Contents

List of Figures

List of Tables

1 Introduction

A tumor is an abnormal growth of body tissue, it can be cancerous (malignant) or non cancerous (benign). The word cancer refers to a group of various diseases, which have in common an uncontrolled cell division leading to growth of abnormal tissue.

Our understanding of cancer biology and the mechanisms underlying cancer cell growth have progressed tremendously over the past decade. The discovery of potential therapeutic targets has led to the development of successful targeted therapies for treating cancer. At the same time, gene microarrays, proteomics, genome-wide association studies, and next-generation sequencing technologies are providing a landslide of complex, information-rich data begging for analysis. The complexity and enormity of many of these data sets present numerous quantitative challenges in terms of storing, processing, and analyzing the data so as to discover and validate the biological information they contain. This requires careful statistical design and analysis considerations, as well as the development of innovative statistical methods to get the most information from these rich data. These new molecular and genetic approaches promise discovery of new targets for cancer treatment and prevention, markers for early cancer detection and to guide therapy decisions, leading to personalized therapy approaches whereby the patient's cancer can be guided by specific molecular and genetic information measured from his or her own cancer.

Given the quantitative nature of many of the modern biomedical research questions, the role of biostatistics is crucial to the continuing effort in reducing mortality and morbidity due to cancer. This is reflected in the increasing efforts to build strong multidisciplinary teams of statisticians and clinicians and to establish training programs that can prepare the next generation of biostatisticians to work in cancer research. The very final goal is to improve cancer diagnosis and treatment by better understanding the mechanism behind the acquisition of malignant phenotype and the progression of cancer. Cancer is the consequence of a dynamic interplay at different levels (DNA, mRNA and protein). Multilevel studies that try to integrate different types of data have therefore become of great interest. Our project is concerned with the integration of gene expression (mRNA) and DNA data.
Gene expressions measure the abundance of a set of mRNA transcripts in a specific tissue. Techniques used for these measurements include, for instance, microarrays. At DNA level, many different kinds of aberration can occur and, for this reason, many different methods have been developed to detect them. Here we focus on methods employed in *molecular genomic studies*, capable of single pair base resolution. Among these, a technique well suited for cancer studies is *Comparative Genomic Hybridization* (CGH), a method able to detect copy number changes. This technique has a relatively high resolution and can span a large part of the genome in a single experiment.

1.1 Related works

In the last decade many publications address the problem of detecting genetic aberrations, in different types of cancer, but only few methods were developed to integrate gene expression measurement with copy number variation data. Although many statistical and computational methods for integrating different types of data have been recently developed, only few of them focuses on the integration of DNA and RNA data and, among them, even fewer concentrate on the integration of CGH and gene expression data.

One of the first work on investigating the direct association between the two data in breast cancer cell lines and tissue samples was the one of Pollack et al. [2002], based mainly on descriptive statistics. Another example could be the one of Van Wieringen and Van de Wiel [2009], based on non parametric tests to study whether whether the estimated CNVs at DNA level would induce differential gene expression at RNA level. In their article Richardson et al. [2010] carry out sparse multivariate analysis developing a framework of hierarchically related sparse regressions to model the associations; they propose to model the relationship in a hierarchical fashion, first associating each response with a small subset of predictors via a subset selection formulation, and then linking the selection indicators in a hierarchical manner. Another article of great interest is the one of Monni and Tadesse [2009], they present a stochastic algorithm that searches for sets of covariates associated with sets of correlated outcomes. Last Choi et al. [2010] develop a double-layered mixture model (DLMM) that simultaneously scores the association between paired copy number and gene expression data using related latent variables in the two data sets.

1.2 Contributions

The proposed method reflects my continuing interest in the development of novel Bayesian methodologies for the analysis of data that arise in genomics. Novel methodological questions are now being generated in Bioinformatics and require the integration of different concepts, methods, tools and data types. The proposed modeling approach is general and can be readily applied to high-throughput data of different types, and to data from different cancers and diseases.

A single mutation is not enough to trigger cancer, as this is the result of a number of complex biological events. Thus, discovering amplification of oncogenes or deletion of tumor suppressors are important steps in elucidating tumor genesis. Delineating the association between gene expression and

CGH data is particularly useful in cancer studies, where copy number aberrations are widespread, due to genomic instability.
This project focuses on the development of an innovative statistical model that integrates gene expression and genetics data. Our approach explicit models the relationship between these two types of data, allowing for the quantification of the effect of the genetic aberrations on the gene expression levels. The proposed model assumes that gene expression levels are affected by copy number aberrations in corresponding and adjacent segments and also allows for the possibility that changes in gene expression may be due to extraneous causes other than copy number aberrations. It allows, at the same time, to model array CGH data to learn about genome-wide changes in copy number considering information taken from all the samples simultaneously.

2

Gene Expression and CGH data

The *central dogma of biology*:

$$DNA \xrightarrow{\text{transcription}} RNA \xrightarrow{\text{translation}} proteins,$$

is at the foundation of any living being. DNA, RNA and proteins are all intensively active, in a very complex and coordinated way to regulate those fundamental mechanisms for life. Trying to explain this dogma in a very short and simplistic manner, using a metaphor, we can say that DNA is a book of instruction that every cell has inside. The alphabet used is very simple and contains only four letters: A, T, G and C. It is used to construct sentences (genes), expressed disguised as proteins. To complete this metaphor, cells look through this book and read, at the right moment, only those genes that provide for certain proteins. RNA solve its role during the Protein bio synthesis.

In April 2003 the International Sequencing Consortium announced that the Human Genome Project had been completed, 99% of the human genome had been sequenced, leading people to talk about a post-genomic era. This event had change deeply the conception of medicine and biology and efforts moved from the sequencing of human genome to the harvesting of the fruits hidden in the genomic *book*.

Before the advent of new technologies developed during this new era biologist, Molecular Biology was indeed based on *one experiment one gene*

criteria, a useful feature to clarify single biological processes, but completely insufficient to study how the entire organism works. This was overcome with the advent of microarray technology, broadly studied in the last decade.
In our work we specifically dealt with arrayCGH[1] technology, that measures DNA copy numbers and microarrays that measure RNA expressions. Many models and studies on those two kind of data came out in the last years , but only few methods that integrate them had been developed. In the next sections I will introduce the basis of the molecular biology as well as an introduction to these kind of data to better understand the framework, describe technologies behind and mention the most important methods that relate to our work and that are available for such kind of data.

2.1 Molecular Biology

Cell is the basic structural and functional unit of all known living organisms, the smallest unit of life. It was discovered by Robert Hooke in 1665 and is often called the building block of life (see Albert et al. [2002]). Problems at cells level, especially problems in the genetic code, can cause most diseases. Every organism is made by one or more cells, vital functions of an organism occur within them and they contain all the hereditary information needed to regulate cell functions and to transmit information to the next generation. In fact all the information are contained in the DNA of each individual cell of an organism.

2.1.1 DNA's structure

DNA (deoxyribonucleic acid) is a big molecule and is one of the nucleic acids[2] (the other one is RNA). As the name says, these acids find their place in the nucleus of cells. They were discovered in 1869 by Friedrich Miescher, and are very important in recent medical and biological research (see, for

[1] CGH stands for "Comparative Genomic Hybridization".
[2] Nucleic acids are very long chain of a single or double nucleotidic strands' sub units.

instance, Dahm [2005]). The basic units that frame nucleic acids are the nucleotides, which are composed of a five-carbon sugar (either ribose or two-deoxyribose), a nucleobase (either a pyrimidine, with a single ring structure, or a purine, with a double ring structure) and a phosphate group. DNA is made of four different types of nucleobases, and they are formed by a deoxyribose (five-carbon sugar), a phosphate group and one of the following nucleobases: adenine (A), cytosine (C), guanine (G), thymine (T), see Figure 2.1.

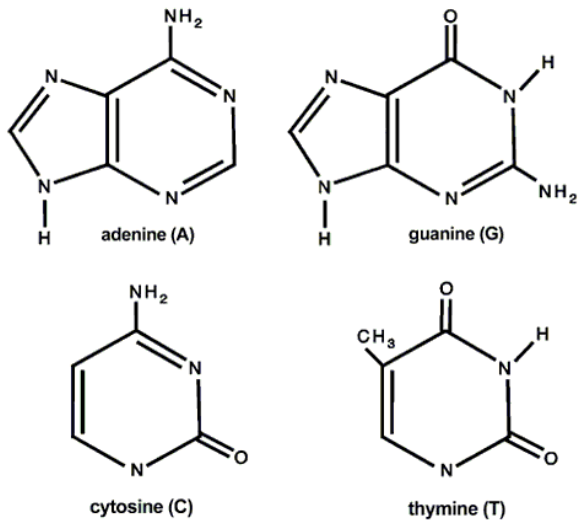


Figure 2.1: thymine (T), adenine (A), cytosine (C), guanine (G)

The abundance of the four nucleobases depends on the organism considered, anyway, because thymine pair with adenine and guanine pair with cytosine, in any species the number of guanine molecules is equal to the number of cytosine molecules, as well as the number of thymine molecules is equal to the number of adenine molecules. A DNA molecule is made of two long polymers that forms the famous double helical structure. This

structure were first discovered by James D. Watson and Francis Crick, (see Watson and Crick [1953]). The two nucleobases are connected by hydrogen bonds, a double one for adenine and thymine and a triple one for cytosine and guanine (that is the cause of the match). However, along the strands, the order of the nucleobases can vary very much: DNA molecules present, at the same time, a regular aspect and a changeable aspect.
The two strands run in opposite directions to each other and are therefore anti-parallel. Usually to specify strands' orientation we use the terms 3'-5': one of the chain follows the direction 3'-5', the other one 5'-3'. To give a simple example consider the sequence 3'-ATCCGTA-5', its complement is 5'-TAGGCAT-3'.

2.1.2 Chromosomes

DNA separate in different chromosomes. In human cells there are 23 pairs of chromosomes, 22 of which are non sex (autosomal) and the 23rd is a sex chromosome. One of each of this pair is inherited from the mother while the other one from the father. In female all the 23 pairs match, thus in human females there are two copies of the genomic code, while in men only 22 pairs match. The ends of chromosomes are called telomers, identified as p if it corresponds to the short arm whereas called q if it corresponds to the long arm.
Chromosomes contains genetic information. An amazing aspect of the human genome is that of the 3.2 billion nucleotides about 99.9% is the same between one individual and another. This means that only 0.1% of the entire sequence makes a person unique. This small amount determines attributes like how a person look, the disease he or she develops. This variations bring differences at genotype level.
The genotype (e.g. see Churchill [1974]) concern the inherited instructions the organism carries within its genetic code. It refers to the set of genes that constitute the DNA of an organism. On the other hand, the term phenotype refers to any observable characteristic or trait of an organism, such as its

morphology, development, biochemical or physiological property, behavior and product of behavior. Only genotype is not sufficient to define the phenotype, but it interacts with (external or internal) environment. Thus two individuals with the same genotype (for instance monozygotic twins) not necessarily share the same phenotype: this could also be explained through epigenetic mechanisms. Phenotypic variation, due to underlying heritable genetic variation, is a fundamental prerequisite for evolution by natural selection.

2.1.3 Protein biosynthesis

During DNA duplication the double helix unwind and the base pairs separate. Then the free nucleotides in cells can pair on the separate base, creating a new strand. In other moments of cell's life, DNA is winded in a way that only enzymes[3] can access only to certain genes. A gene is a specific segment of the DNA molecule, that contains all the codifying information necessary to instruct a cell to synthesize a specific product (such as an RNA molecule or a protein). The way between genes and proteins consist of two different stages: transcription and translation. The first one take place in the nucleus: codons[4] of a gene are copied into messenger RNA by RNA polymerase[5]. This RNA copy is then decoded by a ribosome that reads the RNA sequence by base-pairing the messenger RNA to transfer RNA, which carries amino acids. Since there are 4 bases in 3-letter combinations, there are 64 possible codons (4^3 combinations). These encode the twenty standard amino acids, giving most amino acids more than one possible codon. There are also three 'stop' or 'nonsense' codons signifying the end of the coding region: these are the TAA, TGA and TAG codons. These amino acids forms polymers chain, which in turn join to make proteins.

[3]Biological molecules that catalyze chemical reactions.
[4]Codons are sequence of three nucleotides.
[5]An enzyme that produces RNA.

2.2 Gene Expression

Gene expression is that biological process by which information from a gene is used in the synthesis of a functional gene product. Regulate gene expression means to control the amount and timing of appearance of this functional product. the Central Dogma of Molecular Biology make clear that if, in different cells, different genes are expressed (copied into RNA), different proteins will be produced, thus different types of cells will emerge. Even if all cells in our body have the same genes, cells differentiate[6] their composition, structure and function activating different genes. Also a small mutation or the influence of the environment where the organism live in, can cause highly expression of a gene in one person and a very low one in another. An example of this could be the production of dark pigment in the skin of a person, after a long exposure to the rays of the sun, pigment produced by the expression of the melanin gene.

2.2.1 Quantifying gene expression: Microarrays

Gene expression microarrays are powerful tools to measure the abundance of mRNA[7]. These arrays tell to the scientists how much RNA a gene is making. This is really important because when a gene is expressed it produces RNA which help with the production of the final protein coded for by the gene itself. Microarrays' technology had become an essential tool that many biologists use to observe the wide genomic expression of the amount of genes in a specific organism. The revolutionary aspect is just that they allow to to measure the expression of every single gene in the whole human genome, so scientists can quickly point out differences in the patterns of two different subgroups of individual. On the other hand of the spectrum there are the blotting methods that could only measure one or few genes at a time, through a slow and tedious process.

[6]Each cell use just a small fraction of its genes

[7]Messenger RNA is produced during transcription.

Typically a microarray is a slide with DNA molecules fixed on it in a specific place called *spot*. A single slide can be made by thousands spot and each spot can contain millions of the same DNA molecules, that corresponds to a single gene. Each spot is imprinted on a slide by a robot or are synthesized using the process of photolithography. Microarrays may be used to measure gene expression in many ways, but one of the most popular applications is to compare expression of a set of genes from a cell maintained in a particular condition (condition A) to the same set of genes from a reference cell maintained under normal conditions (condition B). The outcome of this application is a colored slide with colors that must be interpreted and quantified. Assuming that genes in condition A were marked with a red dye and those in condition B with a green dye, if a gene in condition A was very abundant, compared to that in condition B, the corresponding spot would be red. If it was the other way, the spot would be green. . If the gene was expressed to the same extent in both conditions, one would find the spot to be yellow, and if the gene was not expressed in both conditions, the spot would be black. Thus, what is seen at the end of the experimental stage is an image of the microarray, in which each spot that corresponds to a gene has an associated fluorescence value representing the relative expression level of that gene. The intensity of the color is then transformed into a number that usually corresponds to the log-ratio of the expression of the gene in condition A and in condition B. Clearly this application is very useful if, and only if, we want to compare the expression in two different condition, but the quantification of gene expression in a single condition is always possible.

Affymetrix is one of the most famous platform. The data i will use are obtained using this platform. To determine expression levels, gene expression microarrays use the natural attraction between the DNA and RNA target molecules. They use the natural binding between the four basis. In fact RNA is composed of the same four basis, with uracil (U) in place of thymine. Because of hydrogen bonds that bind the couples A-T (double)

and C-G (triple), in RNA we have exactly the same pairing system (with U instead of T). Unlike DNA, RNA appears in a single strand and this allows it to bind easily to any other single stranded sequence (both DNA or RNA). Two strands (one of DNA and one of RNA) that matches are said to be complementary and stick to each other: even a single base that not match its partner could keep a single stranded sequence from sticking to another. This base attraction is known as *hybridization*. Microarray use hybridization to identify RNA sequences in a sample and to establish which genes are expressed by that individual and the abundance of this expression. To illustrate, in details, how this technology works, we will focus on just the measurement of the expression of just one gene.

First of all a DNA strand is build, a probe, onto a surface glass chip. Even if genes are made by hundreds of thousand base, the probes are generally shorter. Thus the choice of the probes to print must face the problem of the trade of between finding sequence that will be unique to the gene of interest and affordability. Affymetrix decided for a length of 25 base for each probe. Scientists must compare the 25 base probe sequence to the rest of human genome, to make sure that it does not match anywhere else, in order that, when an RNA molecule binds to the probe, it is clear that the gene is expressed.

Once a probe is designed to measure expressed RNA, RNA must be extracted from the biological sample (blood, tumor, etc...). The subsequent step is to copy it millions of times. The process used to that aim is called PCR[8], after this amplification it is more easy to detect RNA on the array. While RNA is copied, molecules of biotin are attached to each strand, acting as molecular glue for fluorescent molecules. This prepared RNA sample is then washed over the array to allow the hybridization. At this point scientists use fluorescent molecules that stick to the biotin, making RNA glow in the dark. Last a laser light is shone on the array, cause the stain to glow

[8]Polymerase Chain Reaction is a biochemical technology in molecular biology to amplify a single or a few copies of a piece of DNA across several orders of magnitude, generating thousands to millions of copies of a particular DNA sequence.

and producing an image where the intensities of different colors represent different level of expression. When a gene is highly expressed many RNA molecules stick to the probe causing the probe to shine brightly when the laser hits in, and viceversa. Before analyzing this kind of data pre processing and normalization of the data set are necessary.
Gene expression had been widely studied and lots of different methodologies was being developed to extract meaningful information. Those methods goes from analysis of variance to mixed models, from multiple testing to cluster analysis, empirical Bayes and fully Bayesian methods, functional data analysis and networks.
Our final data set is a matrix of intensity values for each gene and individual, we run our code after an appropriate data cleaning, see section 4.4.1.

2.3 CGH

Comparative Genomic Hybridization (CGH) is a method designed for identifying chromosomal segments with copy number aberration. In Kidd et al. [2008] it was estimated that approximately 0.4% of the genomes of unrelated people differ with respect to copy number. CGH is a powerful method and gives best results when combined with microarrays, the so called array CGH. This technique use microarrays consisting of thousands or million of genomic targets (probes) that are spotted on a glass surface, with a resolution of the order in the range 1 MB (one million base pairs) for BAC (bacterial artificial chromosome) to 50-100 kb (kilo base pairs). Although numerous platforms have been developed to support array CGH studies, they all revolve around the common principle of detecting copy number alterations between two samples. As in measuring gene expression, a sample of interest is labeled with a dye and then mixed with the reference sample labeled with a different dye. The mixed sample obtained is hybridized and the intensity of both colors is then measured through an imaging process. The final quantity is expressed as the $\log_2$ ratio of the two intensities. The expected copy number

number of each segment of DNA is equal to two because in the human body females have 23 matched pairs of chromosomes, thus the intensity ratio is determined by the copy number of the DNA in the test sample. If the test sample has no copy number aberrations the $\log_2$ of the intensity ratio is theoretically equal to zero. Similarly when there is a single copy gain this leads to $\log_2 \frac{3}{2} \simeq 0.58$ and multiple copy gains refers to the sequence $log_2\frac{4}{2}$, $log_2\frac{5}{2}$, etc.. On the other hand a single copy loss leads to the value $log_2\frac{1}{2} = -1$ and the loss of both copies to $-\infty$ (usually a large negative value is observed). Figure 2.2 shows an example of this kind of data.

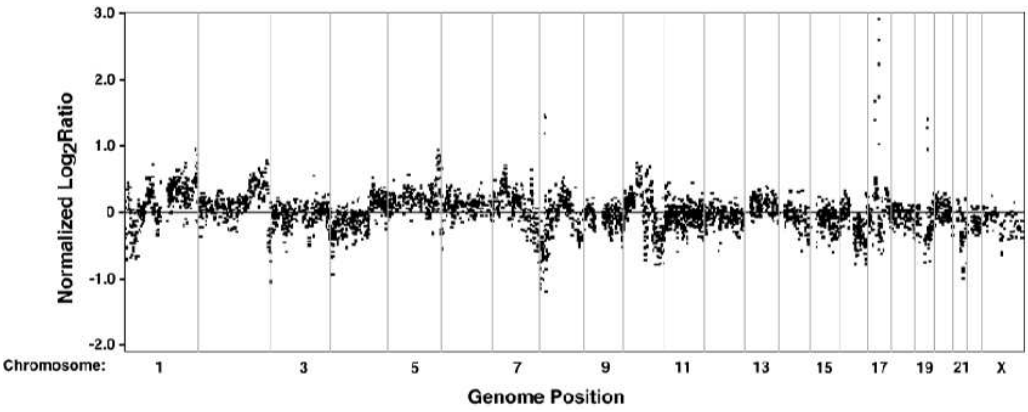


Figure 2.2: CGH mesurement plotted against the BAC genomic position..

2.4 Genomic microarrays

This section is not a complete list of available techniques and is based on the article by Lockwood et al. [2006]. cDNA microarrays (originally used

in gene expression profiling) introduce the genome wide approach to array CGH. The advantage of this technique is that high level amplifications and deletions can be directly correlated to expression changes using the same platform. The disadvantage is that only exonic (codifying) regions of the genome are covered, making alterations to promoter regions and other protein binding sites undetectable. The main advantage of genome wide-arrays (LICs[9] and BACs) are that they provide robust targets for sensitive detection of hybridization signals. BAC is not limited to loci annotated with genes. Because the size of the arrayed elements provide a higher signal to noise ratio, BAC-based platforms allow highly sensitive and reproducible detection of a wide range of copy number changes including single copy number gains or loss, homozygous deletions and high level amplifications.

