subgroups manifest different copy numbers. That is, with prior probability $(1-\pi)$, the random distribution G comes from a DP with a degenerated centering measure F_0, where $\mu_{k1} = \mu_{k2}$ almost surely. With prior probability π, the random distribution G comes from a DP with a centering measure that obeys a bivariate normal law, for which $\mu_{k1} \neq \mu_{k2}$ almost surely (see the Appendix for a simple proof). Therefore, introducing a latent indicator $z_k = I(\mu_{k1} \neq \mu_{k2})$ and assuming $\Pr(z_k = 1) = \pi$, we can rewrite the prior for $\boldsymbol{\mu}$ as

$$\boldsymbol{\mu}_k \mid z_k, G_0, G_1 \overset{\text{ind.}}{\sim} G_{z_k}, \quad \text{with } z_k \overset{\text{ind.}}{\sim} \mathrm{Ber}(\pi) \text{ and } G_r \overset{\text{ind.}}{\sim} \mathcal{DP}(a_r, F_r), \tag{2.2}$$

for $k = 1, \dots, K$ and $r = 0, 1$. Note that (2.1) and (2.2) define a Dirichlet process mixture model (e.g., MacEachern and Müller, 1998).

Due to the discrete nature of the DP prior, some of the $\boldsymbol{\mu}_k$'s will be identical. In summary, the mean copy numbers of segment k for the two disease subtypes, μ_{k1}'s and μ_{k2}, will be clustered in two ways: those segments with the same population copy number levels across the chromosome probes, and those segments with the same population copy number levels across the two disease subtypes.

The model specification is completed with the following prior constructions. For the random effects, m_{lj}, in (2.1), which account for the heterogeneity in the segment means across samples, we assume that

$$m_{lj} \overset{\text{ind.}}{\sim} \mathrm{N}(0, \tau_j^2), \quad \text{with } \tau_j^2 \overset{\text{i.i.d.}}{\sim} \mathrm{IGa}(\alpha_\tau, \beta_\tau). \tag{2.3}$$

The sample variance is also assigned a prior distribution of the form $\sigma_\epsilon^2 \sim \text{IGa}(\alpha_\sigma, \beta_\sigma)$. Finally, for the precision parameter of the Dirichlet processes in (2.2), we further assume $a_r \overset{\text{i.i.d.}}{\sim} \text{Ga}(a_\alpha, b_\alpha)$, for $r = 0, 1$.

2.3 Posterior computation

The likelihood function is the density of the joint distribution of $\mathbf{Y} = \{Y_{ij}\}$, given by

$$
\begin{aligned}
f(\mathbf{y} \mid \boldsymbol{\mu}, \mathbf{m}) = {} & \left(2\pi\sigma_\epsilon^2\right)^{(-1/2)\sum_{j=1}^{J} n_j} \\
& \times \exp\left[-\frac{1}{2\sigma_\epsilon^2}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left\{y_{ij} - \sum_{k=1}^{K}\mu_{kg_j} I(i \in \Omega_k)\right.\right. \\
& \left.\left. - \sum_{l=1}^{L_j} m_{lj} I(i \in \Delta_{lj})\right\}^2\right].
\end{aligned}
$$

We introduce some notation that will be useful in characterizing the posterior:

$$
\begin{aligned}
& s_{kj} = \sum_{i=1}^{n_j} I(i \in \Omega_k), \qquad s_{lj} = \sum_{i=1}^{n_j} I(i \in \Delta_{lj}), \\
& s_{klj} = \sum_{i=1}^{n_j} I(i \in \Omega_k) I(i \in \Delta_{lj}), \qquad (2.4) \\
& s_{kj}^y = \sum_{i=1}^{n_j} y_{ij} I(i \in \Omega_k), \quad \text{and} \quad s_{lj}^y = \sum_{i=1}^{n_j} y_{ij} I(i \in \Delta_{lj}).
\end{aligned}
$$

We now report the conditional posterior distributions needed to perform Markov chain Monte Carlo simulations.

1. *Update* $(\boldsymbol{\mu}_k, z_k)$. We will jointly update $\boldsymbol{\mu}_k$ and z_k. Based on the Polya urn representation of the DP prior (see the Appendix), we can derive the prior conditional distribution for the pair $(\boldsymbol{\mu}_k, z_k)$, given all other pairs $(\boldsymbol{\mu}_j, z_j)$'s for $j \neq k$, as

$$
\begin{aligned}
& (\boldsymbol{\mu}_k, z_k) \mid \boldsymbol{\mu}_{-k}, \mathbf{z}_{-k} \\
& \quad \sim \frac{\pi I(z_k = 1)}{a_1 + K_1}\left\{a_1 \text{N}_2(\boldsymbol{\mu}_k \mid 0, \boldsymbol{\Lambda}_1) + \sum_{j \neq k}^{K} \delta_{\boldsymbol{\mu}_j}(\boldsymbol{\mu}_k) I(z_j = 1)\right\} \\
& \qquad + \frac{(1-\pi) I(z_k = 0)}{a_0 + K_0}\left\{a_0 \text{N}\left(\mu_{k1} \mid 0, \lambda_0^2\right) I(\mu_{k1} = \mu_{k2})\right. \\
& \qquad \left. + \sum_{j \neq k}^{K} \delta_{\boldsymbol{\mu}_j}(\boldsymbol{\mu}_k) I(z_j = 0)\right\},
\end{aligned}
$$

where $K_r = \sum_{j \neq k}^{K} I(z_j = r)$, $r = 0, 1$ such that $K_0 + K_1 = K - 1$. Therefore, the posterior conditional for $(\boldsymbol{\mu}_k, z_k)$ is given by