Another technique recruited for copy number assessment in CGH experiments are arrays of photolithographically synthesized short oligonucleotides (21-25 nucleotides in length) originally designed for detecting single nucleotide polymorphisms (SNPs). A method known as whole-genome sampling assay (WGSA) is an example. In this technique even if the sample is reduced and is not able to represent the entire genome, the probability of cross hybridization to multiple short oligonucleotide targets on the array is reduced. The strength of this strategy is its ability to relate copy number and allelic status at selected loci. The performance of this technique compared with the conventional BAC have been studied by Bignell et al. [2004] using cancer cell lines. Zhao et al. [2004] compared SNP, cDNA and BAC arrays. The BAC arrays showed the highest signal to noise ratio, making them better suited to detect single copy variations. However SNP arrays allow copy number and genotype changes to be measured in a single experiment.

The methods previously described allow copy number changes to be assessed on a genomic-wide scale, but the coverage of the arrayed elements can vary greatly. This leads to large gaps where no information is obtainable. To fully

[9]Large Insert Clones is a Marker-based technique that sample genome at megabase intervals, typically covering about 10% of the genome.

understand the alterations occurring in various diseases, probes that cover the entire genome are required. Submegabase resolution tiling set (SMRT) array can supply for this requirement. Like other large insert clone-based approaches, it yields high signal to noise ratio due to the hybridization sensitivity of the BACs to their corresponding genome targets. In contrast to marker based approaches, the overlapping arrangement of the BAC clones abrogates the need to infer genetic events between marker clones and the redundancy provides confirmation of copy number status at each locus. The tiling nature also increases the probability of detecting micro alterations that may fall between marker probes in other array platforms. The major consideration in interpreting whole genome BAC array data is the fact that some clone map to multiple places in the genome due to cross hybridization to highly homologous sequences.

The choice of platform technology for an array CGH study primarily depends on the type of samples being analyzed and the level of detail desired. A major consideration in selecting an array platform is sample requirement: DNA quality may be compromised in formalin-fixed, paraffin-embedded archival specimens. Large insert clone arrays efficiently capture signals from samples of low DNA quantity and quality for genome-wide analysis, while oligonucleotide and small PCR fragments could facilitate more detailed investigation at selected regions, when DNA quality and quantity are not limiting.Amplification techniques have proven effective in increasing hybridization signal strength and limiting noise through the reduction of sample complexity at the expense of genomic coverage. A problem of these techniques is the variability in results using the same sample. Another consideration for array selection is the tissue heterogeneity in a sample, because it affects detection sensitivity of copy number changes. Increasing the number of measurements over a genomic distance could provide more data points within a segmental alteration, thereby increasing the probability of detection. Thus the use of tiling path resolution arrays should be considered in analyzing heterogeneous tissue samples. Another important consideration is

the selection of reference DNA. Common examples are using a sex-matched reference, sex mismatched; a reference obtained from a pool of individuals or using a reference from a single individual.

2.5 Applications

Even if the most frequent application is in cancer's studies, other application are always possible. In this section will be reviewed the use of array CGH in measuring copy number status in cancer, genetic diseases and in evolutionary comparisons. Initially in cancer studies focused on specific regions of tumor genomes, later they expand to the entire chromosome arms. In terms of genomic-wide approaches the ones that yielded much information on the genomic landscape of a variety of cancers were cDNA microarrays and interval LIC arrays.
In inherited diseases it was shown the presence of segmental duplications and deletions. The discovery of such alterations have been facilitated by the advance in array-based technique. Megabase interval genomic arrays have been instrumental in delineating regions affected in many genetic diseases. In cytogenetically normal patients that exhibit mental retardation and dysmorphisms have been discovered submicroscopic chromosomal deletions and duplication. In addition copy number changes were refined in Cri-du-chat syndrome, congenital diaphragmatic hernia and Prader-Willi Syndrome.
Array CGH have been used in the characterization of large scale DNA variations. In Iafrate et al. [2004] it was shown that 14 large-scale copy number variations were located near loci associated with cancer or genetic disease, suggesting that certain individuals may have higher susceptibility to disease than others. Other studies illustrate that copy number variations contains genes that have been implicated in cell growth and other functions.
Last array CGH technology has been employed for use in inter species comparisons. In a comparison between human genome and four great apes were discovered 63 sites of DNA copy variations among 2460 studied. A signif-

icant number of these sites existed in interstitial euchromatin (see Locke et al. [2003]). In Fortna et al. [2004] they use a cDNA array CGH approach, over 29 000 human genes across five hominoid species (human, bonobo, chimpanzee, gorilla and orangutan) were compared leading to the identification of more than 800 genes that gave genetic signatures unique to a specific hominoid lineage. Moreover, there was a more pronounced difference between copy number increases and decreases in humans and a number of genes amplified are thought to be involved in the structure and function of the brain.

3 Statistical methods

3.1 Gene expression in statistics

Gene expression have been widely studied in statistics. These studies could be clustered in two principal groups: class discovery and class comparison. The aim of the first one is to find groups of genes that could be related and is based on machine learning technique, also known as pattern recognition techniques. The genomic analysis starts with low level operations such as normalization or filtering, and ends with high level ones such as clustering or other pattern recognition techniques. Pattern recognition could be supervised or unsupervised. When it is supervised some links that are prior known are inserted as known and fixed, while the unsupervised analysis progress without this kind of knowledge. Cluster analysis is one example of pattern recognition. This technique explicitly identify underling scheme beneath a data set, assuming that it exist, and must be validate from a statistical and scientific point of view. This technique had been widely studied in different fields of application and is not our intention to go more deep in technicalities such as the definition of the metric or the choice of the algorithm to be used, but we want just to give an idea of the justification in the use of these techniques in genomics. To that aim is useful to introduce some biological notion. Co-expression is that phenomenon for which genes show

similar pattern of expression in a variety of of conditions. Cluster analysis detects sets of co-expressed genes, thus cluster analysis is important as long as this phenomenon is well founded. Co-expression is justified both from a mechanistic and an empiric point of view. If two or more genes are involved in the same process, they will be expressed at the same time. An example are drugs that are removed from the body through the united action of some enzymes. First enzymes convert external products in reaction intermediate that are then conjugate with soluble groups facilitating their elimination through kidney.
The aim of class comparison is to test if a class of genes is linked to a certain effect (e.g. identification of genes related to a certain pathology). An example could be: "Is gene expression in mice exhibit in condition A and B different?". Techniques such as t-test, ANOVA, logistic regression and survival analysis are examples of methods used to give an answer. A well known Bayesian example is the one of Parmigiani et al. [2002] called *probability of expression model* (POE). It is a 3 components mixture model that estimate for each gene the probability of belonging to one of the three latent categories, over-, under- or normal expression.

3.2 Comparative Genomic Hybridization in statistics

When studying CGH data, investigators are interested in finding out:

- which regions of DNA have copy number aberrations;
- how many copies are lost or gained.

Different methods have been proposed (Olshen et al. [2004], Sen and Srivastava [1975], Fridlyand et al. [2004], Baladandayuthapani et al. [2010], Broët et Richardson [2006], Guha et al. [2008], Du et al. [2010], Hodgson et al. [2001]). Most used methods could be grouped in two categories:

calling methods and segmentation methods. Segmentation methods, also known as segmentation methods, seek to identify breakpoints that separate contiguous regions of common means and estimates these means. Olshen et al. [2004] developed a method called circular binary segmentation to translate noisy intensity measurements into regions of equal copy number, that extend the frequentist solution proposed by Sen and Srivastava [1975] to detect change in mean. The circular binary segmentation recursively detects pairs of change points to identify chromosomal segments with altered copy number. Bayesian approaches typically use a joint prior on the configuration of possible change points and the associated parameters. In Baladandayuthapani et al. [2010] they propose a hierarchical Bayesian random segmentation approach that detect recurrent copy number aberration across multiple samples. Other approaches include clustering based approaches to combine similar segments.

Calling methods model aCGH profile at clone level and call the states of each probe as loss, gain or neutral. For example Broët et Richardson [2006] propose a three state spatial mixture model, where the spatial correlation is introduced in the weights of the mixture model through a random Markov field. For our purpose it is essential to describe the model proposed by Guha et al. [2008] in detail. In their work they consider only one sample at a time. They denote with $L_1, \ldots, L_n$ the DNA fragments and with Y_k the normalized $\log_2$ ratio observed at clone L_k ($k = 1, \ldots, n$). They therefore introduce a latent variable called copy number state s_k associated with each clone L_k that assumes values in the set $\{1, 2, 3, 4\}$:

1. $s_k = 1$ representing a copy number loss (less than two copies of the sequence in the fragment L_k);

2. $s_k = 2$ representing a copy-neutral state (exactly two copies of the sequence in the fragment L_k);

3. $s_k = 3$ representing a single copy gain (exactly three copies of the sequence in the fragment L_k);

4. $s_k = 4$ representing a multiple copy gain (more than three copies of the sequence in the fragment L_k);

They denote with μ_j $(j = 1, \ldots, 4)$ the expected $\log_2$ ratio of all clones L_k for which $s_k = j$ and assume that the normalized $\log_2$ ratio are distributed as $Y_k \overset{ind}{\sim} N(\mu_{s_k}, \sigma^2_{s_k})$. Then the dependence of the neighboring clones is modeled using a hidden Markov model (HMM)[1]. In this context assuming an HMM means that the probability of a clone to be in a certain state depends only on the state of the previous one and not on the entire history of the chain, this implies that the conditional probabilities $P(s_{k+1}|s_k, \ldots, s_1) = P(s_{k+1}|s_k) = a_{s_k s_{k+1}}$, where the element $a_{s_k s_{k+1}}$ is taken from a 4×4 matrix $\boldsymbol{A}$ of stationary transition probabilities. They assume that the elements of this matrix are strictly positive, thus the hidden Markov process is a periodic, irreducible and its four states are positive recurrent. The unique stationary distribution of $\boldsymbol{A}$, denoted by $\pi_A = (\pi_A(1), \pi_A(2), \pi_A(3), \pi_A(4))$ (where $\pi_A(i)$ is strictly positive for state $i = 1, \ldots, 4$), obtained as the normalized left eigenvector of the matrix associated with eigenvalue 1, is assumed to be the distribution of the first clone. In this way they uniquely determines the joint likelihood of a given sequence. Figure 3.1 shows an example of results obtained applying this model.

Calling methods shown so far assume the number of states as a fixed number. Using infinite hidden Markov model (iHMM) this assumption could be overcome. An example of model without assumption of the number of states is the one of Du et al. [2010]. In their paper they first give an overview of hidden Markov model with Dirichlet priors (HMM with Dirichlet distribution prior, infinite HMM with HDP prior, Sticky HMM with HDP prior), and then present their sticky hidden Markov model of comparative genomic hybridization.

[1]For a description of the HMM look at the next section

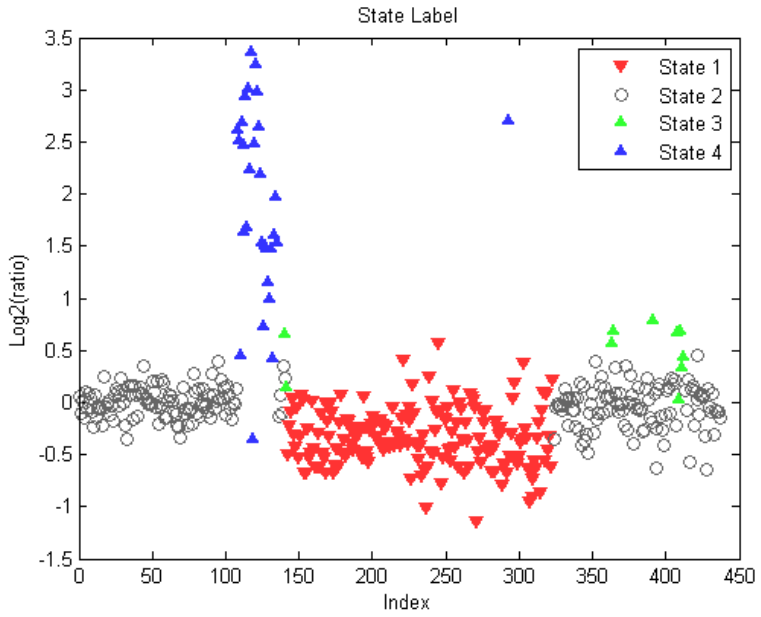


Figure 3.1: The normalized log2 ratio plotted against the position index with state labels.

3.3 Hidden Markov Models

This section is a basic introduction to Markov process and, in particular, to the special case of the hidden Markov process. A Markov process could be thought as a process without memory, where the prediction of future is based only on the present state, ignoring the whole process' full history. Given a set of state $\mathbf{S} = \{s_1, s_2, \ldots, s_{|\mathbf{S}|}\}$, where $|\mathbf{S}|$ is the cardinality of the set $\mathbf{S}$, a series over time can be observed: $\boldsymbol{x} \in \mathbf{S}^T$, where T is the length of the series. Consider, as example, a weather system, where the states are three climate condition: $\boldsymbol{x} = \{sun, cloud, rain\}$, with $|\mathbf{S}| = 3$. Going on with the example, consider a realization of the weather in five days: $\{x_1 = sun, x_2 = cloud, x_3 = cloud, x_4 = rain, x_5 = cloud\}$, with $T = 5$. This is a typical example of output of a random process over time. If we do

not put any further assumption, state s at time t could be a function of any number of variables, including all the states from times 1 to $t-1$ or even many others that we did not consider. To make more tractable this time series, two assumption, called Markov assumption, are introduced and drive us to the Markov process. The first assumption say that the probability of being in a state at time t depends only on the state at time $t-1$. Formally:

$$P(x_t|x_{t-1}, x_{t-2}, \ldots, x_1) = P(x_t|x_{t-1}),$$

The intuition beyond this assumption is that state t summarize enough information of the past to reasonably predict the future. In our example the climate condition of yesterday gives a strong idea of what weather will be today. The second assumption say that the distribution over the next state given current state does not change over time, the probability of having rain given clouds yesterday is always the same. Formally:

$$P(x_t|x_{t-1}) = P(x_2|x_1); t \in 2, \ldots, T$$

Conventionally, it is also assumed that there is an initial state and the initial observation $x_0 \equiv \boldsymbol{p}_0$, where $\boldsymbol{p}_0$ is the initial probability distribution over states at time 0.
To identify a Markov process we need to define a transition matrix and an initial probability vector. In our example we could give equal probabilities to the states at time 0, $\boldsymbol{p}_0 = [0.33, 0.33, 0.33]$, and use a transition matrix $\boldsymbol{A}$ defined as follows:

$$\boldsymbol{A} = \begin{bmatrix} & sun & cloud & rain \\ sun & 0.8 & 0.1 & 0.1 \\ cloud & 0.2 & 0.6 & 0.2 \\ rain & 0.1 & 0.2 & 0.7 \end{bmatrix}$$

These numbers show intuition that weather is self-correlated: it's more probable to stay in the same condition (diagonal elements) and some transition are more probable than others. Now we can calculate the probability of our sequence.

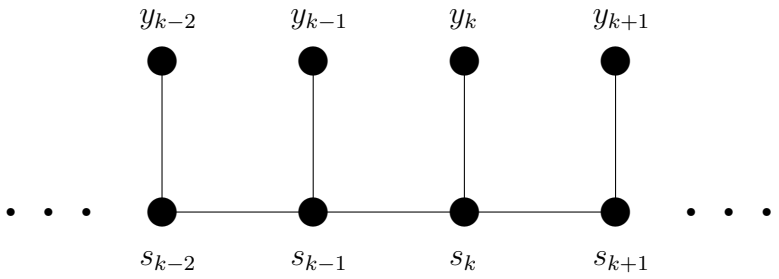
$$
\begin{aligned}
P(\boldsymbol{x}) &= P(x_T, x_{T-1}, x_{T-2}, \ldots, x_1; \boldsymbol{p}_0, A) \\
&= P(x_T, x_{T-1}, x_{T-2} \ldots, x_2; A) P(x_1; \boldsymbol{p}_0) \\
&= P(x_T | x_{T-1}, x_{T-2} \ldots, x_1; A) P(x_{T-1} | x_{T-2}, \ldots, x_1; A) \ldots P(x_2 | x_1; A) P(x_1; \boldsymbol{p}_0) \\
&= P(x_T | x_{T-1}; A) P(x_{T-1} | x_{T-2}) \ldots P(x_2 | x_1; A) P(x_1; \boldsymbol{p}_0) \\
&= P(x_1; \boldsymbol{p}_0) \prod_{t=2}^{T} P(x_t | x_{t-1}; A) \\
&= P(x_1; \boldsymbol{p}_0) \prod_{t=2}^{T} A_{x_{t-1} x_t} \\
&= P(x_1 = sun, x_2 = cloud, x_3 = cloud, x_4 = rain, x_5 = cloud) \\
&= P(sun | \boldsymbol{p}_0) P(cloud | sun) P(cloud | cloud) P(rain | cloud) P(cloud | rain) \\
&= 0.33 \times 0.1 \times 0.6 \times 0.2 \times 0.2 \\
&= 0.000792.
\end{aligned}
$$

This model is a nice abstraction of time series, but fails to capture a very common scenario. What happens if we cannot observe the states themselves, but only a probabilistic function of those states? Consider, for example, a situation proposed in Jason Eisner [2002]:

> You are a climatologist in the year 2799, studying the history of global warming. You can't find any records of Baltimore weather, but you do find my (Jason Eisner's) diary, in which I assiduously recorded how much ice cream I ate each day. *What can you Figure out from this about the weather that summer?*

To explore this scenario a hidden Markov model could be used. We do not observe directly the weather, but we observe an outcome generated by each day (in this example the number of ice cream). To model the probability of generating an output observation as a function of the hidden state, we make the output independence assumption and define $P(y_t =$

$v_k|z_t = j) = P(y_t = v_k|y_1, \ldots, y_T, x_1, \ldots, x_T) = b_{jk}$. Matrix $\mathbf{B}$ encodes the probability of our hidden state generating output v_k given that the state at the corresponding time was s_j. Note that, in this scenario, the value of the states are unobserved. There are three fundamental questions we might ask of an HMM. What is the probability of an observed sequence, what is the most likely configuration of states that generate an observed sequence and how can we learn on the parameters $\boldsymbol{A}$ and $\mathbf{B}$, given some data. To give answers to this questions, different models have been developed. Figure below shows a graphical representation of a HMM.



3.4 Bayesian variable selection

This section is a brief introduction to the theory of the variable selection, I will focus on the method developed by George and McCulloch [1993] known as Stochastic Search Variable Selection (SSVS), introduced in the linear regression framework, adapted by many other authors to other model settings, see for instance George and McCulloch [1997], Smith and Kohn [1996] and Sha and al. [2004].

A main issue in building up a regression model is the selection of the regressors to include. Given a dependent variable Y and a set of possible predictors $X_1, \ldots, X_p$, the issue is to find a subset of predictors $X_1^*, \ldots, X_q^*$ that best fit the model $Y = X_1^*\beta_1^* + \ldots + X_q^*\beta_1^*$. Different methods have been proposed, such as AIC, BIC, Cp, based on a comparison of all 2^p possible models. Unfortunately when p is large the computational cost could

be prohibitive. Typically practitioners use heuristic methods to reduce the number of potential subset to be investigated. Examples are stepwise procedures, such as backward and forward algorithms, which include or exclude variables based on R^2 considerations. A special case is the small n large p context, where most of the standard methods cannot be used, note, for example, that the matrix $X^T X$ is not invertible. To overwhelm these problems George and McCulloch [1993] proposed a Stochastic Search Variable Selection method, based on embedding the entire regression setup in a hierarchical Bayes normal mixture model, where latent variables were used to identify subset choices. In the original methods a Gibbs sampler is used to indirectly sample from the posterior distribution on this set of possible subset choices, while, in most of the following methods, a Metropolis-Hastings is used. Clearly best models are those with higher probability. This procedure use MCMC to obtain a sample from the posterior distribution quickly and efficiently, in a high-dimensional framework. Starting from the canonical regression setup:

$$\boldsymbol{Y}|\boldsymbol{\beta}, \sigma^2 \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector, $\boldsymbol{X}$ is an $n \times p$ matrix ($\boldsymbol{X} = [\boldsymbol{X_1}, \ldots, \boldsymbol{X_p}]$), $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]'$ ($1 \times p$ vector) and σ^2 is a scalar. In this setting $\boldsymbol{\beta}$ and σ^2 are considered unknown and $\boldsymbol{Y}$ and $\boldsymbol{X}$ are our data. Selecting predictors is equivalent to set equal to 0 those element of $\boldsymbol{\beta}$ corresponding to the non selected predictors. It is assumed that $\boldsymbol{X_1}, \ldots, \boldsymbol{X_p}$ contains no variables that would be included in every possible model, this is justified from a Bayesian prospective as initially integrating out those coefficients that belongs to those variable that are included in every possible model. For example if an intercept was to be included in every model (that is the usual case), then $\boldsymbol{1_p} = [1, \ldots, 1]$ should be excluded from the set of potential predictors and $\boldsymbol{Y}$ and $\boldsymbol{X_i}$ should be centered. In their work those two authors build up a prior on each β_i that is a mixture of two normal distributions with

different variances. The first one has most of its mass around zero, while the second one has its mass spread out over possible values. This setup is similar to the "spike and slab" mixture of Mitchell and Beauchamp [1988], but they put a probability mass on $\beta_i = 0$ instead. Introduce the latent variable γ_i that could only assume two values $(0, 1)$, this normal mixture prior on each β_i is represented by:

$$\beta_i|\gamma_i \sim (1-\gamma_i)N(0, \tau_i^2) + \gamma_i N(0, c_i\tau_i^2).$$

The hyper parameter τ_i is set small while c_i is set large so that $N(0, \tau_i^2)$ has its mass around zero, while $N(0, c_i\tau_i^2)$ is diffuse. With this setting when $\gamma_i = 0$ β_i will assumes very small values clustered around zero that it could be estimated as zero, whereas when $\gamma_i = 1$ betas are dispersed and corresponds to an important predictor (see figure 3.2).

This model implies that β_is are independent given the $\boldsymbol{\gamma}$ vector. The prior distribution on each γ_i is a Bernoulli and they are independently distributed, so the prior probability on the $\boldsymbol{\gamma}$ vector is simply the product of p independent Bernoulli:

$$P(\boldsymbol{\gamma}) = \prod_{i=1}^{p} p_i^{\gamma_i}(1-p_i)^{(1-\gamma_i)}.$$

where p_i is the prior probability of γ_i to be equal to one. This probability may be thought as the prior probability of a generic regressor X_i to be included in the model: setting the parameter p_i is equivalent to set the sparsity of the model by defining the a priori expected number of significant regressors. The hierarchical model is then completed by defining the prior on the residual variance σ^2. For this purpose the conjugate invers gamma prior is used:

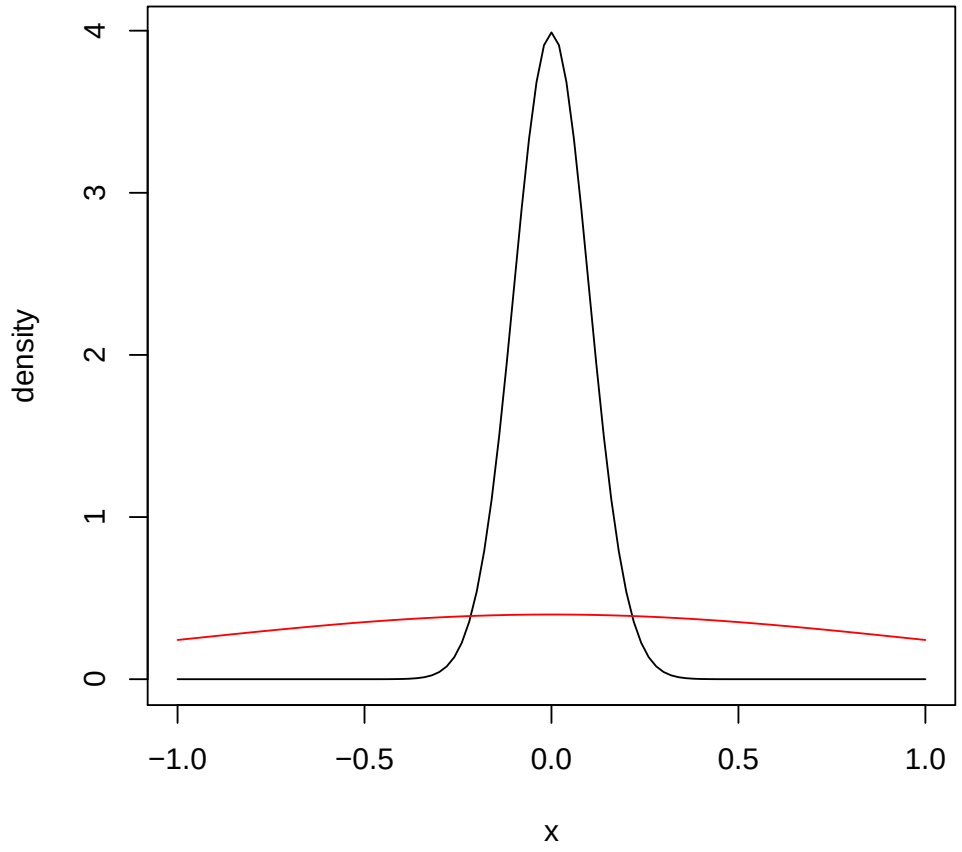


Figure 3.2: An example of overlapped $N(0, \tau_i^2)$ and $N(0, c_i\tau_i^2)$ densities.

$$\sigma^2|\boldsymbol{\gamma} \sim IG(\frac{\nu_\gamma}{2}, \frac{\nu_\gamma\lambda_\gamma}{2}),$$

which is equivalent to $\frac{\nu_\gamma\lambda_\gamma}{\sigma^2} \sim \chi^2_{\nu_\gamma}$. Note that, in this configuration, ν_γ and λ_γ may depend on $\boldsymbol{\gamma}$ to incorporate dependence between $\boldsymbol{\beta}$ and σ^2. For instance one could expect that σ^2 would decrease as the dimension of $\boldsymbol{\beta}$ increased.

Embedding the normal linear model in the hierarchical mixture model allows to obtain the marginal posterior distribution of $\boldsymbol{\gamma}$ as $P(\boldsymbol{\gamma}|\boldsymbol{Y}) \propto P(\boldsymbol{Y}|\boldsymbol{\gamma})P(\boldsymbol{\gamma})$, and it can be used to updates the probabilities on each of the 2^p possible values of $\boldsymbol{\gamma}$. This allows to identify those models that are the "best", i.e. those models that have the higher posterior probabilities, and therefore, that are most supported by the data and the prior distribution. The main target of the Gibbs sampler is to generate a sequence of $\boldsymbol{\gamma}$ values $(\boldsymbol{\gamma^{(0)}}, \boldsymbol{\gamma^{(1)}}, \ldots, \boldsymbol{\gamma^{(m)}})$ which converges in distribution to $P(\boldsymbol{\gamma}|\boldsymbol{Y})$. A crucial observation is that the sequence generated by SSVS, with high probability, contains exactly information relevant to variable selection. This is due to the fact that those $\boldsymbol{\gamma}$ with highest probability will also appear more frequently, while those that appear infrequently or not at all are simply not of interest and can be discarded. Therefore to find the most probable models it is not necessary to explore the whole distribution: many models have small posterior probability and can be ignored. This is also due to the idea of *sparsity*: many of the possible relations can be practically considered as zero. In the small n large p context the true model is always considered as *sparse*. This sequence is embedded in the auxiliary Gibbs sequence of the full sequence of parameter values:

$$\boldsymbol{\beta^{(0)}}, \boldsymbol{\gamma^{(0)}}, \boldsymbol{\sigma^{(0)}}, \boldsymbol{\beta^{(1)}}, \boldsymbol{\gamma^{(1)}}, \boldsymbol{\sigma^{(1)}}, \ldots, \boldsymbol{\beta^{(j)}}, \boldsymbol{\gamma^{(j)}}, \boldsymbol{\sigma^{(j)}}, \ldots, \boldsymbol{\beta^{(m)}}, \boldsymbol{\gamma^{(m)}}, \boldsymbol{\sigma^{(m)}}$$

an ergodic Markov chain generated by the full conditional distributions $P(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\gamma}, \boldsymbol{Y})$, $P(\sigma^2|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{Y})$ and $P(\gamma_i|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma_{-i}}, \boldsymbol{Y})$, where $\boldsymbol{\gamma_{-i}}$ is the vector $\boldsymbol{\gamma}$ without position i, $\boldsymbol{\gamma_{-i}} = [\gamma_1, \gamma_2, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_p]$. Notice that the full conditional[2] of γ does not depend on $\boldsymbol{Y}$ and on σ (this imply a fast update) and the full conditional of σ does not depend on $\boldsymbol{\gamma}$, leading to the following simplifications:

- $P(\sigma^2|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{Y}) = P(\sigma^2|\boldsymbol{\beta}, \boldsymbol{Y})$

[2]This simplification results from the hierarchical structure where $\boldsymbol{\gamma}$ affects $\boldsymbol{Y}$ only through β.

- $P(\gamma_i|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma_{-i}}, \boldsymbol{Y}) = P(\gamma_i|\boldsymbol{\beta}, \boldsymbol{\gamma_{-i}})$.

One of the extension of this model is the one proposed by Smith and Kohn [1996]. Those two authors use the variable selection procedure in the semiparametric additive models context. They implicitly introduce the so called *spike and slab* prior for the regression coefficients, that will be explicitly introduced by George and McCulloch [1997]:

$$\beta_i|\gamma_i, \sigma^2, \boldsymbol{X} \sim (1-\gamma_i)\delta_0 + \gamma_i N(0, \sigma^2 c x_{ii}),$$

where δ_0 represents the Dirac delta function pointed in 0 and x_{ii} is the ith element of the diagonal of $(\boldsymbol{X'X})^{-1}$. On the vector of the selected regression coefficients they specify a g-prior $\boldsymbol{\beta}|\boldsymbol{\gamma}, \sigma^2 \sim N(0, \sigma^2 c(\boldsymbol{X'X})^{-1})$, while the non selected β's are simply excluded from the model. This setting allows to integrate out $\boldsymbol{\beta}$ and σ^2 from the model, leading to a faster computing algorithm. Like George and McCulloch [1993] they uses a Gibbs sampler, but they only need to sample $\boldsymbol{\gamma}$, that is equivalent to explore the model space.
A similar parametrization was used by Brown et al. [1998a]. The novelty introduced by those authors is the procedure adopted for updating $\boldsymbol{\gamma}$ after having integrated out parameters $\boldsymbol{\beta}$ and σ^2. This procedure is faster then the one proposed by Smith and Kohn [1996] and consist in starting from a randomly chosen $\boldsymbol{\gamma}$ and then it moves through a sequence of further values, with each step in the sequence having an element of randomness. A new candidate $\boldsymbol{\gamma}$ is generated at each point of the sequence by randomly modifying the current one. If this new candidate has a higher probability than the current one, then we move to it. The move is still possible even if the new candidate has a lower probability, but it must be accepted with a certain probability. The new candidate $\boldsymbol{\gamma^{i+1}}$ is generated from the current one $\boldsymbol{\gamma^i}$ by one of two types of moves:

- **Adding or deleting** Select one of the p covariates at random. If it is

already included in the model, delete the variable; if it is not currently in the model, add it to the model. In this way the new candidate $\boldsymbol{\gamma^{i+1}}$ differs from the previous $\boldsymbol{\gamma^{i}}$ in just one of its entries;

- **Swapping** At the same time choose at random one of the included covariates and one of the excluded ones. Then swap their values: exclude the previously included variable and include the previously excluded variable. Thus the new candidate $\boldsymbol{\gamma^{i+1}}$ differs from the previous one in two of its entries.

The new candidate $\boldsymbol{\gamma^{i+1}}$ is then accepted with probability:

$$min[\frac{g(\boldsymbol{\gamma^{i+1}})}{g(\boldsymbol{\gamma^{i}})}, 1]$$

with $g(\boldsymbol{\gamma}) = P(\boldsymbol{\gamma})P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\gamma})$, where $P(\boldsymbol{\gamma})$ is the prior on $\boldsymbol{\gamma}$ and $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\gamma})$ is the likelihood. The above formula is obtained considering that the proposal distribution. Furthermore, at each iteration the first kind of move is chosen with probability ϕ and the second one with the reminder probability $(1-\phi)$. In this scheme the parameter ϕ must be chosen. The value proposed by Brown et al. [1998a] is 0.5 but other values are always possible.
When it's possible to integrate out all the parameters but $\boldsymbol{\gamma}$ this Metropolis algorithm is preferred to the Gibbs sampler because it allows a faster exploration of the space of the relevant models. Over the years a large amount of MCMC schemes have been proposed to achieve a faster exploration of the relevant models. Recently Bottolo and Richardson [2010] proposed a sampling algorithm based upon Evolutionary Monte Carlo with a parallel tempering approach to explore the model space faster. This algorithm overcomes the known difficulties faced by MCMC schemes when attempting to sample a high dimension multimodal space.
The previous examples are all based on regression with only one response variable. Brown et al. [1998b] generalized the idea of SSVS to multivariate regression models with q response variables. To define the SSVS procedure

in this context is necessary an introduction to matrix variate distribution. Following the notation introduced by Dawid [1981] $\boldsymbol{Y} - \boldsymbol{M} \sim N(\boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ represent a $n \times q$ normal matrix-variate distribution where $\boldsymbol{M}$ indicates the mean and $\gamma_{ii}\boldsymbol{\Sigma}$ and $\sigma_{jj}\Gamma$ indicate the covariance matrices of respectively $i-th$ row and $j-th$ column. This notation has the advantage of preserving the matrix structure without the need to string by row or column as a vector. Conditionally on parameters $\boldsymbol{a}$, $\boldsymbol{B}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}$ the standard multivariate normal regression model is defined as

$$\boldsymbol{Y} - \mathbf{1}\boldsymbol{a'} - \boldsymbol{X}\boldsymbol{B} \sim N(\mathbf{I}_n, \boldsymbol{\Sigma}),$$

with $n \times q$ random matrix $\boldsymbol{Y}$, $\mathbf{1}$ an $n \times 1$ vector of 1s, $1 \times q$ vector of intercepts $\boldsymbol{a}$, $n \times p$ model matrix $\boldsymbol{X}$ regarded as fixed and $\boldsymbol{B}$ the $p \times q$ matrix of regression coefficients. Then special forms of prior distribution for parameters $\boldsymbol{a}$, $\boldsymbol{B}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}$ are given

$$\begin{aligned} \boldsymbol{a'} - \boldsymbol{a_0} &\sim N(h, \boldsymbol{\Sigma}) \\ \boldsymbol{B} - \boldsymbol{B_0} &\sim N(\boldsymbol{H}, \boldsymbol{\Sigma}) \\ \boldsymbol{\Sigma} &\sim IW(\delta, \boldsymbol{Q}) \end{aligned}$$

and those three parameters can be integrated out from the model. Note that Brown et al. [1998b] specify a latent $p \times 1$ vector indicator for the inclusion of the covariates. In this setting if the $j-th$ element of the vector is equal to 1, then the $j-th$ covariate is significant for all the q response variables. As a consequence is not possible to define different sets of of significant covariates for different response variables. Integrating out the three parameters, jointly with a QR deletion-addition algorithm in the calculation of the marginal likelihood, leads to a very efficient Gibbs MCMC scheme for posterior inference. Moreover, Brown et al. [1998b] use the *model averaging* idea of Madigan and York [1995] for prediction of new observations Y_f.

Procedure based on the predictive distribution $p(Y_f|\boldsymbol{Y}, \boldsymbol{X_f})$ and exploits the conjugacy of the model; after integrating out $\boldsymbol{a}$, $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ it is possible to calculate Y_f as weighted mean of the expected values of $p(Y_f|\boldsymbol{Y}, \boldsymbol{X_f})$ given different configurations of $\boldsymbol{\gamma}$, using as weight the posterior probabilities of these configurations. According to the posterior probabilities only the best k configurations are used for prediction.
Scott and Berger [2010] studied the multiplicity correction effect of standard Bayesian variable selection priors in the linear regression. Their first goal is to clarify when, and how, multiplicity correction is automatic in Bayesian analysis, and to contrast this multiplicity correction with the Bayesian Ockham's-razor effect. These two authors find that multiplicity issues are particularly relevant when researchers have little reason to suspect one model over another, and simply want the data to flag important covariates from a large pool. In such cases Bayesian variable selection must be used as an exploratory tool. Then they focus their attention in the comparison of empirical-Bayes and fully Bayesian approaches to multiplicity correction in variable selection. They found considerable differences between the results of the two approaches and suggest that considerable care must be be taken with the empirical-Bayes approach in variable selection.
Last Bayesian variable selection have been studied in logit and probit models for binary and multinomial outcomes. Some examples are the models proposed by Sha and al. [2004], Holmes and Held [2006] and Albert and Chib [1993].

3.5 Integration, some existing methods

The main goal of our method is to find groups of DNA fragments that possibly affects the expression of one or more genes. Not many techniques to integrate CGH and gene expression data have been developed; here we briefly show the ones of Chin et al. [2006], Choi et al. [2010], Richardson et al. [2010], Monni and Tadesse [2009], **?**,Cai and al. [2011] and Yin and Li

[2011].

One of the first works on these data is the one of Chin et al. [2006]. They explore the role of genome copy number aberrations in breast cancer pathophysiology by identifying association between recurrent CNAs, gene expression and clinical outcome. Using an unsupervised clustering approach they shows that the recurrent CNAs differ between tumor sub types defined by expression pattern and that stratification of patients according to outcome can be improved by measuring both expression and copy number, especially high level amplification.

Choi et al. [2010] developed a double-layered mixture model (DLMM) that simultaneously scores the association between paired copy number and gene expression data using related latent variables in the two datasets. The method assigns high scores to elevated or reduced measurements only if the expression changes are co-observed consistently across samples with copy number aberration. However, Choi et al. [2010] consider only copy number-associated changes in gene expression levels. In other terms, the definition of over or under expression is relative to the distribution of expression values in samples with no aberrant copy numbers. Thus, even if a gene is highly expressed in many samples, this gene will not be considered as over-expressed as long as this is not related to a concordant amplification. This feature may not be optimal, as the investigation of gene expression changes that appear to be independent of concurrent amplifications may also be of interest. Furthermore, in the model proposed by Choi et al. [2010], gene expression levels are affected only by copy number aberrations occurring on the same segment of DNA. Indeed, it may be expected that expression levels and CGH aberration of adjacent segments in the DNA may not be independent. Hence, a more realistic model of gene-gene interaction may be preferable.

Richardson et al. [2010] consider the generic task of building efficient regression models for sparse multivariate analysis of high dimensional data sets, where both the number of responses and of predictors are large with respect to the sample size. The define q regression equations as $\boldsymbol{y_k} =$

$\alpha_k \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta_k} + \boldsymbol{\epsilon_k}, \quad k = 1, \ldots, q$ where $\boldsymbol{\epsilon_k} \sim N_n(\mathbf{0}, \sigma_k^2 \mathbf{I}_n)$. Note that in this definition every regression equation has its own intercept α_k and error variance σ_k^2. As many other authors they introduce the latent binary vector $\boldsymbol{\gamma_k} = [\gamma_{k1}, \ldots, \gamma_{kj}, \ldots, \gamma_{kp}]'$ for each equation where $\gamma_{kj} = 0$ if $\beta_{kj} \neq 0$ and $\gamma_{kj} = 0$ if $\beta_{kj} = 0$, $j = 1, \ldots, p$. Considering at the same time all q regression, the $q \times p$ latent binary matrix $\boldsymbol{\Gamma} = (\boldsymbol{\gamma_1}, \ldots, \boldsymbol{\gamma_k}, \ldots, \boldsymbol{\gamma_q})'$ can be obtained. then they assume independence on the q regression, given $\boldsymbol{\Gamma}$, therefore the likelihood becomes:

$$\prod_{k=1}^{q} \left(\frac{1}{2\pi\sigma_k^2} \right)^{1/2} exp\left\{ -\frac{1}{2\sigma_k^2} \left(\boldsymbol{y_k} - \alpha_k \mathbf{1}_n - \boldsymbol{X_{\gamma k}} \boldsymbol{\beta_{\gamma k}} \right)' \left(\boldsymbol{y_k} - \alpha_k \mathbf{1}_n - \boldsymbol{X_{\gamma k}} \boldsymbol{\beta_{\gamma k}} \right) \right\},$$

where $\boldsymbol{\beta_\gamma k}$ is the non-zero vector of regression coefficients of the k-th regression and, similarly, $\boldsymbol{X_{\gamma k}}$ is the design matrix with columns corresponding to $\gamma_k j = 1$. They follows a g-priors representation for the regression coefficient, assuming that $\boldsymbol{\beta_{\gamma k}} | \boldsymbol{\gamma_k}, g, \sigma_k^2 \sim N_{p_{\gamma k}}\left(\mathbf{0}, g(\boldsymbol{X'_{\gamma k}} \boldsymbol{X_{\gamma k}})^{-1} \sigma_k^2\right)$, where $p_{\gamma k} \equiv \boldsymbol{\gamma'_k} \mathbf{1_p}$ the number of non-zero elements in $\boldsymbol{\gamma_k}$. To increase flexibility $g \sim InvGam(a_g, b_g)$. Note that g is the level of shrinkage and it is common for all q regression equations, thus g is one of the parameters that links the q regressions. Prior specification is then completed by assigning a Bernoulli prior on the latent binary indicators, $p(\gamma_{kj} | \omega_{kj}) = \omega_{kj}^{\gamma_{kj}} (1 - \omega_{kj})^{1 - \gamma_{kj}}$, with $k = 1, \ldots, q, j = 1, \ldots, p$. A crucial point of their model are the prior probability for $\boldsymbol{\Gamma}$, i.e. to model the matrix

$$\boldsymbol{\Omega} = \begin{bmatrix} \omega_{11} & \ldots & \omega_{1j} & \ldots & \omega_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \omega_{k1} & \ldots & \omega_{kj} & \ldots & \omega_{kp} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \omega_{q1} & \ldots & \omega_{qj} & \ldots & \omega_{qp} \end{bmatrix}$$

This matrix is so important because it controls sparsity of the model and here strength between the responses can be included. Richardson et al.

[2010] propose three different strategies:

1. $\omega_{kj} = \omega_k$ with $\omega_k \sim Beta(a_{\omega_k}, b_{\omega_k})$;
2. $\omega_{kj} = \omega_j$ with $\omega_j \sim Beta(a_{\omega_j}, b_{\omega_j})$;
3. $\omega_{kj} = \omega_k \times \rho_j$ with $\omega_k \sim Beta(a_{\omega_k}, b_{\omega_k}), \rho_j \sim Gamma(c_{\rho_j}, d_{\rho_j}), 0 \leq \omega_{kj} \leq 1$.