$$
\begin{aligned}
f(\boldsymbol{\mu}_k, z_k \mid \mathbf{y}, \text{rest}) = {} & q_{00}\mathrm{N}(\mu_{k1} \mid B_0, C_0) I(\mu_{k1} = \mu_{k2}) I(z_k = 0) \\
& + \sum_{j \neq k}^{K} q_{0j} \delta_{\mu_j}(\boldsymbol{\mu}_k) I(z_j = 0) I(z_k = 0) \\
& + q_{10}\mathrm{N}_2(\mu_k \mid B_1, C_1) I(z_k = 1) \\
& + \sum_{j \neq k}^{K} q_{1j} \delta_{\mu_j}(\boldsymbol{\mu}_k) I(z_j = 1) I(z_k = 1),
\end{aligned} \tag{2.5}
$$

where

$$
\begin{aligned}
q_{00} &= \frac{a_0(1-\pi)A_0}{(a_0 + K_0)Q}, & q_{0j} &= \frac{(1-\pi)D_{0j}}{(a_0 + K_0)Q}, \\
q_{10} &= \frac{a_1 \pi A_1}{(a_1 + K_1)Q}, & q_{1j} &= \frac{\pi D_{1j}}{(a_1 + K_1)Q}
\end{aligned}
$$

with Q the normalizing constant such that $q_{00} + q_{10} + \sum_{j \neq k}^{K}(q_{0j} + q_{1j}) = 1$, and

$$
\begin{aligned}
A_0 &= \phi\left\{ \left(\frac{\bar{y}_{k1}}{\psi_{k11}} + \frac{\bar{y}_{k2}}{\psi_{k22}} \right) \left(\frac{1}{\psi_{k11}} + \frac{1}{\psi_{k22}} \right)^{-1} \,\middle|\, 0, \left(\frac{1}{\psi_{k11}} + \frac{1}{\psi_{k22}} \right)^{-1} + \lambda_0^2 \right\}, \\
B_0 &= \left(\frac{\bar{y}_{k1}}{\boldsymbol{\Psi}_{k11}} + \frac{\bar{y}_{k2}}{\psi_{k22}} \right) C_0, \\
C_0 &= \left(\frac{1}{\psi_{k11}} + \frac{1}{\psi_{k22}} + \frac{1}{\lambda_0^2} \right)^{-1}, \\
D_{0j} &= \phi(\mu_{j1} \mid \bar{y}_{k1}, \psi_{k11}) \phi(\mu_{j1} \mid \bar{y}_{k2}, \psi_{k22}), \qquad \text{for } j \neq k, \\
A_1 &= \phi_2(\bar{\mathbf{y}}_k \mid 0, \boldsymbol{\Psi}_k + \boldsymbol{\Lambda}_1), \\
\mathbf{B}_1 &= \begin{pmatrix} \dfrac{\bar{y}_{k1}}{\boldsymbol{\Psi}_{k11}} \left(\dfrac{1}{\boldsymbol{\Psi}_{k11}} + \dfrac{1}{\lambda_1^2} \right)^{-1} \\ \dfrac{\bar{y}_{k2}}{\boldsymbol{\Psi}_{k22}} \left(\dfrac{1}{\boldsymbol{\Psi}_{k22}} + \dfrac{1}{\lambda_2^2} \right)^{-1} \end{pmatrix}, \\
\mathbf{C}_1 &= \operatorname{diag}\left\{ \left(\frac{1}{\psi_{k11}} + \frac{1}{\lambda_1^2} \right)^{-1}, \left(\frac{1}{\psi_{k22}} + \frac{1}{\lambda_2^2} \right)^{-1} \right\}, \\
D_{1j} &= \phi_2(\boldsymbol{\mu}_j \mid \bar{\mathbf{y}}_k, \boldsymbol{\Psi}_k), \qquad \text{for } j \neq k,
\end{aligned}
$$

where ϕ and ϕ_2 are the univariate and bivariate normal density functions, respectively. Finally,

$$\bar{\mathbf{y}}_k = \begin{pmatrix} \dfrac{\sum_{j=1}^{J} I(g_j = 1)(s_{kj}^{y} - \sum_{l=1}^{L_j} m_{lj} s_{klj})}{\sum_{j=1}^{J} I(g_j = 1) s_{kj}} \\ \dfrac{\sum_{j=1}^{J} I(g_j = 2)(s_{kj}^{y} - \sum_{l=1}^{L_j} m_{lj} s_{klj})}{\sum_{j=1}^{J} I(g_j = 2) s_{kj}} \end{pmatrix}$$

and

$$\boldsymbol{\Psi}_k = \begin{pmatrix} \dfrac{\sigma_\epsilon^2}{\sum_{j=1}^{J} I(g_j = 1) s_{kj}} & 0 \\ 0 & \dfrac{\sigma_\epsilon^2}{\sum_{j=1}^{J} I(g_j = 2) s_{kj}} \end{pmatrix}$$

for $k = 1, \ldots, K$, with s_{kj} and s_{kj}^{y} as given in (2.4).