The first strategy assumes that the underlying selection probabilities for each response may be different and arise from independent Beta distribution. In the second model ω_j quantifies the probability for each predictor to be included in all the regression. The third one uses a shared column effect ρ_j to moderate the underlying selection probability ω_k specific to the k-th regression in a multiplicative fashion. They also show that the third model is the one with better performance allowing an excellent separation between hot spot[3] and background.

Monni and Tadesse [2009] proposed a stochastic partitioning method to associate responses and covariates, both much larger in number than the sample size. They present a stochastic algorithm that searches for sets of covariates associated with sets of of correlated outcomes. Their model combines the ideas of mixture models, regression models and variable selection identify group structures and key relationships in high-dimensional data sets. To that aim they construct a Markov chain in the space of pairwise partition of the set of regressors and of the sets of responses. Each element of the partition is then a pair of subsets, the first one composed of covariates and the other one composed of their correlated outcomes. They impose (because of the asymmetric role of predictors and responses) that each outcome should belong to one and only one pair, while covariates could belong to more than one element of the partition.

Now consider a data set of N independent samples with p covariates and q outcomes. In order to identify sets of outcome related to a set of predictor,

[3] *Hot spot* refers to predictor associated with many responses.

they consider partitions of the variables into sets of pairs $S = (X_I, Y_J)$, with $I \subset \{1, \ldots, p\}$ and $J \subset \{1, \ldots, q\}$. In their paper Monni and Tadesse [2009] called a partition of the data as a *configuration*, its pairs as the *component* of the configuration and the number of the latter as the *length of the configuration*. Assuming independence among outcomes in distinct components of a given partition, they define the probability of each configuration as the product of the probabilities of its components. Then they consider a multivariate Gaussian mixture model with an unknown number of components, where the mean and the scale of each component are determined by a regression model on a subset of predictors. Consider the distribution of the outcomes $Y_{t_1}, \ldots, Y_{t_{n_k}}$ of a component $S_k = (m_k, n_k)$, it is assumed to be:

$$Y_{ji}|S_k \overset{iid}{\sim} N(\alpha_j + \mu_k, \sigma_k^2),$$

where $j = t_1, \ldots, t_{n_k}$ indices outcomes that belong to that component, $i = 1, \ldots, N$ indices samples, σ_k^2 is the component specific variance, the location of the distribution is split in two parts α_j and $\mu_k = g_k(X_{s_1}, \ldots, X_{s_{m_k}})$. Substantially they are fitting a mixture of regression models, where the effects of the regressors on the response are the same within a component, but vary from one component to another. Writing the regression model as:

$$Y_{ji} = \alpha_j + \sum_{r=1}^{m_k} \beta_{ks_r} X_{s_r i} + \epsilon_{ji}, \qquad \epsilon_{ji} \sim N(0, \sigma_k^2)$$

the likelihood for S_k is then given by:

$$\phi(m_k, n_k) = (2\pi\sigma_k^2)^{-n_k N/2} exp\left\{ -\frac{1}{2\sigma_k^2} \sum_{i=1}^{N} \sum_{j=1}^{n_k} \left(Y_{t_j i} - \alpha_{t_j} - \sum_{r=1}^{m_k} \beta_{ks_r} X_{s_r i} \right)^2 \right\}.$$

Note that component of type $(0, f)$ are distributed as $N(\alpha_j, \sigma_k^2)$ since no

regressors are associated with f response variables. Moreover components of type $(v, 0)$ are equivalent to those of type $(1, 0)$, since X is viewed as a fixed covariate matrix and the corresponding likelihood is equal to 1.
The prior probability distribution for the model parameters are then specified; conjugate priors are chosen and the component parameters are integrated out. They indicate with $\theta'_k = (\alpha_{t_1}, \ldots, \alpha_{t_{n_k}}, \beta_{s_1}, \ldots, \beta_{s_{m_k}})$ the $(n_k + m_k)$ vector of regression coefficients, then choose:

$$\begin{aligned} \theta_k &\sim N(\theta_{0k}, H_0\sigma_k^2) \\ \sigma_k^2 &\sim IG(\sigma_0^2, \nu) \end{aligned}$$

where $\theta'_{0k} = (\alpha_{0t_1}, \ldots, \alpha_{0t_{n_k}}, \beta_{0s_1}, \ldots, \beta_{0t_{m_k}})$, $H_0 = \mathsf{diag}(h_0\mathbf{1}_{n_k}, h\mathbf{1}_{m_k})$. In this parametrization H_0 controls the strength of the prior information on the regression coefficients with larger values of h_0 and h corresponding to a wider spread around θ_{0k}. Last the prior probability of each configuration is given by:

$$p((m_1, n_1) \ldots (n_k, m_k)) \propto \prod_{k=1}^{K} \rho^{m_k n_k},$$

with $0 < \rho < 1$. This way large components are a priori penalized with stronger penalty as ρ decreases. Monni and Tadesse [2009] proposed two different MCMC algorithm, the first one very simple and the second one is an extension of the first one using parallel tempering.
Scott-Boyer and al. [2012] studied eQTL mapping in a regression framework. Let $g = 1, \ldots, G$ denotes a particular gene or trait, $i = 1, \ldots, n$ denotes a particular strain or individual and $j = 1, \ldots, S$ denotes a particular SNP. The model is then defined by:

$$y_{ig} = \mu_g + \sum_{j=1}^{S} x_{ij}\gamma_{jg}\beta_{jg} + \epsilon_{ig},$$

where y_{ig} is the expression level of gene g for the individual strain i, μ_g is the overall mean expression level of gene g, x_{ij} represents the genotype at locus j for strain i, β_{jg} is the effect size of SNP j on gene g, γ_{jg} is the binary inclusion indicator and ϵ_{ig} is an error term, assumed to be Gaussian with gene specific variance σ_g^2. They put a Bernoulli prior distribution over each γ_{jg} with parameter ω_{jg} that represents the inclusion probability. On this parameter they put a mixture probability, in order to reduce the false discovery rate, as follows:

$$\omega_{ig} \sim p_j\delta_0(\omega_{ig}) + (1 - p_j)Beta(a_j, b_j)(\omega_{ig}),$$

where δ_0 is the delta Dirac centered in zero, p_j, represent the probability that ω_{jg} is zero and is identical for all genes. On p_j they put a $Beta(a_0, b_0)$ distribution, on a_j an $Exp(\lambda_a)$ and on b_j an $Exp(\lambda_b)$. They assume $\mu_g \sim N(m_g, \tau_g^2)$ where m_g and τ_g are the empirical mean and standard deviation of gene expression g. Last β_{jg} is expressed by a mixture as follows:

$$\beta_{jg} = \gamma_{jg}N(0, v_{jg}^2) + (1 - \gamma)\delta_0,$$

with $v_{jg}^2 = c(x_j'x_j)^{-1}\sigma_g^2$, where c is a scaling factor parameter fixed at S, $(x_j'x_j)^{-1}$ mimics the regression variance, leading to the g-prior and $\sigma_g^2 \sim IG(\frac{1}{2}, \frac{1}{2})$ and can be integrated out. The authors underline that this model has two clear advantages. First it can deal with a large number of genes at a time (that facilitates the detection of hotspots), and second each gene expression/trait has its own inclusion indicator γ_{jg}, with inclusion probability parameters not considered common for all SNP positions nor

supposed identical for all genes but depending on the SNP positions as in Richardson et al. [2010].

Cai and al. [2011] developed a method to estimate the conditional independent relationships among a set of genes adjusting for possible genetic effects, as well as the genetic architecture that influences gene expression. They realized a covariance-adjusted precision matrix estimation (CAPME) method using constrained l_1 minimization, implemented by linear programming. Let p be the number of outcomes, q the number of regressors and n the number of samples, they build up the following model:

$$\boldsymbol{y} = \Gamma_0 \boldsymbol{x} + \boldsymbol{z},$$

where $\boldsymbol{y} = (y_1, \ldots, y_p)'$ is a random vector denoting expression levels, $\boldsymbol{x} = (x_1, \ldots, x_q)'$ is a random vector describe the coding for q markers, Γ_0 is a $p \times q$ unknown coefficients matrix, $\boldsymbol{z}$ is a $p \times 1$ normal vector with mean zero, covariance matrix $\Sigma_0 = (\sigma_{ij}^0)$ and precision matrix $\Omega_0(\omega_{ij}^0) = \Sigma_0^{-1}$. They further assume that $\boldsymbol{x}$ and $\boldsymbol{z}$ are independent and that they have n independent identically distributed observations $(\boldsymbol{x}_k, \boldsymbol{y}_k)$ $(k = 1, \ldots, n)$ from the previous model. Both matrices Γ_0 and Ω_0 are expected to be sparse, and Ω_0 as an interpretation of conditional dependency and can be used to construct a conditional dependency graph. For example a generic edge between y_i and y_j is excluded if and only if the corresponding z_i and z_j are conditionally independent given all other z_k's. Since $\boldsymbol{z}$ follows a multivariate normal distribution, the conditional independence of z_i and z_j leads to $\omega_{ij} = 0$. The authors are interested in both estimation of Ω_0 and Γ_0, and use a two step estimation: First they estimate Γ_0 by solving a linear programming problem, then using this estimated value they estimate Ω_0 again solving an optimization problem and they iterate until convergence.

A similar approach was developed by Yin and Li [2011], but in their model they are most interested in the estimation of the covariance structure and is of less importance for our purposes.

4 Integrating CHG and Gene expression data

In this chapter we describe the model we develop for the integration of genetical genomics data. We first describe the model build up on the covariates of the regression and then we focus on the variable selection model. After a brief overview on the posterior inference, we focus on the simulations we implemented. Finally we apply our method to real data.

4.1 Hierarchical Model

Our proposed modeling strategy starts with the formulation of a hierarchical model, integrating gene expression levels with genetic data, that includes measurement errors and mixture priors for variable selection. We couple this model with a hidden Markov model (HMM) on the genetic covariates. Our approach utilizes prior distributions that cleverly incorporate dependencies among selected variables. It also incorporates stochastic search variable selection techniques within an inferential scheme that allows to select associations among genomic and genetic variables while simultaneously inferring the hidden states of the HMM. The graphical formulation of the model is illustrated in Figure 4.1 and its major components are described below. We also summarize the hierarchical formulation of our full model in Figure 4.2 at the end of this Section.

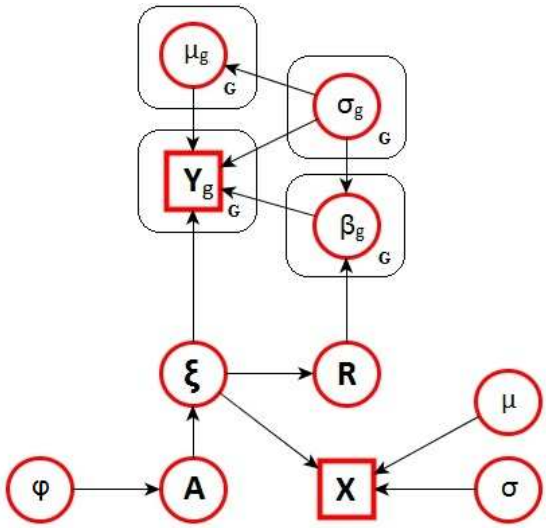


Figure 4.1: Graphical formulation of the proposed probabilistic model described in Section 2.

4.1.1 Measurement Error Model for Genetical Genomic Data

Let Y_{ig}, for $g = 1, \ldots, G$, denote the gene expression measurements and X_{im} the CGH measurement, i.e., the normalized $\log_2$ ratio observed at the m-th probe for sample i, with $m = 1, \ldots, M$ and $i = 1, \ldots, n$. We incorporate measurement errors in the formulation of our model by introducing latent variables ξ_{im} representing the copy number state, i.e., loss, gain or neutral, of the m-th probe in the i-th sample,

$\xi_{im} = 1$ for copy number loss (less than two copies of the fragment);

$\xi_{im} = 2$ for copy-neutral state (exactly two copies of the fragment);

$\xi_{im} = 3$ for a single copy gain (exactly three copies of the fragment);

$\xi_{im} = 4$ for multiple copy gains (more than three copies of the fragment).

The rationale behind this modeling choice is that, at the genomic level, the expression of a gene is affected by the copy number state of a given clone, with the CGH measurement representing a surrogate of this effect on a continuous scale.

Let $\boldsymbol{Z} = [\boldsymbol{Y}, \boldsymbol{X}]$ denote the $(n \times (G + M))$ matrix including all the data measurements and let $\boldsymbol{\xi} = [\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_M]$ be the $(n \times M)$ matrix of the categorical latent variables. We assume conditional independence of the gene measurements, conditionally upon the copy number states, that is, $\mathbf{Y}_i \perp \mathbf{Y}_j | \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M$, and of the CGH measurements, conditionally upon their states, that is, $\mathbf{X}_i \perp \mathbf{X}_j | \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M$. We consequently factorize the likelihood as

$$f(\mathbf{Z}|\boldsymbol{\xi}) = \prod_{g=1}^{G} f(\mathbf{Y}_g|\boldsymbol{\xi}) \prod_{m=1}^{M} f(\mathbf{X}_m|\boldsymbol{\xi}_m). \tag{4.1}$$

Monni and Tadesse [2009] and Richardson et al. [2010] have suggested linear regression models that relate the gene expression levels to the CGH data. For the conditional model, we therefore assume $f(\mathbf{Y}_g|\boldsymbol{\xi}) \sim N(\boldsymbol{\xi}\boldsymbol{\beta}_g, \sigma_g \mathbf{I}_n)$. This model formulation is equivalent to a system of G linear regression models of the type

$$\begin{aligned} \mathbf{Y}_1 &= \mathbf{1}_n \mu_1 + \boldsymbol{\xi}\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1 \\ \mathbf{Y}_2 &= \mathbf{1}_n \mu_2 + \boldsymbol{\xi}\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2 \\ &\vdots \\ \mathbf{Y}_G &= \mathbf{1}_n \mu_G + \boldsymbol{\xi}\boldsymbol{\beta}_G + \boldsymbol{\epsilon}_G, \end{aligned} \tag{4.2}$$

with $\mu_1, \ldots, \mu_G$ gene-specific intercepts and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G$ vectors of regressions coefficients. Model (4.1) is completed with an appropriate choice for the

distribution of the CGH measurements, that is a model for $\boldsymbol{X}|\boldsymbol{\xi}_m$. For this we employ hidden Markov models.

4.1.2 Hidden Markov Model on the Genetic Covariates

CGH data are "state persistent", meaning that copy numbers gains or losses at a region are often associated to an increased probability of gains and losses at a neighboring region. We therefore aim at a model that enforces clustering of the measurements into homogeneous groups, reflecting different levels of amplification/deletion as captured by the (hidden) states of the CGH clones. Several methods have been proposed in the literature for such purpose. Here, we adapt the popular model proposed by Guha et al. [2008], that uses hidden Markov models and 4 hidden states, and apply it independently to each sample. Other models, such as those of Fox et al. [2007] and Du et al. [2010], that consider the number of possible states as a random variable, may be similarly adapted to our setting.

We assume the CGH measurements independently normally distributed, conditionally on their copy number states,

$$X_{im}|(\xi_{im} = j) \overset{iid}{\sim} N(\mu_j, \sigma_j^2), \tag{4.3}$$

with μ_j and σ_j^2 representing the expected $\log_2$ ratio and the variance of all probes in state j, for $j = 1, \ldots, 4$. We then capture dependence with an hidden Markov model which assumes that the probability of a probe to be in a certain state depends only on the state of the previous one along the DNA sequence. This implies that the conditional probability of neighboring probes to be in a certain state is

$$P(\xi_{i(m+1)}|\xi_{i1}, \ldots, \xi_{im}) = P(\xi_{i(m+1)}|\xi_{im}) = a_{\xi_{ik}\xi_{i(k+1)}}. \tag{4.4}$$

with $\mathbf{A} = (a_{hj})$, for $h, j = 1, \ldots, 4$, forming the matrix of transition probabilities, with strictly positive elements. This matrix has a unique stationary distribution π_A. We also assume that the state for the first CGH probe, ξ_{i1}, is distributed as π_A. Finally, we assume independent Dirichlet priors, $Dir(\phi)$, over the rows of the matrix $\boldsymbol{A}$. We complete the model, following Guha et al. [2008], by choosing truncated prior distribution for the parameters μ_j in (4.3), reflecting their ordering, as

$$\mu_j \begin{cases} \mu_1 \sim N(-1, \tau_1^2) I_{\{\mu_1 < upp_{\mu_1}\}} \\ \mu_2 \sim N(0, \tau_2^2) I_{\{low_{\mu_2} < \mu_2 < upp_{\mu_2}\}} \\ \mu_3 \sim N(0.58, \tau_3^2) I_{\{low_{\mu_3} < \mu_3 < upp_{\mu_3}\}} \\ \mu_4 \sim N(1, \tau_4^2) I_{\{\mu_4 > low_{\mu_4}\}}. \end{cases}$$

The low boundary for μ_4 avoids that a large number of single copy gains can be erroneously classified as multiple copy gains. Similarly, we choose truncated distributions for the σ_j^{-2}

$$\sigma_j \begin{cases} \sigma_1^{-2} \sim Ga(b_1, l_1) I_{\{\sigma_1^{-2} > upp_{\sigma_1}\}} \\ \sigma_2^{-2} \sim Ga(b_2, l_2) I_{\{\sigma_2^{-2} > upp_{\sigma_2}\}} \\ \sigma_3^{-2} \sim Ga(b_3, l_3) I_{\{\sigma_3^{-2} > upp_{\sigma_3}\}} \\ \sigma_4^{-2} \sim Ga(b_4, l_4) I_{\{\sigma_4^{-2} > upp_{\sigma_4}\}}. \end{cases}$$

The choice of the truncation $\sigma_j^{-2} > 6$ is a mild assumption, and it is equivalent to setting $\sigma_j < 0.41$.

4.1.3 Prior Model for Variable Selection

For each gene we wish to find a parsimonious list of CGH aberrations that affect the gene expression measurements with high confidence. This is equivalent to infer which elements of the vector $\boldsymbol{\beta}_g$ in (4.2) are non-zero, a classical variable selection problem. The resulting "network" of gene-CGH associations can be encoded by a $(G \times M)$ matrix $\boldsymbol{R}$ of binary elements. Specifically,

for gene g and probe m, the value $r_{gm} = 1$ indicates that the corresponding coefficient β_{gm} is significant, and therefore included in the g-th regression in (4.2). Otherwise, $r_{gm} = 0$ indicates that the corresponding regression coefficient is zero. The regression coefficient parameters are then stochastically independent, given $\mathbf{R}$, and have the following mixture prior distribution

$$\pi(\beta_{gm}|r_{gm}, \sigma_g^2) = r_{gm} N(0, c_\beta \sigma_g^2) + (1 - r_{gm})\delta_0(\beta_{gm}), \tag{4.5}$$

with $\delta_0(\cdot)$ a point mass at zero. Prior of type (4.5) are known as a spike-and-slab prior in the Bayesian variable selection literature, see George and McCulloch [1997] for univariate linear regression models, and Brown et al. [1998b, 2002] and Sha and al. [2004] for multivariate regression models. The prior model is completed with a Gamma prior on σ_g^{-2}, that is $\sigma_g^{-2} \sim Ga(\frac{\delta}{2}, \frac{d}{2})$, and a Normal distribution on the intercepts, $\mu_g|\sigma_g^2 \sim N(0, c_\mu \sigma_g^2)$, with δ, d and c_μ to be chosen. Note that assumptions on the marginal distribution of $(\mathbf{X}_1, \ldots, \mathbf{X}_M)$ do not affect the association network, which is fully encoded by $\mathbf{R}$. Mixture priors for variable selection have been employed in genomic applications to infer biological networks of high dimensionality, see for example Jones and al. [2005], Richardson et al. [2010] and Stingo et al. [2010]. The variable selection formulation we adopt follows Stingo et al. [2010] and overcomes the somehow rigid structure of the model in Brown et al. [1998b], which does not allow to select different predictors for different responses. See also Monni and Tadesse [2009] for an approach based on partition models.

Lastly, we describe our prior choice for the elements r_{gm}'s of the matrix $\boldsymbol{R}$. For this, we incorporate information about the dependence structure among states of adjacent CGH aberrations. More precisely, we enforce a dependence structure among the r_{gm} indicators as follows: we assume that the probability of selection at location m depends on the state of local aberrations at the adjacent positions $m-1$ and of $m+1$. In particular, we want to exploit dependence if $\boldsymbol{X_m}$ and $\boldsymbol{X_{(m-1)}}$ (or $\boldsymbol{X_{(m+1)}}$ or both) share the

same copy number state, i.e. if there is no change in state, since arguably the effect of $\boldsymbol{\xi_m}$ on $\boldsymbol{Y}_g$ should be more likely the more probes share the same state, i.e., the more persistent the state. We express this as a conditional mixture prior distribution of the type

$$\pi(r_{gm}|r_{g(m-1)}, r_{g(m+1)}, \boldsymbol{\xi}, \pi_1) = \gamma[\pi_1^{r_{gm}}(1-\pi_1)^{(1-r_{gm})}] + \sum_{j=1}^{2} \omega_j I_{\{r_{gm}=r_{g(m+(-1)^j)}\}}, \tag{4.6}$$

with $\gamma \in [0,1]$ and where we impose the constraint $\sum_{j=1}^{2} \omega_j = (1-\gamma)$. Note that for the first and last segment respectively ω_1 and ω_2 are equal to zero.