2. *Update* m_{lj}. For the specific random effects m_{lj}, the conditional posterior is given by

$$f(m_{lj} \mid \mathbf{y}, \text{rest}) = \text{N}\big(m_{lj} \mid m^*, \tau_*^2\big), \tag{2.6}$$

where

$$m^* = \frac{s_{lj}^{y} - \sum_k \mu_{kg_j} s_{klj}}{\sigma_\epsilon^2 (s_{lj}/\sigma_\epsilon^2 + 1/\tau_j^2)} \quad \text{and} \quad \tau_*^2 = \left(\frac{s_{lj}}{\sigma_\epsilon^2} + \frac{1}{\tau_j^2}\right)^{-1},$$

for $l = 1, \ldots, L_j$ and $j = 1, \ldots, J$, where s_{lj}, s_{lj}^{y} and s_{klj} are as given in (2.4).

3. *Update* σ_ϵ^2. For the measurement error variance σ_ϵ^2, the conditional posterior is given by

$$f\big(\sigma_\epsilon^2 \mid \mathbf{y}, \text{rest}\big) = \text{IGa}\big(\sigma_\epsilon^2 \mid \alpha_\sigma^*, \beta_\sigma^*\big), \tag{2.7}$$

where

$$\alpha_\sigma^* = \alpha_\sigma + \frac{1}{2} \sum_{j=1}^{J} n_j$$

and

$$\beta_\sigma^* = \beta_\sigma + \frac{1}{2} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left\{ y_{ij} - \sum_{k=1}^{K} \mu_{kg_j} I(i \in \Omega_k) + \sum_{l=1}^{L_j} m_{lj} I(i \in \Delta_{lj}) \right\}^2.$$

4. *Update* τ_j^2. For the variance of the specific random effects τ_j^2, the conditional posterior depends on only m_{lj}^2 and is given by

$$f(\tau_j^2 \mid m_{kj}) = \mathrm{IGa}(\tau_j^2 \mid \alpha_\tau^*, \beta_\tau^*), \tag{2.8}$$

where

$$\alpha_\tau^* = \alpha_\tau + \frac{L_j}{2} \quad \text{and} \quad \beta_\tau^* = \beta_\tau + \frac{1}{2}\sum_{l=1}^{L_j} m_{lj}^2.$$

5. *Resampling* $\boldsymbol{\mu}_k$. As is customary when dealing with almost surely discrete random measures, as in the case for the Dirichlet process, an acceleration step is required to resample the cluster locations (e.g., Bush and MacEachern, 1996). If we denote by $\boldsymbol{\mu}_1^*, \dots, \boldsymbol{\mu}_H^*$ the distinct values of the $\boldsymbol{\mu}_k$'s, and by $z_1^*, \dots, z_H^*$ the corresponding latent indicators, then each $\boldsymbol{\mu}_h^*$, conditional on the cluster configuration (c.c.), needs to be resampled from

$$f(\mu_{h1}^* \mid \mathbf{y}, \text{c.c.}, z_h^* = 0, \text{rest}) = \mathrm{N}(\mu_{h1}^* \mid B_0^*, C_0^*), \tag{2.9}$$

where

$$B_0^* = \sum_{\{k\,:\,\boldsymbol{\mu}_k = \mu_h^*\}} \left(\frac{\bar{y}_{k1}}{\psi_{k11}} + \frac{\bar{y}_{k2}}{\psi_{k22}}\right) C_0^*$$

and

$$C_0^* = \left\{\sum_{\{k\,:\,\boldsymbol{\mu}_k = \mu_h^*\}} \left(\frac{1}{\psi_{k11}} + \frac{1}{\psi_{k22}}\right) + \frac{1}{\lambda_0^2}\right\}^{-1},$$

if $z_h^* = 0$ and setting $\mu_{h2} = \mu_{h1}$; or from

$$f(\boldsymbol{\mu}_h^* \mid \mathbf{y}, \text{c.c.}, z_h^* = 1, \text{rest}) = \mathrm{N}_2(\boldsymbol{\mu}_h^* | \mathbf{B}_1^*, \mathbf{C}_1^*), \tag{2.10}$$

where

$$\mathbf{B}_1^* = \begin{pmatrix} C_{111}^* \displaystyle\sum_{\{k\,:\,\boldsymbol{\mu}_k = \mu_h^*\}} \frac{\bar{y}_{k1}}{\psi_{k11}} \\ C_{122}^* \displaystyle\sum_{\{k\,:\,\boldsymbol{\mu}_k = \mu_h^*\}} \frac{\bar{y}_{k2}}{\psi_{k22}} \end{pmatrix}$$

and

$$\mathbf{C}_1^* = \mathrm{diag}\left\{\left(\sum_{\{k\,:\,\boldsymbol{\mu}_k = \mu_h^*\}} \frac{1}{\psi_{k11}} + \frac{1}{\lambda_1^2}\right)^{-1}, \left(\sum_{\{k\,:\,\boldsymbol{\mu}_k = \mu_h^*\}} \frac{1}{\psi_{k22}} + \frac{1}{\lambda_2^2}\right)^{-1}\right\},$$

if $z_h^* = 1$.

6. *Update a_r*. Finally, for the precision parameters a_r, $r = 0, 1$ in the Dirichlet processes, we know from Antoniak (1974) that the conditional posterior distribution of a_r depends on only the number of distinct $\boldsymbol{\mu}_k$'s and the sample size, that is,

$$f(a_r \mid H_r, K_r) \propto \frac{\Gamma(a_r)}{\Gamma(a_r + K_r)} a_r^{H_r + \alpha_a - 1} e^{-\beta_a a_r}, \tag{2.11}$$

where $K_r = \sum_{k=1}^{K} I(z_k = r)$ and H_r is the number of distinct values $\boldsymbol{\mu}_k$'s such that $z_k = r$, for $r = 0, 1$.

The precision parameter a_r plays a very important role. It largely affects the number of clusters in the $\boldsymbol{\mu}_k$'s. A small value of a_r implies fewer clusters, whereas a large value results in many clusters. Since for the arrayCGH data we anticipate having a relatively small number of segments per chromosome, we will consider relatively informative priors in such a way that they assign most of the mass to small values of a_r.

Sampling from each of the previous conditional posterior distributions is straightforward, since almost all of them have a standard form. The exception is the updating step 6, for which we would require a Metropolis–Hastings step (Tierney, 1994). The main challenge lies in the speed of computation for large data sets, which we have. Programming language such as R will not scale. Instead, we used Fortran, a low-level but much faster language for coding. The computing speed is much improved.

2.4 Calling differential copy number alterations

There are several quantities of interest that we want to focus on in order to achieve our inferential objectives. We break down these quantities into two categories, those for population-level inference, and those for sample-specific inference. The key parameters of interest are $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2})$, z_k, and m_{lj}.

To obtain summaries at the population level, for each population segment $k = 1, \ldots, K$, we compute marginal posterior probabilities for each disease subtype, say $p_1 = \mathrm{P}(|\mu_{k1}| \geq d_1 \mid \text{data})$ and $p_2 = \mathrm{P}(|\mu_{k2}| \geq d_2 \mid \text{data})$, for given values of d_1 and d_2. Higher values of these probabilities will imply a marginal CNA for each subtype.

Moreover, to determine differential CNAs across disease subtypes, we compute the joint posterior probabilities,

$$p_z = \mathrm{P}\big(\{|\mu_{k1}| \geq d_1 \text{ or } |\mu_{k2}| \geq d_2\}, \{z_k = 1\} \mid \text{data}\big),$$

for $k = 1, \ldots, K$. Higher values of these probabilities indicate that segment k has CNAs in any of the disease subtypes and there are differential copy numbers across the two subtypes. The different combinations given in the previous description of the probability may also produce alternative inferences. The threshold to determine high probability values for p_1, p_2 and p_r is set according to a prespecified FDR.

For the sample-specific inference, for each sample j and segment l, the segment-specific mean copy number is $(\mu_{k,g_j} + m_{l,j})$, in which the population segment k overlaps with the sample-specific segment. Note that there can be several population segments that are embedded in the same sample-specific segment. When this is the case, we simply report inference according to the segments defined by the population segments.

3 Data analysis

In this section, we consider two analyses, one with simulated data under different scenarios to test our model, and the other with real data obtained from a breast cancer study conducted at MD Anderson Cancer Center.

3.1 Simulated data

We implemented a simulation study in order to evaluate the operating characteristics of our approach. We simulated samples with $n = 1000$ probes, with ordered locations ranging from 1 to n. For group $g = 1$, we considered four regions of shared aberrations around locations $\{200, 400, 600, 800\}$, alternating gain and loss. Group $g = 2$ contains only two regions of aberration at locations $\{600, 800\}$, identified as a copy number gain and loss, respectively. We randomly generated aberration widths from a $\text{Ga}(2.5, 0.05)$ distribution that has a mean of 50 and 99% interval $(5, 168)$, which shows a large variability and accommodates both large and short segments. We took the level of the profiles for each probe to be zero for the neutral zones and to be a positive/negative random value $\text{Un}(0.1, 0.25)$ for the gain/loss zones, respectively.

We then added white noise to these mean profiles. We generated random errors from $\text{N}(0, \sigma^2)$, with $\sigma^2 \in \{0.1, 0.3\}$ to show low and high levels of noise in the $\log_2$ ratios. We generated 100 profiles, 50 from each group. To test our model under different conditions, only a percentage $\omega 100\%$ of the profiles presented the shared aberrations; the remainder $(1 - \omega)100\%$ were all neutral, showing only white noise around zero. We took three prevalence levels, $\omega \in \{1, 0.6, 0.3\}$. Therefore, we have a total of 6 different scenarios. Scenarios 1 to 3 have low noise with 100%, 60% and 30% prevalence levels, respectively, and scenarios 4 to 6 have high noise with 100%, 60% and 30% prevalence levels, respectively.

Figure 2 shows four profiles for the low (left column) and high (right column) noise levels. We can see that for the high noise profiles, it is very difficult to distinguish (visually) the aberration zones. In the same figure, we present group 1, with four aberration zones (top row) and group 2, with only two aberration zones (bottom row).

To obtain the sample-specific partitions $\{\Delta_{lj}\}$, we ran the CBS algorithm with the default tuning parameter $\alpha = 0.01$. We fitted our model with the following prior specifications: $\lambda_0^2 = \lambda_1^2 = \lambda_2^2 = 100$ to induce flat centering measures, and