We assume that the value of r_{gm} is drawn independently of the adjacent configurations with probability γ, whereas, with probability $(1-\gamma)$, it coincides with one (or both) of the adjacent values in the $\boldsymbol{R}$ matrix. We note that (4.6) reduces to independent priors of type $r_{gm} \sim Bern(\pi_1)$ in the case $\gamma = 1$. For each DNA segment we define the parameters γ, ω_1 and ω_2 as follows:

$$\gamma = \frac{\alpha}{\alpha + s_{(m-1)m} + s_{m(m+1)}},$$

$$\omega_1 = \frac{s_{(m-1)m}}{\alpha + s_{(m-1)m} + s_{m(m+1)}}, \quad \omega_2 = \frac{s_{m(m+1)}}{\alpha + s_{(m-1)m} + s_{m(m+1)}} \tag{4.7}$$

with $s_{(m-1)m} = (\frac{\exp\{1-\frac{d_m}{D}\}-1}{\exp\{1\}-1})\frac{1}{N}\sum_{i=1}^{N} I_{\{\xi_{i(m-1)}=\xi_{im}\}}$ and with α a value to be chosen. Quantities $s_{(m-1)m}$ and capture the average empirical frequencies of change points along the sequence of the copy number states, across samples. Construction (4.6) and (4.7) allows us to incorporate spatial dependence along the genome sequence into the prior model, as the value of r_{gm} depends on the persistence of a particular state in its neighborhood. In particular,

the case $s_{(m-1)m} = s_{m(m+1)} = 0$ reduces to the independent prior $r_{g(m)} \sim Bern(\pi_1)$, while larger non-negative values of either $s_{(m-1)m}$ or $s_{m(m+1)}$ lead to smaller γ values and larger ω_1 and ω_2 values. The prior probability of $r_{gm} = 1$ therefore increases if $r_{g(m-1)}$ (or $r_{g(m+1)}$) is equal to one and if in at least one sample there is no change point between states of probes m and $m-1$ (or $m+1$). The parameter α captures the relative strength of the dependence structure. In particular, $\alpha = 0$ implies $\gamma = 0$, while $\alpha \to \infty$ leads to $\gamma = 1$, that is the independent prior.

We complete prior (4.6) by further imposing a Beta hyperprior, $\pi_1 \sim Beta(e, f)$, and integrating π_1 out we obtain

$$\begin{aligned} \pi(r_{gm}|r_{g(m-1)}, r_{g(m+1)}, \boldsymbol{\xi}) &= \gamma \frac{\Gamma(e+f)\Gamma(e+r_{gm})\Gamma(f+1-r_{gm})}{\Gamma(e+f+1)\Gamma(e)\Gamma(f)} \\ &\quad + \sum_{j=1}^{2} \omega_j I_{\{r_{gm}=r_{g(m+(-1)^j)}\}}. \end{aligned} \tag{4.8}$$

Figure 4.2 summarizes the hierarchical formulation of our full model.

In the case study of Section 4.4.3 we also investigate a refined version of our prior model (4.6) and (4.7) by taking into account physical distances between CGH clones. For this we define

$$s_{(m-1)m} = \frac{1}{n} \sum_{i=1}^{n} I_{\{\xi_{im}=\xi_{i(m-1)}\}} \left(\frac{\exp\{1 - \frac{d_m}{D}\} - 1}{\exp - 1} \right) \tag{4.9}$$

with d_m the distance between the $(m-1)$-th and m-th CGH probes and D a quantity to be chosen (some instances could be: length of the chromosome, length of DNA, ...). This formulation allows us to further weigh quantities s_1 and s_2 by the distances between adjacent probes. Similar non-linear weights have been used for example by Wang and al. [2008]. If two adjacent segments overlap (i.e. $d_m = 0$), then the weight is zero. At the maximal distance (i.e. $d_m = D$) the weight assumes its maximum value, that is one. We note that distances for probes that occur at the sequence boundaries are set to zero by default.

Likelihood:

$$
\begin{array}{lcl}
f(\mathbf{Z}|\boldsymbol{\xi}) & = & \prod_{g=1}^{G} f(\mathbf{Y}_g|\boldsymbol{\xi}) \prod_{m=1}^{M} f(\mathbf{X}_m|\boldsymbol{\xi}_m), \\
f(\mathbf{Y}_g|\boldsymbol{\xi}) & \sim & N(\boldsymbol{\xi}\boldsymbol{\beta}_g + \mathbf{1}_n\mu_g, \sigma_g^2\mathbf{I}_n), \\
f(X_{im}|(\xi_{im}=j)) & \sim & N(\mu_j, \sigma_j^2), \\
P(\xi_{i(m+1)}=i|\xi_{im}=j) = a_{ij}. & &
\end{array}
$$

Model parameters:

Conditional model			*Marginal Model*		
$\beta_{gm}\|r_{gm},\sigma_g^2$	$\sim$	$r_{gm}N(0,c_\beta\sigma_g^2)+(1-r_{gm})\delta_0(\beta_{gm})$	$\boldsymbol{A}_{j.}$	$\sim$	$Dir(\boldsymbol{\phi})$
$\mu_g\|\sigma_g^2$	$\sim$	$N(0,c_\mu\sigma_g^2)$	μ_j	$\sim$	$N(\delta_j,\tau_j)\mathbf{I}_j$
σ_g^{-2}	$\sim$	$Ga(\frac{\delta}{2},\frac{d}{2})$	σ_j^{-2}	$\sim$	$Ga(b_j,l_j)\mathbf{I}_j$

Variable selection parameters:

$$
\begin{array}{rcl}
\pi(r_{gm}|r_{g(m-1)},r_{g(m+1)},\boldsymbol{\xi},\pi_1) & = & \gamma[\pi_1^{r_{gm}}(1-\pi_1)^{(1-r_{gm})}] + \sum_{j=1}^{2}\omega_j I_{\{r_{gm}=r_{g(m+(-1)^j)}\}} \\
\pi_1 & \sim & Beta(e,f)
\end{array}
$$

Fixed Hyperparameters:

$$c_\mu, c_\beta, \delta, d, e, f, \alpha, \delta_j, \tau_j, b_j, l_j, \boldsymbol{\phi}$$

Figure 4.2: Hierarchical formulation of the proposed probabilistic model.

4.2 Posterior inference

Our primary interest lies in the estimation of the association matrix $\mathbf{R}$ and of the matrix $\boldsymbol{\xi}$. For this we design a Markov chain Monte Carlo algorithm. Our marginal likelihood, integrating out $\boldsymbol{\mu}$, $\boldsymbol{\beta_g}$ and σ_g^2, is

$$f(\boldsymbol{Y}_g|\boldsymbol{\xi},\boldsymbol{R}) = \frac{(2\pi)^{-\frac{n}{2}}(\frac{c_\mu}{c_\mu+n})^{\frac{1}{2}}(c_\beta)^{\frac{k_g}{2}}\Gamma(\frac{n+\delta}{2})(\frac{d}{2})^{\frac{\delta}{2}}}{|\boldsymbol{U}_g|^{\frac{1}{2}}\Gamma(\frac{\delta}{2})(\frac{d+q_g}{2})^{(\frac{n+\delta}{2})}}, \tag{4.10}$$

with

$$\boldsymbol{U}_g = c_\beta \mathbf{I}_{k_g} + \boldsymbol{\xi}_R' \boldsymbol{H}_n \boldsymbol{\xi}_R;\ q_g = \boldsymbol{Y}_g' \boldsymbol{H}_n \boldsymbol{Y}_g - \boldsymbol{Y}_g' \boldsymbol{H}_n \boldsymbol{\xi}_R \boldsymbol{U}_g^{-1} \boldsymbol{\xi}_R' \boldsymbol{H}_n \boldsymbol{Y}_g;\ \boldsymbol{H}_n = \mathbf{I}_n - \frac{\mathbf{1}_n \mathbf{1}_n'}{n+c_\mu}$$

and where k_g is the number of significant regressors for the gth regression. See Appendix A for details of this derivation. Our MCMC algorithm consists of the following steps:

- Update $\boldsymbol{R}$ via a Metropolis step. Here we first select n_g genes at random using a geometric distribution (see Appendix B for details). Then, for each selected gene, we either add/delete one element (with probability ρ) or swap two elements of the row of $\boldsymbol{R}$ that corresponds to that particular gene. An A/D step consists in choosing at random an element and changing its value (if it is a 0 it becomes a 1 and vice versa), while the swap step consists in swapping the position of a 1 and a 0 elements. In this step we do not consider genes whose probes have more than 90% of the samples estimated in state 2. This constraint is biologically justified, as CGH probes that are in neutral state do not exhibit chromosomal aberration and would therefore not be associated with changes in mRNA transcript abundance. The change proposed for each gene is then accepted with probability

$$min[\tfrac{f(\mathbf{Y}|\boldsymbol{\xi},\boldsymbol{R}^{new})\pi(\boldsymbol{R}^{new}|\boldsymbol{\xi})}{f(\mathbf{Y}|\boldsymbol{\xi},\boldsymbol{R}^{old})\pi(\boldsymbol{R}^{old}|\boldsymbol{\xi})}, 1]. \tag{4.11}$$

Because All moves are symmetric, the proposal distribution does not appear in the previous ratio. Details on the functional form of $\pi(\boldsymbol{R}|\boldsymbol{\xi})$ are given in Appendix B.

- Update $\boldsymbol{\xi}$ via a Metropolis step. This step consists in choosing at random a column of $\boldsymbol{\xi}$ (say m) and update the value of n_m of its elements, selected at random using a geometric distribution (see Appendix B for details). For each element, a candidate state is sampled using the current transition matrix A (i.e., we propose ξ_{im}^{new} based on $\xi_{i(m-1)}^{old}$) and the proposal is accepted with probability:

$$min[\frac{\pi(\boldsymbol{R}|\boldsymbol{\xi}^{new})f(\boldsymbol{X}|\boldsymbol{\xi}^{new})f(\boldsymbol{Y}|\boldsymbol{\xi}^{new},\boldsymbol{R})\pi(\boldsymbol{\xi}^{new}|\boldsymbol{\xi}^{old},\boldsymbol{A})q(\boldsymbol{\xi}^{old}|\boldsymbol{\xi}^{new})}{\pi(\boldsymbol{R}|\boldsymbol{\xi}^{old})f(\boldsymbol{X}|\boldsymbol{\xi}^{old})f(\boldsymbol{Y}|\boldsymbol{\xi}^{old},\boldsymbol{R})\pi(\boldsymbol{\xi}^{old}|\boldsymbol{\xi}^{old},\boldsymbol{A})q(\boldsymbol{\xi}^{new}|\boldsymbol{\xi}^{old})}, 1]. \qquad (4.12)$$

  See Appendix B for details on how to calculate the various terms of the acceptance probability.

- Update μ_j, for $j = 1, \ldots, 4$, via a Gibbs step. Here we generate

$$\mu_j|\boldsymbol{X}, \boldsymbol{\xi}, \sigma_j \sim N(\eta_j, \theta_j^{-2})\mathbf{I}_j,$$

  with precisions $\theta_j = \tau_j^{-2} + n_j\sigma_j^{-2}$ and weighted means $\eta_j = \theta_j^{-2}(\delta_j\tau_j^{-2} + \bar{X}_j n_j \sigma_j^{-2})$, where $n_j = \sum_{m=1}^{M}\sum_{i=1}^{n}\mathbf{I}_{\{\xi_{im}=j\}}$, $\bar{X}_j = \frac{1}{n_j}\sum_{m=1}^{M}\sum_{i=1}^{n}X_{im}\mathbf{I}_{\{\xi_{im}=j\}}$ and $\mathbf{I}_j$ indicates the truncation.

- Update σ_j, for $j = 1, \ldots, 4$, via a Gibbs step. We generate

$$\sigma_j|\boldsymbol{X}, \boldsymbol{\xi}, \mu_j \sim IG(\alpha_j + \frac{n_j}{2}, \beta_j + \frac{V_j}{2})\mathbf{I}_j,$$

  where $n_j = \sum_{m=1}^{M}\sum_{i=1}^{n}\mathbf{I}_{\{\xi_{im}=j\}}$, $V_j = (X_{im} - \mu_j)^2\mathbf{I}_{\{\xi_{im}=j\}}$ and $\mathbf{I}_j$ indicates the truncation.

- Update $\boldsymbol{A}$ via a Metropolis step. We generate a new value for each row of $\boldsymbol{A}$ as $\boldsymbol{A}_{.j}^{new}|\boldsymbol{\xi} \sim Dir(\phi + o_{h1}, \phi + o_{h2}, \phi + o_{h3}, \phi + o_{h4})$, where

$o_{hk} = \sum_{i=1}^{n} \sum_{m=1}^{M-1} \mathbf{I}_{\{\xi_{im}=h, \xi_{i(m+1)}=k\}}$, and accept the proposed value with probability

$$min[1, \prod_{i=1}^{n} \frac{\pi_{A^{new}}(\xi_{i1})}{\pi_{A^{old}}(\xi_{i1})}], \tag{4.13}$$

where π_A denotes the stationary distribution of the transition matrix $\boldsymbol{A}$.

Given the MCMC output, we first perform inference on $\mathbf{R}$ by estimating the marginal posterior probability of inclusion for the single elements, counting how many times each position was set equal to one, after burn-in. The final selection is then made by looking at those elements of $\mathbf{R}$ that have marginal posterior probability greater than a certain threshold. We then estimate $\boldsymbol{\xi}$ by calculating, for each position, the most frequent state value. Our MCMC output also allows us to do inference on the transition matrix $\mathbf{A}$ and the means and variance components.

4.3 Simulation Studies

We generate an $n \times M$ matrix $\boldsymbol{\xi}$ as follows:

- We set all the elements of the matrix $\boldsymbol{\xi}$ equal to 2.
- We select[1] $L < M$ columns and, for each column, we randomly generate all its values using a transition matrix of the form

$$\boldsymbol{A} = \begin{pmatrix} 0.7500 & 0.1800 & 0.0500 & 0.020 \\ 0.4955 & 0.0020 & 0.4955 & 0.007 \\ 0.0200 & 0.1800 & 0.7000 & 0.010 \\ 0.0001 & 0.3028 & 0.1000 & 0.597 \end{pmatrix}$$

[1] We select some groups of adjacent columns and some random columns.

- We then randomly select additional $(M - L)/2$ columns and, for each column, we randomly generate 10% of its values according to the same transition matrix above.

We generate the copy number state of the first probe by sampling from the initial probability vector $\pi_{\boldsymbol{A}}$, which we obtain, following Guha et al. [2008], as the unique stationary distribution of the transition matrix A, defined as the normalized left eigenvector of the matrix associated with eigenvalue 1. We fix the values of μ_j and σ_j, $j = 1, \ldots, 4$, to $\mu_1 = -0.65, \mu_2 = 0, \mu_3 = 0.65, \mu_4 = 1.5$ and $\sigma_1 = 0.1, \sigma_2 = 0.1, \sigma_3 = 0.1, \sigma_4 = 0.2$, and then generate the data matrix $\boldsymbol{X}$ of the CGH profiles by sampling each CGH probe from a Normal distribution with mean and variance corresponding to the state it belongs to. Next we select l significant β's among the L DNA segments and generate them as $\beta \sim N(\beta_0, 0.3^2)$, picking the sign at random. We generate the error term as $\epsilon \sim N(0, \sigma_\epsilon^2)$ and the intercept as $\mu_g \sim N(0, \sigma_{\mu_g}^2)$, with $\sigma_{\mu_g} = 0.1$, and, finally, set $Y_{ig} = \mu_g + \beta X_{ig} + \epsilon$, with $g = 1, \ldots, G$. Figure 4.3 depicts the simulated data for $G = 100, M = 1,000, L = 250, l = 20$ and $\sigma_\epsilon^2 = .01$, for one sample.

For hyperparameter settings, a vague prior is assigned to the intercept parameters by setting c_μ to a large value tending to ∞. The hyperparameter c_β in the prior on the regression coefficients determines, together with the hyperparameters of the prior on R, the amount of shrinkage in the model. We follow the guidelines provided by Sha and al. [2004] and specify this parameter in the range of variability of the data so as to control the ratio of prior to posterior precision. Specifically, we set $c_\mu = 10^6$ and $c_\beta = 10$. Also, we specify vague priors on the error variances by setting $\delta = 3$, the minimum value such that the prior expectation exists, and choosing d so that the expected value of the variance parameter is comparable in size to a small percentage of the expected error variances of the standardized responses (we chose 5% for the results reported in the paper). Some sensitivity to the hyperparameters of the prior on $\boldsymbol{R}$, and specifically the Beta hyperprior on π_1, is to be expected. In the simulations reported below we consider the

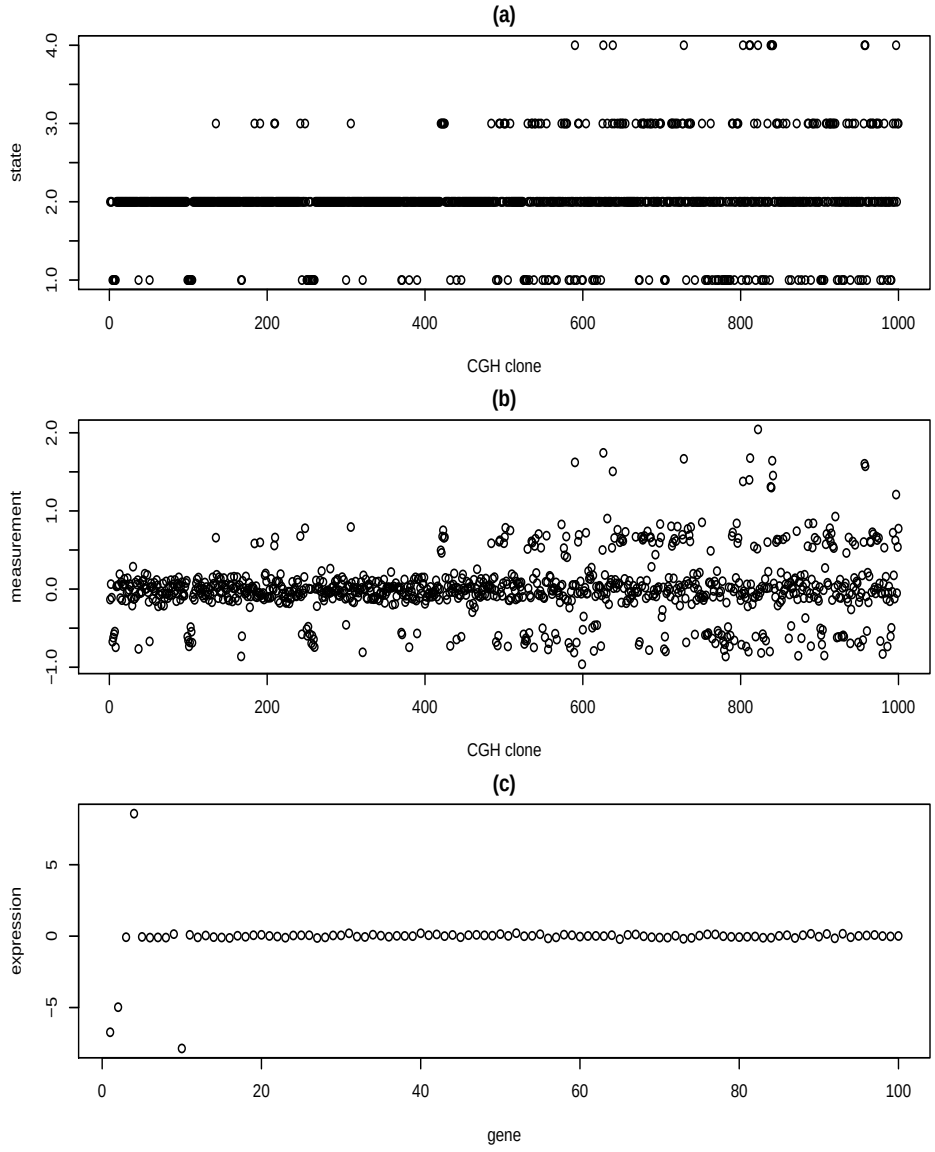


Figure 4.3: Simulated data: Example of simulated $\boldsymbol{\xi}_i, \mathbf{X}_i$ and $\mathbf{Y}_i$, from top to bottom, respectively, for $G = 100, M = 1,000, L = 250, l = 20$ and $\sigma_\epsilon^2 = .01$, for one sample ($n = 1$).

dependent prior model (4.8), where we set $\pi_1 \sim Beta(e, f)$ with $e = .001$ and $f = .999$, and look at performances for different values of α. We also consider the simpler independent prior, corresponding to $\alpha \to \infty$. As for the prior settings of the HMM model, we again follow Guha et al. [2008] and use a very similar hyperparameter specification. Specifically, we set $\mu_j \sim N(\delta_k, \tau_k^2) \cdot \mathbf{I}_{\{low_\mu < \mu < upp_\mu\}}$, for $j = 1, \ldots, 4$, with $\delta_k = [-1, 0, 0.58, 1]$, $\tau_k = [1, 1, 1, 2]$, $low_\mu = [-\infty, -0.1, 0.1, \mu_3 + \sigma_3]$ and $upp_\mu = [-0.1, 0.1, 0.73, \infty]$. Also, we use $\sigma_j^{-2} \sim Ga(b_j, l_j) \cdot \mathbf{I}_{\{\sigma < upp_{\sigma_j}\}}$ with $b_j = 1$, $l_j = 1$ and $upp_\sigma = [.41, .41, .41, 1]$. Finally, we assume each row of the transition matrix as independently distributed according to a Dirichlet $D(\phi_{i1}, \phi_{i2}, \phi_{i3}, \phi_{i4})$ with $\phi = [\mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1}]$.

When running the MCMC chain, we sample the initial values for μ_j and σ_j from their respective priors, while $\boldsymbol{\xi}$ is initialized by fixing three thresholds and considering $\xi_{im} = j$ $(j = 1, \ldots, 4)$ if $X_{im} > T_j$ with $\boldsymbol{T} = [-\infty, -0.5, 0.29, 0.79]$. The initial value of $\boldsymbol{A}$ is derived from $\boldsymbol{\xi}$ by counting the number of transitions at each position and dividing by the row total. We set the initial $\boldsymbol{R}$ as a matrix with all positions equal to zero. We set the probability of an A/D move to $\rho = 0.5$. All results we report here were obtained by running MCMC chains with 500,000 iterations and a burn-in of 350,000. We assessed convergence by inspecting the MCMC sample traces for all parameters, see Figure 4.4 for an example.

4.3.1 Inference on $\boldsymbol{R}$ and $\boldsymbol{\xi}$

In a first scenario (*simulated scenario 1*), we set $G = 100$ and $(n; M; L; l) = (100; 1,000; 250; 20)$. We then generated the $l = 20$ non-zero β's from a $N(2, 0.3^2)$, except for six of them which we sampled from a $N(.5, 0.3^2)$. We repeated the simulation for two different values of the variance of the error term, that is, $\sigma_\epsilon^2 = .01, .25$. We start by summarising the inference on the association matrix $\mathbf{R}$. We investigate the effect of the dependent prior (4.6) by running chains for different values of α, that is, $\alpha = (5, 10, 50, 100, \infty)$. Figure 4.5 shows marginal posterior probabilities of inclusion for the ele-

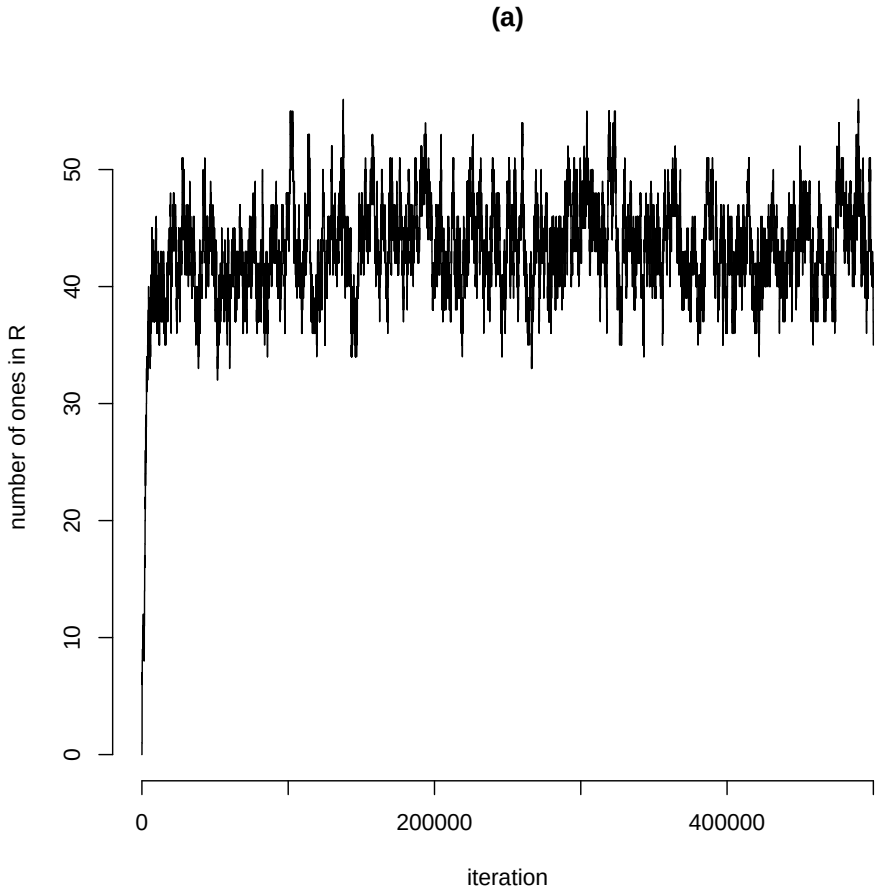


Figure 4.4: Simulated data: Example of trace plots for $\boldsymbol{R}$ for one MCMC run on simulated scenario 1.

ments r_{gm} of the matrix $\boldsymbol{R}$, for the case $\sigma^2_\epsilon = .01$ and the five α values. The figure show that, as α decrease, for some group of links their posterior probabilities increase, while for some other they decrease. Table 4.1 shows the results in terms of false positives (FP), false negatives (FN), sensitivity and specificity, obtained with a threshold of .5 on the marginal posterior probability of inclusion shown in Figure 4.5. With the term false positives we indicate the number of positions estimated as significant which are not significant in the true $\boldsymbol{R}$ matrix. With false negatives we indicate the number of positions estimated as not significant which are significant in the true $\boldsymbol{R}$ matrix. We calculate sensitivity as the number of false positives divided by (l) and specificity as the number of true negatives divided by $(G \times M - l)$. We notice that lower values of α, enforcing more dependence among CGH probes, lead to lower numbers of FN but increased numbers of FP, leading to improved sensitivity and only slightly worse specificity. The small variability in specificity is due to the fact that the number of true negatives (TN) is always a very large value. Results are similar for the two different error variance values we considered, although, as expected, the model works better when the error variance is smaller (see Table 4.1).

In order to investigate the effect of the choice of the threshold on the marginal probabilities of inclusion, Figure 4.6 shows numbers of FP and FN obtained by considering different threshold values, calculated as a grid of equispaced points in the range $[.07, 1]$. The 4 panels clearly show that the dependent prior outperforms the independent one regardless of the choice of the threshold value.

As for inference on $\boldsymbol{\xi}$, Table 4.2 shows the numbers of misclassified elements of $\boldsymbol{\xi}$ for $\alpha = (5, 10, 50, 100, \infty)$. It seems that there is no effect on the choice of α on the misclassification rate, i.e. a better estimation in the variable selection do not lead to a better estimation in the classification of the X. Our inferential scheme also allows us to look at the distribution of the misclassification values of ξ over the four possible states. For example, for $\sigma^2_\epsilon = .01$ and $\alpha = (5, \infty)$ we obtained

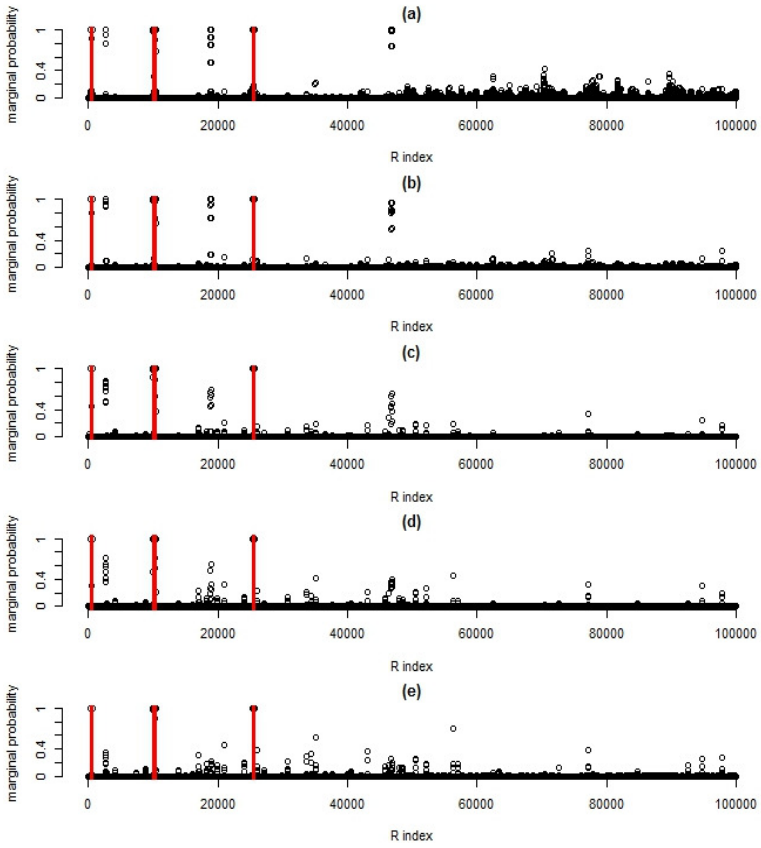


Figure 4.5: Simulated scenario 1 with $\sigma_\epsilon^2 = .01$: Marginal posterior probability of inclusion of the elements r_{gm} of the association matrix $\boldsymbol{R}$. Plots refer to prior model (4.6) with (a) $\alpha = 5$, (b) $\alpha = 10$, (c) $\alpha = 50$, (d) $\alpha = 100$ and (e) $\alpha = \infty$ (independent prior).

Scenario 1 $\sigma^2_\epsilon = 0.01$	$\alpha = 5$	$\alpha = 10$	$\alpha = 50$	$\alpha = 100$	$\alpha = \infty$ (indep prior)
False Positives	24	22	14	7	2
False Negatives	1	1	3	3	5
Sensitivity	0.95	0.95	0.85	0.85	0.75
Specificity	0.99975	0.99978	0.99985	0.99992	0.99997
Scenario 1 $\sigma^2_\epsilon = 0.25$	$\alpha = 5$	$\alpha = 10$	$\alpha = 50$	$\alpha = 100$	$\alpha = \infty$ (indep prior)
False Positives	32	24	8	1	0
False Negatives	2	3	5	5	6
Sensitivity	0.9	0.85	0.75	0.75	0.7
Specificity	0.99968	0.99976	0.99992	0.99999	1
Scenario 2 $\sigma^2_\epsilon = 0.01$	$\alpha = 5$	$\alpha = 10$	$\alpha = 50$	$\alpha = 100$	$\alpha = \infty$ (indep prior)
False Positive	11	10	6	5	3
False Negative	2	2	5	6	12
Sensitivity	0.9	0.9	0.75	0.7	0.4
Specificity	0.99983	0.99990	0.99994	0.99995	0.99997
Scenario 2 $\sigma^2_\epsilon = 0.25$	$\alpha = 5$	$\alpha = 10$	$\alpha = 50$	$\alpha = 100$	$\alpha = \infty$ (indep prior)
False Positive	25	8	3	5	2
False Negative	6	5	8	10	14
Sensitivity	0.7	0.75	0.6	0.5	0.3
Specificity	0.99975	0.99992	0.99997	0.99995	0.99998

Table 4.1: Simulated scenarios 1 and 2: Results on false positives, false negatives, sensitivity and specificity for the dependent prior model (4.6) and the independent case ($\alpha = \infty$) obtained with a threshold of 0.5 on the marginal posterior probability of inclusion on $\boldsymbol{R}$.

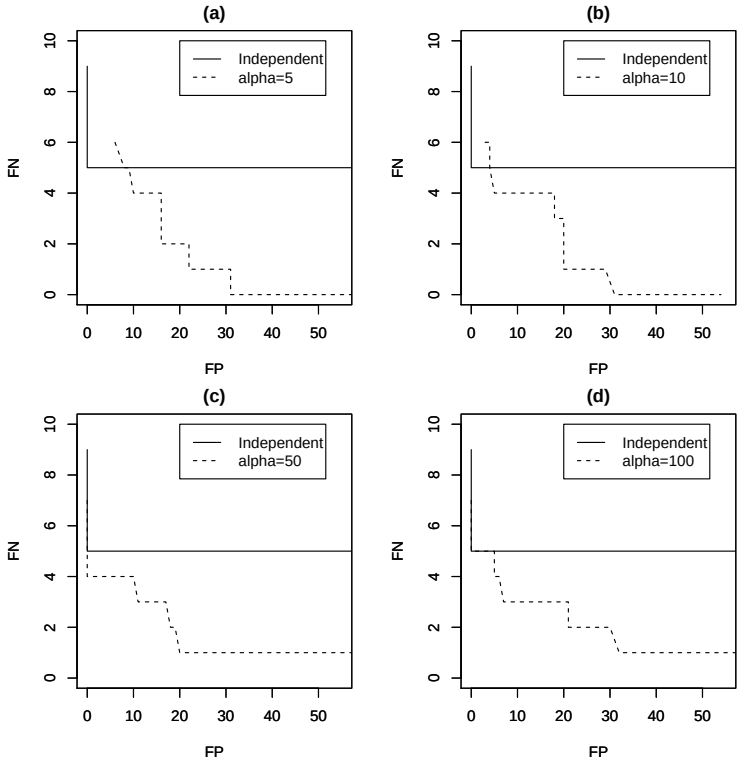


Figure 4.6: Simulated scenario 1 with $\sigma^2_\epsilon = .01$: Numbers of FP and FN obtained by considering different thresholds on the marginal probabilities of inclusion of Figure 4.5. Threshold values are calculate as a grid of equispaced points in the range $[.07, 1]$. Plots refer to prior model (4.6) with different values of α.

# missclassified (percent)	$\alpha = 5$	$\alpha = 10$	$\alpha = 50$	$\alpha = 100$	$\alpha = \infty$ (indep prior)
Scenario 1 $\sigma^2_\epsilon = 0.01$	79 (0.079%)	75 (0.075%)	72 (0.072%)	75 (0.075%)	70 (0.07%)
Scenario 1 $\sigma^2_\epsilon = 0.25$	75 (0.075%)	70 (0.07%)	86 (0.086%)	70 (0.07%)	76 (0.076%)
Scenario 2 $\sigma^2_\epsilon = 0.01$	62 (0.062%)	68 (0.068%)	59 (0.059%)	70 (0.07%)	66 (0.066%)
Scenario 2 $\sigma^2_\epsilon = 0.25$	62 (0.062%)	71 (0.071%)	66 (0.066%)	74 (0.074%)	72 (0.072%)

Table 4.2: Simulated scenarios 1 and 2: Results on $\boldsymbol{\xi}$ as number of misclassified elements, for the dependent prior model (4.6) and the independent case ($\alpha = \infty$).

$\alpha = 5$

	1	2	3	4
1	12606	19	0	0
2	19	73451	13	0
3	0	12	12663	6
4	0	0	10	1201

$\alpha = \infty$

	1	2	3	4
1	12612	17	0	0
2	13	73454	15	0
3	0	11	12661	4
4	0	0	10	1203

with columns referring to the true states and rows to the estimated ones. Note the large diagonal elements, representing the number of correct classifications, and the relative small numbers of misclassified samples. Also, all misclassified cases occur between adjacent classes (elements at the right or left of the diagonal). Similar results were obtained in all the other cases.

In a second scenario (*simulated scenario 2*) we induced dependence among regression coefficients by selecting two clusters of adjacent β's and generating them as $\beta \sim N(.5, 0.3^2)$. We again set $\sigma^2_\epsilon = .01, .25$. Tables 4.1 shows the results of the inference on $\boldsymbol{R}$. As expected, the dependent prior works better in terms of FP than in the previous simulated scenario, and has similar results in terms of FN, even if the β's used to generate the data are in general smaller. The independent prior instead shows worse performances

than in the previous simulated scenario, particularly in terms of FN. Figure 4.8 shows the numbers of FP and FN obtained by considering different threshold values, calculated on the same grid as for Figure 4.6. Again, results show how the dependent prior outperforms the independent one. With respect to Figure 4.6 we notice that results for the dependent priors show numbers of FN that decrease more rapidly for larger threshold values.

The number of misclassification in terms of ξ that can be red on Table 4.2, shows an error rate of about 0.07%, that is a very good result. The errors are in the range $[62, 86]$ and seem to be randomly distributed. Looking at the distribution of the misclassification values of ξ over the four possible states, for $\sigma^2_\epsilon = .01$ and $\alpha = (5, \infty)$ we obtained

$\alpha = 5$

	1	2	3	4
1	12574	7	0	0
2	18	73623	12	0
3	0	6	12561	6
4	0	0	13	1207

$\alpha = \infty$

	1	2	3	4
1	12544	6	0	0
2	21	73619	9	0
3	0	11	12557	5
4	0	0	20	1208

We also looked at the results when increasing the size of the simulated data. Performances of our model and priors were very consistent with those we have reported above. For example, for $(n; M; L; l) = (200; 2,000; 500; 50)$ and by generating the β's independently, similarly to scenario 1, with $\sigma^2_\epsilon = .01$, the dependent prior with $\alpha = 100$ led to $(FP; FN; Sensitivity; Specificity) = (0; 5; 0.9, 1)$, for inference on $\boldsymbol{R}$. The model also incorrectly classified 331 elements of the $\boldsymbol{\xi}$ matrix, corresponding to a 0.08275% misclassification rate.

4.3.2 Inference on HMM parameters

Our MCMC algorithm also allows inference on the parameters of the HMM, that is the transition matrix $\mathbf{A}$ and the mean and variance parameters of model (4.3). For example, Table 4.3 shows results for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ for scenario 1, with the smaller error variance value and using the independent prior. The estimated values are all very close to the true ones, with the exception

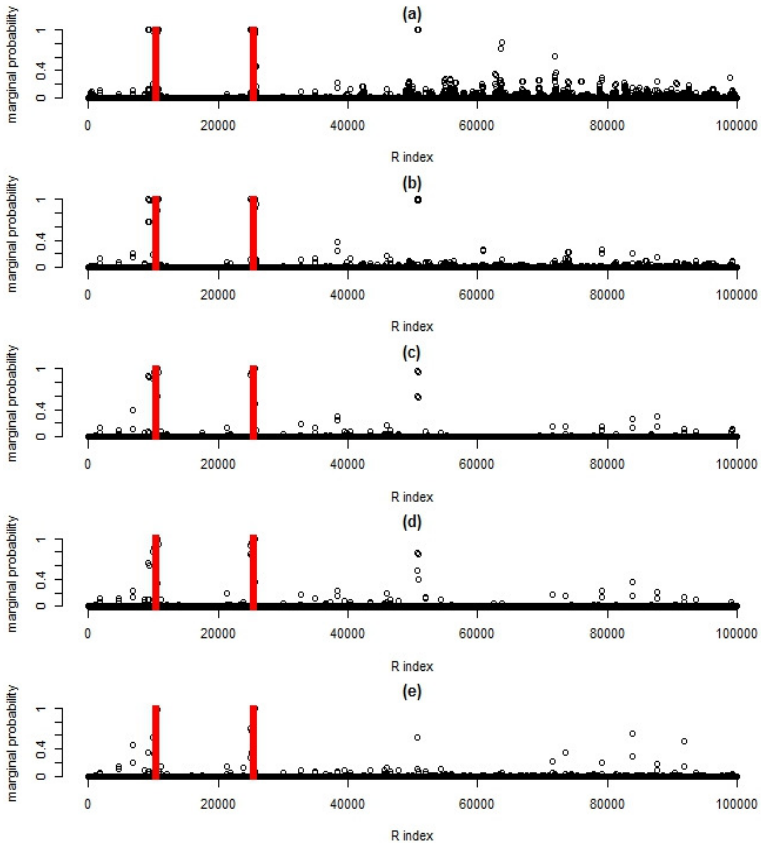


Figure 4.7: Simulated scenario 2 with $\sigma_\epsilon^2 = .01$: Marginal posterior probability of inclusion of the elements r_{gm} of the association matrix $\boldsymbol{R}$. Plots refer to prior model (4.6) with (a) $\alpha = 5$, (b) $\alpha = 10$, (c) $\alpha = 50$, (d) $\alpha = 100$ and (e) $\alpha = \infty$ (independent prior).

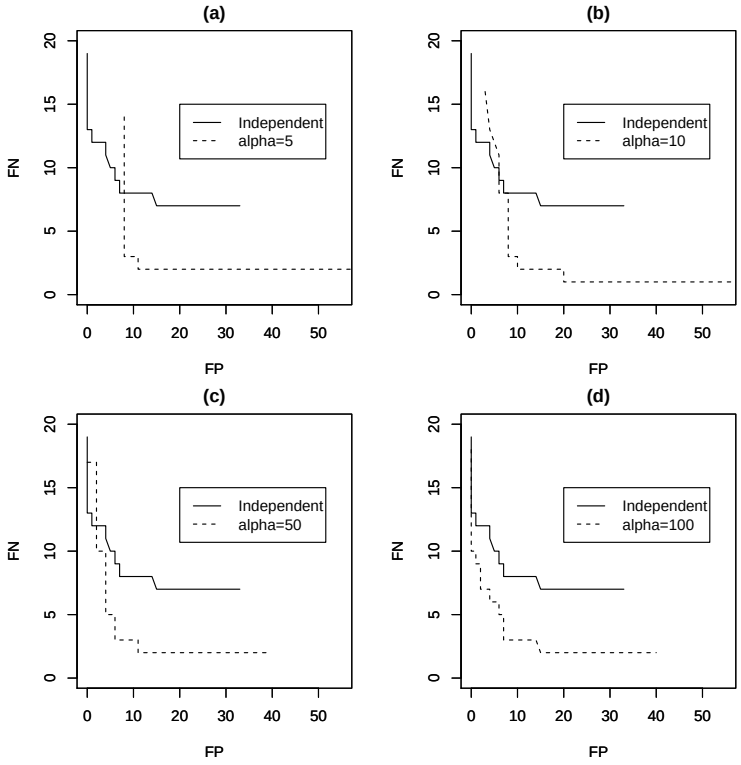


Figure 4.8: Simulated scenario 2 with $\sigma^2_\epsilon = .01$: Numbers of FP and FN obtained by considering different thresholds on the marginal probabilities of inclusion on $\boldsymbol{R}$. Threshold values are calculate as a grid of equispaced points in the range $[.07, 1]$. Plots refer to prior model (4.6) with different values of α.