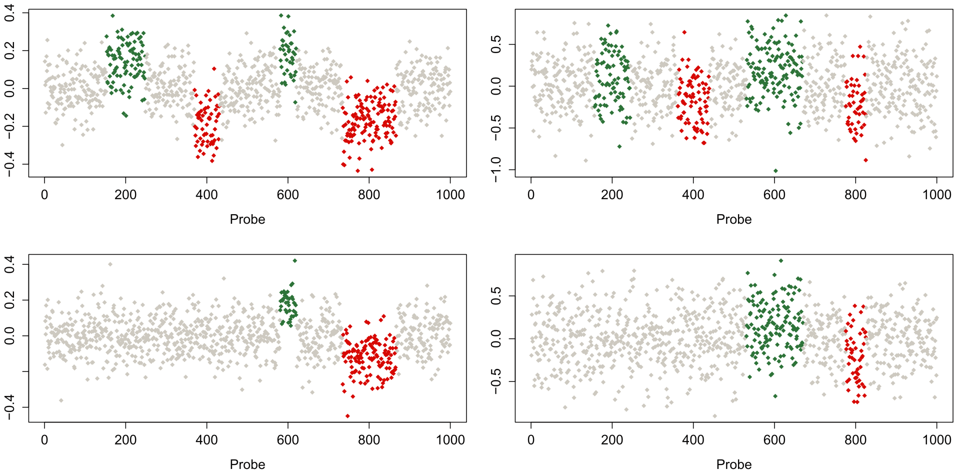

Figure 2 *Simulated aCGH data. No copy number alterations (white dots), amplifications (green/medium grey dots) and deletions (red/dark grey dots). Group* 1 *(top row) and group* 2 *(bottom row). Low noise level (left column) and high noise level (right column).*

$(\alpha_a, \beta_a) = (1, 1)$ as a relatively informative prior to induce a small a_r, and thus a low number of point masses in the Dirichlet processes. For the inverse gamma prior on the sampling variance σ_ϵ^2, we took $(\alpha_\sigma, \beta_\sigma) = (2, 1)$ to be a little informative. The crucial parameter in the model is the variance of the segment-specific random effects, τ_j^2. Large values of τ_j^2 would make the sample-specific effects capture most of the variability of the data, leaving little information for the population mean. On the other hand, if τ_j^2 is small, the variability of the data is shared between the population effects and the sample-specific effects. In fact, if we choose $(\alpha_\tau, \beta_\tau) = (2, 1)$, the logarithm of the pseudo marginal likelihood (LPML) statistic (Geisser and Eddy, 1979) for scenario 1 is $88, 686$; whereas if $(\alpha_\tau, \beta_\tau) = (3, 0.01)$, the LPML is $80, 211$. Although the fitting of the individual samples is better with the former choice, we prefer the latter because it produces better estimates for both the population and individual samples. In all cases, we ran the Gibbs sampler for 10,000 iterations with a burn-in of 1000, keeping every other draw after burn-in for computing the estimates. The Markov chain converged quickly and mixed well.

For calling differential CNAs, we took FDR $= 5\%$, with thresholds $d_1 = d_2 = d$. Since we have different levels of prevalence of aberrations in the samples, in the different scenarios, it is more difficult to call a CNA. To be fair, we took $d = 0.10, 0.05, 0.03$ as threshold values for the 100%, 60% and 30% prevalence levels, respectively.

Figure 3 presents the CNA calls for the different scenarios. In each panel, we show three rectangles, with the x-axis indicating the probe location. The first two rectangles correspond to the marginal CNAs of groups 1 and 2 called from p_1 and p_2. The third rectangle indicates the regions along the chromosome where

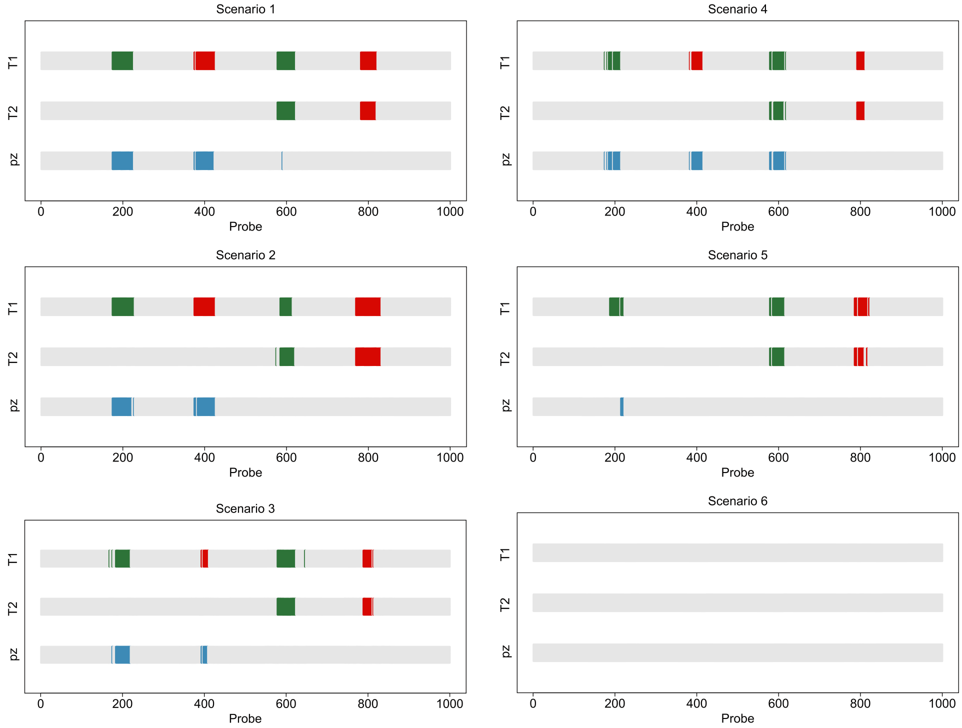

Figure 3 *Simulated aCGH data. Calls of marginal CNAs and differences between the two groups* ($T1$ *and* $T2$)*. Copy number gain* (*green/medium grey*), *loss* (*red/dark grey*) *and differential CNAs across groups* (*blue/light grey*). *Low noise level* (*left column*) *and high noise level* (*right column*). *Prevalence levels*: 100% (*top row*), 60% (*middle row*) *and* 30% (*bottom row*).

there are CNA differences across the two chromosomes, called from p_z. As we can see from this figure, with a low level of noise in the data, our model is able to detect the regions of aberration in each group, as well as the regions of CNA differences across the two groups, for the three prevalence scenarios.

Now, looking at the right column in Figure 3, which corresponds to the scenario of a high noise level, our model is able to detect most marginal regions of aberrations for the cases with 100% and 60% prevalence; however, it is not able to detect any of the aberrations in the cases with low prevalence. This is reasonable because, given 30% aberration prevalence in the samples combined with a high noise level, the findings are essentially white noise. For the 100% aberration prevalence in the samples (top right panel in Figure 2), we notice that even though our model correctly detects the marginal regions of aberrations in each group, it also detects a difference in the levels of the second region of amplifications, denoted with a blue/light grey segment aligned with the two green/medium grey segments in groups 1 and 2. This false discovery is also due to the high level of noise present in the data.

To study the dependence of our model to the sample specific partitions $\{\Delta_{lj}\}$, we repeated the analysis with other two values of the tuning parameter α in the CBS algorithm, say 0.001 and 0.05. With a smaller value of α, CBS detects less changing points, whereas with a larger value, more changing points are detected. Results (not included here) showed that the impact of the partitions in the inference is almost null, perhaps it is preferred a partition with more segments (larger α in the CBS) than another with few segments, specially when the level of prevalence is low.

3.2 Breast cancer data

At the University of Texas MD Anderson Cancer Center, we conducted arrayCGH experiments using samples from 122 patients with breast cancer. For each sample, we used an Agilent HG 4x44K array with 42,416 unique probes. As a result, the raw data contained a matrix of 42,416 × 122 $\log_2$ ratios. The tumor samples we analyzed represented 60 patients with ER+ breast cancer, 11 patients with PR+ breast cancer, and 51 patients with TN breast cancer. Given the reduced number of patients with the subtype PR+, we concentrated on comparing the other two subtypes, ER+ and TN. It is common practice to analyze each chromosome separately since it is rare to see cross-chromosomal CNAs. Therefore, we split the data based on the chromosomes and analyzed each of them separately.

To prepare for Bayesian inference, we preprocessed the arrayCGH data, which included a global normalization process to center the sample mean for each of the 111 samples. Analogous to the simulated data, we obtained sample-specific partitions $\{\Delta_{lj}\}$ by running the CBS algorithm with a tuning parameter of $\alpha = 0.01$. We used the same prior specifications that we used in the simulated data analysis. We ran the Gibbs samplers for 10,000 iterations, with a burn-in of 1000, keeping every other draw. Again, for calling a differential CNA, we took FDR $= 5\%$, with thresholds $d_1 = d_2 = 0.2$ for all chromosomes. We found CNA differences between the two cancer subtypes in 16 of the 23 chromosomes; predominantly in chromosomes 3–7, 9–12, 14–19 and 23.

In Figures 4 and 5, we present marginal CNAs for the two cancer subtypes (ER+ and TN) and copy number differences across the two subtypes.

Curtis et al. (2012) provided what is perhaps the most comprehensive report on genomic architecture for breast cancer, based on genomics findings from a study of 2000 breast tumors. We compared our statistical inference with the findings in that article, and report the results below.

An ER+ subgroup of breast cancer found in Curtis et al. (2012) is uniquely marked by 11q deletion. This subgroup of patients exhibited a steep mortality rate and elevated hazard ratios in the findings of Curtis et al. The top panel of our Figure 4 clearly shows the deletion to chromosome 11 in the second half, marked by the red bars. These deletions are not present in the TN subgroup, echoing the findings of Curtis et al. (2012). Furthermore, the green/medium grey bars in the

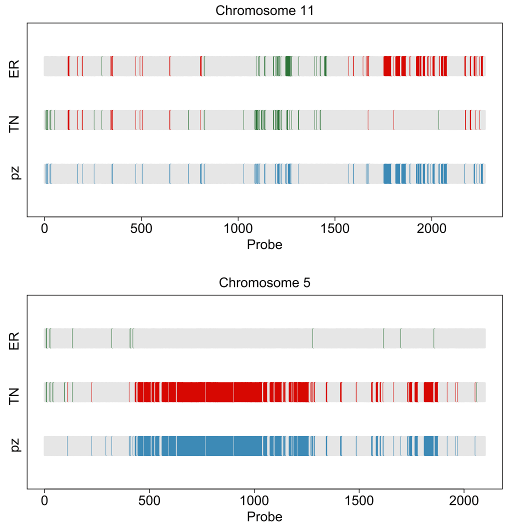

Figure 4 *Calls of marginal CNAs and differences between the two cancer subtypes. Copy number gain (green/medium grey), loss (red/dark grey) and differential CNAs across groups (blue/light grey). Chromosomes* 11 (*top*) *and* 5 (*bottom*).