	μ_1	μ_2	μ_3	μ_4
True	-0.64974	0.00039	0.64955	1.51394
Estimated	-0.64968	0.00041	0.64932	1.50810
	σ_1	σ_2	σ_3	σ_4
True	0.10113	0.09979	0.10049	0.19928
Estimated	0.10199	0.09993	0.10083	0.21072

Table 4.3: Simulated scenario 1 with $\sigma_\epsilon^2 = .01$: Results on the estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ with the independent prior ($\alpha = \infty$).

of σ_4 which shows a slight overestimation. We obtain similar results in all simulations we considered.

As for the estimation of the transition matrix, we obtained good for the elements that corresponds to high probabilities, but are imprecise for very little values.

$$\begin{bmatrix} 0.3513 & 0.6183 & 0.0216 & 0.0088 \\ 0.1101 & 0.7762 & 0.1118 & 0.0019 \\ 0.0089 & 0.6209 & 0.3267 & 0.0436 \\ 0 & 0.6000 & 0.0586 & 0.3414 \end{bmatrix} \begin{bmatrix} 0.3269 & 0.6412 & 0.0217 & 0.0102 \\ 0.1026 & 0.7826 & 0.1037 & 0.0110 \\ 0.0079 & 0.6204 & 0.3015 & 0.0702 \\ 0.0009 & 0.6015 & 0.0602 & 0.3374 \end{bmatrix}$$

4.3.3 Sensitivity analysis

When looking at the sensitivity of the results to our prior choices we focused in particular on c_β, e, f, α). We follow Guha et al. [2008] as a guideline to choose the truncation on the means and variances of the hidden Markov model, and for the hyperparameter settings on the transition matrix. They performed sensitivity analysis on the truncation values around the mean of the second state and find out that the false discovery rate in terms of $\boldsymbol{\xi}$ is robust to the choice of this values in the interval $[0.05, 0.15]$. Same results for the choice of parameters $\boldsymbol{\phi}$ on the Dirichlet when they are very small compared to the sample size. On the third simulation we look what happens using different values for the hyperparameters $(e;f)$=(.01;.99) using

$(e;f)$	(.01;.99)		(.001;.999)	
α	10	100	10	100
False Positive	9	4	10	5
False Negative	2	6	2	6
Sensitivity	0.9	0.7	0.90	0.7
Specificity	0.99991	0.99996	0.99990	0.99995

Table 4.4: Sensitivity (e,f)

the dependent prior with two different values of $\alpha = 10, 100$. Table 4.4 shows the results in terms of $\boldsymbol{R}$ comparing them with the results obtained using the same settings of α and with $(e;f)$=(.001;.999).

4.4 Real data analysis

4.4.1 NCI-60 Data

The model we described is specific for copy number aberrations data on the X, while on the Y any kind of data that could be affected by variation at DNA level can be used, in principle. In this section we consider gene expression data. Our data arises from a well known public database: Cellminer[2]. This web application facilitates systems biology through the retrieval and integration of the molecular and pharmacological data sets for the NCI-60 cell lines. On the website there is the chance to download data at DNA, RNA or Protein levels, as well as view meta data on the cell lines, download drug data, get a list of mutations found in 24 known, important human cancer genes and access to several done analysis.
The NCI-60 is a set of 60 human cancer cell lines derived from diverse tissues: brain, blood and bone marrow, breast, colon, kidney, lung, ovary, prostate and skin.
There are many different data available for DNA, RNA and proteins, both

[2]http://discover.nci.nih.gov/cellminer/home.do

raw and normalized data set. We consider the normalized data sets, focusing on aCGH Agilent 44K, DNA level, and Affy HG-U133(A), RNA level, using the RMA normalization method (available only for RNA). Both data sets are made by more than 44,000 variables, thus a reduction in the dimensionality is needed. To that aim we first reduce the number of samples by one because sample 40 had all missing data at DNA level. Then we impute the remaining missing data using the k-nearest-neighbours, setting $k = 5$. Finally for RNA data we perform an ANOVA and we select those probes with corresponding adjusted p-values lower than 0.01. For DNA data we must first select genes that belong all to the same chromosome (we choose chromosome 8) and then perform an ANOVA and select those probes with adjusted p-values lower than 0.2. In this way the final data set for RNA is made by 59 samples and 3296 probes, while the one for DNA is made by 59 samples and 89 probes.

4.4.2 Parameter settings

Before running the code we need to set all the hyperparameters of the model. Parameter setting is very similar to that used in the simulation analysis, except for some parameters. We set $c_\mu = 10^6$ and $c_\beta = 10$, following the same guidelines we used in the simulations. Again we follow the same setting for the error variances, specifically we set $\delta = 3$, and choosing d so that the expected value of the variance parameter is comparable in size to a small percentage (5%) of the expected error variances of the standardized responses. In the variable selection framework we set the two hyperparameters of the Beta hyperprior on the parameter of the Bernoulli in such a way that the expected number of included links is 1% of the total. Specifically we set $e = 0.01$ and $f = 0.99$. For α[3] we will show results for two different values, $\alpha = 25$ and ∞. As prior settings of the HMM, we follow again what Guha et al. [2008] did. Specifically, we set $\mu_j \sim N(\delta_k, \tau_k^2) \cdot \mathbf{I}_{\{low_\mu < \mu < upp_\mu\}}$, for $j = 1, \ldots, 4$, with

[3]See section 4.4.4.

$\delta_k = [-1, 0, 0.58, 1]$, $\tau_k = [1, 1, 1, 2]$, $low_\mu = [-\infty, -0.1, 0.1, \mu_3 + \sigma_3]$ and $upp_\mu = [-0.1, 0.1, 0.58, \infty]$. Also, we use $\sigma_j^{-2} \sim Ga(b_j, l_j) \cdot \mathbf{I}_{\{\sigma < upp_{\sigma_j}\}}$ with $b_j = 1$, $l_j = 1$ and $upp_\sigma = [.41, .41, .41, 1]$. Finally, we assume each row of the transition matrix as independently distributed according to a Dirichlet $D(\phi_{i1}, \phi_{i2}, \phi_{i3}, \phi_{i4})$ with $\phi = [\mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1}]$. We also set the probability of an A/D step $\rho = 0.5$.
We then run 100,000 iterations with burn in of 50,000, and set the probability of the geometric on the number of genes to be updated at each iteration on the $\boldsymbol{R}$ matrix $\pi_R = 0.1$, and the one on the number of samples to be updated at each iteration on the $\boldsymbol{\xi}$ matrix $\pi_\xi = 0.3$.

4.4.3 Results

In this section we show results both in terms of variable selection, that is our primary aim, and in terms of CGH data classification.

Variable selection results

Before looking at the results, we assess convergence by inspecting the MCMC sample traces for all parameters, see Figure 4.9 for an example. As common practice for the posterior inference on the coefficients of the model, we perform it on the marginal posterior probability of each of them, and not on the posterior probability of the entire model. This is due to the fact that because of the huge number of potential coefficients the weight of just one of them toward the probability of the entire model could be very small.

Figure 4.10 shows the posterior probability values for all the possible links. This figure shows a very huge number of links that have very little values, and just for some of them the posterior probability rise up. Since looking at this plots do not give a clear suggestion on the threshold to be chosen, we decided to set the threshold to 0.07. The rationale behind this choice is that with this threshold we have about 100 genes for which at least one CGH probe is estimated as significant.

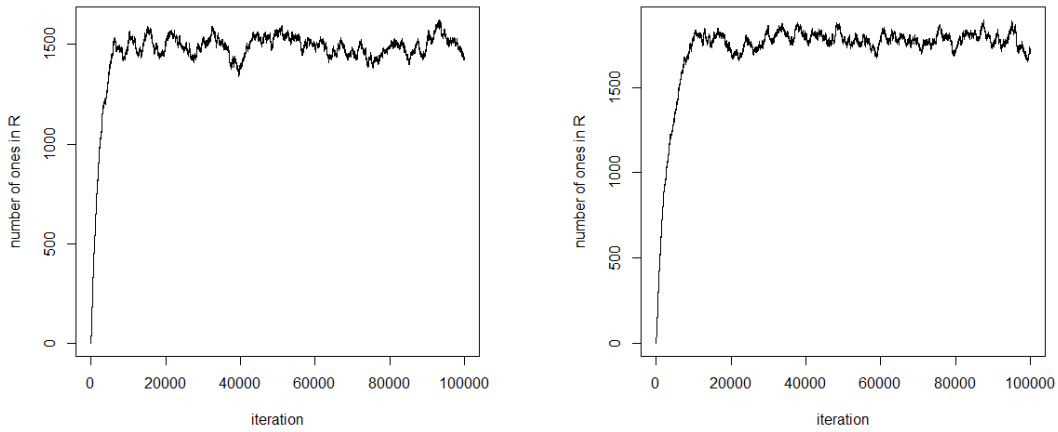


Figure 4.9: Traceplot of the number of included links at each iteration using $\alpha \to \infty$ (left) and $\alpha = 25$ (right).

We use the same threshold to obtain the two Heatmaps in figures 4.11 and 4.12. Specifically we select those genes that have at least one CGH probe selected for at least one of the two values of α. Both figures shows the detection of the so called hotspots: if a CGH probe is significant for one Affymetrix probe it is expected to be significant for some others. At the same time heatmap for $\alpha \to \infty$ shows a tendency of including groups of adjacent CGH probes to be significant for the same Affymetrix probe. This tendency is enforced when using $\alpha = 25$ and is coherent with how we build up our probability on each element of the inclusion matrix.

Choosing the four Affymetrix probes with the higher number of related CGH probes we obtain Figure 4.13. In this figure green ellipses corresponds to non codifying regions, thus we do not have the names of the corresponding genes.

CGH classification results

Looking at the results on the right part of our model, the estimates of the state specific means and variances are respectively $[-0.6373, -0.0057, 0.5003, 1.0291]$

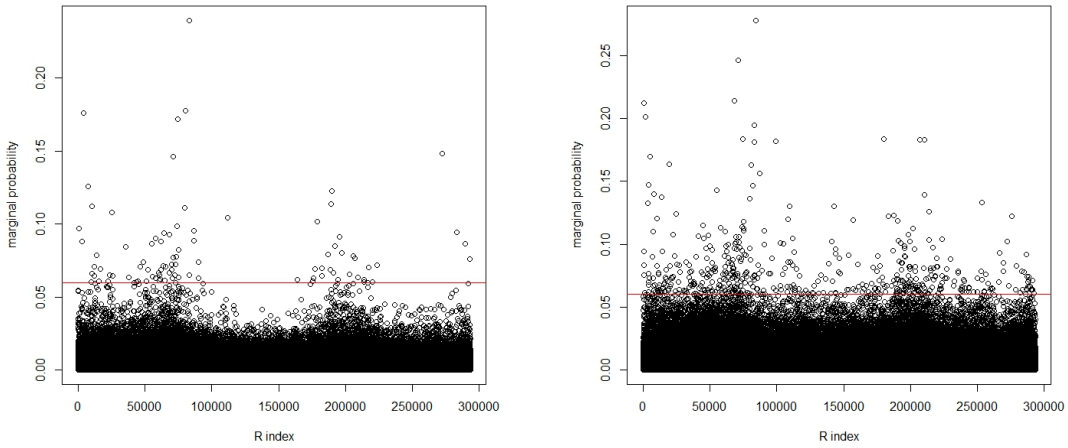


Figure 4.10: Posterior marginal probabilities using $\alpha \to \infty$ (left) and $\alpha = 25$ (right). Red line on probability value 0.06.

and $[0.2092, 0.0897, 0.1227, 0.2683]$. The estimated transition matrix is

$$\begin{pmatrix} 0.9654 & 0.0277 & 0.0057 & 0.0012 \\ 0.0029 & 0.9878 & 0.0077 & 0.0016 \\ 0.0171 & 0.0079 & 0.0257 & 0.9661 \end{pmatrix}$$

Figure 4.14 shows estimated gain ($\boldsymbol{\xi} > 2$) and loss ($\boldsymbol{\xi} = 1$) frequencies along samples for each of the considered CGH probes. Our result is very similar to that obtained using Guha et al. [2008] method. Anyway our method seem to estimate a smaller number of alterations. We believe that this difference is due to the fact that the transition matrix of our model is estimated using the information of all the sample simultaneously, while Guha et al. [2008] method consider only one sample at a time.

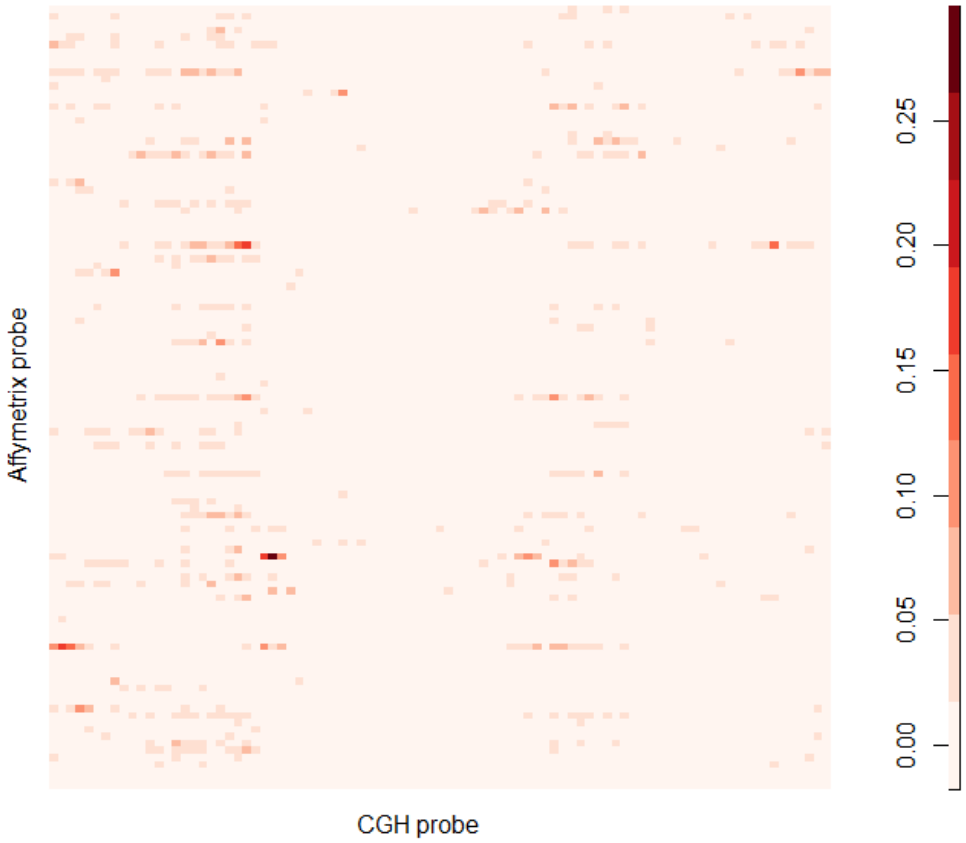


Figure 4.11: Heatmap for $\alpha \to \infty$.

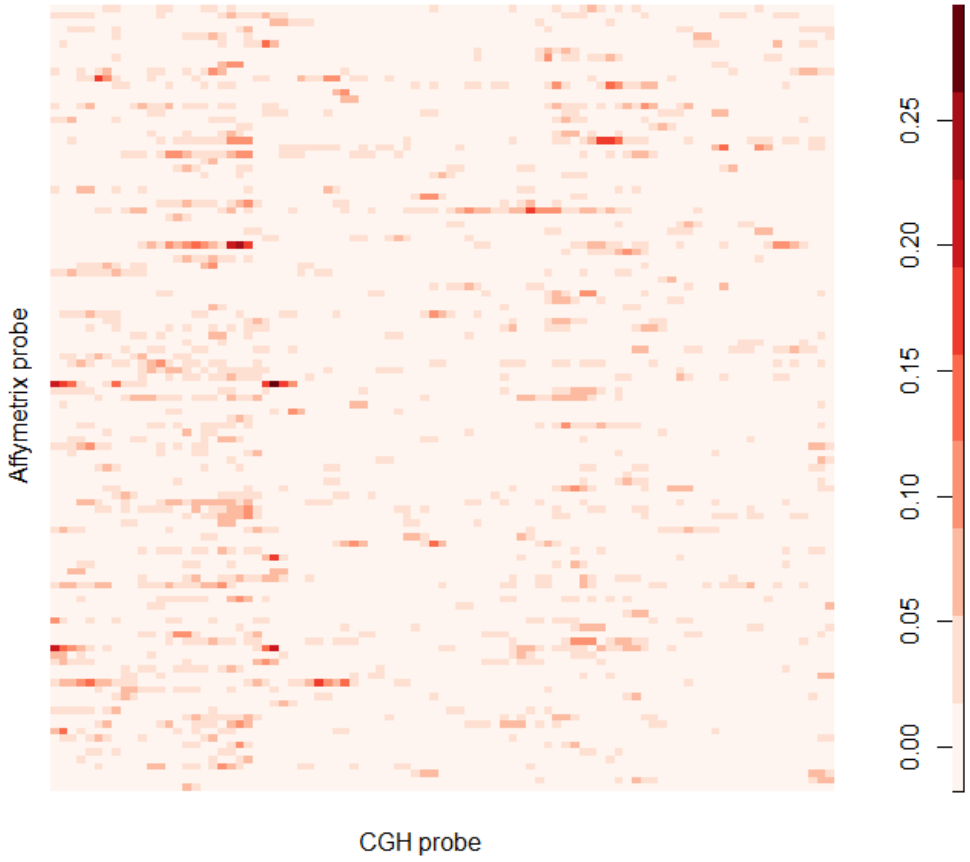


Figure 4.12: Heatmap for $\alpha = 25$.

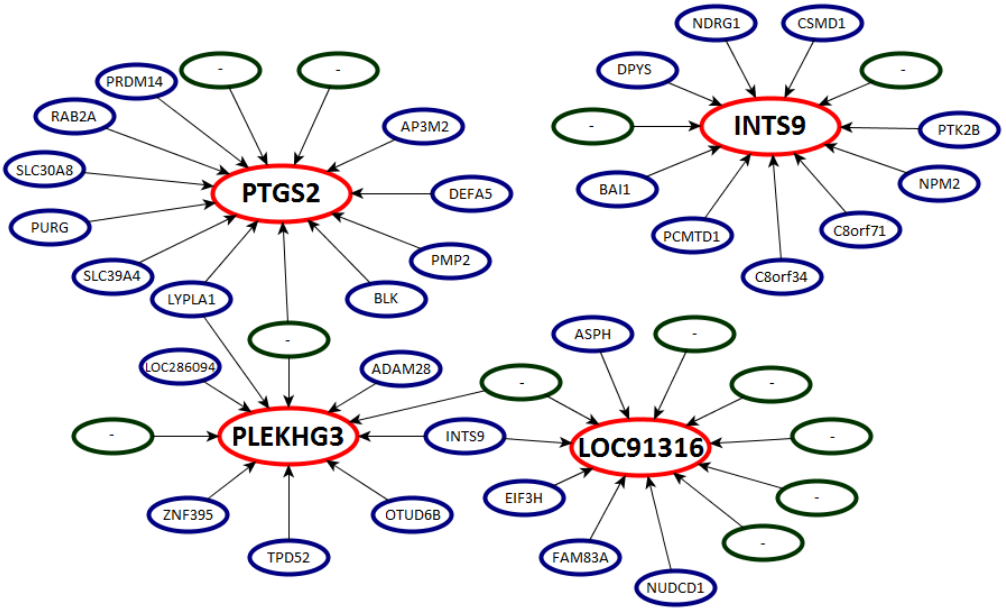


Figure 4.13: Selected links for four Affymetrix genes using a threshold of 0.07.

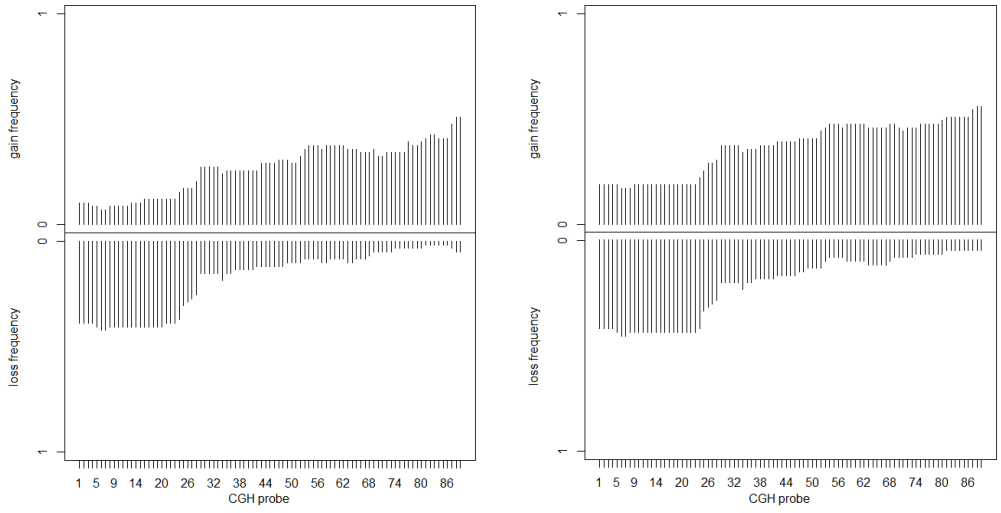


Figure 4.14: Gain/loss estimated frequencies along samples for the 89 CGH probes considered, our method (left) Guha's method (right).