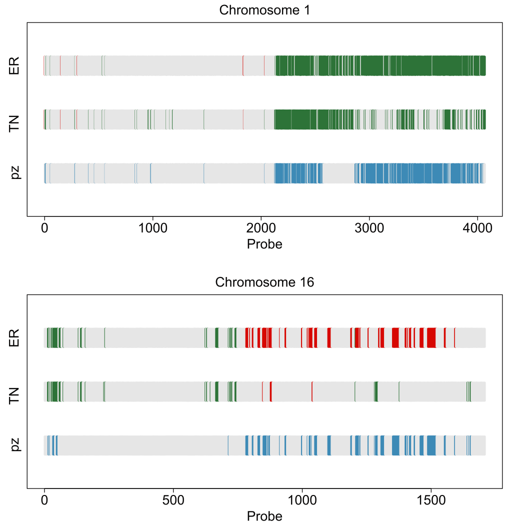

Figure 5 *Calls of marginal CNAs and differences between the two cancer subtypes. Copy number gain (green/medium grey), loss (red/dark grey) and differential CNAs across groups (blue/light grey). Chromosomes* 1 (*top*) *and* 16 (*bottom*).

middle of chromosome 11 indicate copy number gains in this region, which was also reported by Curtis et al. However, these copy number gains are present in both the ER+ and TN groups, making them less interesting for distinguishing the two subgroups. The bottom panel of our Figure 4 shows a large chunk of copy number loss on chromosome 5, which is unique to the TN subgroup. This is one of the major findings of Curtis et al., as well. This is a region containing numerous important signaling molecules and transcription factors, the aberration of which not only affects the genes residing in the region, but those regulated by them. Therefore, this is marked as a trans-influenced region by Curtis et al. (2012).

The two plots in our Figure 5 show a classical 1q gain and 16q loss pattern that is shared by luminal A breast cancer, a subgroup of the ER+ subtype. The combination of copy number gain of 1q and loss of 16q is believed to be a centromere-close translocation (Russnes et al., 2010), which is mainly seen in the luminal A subgroup. In contrast, there is little copy number variation on 16q for the TN subgroup.

We identified several other new findings regarding the copy number variations between the ER+ and TN subgroups. For example, a large region of 15p loss (Figure 6) is identified in the TN subgroup, but not the ER+ subgroup. This has not been reported in the literature. However, in Figure 2 on page 12 in Curtis et al. (2012), a copy number loss at 15q is present. Figure 7 summarizes the differential CNA probabilities between groups for the whole genome. We believe that our findings confirm several major results reported in the literature, while also providing new hypotheses for future validations.

4 Discussion

Determining regions of shared CNAs in different samples is a challenging task and is of great importance for the advance of medical science. In this article, we addressed the problem of determining shared CNAs based on a two step model with the second step based on a semiparametric model. The model is equipped with the ability to identify differences along the genome where two disease subtypes show differential CNAs. This was achieved by considering a mixture distribution for the vector of the population levels, the elements of which were in turn assigned Dirichlet process priors.

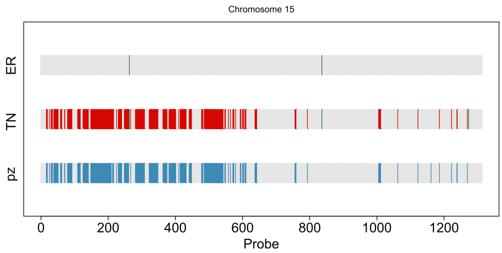

Figure 6 *Calls of marginal CNAs and differences between the two cancer subtypes in chromosome* 15. *Copy number gain* (*green/medium grey*), *loss* (*red/dark grey*) *and differential CNAs across groups* (*blue/light grey*).

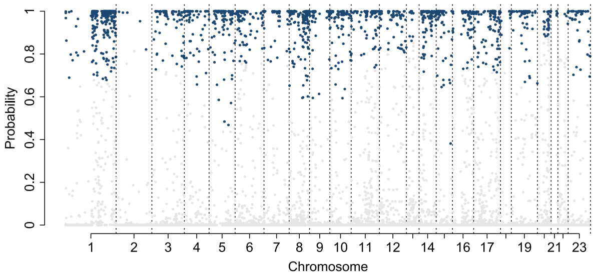

Figure 7 *Differential CNA probabilities between groups for all chromosomes. Blue/grey lines represent significant probabilities controlled by a* 5% *FDR within each chromosome.*

Through simulation studies, we have shown that the proposed model adequately determines the shared aberration regions and detects the differences across the two subgroups. The model was tested under different levels of aberration prevalence and with different degrees of noise. In most of the scenarios we considered, our model worked well. The exception occurred in scenarios with a combination of high noise level and low aberration prevalence, which is an expected finding. We also found out that the sample specific partitions have almost no influence in the final inferences.

Future work includes the extension of our model to compare more than two groups, for which the number of possible combinations then increases dramatically. For example, in the case of three groups, we would have to consider a total of five cases: the three groups as equal, any two as equal, and all as different. This will entail a nontrivial generalization of our mixture prior set-up, a task we will undertake in the future.

Appendix

Stick breaking and Polya urn representation of a $\mathcal{DP}$

In general, a Dirichlet process G, such that $G \sim \mathcal{DP}(a, F)$, is a random discrete measure with precision parameter a and centering measure F. According to Sethuraman (1994), G can be written as a stick breaking representation of the form

$$G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\eta_h}(\cdot),$$

where $\delta_x(\cdot)$ defines a point mass at x, $\eta_h \overset{\text{i.i.d.}}{\sim} F$ and $w_h = v_h \prod_{j<h}(1 - v_j)$, with $v_h \overset{\text{i.i.d.}}{\sim} \text{Be}(1, a)$. Additionally, if $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K \mid G \overset{\text{i.i.d.}}{\sim} G$, then after integrating out the process G, Blackwell and MacQueen (1973) showed that the sequence of $\boldsymbol{\mu}_k$'s

has a Polya urn representation of the form

$$\boldsymbol{\mu}_1 \sim F;$$

$$\boldsymbol{\mu}_k \mid \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{k-1} \sim \frac{a}{a+k-1} F + \frac{1}{a+k-1} \sum_{j=1}^{k-1} \delta_{\boldsymbol{\mu}_j},$$

for $k = 2, \ldots, K$. Denoting by $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$ and by $\boldsymbol{\mu}_{-k} = \boldsymbol{\mu} \setminus \{\boldsymbol{\mu}_k\}$, it can be easily shown that

$$\begin{aligned} \boldsymbol{\mu}_k \mid \boldsymbol{\mu}_{-k} &\sim \frac{a}{a+K-1} F + \frac{1}{a+K-1} \sum_{j=1, j \neq k}^{K} \delta_{\boldsymbol{\mu}_j}, \\ &\sim \frac{a}{a+K-1} F + \frac{1}{a+K-1} \sum_{j=1}^{r} K_j^* \delta_{\boldsymbol{\mu}_j^*}, \end{aligned}$$

where $\boldsymbol{\mu}_j^*$'s are the distinct values of $\boldsymbol{\mu}_k$'s and K_j^*'s are the numbers of repetitions. By the above construction, it is easy to show that when $F = F_0 = \mathrm{N}(0, \lambda_0^2) I(\mu_{k1} = \mu_{k2})$, almost surely $\mu_{k1} = \mu_{k2}$. Alternatively, if $F = F_1 = \mathrm{N}_2(\mathbf{0}, \boldsymbol{\Lambda}_1)$, almost surely $\mu_{k1} \neq \mu_{k2}$.

Acknowledgments

Luis Nieto-Barajas acknowledges support from the National Council for Science and Technology of Mexico (CONACYT), grant 130991. Yuan Ji's research is partly supported by NIH R01 CA132897. Veerabhadran Baladandayuthapani's research was supported by NCI grant R01 CA160736. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* **2**, 1152–1174. MR0365969

Baladandayuthapani, V., Ji, Y., Talluri, R., Nieto-Barajas, L. E. and Morris, J. S. (2010). Bayesian random segmentation models to identify shared copy number aberrations for array CGH data. *Journal of the American Statistical Association* **105**, 1358–1375. MR2796556

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics* **1**, 353–355. MR0362614

Bush, C. A. and MacEachern, S. N. (1996). A semiparametric Bayesian model for randomized block designs. *Biometrika* **83**, 275–285.

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352.

Eilers, P. H. C. and de Menezes, R. X. (2005). Quantile smoothing of array CGH data. *Bioinformatics* **21**, 1146–1153.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230. MR0350949

Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G. and Jain, A. N. (2004). Hidden Markov models approach to the analysis of the array CGH data. *Journal of Multivariate Analysis* **90**, 132–153. MR2064939

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74**, 153–160. MR0529531

Guha, S., Li, Y. and Neuberg, D. (2008). Bayesian hidden Markov modeling of array CGH data. *Journal of the American Statistical Association* **103**, 485–497. MR2523987

Hodgson, G., Hager, J., Volik, S., Hariono, S., Wernick, M., Moore, D., et al. (2001). Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nature Genetics* **929**, 459–464.

Huang, T., Wu, B., Lizardi, P. and Zhao, H. (2005). Detection of DNA copy number alterations using penalized least squares regression. *Bioinformatics* **21**, 3811–3817.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7**, 223–239.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83**, 1023–1032. MR0997578

Newton, M. A., Noueiry, A., Sarkar, D. and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**, 155–176.

Olshen, A. B., Venkatraman, E. S., Lucito, R. and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **4**, 557–572.

Pinkel, D. and Albertson, D. G. (2005). Array comparative genomic hybridization and its applications in cancer. *Nature Genetics* **37**, 11–17.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., et al. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics* **20**, 207–211.

Pollack, J. R., Sorlie, T., Perou, C., Rees, C., Jeffrey, S., Lonning, P., et al. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 12963–12968.

Russnes, H. G., Vollan, H. K., Lingjaerde, O. C., Krasnitz, A., Lundin, P., Naume, B., et al. (2010). Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. *Science Translational Medicine* **2**, 1–13.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650. MR1309433

Shah, S. P., Lam, W. L., Ng, R. T. and Murphy, K. P. (2007). Modeling recurrent DNA copy number alterations in array CGH data. *Bioinformatics* **23**, 450–458.

Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., et al. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics* **29**, 263–264.

Teo, S. M., Pawitan, Y., Kumar, V., Thalamuthu, A., Seielstad, M., Chia, K. S. and Salim, A. (2011). Multi-platform segmentation for joint detection of copy number variants. *Bioinformatics* **27**, 1555–1561.

Tibshirani, R. and Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics* **9**, 18–29.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics* **22**, 1701–1722. MR1329166

Yau, C., Papaspiliopoulos, O., Roberts, G. and Holmes, C. (2011). Bayesian non-parametric hidden Markov models with applications in genomics. *Journal of the Royal Statistical Society, Series B* **73**, 37–57. MR2797735

Zhang, Z., Lange, K. and Sbatti, C. (2012). Reconstructing DNA copy number by joint segmentation of multiple sequences. *BMC Bioinformatics* **13**, 205.

L. Nieto-Barajas
Department of Statistics
ITAM
Rio Hondo 1
Progreso Tizapan
01080 Mexico, D.F.
Mexico
E-mail: lnieto@itam.mx

Y. Ji
Biomedical Informatics
NorthShore University HealthSystem
1001 University Place
Evanston, Illinois 60201
USA
E-mail: yji@health.bsd.uchicago.edu

V. Baladandayuthapani
Department of Biostatistics
University of Texas MDACC
1515 Holcombe Boulevard
Houston, Texas 77030
USA
E-mail: veera@mdanderson.org