4.4.4 Choice of α

Using different values of α could lead to very different results. For example, let's fix $s_{(m-1)m} = s_{m(m+1)} = 0.65$[4]. Fix also the probability of the Bernoulli prior $\pi_1 = 0.001$. In this scenario a value of $\alpha = 5$ leads to the triplet $(\gamma, \omega_1, \omega_2) = (0.7936, 0.1032, 0.1032)$. With this setting consider the probability of a certain link to be included if the two adjacent coefficients[5] are both estimated as significant: this probability goes from 0.001 (independent prior) to 0.2072. At the same time consider the probability of not being included if both the adjacent links are estimated as not significant: it goes from 0.999 (independent) to 0.9992064. Thus we have at the same time a big increase in the probability of inclusion and small decrease in the probability of not inclusion. This could lead to a scenario where the model starts including coefficients and the number of estimated significant links become unlikely huge.
Figure 4.15 shows this effect for a grid of 100 values in the range $[1, 100]$, considering two different values on the probability of inclusion of the Bernoulli, respectively $\pi_1 = 0.001$ and $\pi_1 = 0.1$. The impact on the probability of inclusion is particularly strong when the probability of inclusion on the Bernoulli prior is low. Compare the two sub figures of 4.15 shows, for instance, that the effect described above is less strong when $\pi_1 = 0.1$, but still remains.

4.5 Discussion

We developed a hierarchical Bayesian model to discover sets of CGH probes that could affect gene expression. We first model CGH measurement and group them into four possible categories. Then we use this latent variable to measure similarities between one segment and the following. Therefore using this similarity measure jointly with distance between CGH segment,

[4]Consider that this quantity could assume values in the range $[0, 1]$.
[5]With adjacent coefficient we mean the coefficient between the same gene expression and adjacent CGH probes.

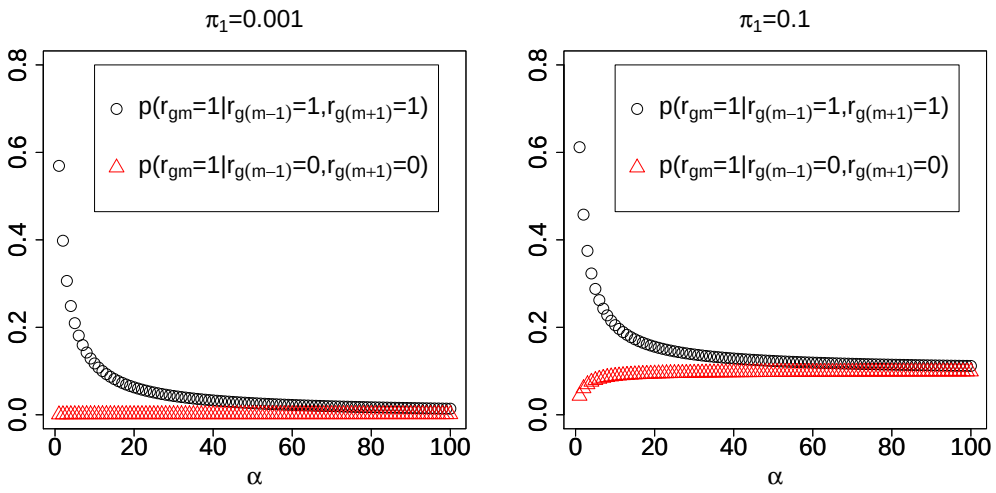


Figure 4.15: Effect of different values of α on the probabilities of inclusion for different values of π_1.

we model the probability of inclusion of each link, enforcing the detection of adjacent groups of covariates that affect the same response. We believe that our model is supported by the results obtained.

Simulation results shown a better performance of our prior on the variable selection model, with respect to a simple Bernoulli prior. Our model outperform the Bernoulli prior in any scenario, and in particular when assuming groups of adjacent CGH probes that affect the same gene at mRNA level, with low value of the β coefficients. At the same time parameter estimation on the hidden Markov model lead to very good results.

Results on real data using the independent prior seems to support the idea behind our model. At the same time the hidden Markov model on the covariates seems to work properly and using more than one sample at a time to estimate parameters seem to lead to even better results.

We aim to compare our results with those that can be obtained applying the model of Monni and Tadesse [2009]. any other methods can be used for a comparison, such as those of Richardson et al. [2010], Scott-Boyer and al. [2012], etc.. (see Section 3.5).

We also want to apply our method using pathway data instead of gene expression data as responses. This could be obtained performing a principal component analysis on genes that belong to the same pathway. In such a way the assumption of independence among responses become more realistic. Moreover we would like to increase the number of covariates, as well as to consider covariates that come from different chromosomes. Lastly we can use a model with dummy variables for the copy number variations effects. When focusing on these novelties we must consider, as always, the computational issue. For instance using dummy variables for copy number variation effects is really simple to implement, but can increase substantially the computational time.

A

Likelihood derivation

We outline here the basic calculations to get to the marginal likelihood (4.10). Our regression model can be expressed as:

$$f(\boldsymbol{Y}_g|\boldsymbol{\xi}_R, \boldsymbol{R}, \sigma_g^2, \boldsymbol{\beta_g}, \mu_g) \sim N(\boldsymbol{\xi}_R\boldsymbol{\beta}_g + \mathbf{1_n}\mu_\mathbf{g}, \sigma_\mathbf{g}^\mathbf{2}\mathbf{I_n})$$

with $\pi(\mu_g|\sigma_g^2) \sim N(0, c_\mu^{-1}\sigma_g^2)$, $\pi(\boldsymbol{\beta}_g|\boldsymbol{R}, \sigma_g^2) \sim N(\mathbf{0}, \frac{\sigma_g^2}{c_\beta}\mathbf{I}_{k_g})$ and $\pi(\sigma_g^2) \sim IG(\delta, \frac{d}{2})$. Integrating out the intercept parameters we obtain:

$$\begin{aligned}
f(\boldsymbol{Y}_g|\boldsymbol{\xi}_R, \boldsymbol{R}, \sigma_g^2, \boldsymbol{\beta}_g) &= \int f(\boldsymbol{Y}_g|\boldsymbol{\xi}_R, \boldsymbol{R}, \sigma_g^2, \boldsymbol{\beta}_g, \mu_g)\pi(\mu_g|\sigma_g^2)d\mu_g \\
&= \int (2\pi\sigma_g^2)^{-\frac{n+1}{2}} c_\mu^{\frac{1}{2}} \exp\{-\tfrac{1}{2\sigma_g^2}(\boldsymbol{Y}_g - \boldsymbol{\xi}_R\boldsymbol{\beta}_g - \mathbf{1}_n\mu_g)' \\
&\quad \times(\boldsymbol{Y}_g - \boldsymbol{\xi}_R\boldsymbol{\beta}_g - \mathbf{1}_n\mu_g)\} \exp\{-\tfrac{c_\mu}{2\sigma_g^2}\mu_g^2\}d\mu_g \\
&= \int (2\pi\sigma_g^2)^{-\frac{n+1}{2}} c_\mu^{\frac{1}{2}} \exp\{-\tfrac{n+c_\mu}{2\sigma_g^2}(\mu_g - \tfrac{\mathbf{1}_n'(\boldsymbol{Y}_g - \boldsymbol{\xi}_R\boldsymbol{\beta}_g)}{n+c_\mu})^2 \\
&\quad -\tfrac{1}{2\sigma_g^2}[(\boldsymbol{Y}_g - \boldsymbol{\xi}_R\boldsymbol{\beta}_g)'(\boldsymbol{Y}_g - \boldsymbol{\xi}_R\boldsymbol{\beta}_g) - (\boldsymbol{Y}_g - \boldsymbol{\xi}_R\boldsymbol{\beta}_g)'\tfrac{\mathbf{1}_n\mathbf{1}_n'}{n+c_\mu} \\
&\quad \times(\boldsymbol{Y}_g - \boldsymbol{\xi}_R\boldsymbol{\beta}_g)]\}d\mu_g \\
&= (2\pi\sigma_g^2)^{-\frac{n}{2}}(\tfrac{c_\mu}{n+c_\mu})^{\frac{1}{2}} \exp\{-\tfrac{1}{2\sigma_g^2}[(\boldsymbol{Y}_g - \boldsymbol{\xi}_R\boldsymbol{\beta}_g)'(\boldsymbol{Y}_g - \boldsymbol{\xi}_R\boldsymbol{\beta}_g) \\
&\quad -(\boldsymbol{Y}_g - \boldsymbol{\xi}_R\boldsymbol{\beta}_g)'\tfrac{\mathbf{1}_n\mathbf{1}_n'}{n+c_\mu}(\boldsymbol{Y}_g - \boldsymbol{\xi}_R\boldsymbol{\beta}_g)]\}
\end{aligned}$$

Integrating out the regression coefficients leads to:

$$\begin{aligned}
f(\boldsymbol{Y}_g|\boldsymbol{\xi}_R,\boldsymbol{R}) &= \int f(\boldsymbol{Y}_g|\boldsymbol{\xi}_R,\boldsymbol{R},\boldsymbol{\beta}_g)\pi(\boldsymbol{\beta}_g|\boldsymbol{R},\sigma_g^2)d\boldsymbol{\beta}_g \\
&= \int (2\pi\sigma_g^2)^{-\frac{n+k_g}{2}}(\tfrac{c_\mu}{n+c_\mu})^{\frac{1}{2}}c_\beta^{\frac{k_g}{2}}\exp\{-\tfrac{1}{2\sigma_g^2}[(\boldsymbol{Y}_g-\boldsymbol{\xi}_R\boldsymbol{\beta}_g)'\boldsymbol{H}_n \\
&\quad \times(\boldsymbol{Y}_g-\boldsymbol{\xi}_R\boldsymbol{\beta}_g)+c_\beta\boldsymbol{\beta}_g'\boldsymbol{\beta}_g]\}d\boldsymbol{\beta}_g \\
&= \int (2\pi\sigma_g^2)^{-\frac{n+k_g}{2}}(\tfrac{c_\mu}{n+c_\mu})^{\frac{1}{2}}c_\beta^{\frac{k_g}{2}}\exp\{-\tfrac{1}{2\sigma_g^2}[\boldsymbol{Y}_g'\boldsymbol{H}_n\boldsymbol{Y}_g+\boldsymbol{\beta}_g'\boldsymbol{U}_g\boldsymbol{\beta}_g \\
&\quad -2\boldsymbol{\beta}_g'\boldsymbol{\xi}_R'\boldsymbol{H}_n\boldsymbol{Y}_g]\}d\boldsymbol{\beta}_g \\
&= (2\pi\sigma_g^2)^{-\frac{n+k_g}{2}}(\tfrac{c_\mu}{n+c_\mu})^{\frac{1}{2}}c_\beta^{\frac{k_g}{2}}|\boldsymbol{U}_g|^{-\frac{1}{2}}\exp\{-\tfrac{1}{2\sigma_g^2}[\boldsymbol{Y}_g'\boldsymbol{H}_n\boldsymbol{Y}_g \\
&\quad -\boldsymbol{Y}_g'\boldsymbol{H}_n\boldsymbol{\xi}_R\boldsymbol{U}_g^{-1}\boldsymbol{\xi}_R'\boldsymbol{H}_n\boldsymbol{Y}_g]\} \\
&= (2\pi\sigma_g^2)^{-\frac{n}{2}}(\tfrac{c_\mu}{n+c_\mu})^{\frac{1}{2}}c_\beta^{\frac{k_g}{2}}|\boldsymbol{U}_g|^{-\frac{1}{2}}\exp\{-\tfrac{1}{2\sigma^2}[q_g]\},
\end{aligned}$$

where $\boldsymbol{H}_n = \mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n'}{n+c_\mu}$, $\boldsymbol{U}_g = c_\beta\mathbf{I}_{k_g} + \boldsymbol{\xi}_R'\boldsymbol{H}_n\boldsymbol{\xi}_R$ and $q_g = \boldsymbol{Y}_g'\boldsymbol{H}_n\boldsymbol{Y}_g - \boldsymbol{Y}_g'\boldsymbol{H}_n\boldsymbol{\xi}_R\boldsymbol{U}_g^{-1}\boldsymbol{\xi}_R'\boldsymbol{H}_n\boldsymbol{Y}_g$. Finally integrating out σ_g^2 we obtain:

$$\begin{aligned}
f(\boldsymbol{Y}_g|\boldsymbol{\xi}_R,\boldsymbol{R}) &= \int f(\boldsymbol{Y}_g|\boldsymbol{\xi}_R,\boldsymbol{R},\sigma_g^2)d\sigma_g^2 \\
&= \int (2\pi\sigma_g^2)^{-\frac{n}{2}}(\tfrac{c_\mu}{n+c_\mu})^{\frac{1}{2}}c_\beta^{\frac{k_g}{2}}|\boldsymbol{U}_g|^{-\frac{1}{2}}\tfrac{(\frac{d}{2})^{\frac{\delta}{2}}}{\Gamma(\frac{\delta}{2})}(\sigma_g^2)^{-(1+\frac{\delta}{2})}\exp\{-\tfrac{q_g}{2\sigma^2}-\tfrac{d}{2\sigma_g^2}\}d\sigma_g^2 \\
&= (2\pi)^{-\frac{n}{2}}(\tfrac{c_\mu}{n+c_\mu})^{\frac{1}{2}}c_\beta^{\frac{k_g}{2}}|\boldsymbol{U}_g|^{-\frac{1}{2}}\int\tfrac{(\frac{d}{2})^{\frac{\delta}{2}}}{\Gamma(\frac{\delta}{2})}(\sigma_g^2)^{-(1+\frac{n+\delta}{2})}\exp\{-\tfrac{d+q_g}{2\sigma_g^2}\}d\sigma_g^2 \\
&= \tfrac{((2\pi)^{-\frac{n}{2}}(\frac{c_\mu}{n+c_\mu})^{\frac{1}{2}}c_\beta^{\frac{k_g}{2}}(\frac{d}{2})^{\frac{\delta}{2}}\Gamma(\frac{n+\delta}{2})}{|\boldsymbol{U}_g|^{\frac{1}{2}}\Gamma(\frac{\delta}{2})(\frac{q_g+d}{2})^{\frac{n+\delta}{2}}},
\end{aligned}$$

which is (4.10).

B

MCMC steps

Here we describe how to compute the quantities involved in the various steps of our MCMC algorithm. During the update of $\boldsymbol{R}$ (and of $\boldsymbol{\xi}$) we first select a list of genes, as rows of $\boldsymbol{R}$ (and a list of samples, as elements of a randomly selected column of ξ) and then update and accept/reject their individual values. For this, we first sample from a geometric distribution with probability p_g (p_n) and add the result to the index of the last selected gene (sample). If the resulting index is greater than G (n), then we discard the new value and stop, otherwise we add the new position to the list of selected genes (samples) and draw a new value from the geometric distribution. For the first draw, we simply consider the result as the position to be updated. The updates on μ_j, σ_j, for $j = 1, \ldots, 4$, and the transition matrix $\mathbf{A}$ follow Guha et al. [2008], though applied to all samples simultaneously.

Updating R

We give details on how to calculate the probability $\pi(\boldsymbol{R}|\boldsymbol{\xi})$ when updating $\boldsymbol{R}$:

$$\pi(\boldsymbol{R}|\boldsymbol{\xi}) = \prod_{g=1}^{G} \pi(r_{g1}|r_{g2}, \boldsymbol{\xi}) \pi(r_{gM}|r_{g(M-1)}, \boldsymbol{\xi}) \prod_{m=2}^{M-1} \pi(r_{gm}|r_{g(m-1)}, r_{g(m+1)}, \boldsymbol{\xi}).$$

When calculating the ratio $\frac{\pi(\boldsymbol{R}^{new}|\boldsymbol{\xi})}{\pi(\boldsymbol{R}^{old}|\boldsymbol{\xi})}$ we need to consider only those quantities whose values change when a single element of $\boldsymbol{R}$ is updated. What follows is the description of the different scenarios that could occur when applying our MCMC update.

- Adding/deleting:
  - If the selected element is not either the first or last marker, three elements change their values: $\pi(r_{gm}|r_{g(m-1)}, r_{g(m+1)}, \boldsymbol{\xi})$, $\pi(r_{g(m-1)}|r_{g(m-2)}, r_{gm}, \boldsymbol{\xi})$ and $\pi(r_{g(m+1)}|r_{gm}, r_{g(m+2)}, \boldsymbol{\xi})$.
  - If the selected element is either marker 1 or M, only two quantities change their values:
    - $\pi(r_{g1}|r_{g2}, \boldsymbol{\xi})$ or $\pi(r_{gM}|r_{g(M-1)}, \boldsymbol{\xi})$;
    - $\pi(r_{g2}|r_{g1}, r_{g3}, \boldsymbol{\xi})$ or $\pi(r_{g(M-1)}|r_{g(M-2)}, r_{gM}, \boldsymbol{\xi})$.
- Swapping:
  - Swap between adjacent elements; four quantities change their values:
    - $\pi(r_{g(m-1)}|r_{g(m-2)}, r_{gm}, \boldsymbol{\xi})$;
    - $\pi(r_{gm}|r_{g(m-1)}, r_{g(m+1)}, \boldsymbol{\xi})$;
    - $\pi(r_{g(m+1)}|r_{gm}, r_{g(m+2)}, \boldsymbol{\xi})$;
    - $\pi(r_{g(m+2)}|r_{g(m+1)}, r_{g(m+3)}, \boldsymbol{\xi})$.
  - Swap between "quasi-adjacent" elements, i.e., two elements that are two marker positions apart. Five quantities get involved (say, for example, that r_{gm} get swapped with $r_{g(m-2)}$):
    - $\pi(r_{g(m-3)}|r_{g(m-4)}, r_{g(m-2)}, \boldsymbol{\xi})$;
    - $\pi(r_{g(m-2)}|r_{g(m-3)}, r_{g(m-1)}, \boldsymbol{\xi})$;
    - $\pi(r_{g(m-1)}|r_{g(m-2)}, r_{gm}, \boldsymbol{\xi})$;
    - $\pi(r_{gm}|r_{g(m-1)}, r_{g(m+1)}, \boldsymbol{\xi})$;
    - $\pi(r_{g(m+1)}|r_{gm}, r_{g(m+2)}, \boldsymbol{\xi})$.

Note that if the swap involves either marker 1 or M then these quantities reduce by one. Equation (4.8) is used to calculate all quantities involved in the steps above.

Updating $\boldsymbol{\xi}$

With this update, when calculating the probability $\pi(\boldsymbol{R}|\boldsymbol{\xi})$ we need to look for changes in the values of γ, ω_1 and ω_2. Suppose we change the value of the kth element, then

- we need to recalculate $\frac{1}{n}\sum_{i=1}^{n} I_{\{\xi_{ik}=\xi_{i(k-1)}\}}$ and $\frac{1}{n}\sum_{i=1}^{n} I_{\{\xi_{ik}=\xi_{i(k+1)}\}}$;
- these quantities result in changes in the values of γ^k, ω_1^k, ω_2^k, γ^{k-1}, ω_1^{k-1}, ω_2^{k-1}, γ^{k+1}, ω_1^{k+1}, ω_2^{k+1};
- we apply equation (4.8) to calculate the new values of $\pi(r_{gk}|r_{g(k-1)}, r_{g(k+1)}, \boldsymbol{\xi})$, $\pi(r_{g(k-1)}|r_{g(k-2)}, r_{gk}, \boldsymbol{\xi})$ and $\pi(r_{g(k+1)}|r_{gk}, r_{g(k+2)}, \boldsymbol{\xi})$.

Equation (4.10) is then used to calculate $f(\boldsymbol{Y}|\boldsymbol{\xi}^{new}, \boldsymbol{R})$ and $f(\boldsymbol{Y}|\boldsymbol{\xi}^{old}, R)$, while $f(x_{im}|\xi_{im})$ is simply the density of a $N(\mu_{\xi_{im}}, \sigma^2_{\xi_{im}})$, calculated in the current values of $\mu_{\xi_{im}}$ and $\sigma^2_{\xi_{im}}$.

Next, we focus on the ratio:

$$\frac{\pi(\boldsymbol{\xi}^{new}|\boldsymbol{\xi}^{old}, \boldsymbol{A})q(\boldsymbol{\xi}^{old}|\boldsymbol{\xi}^{new})}{\pi(\boldsymbol{\xi}^{old}|\boldsymbol{\xi}^{old}, \boldsymbol{A})q(\boldsymbol{\xi}^{new}|\boldsymbol{\xi}^{old})},$$

that can be factorized as

$$\prod_{i=1}^{n} \frac{\pi(\xi^{new}_{im}|\xi^{old}_{i(m-1)}, \xi^{old}_{i(m+1)}, \boldsymbol{A})q(\xi^{old}_{im}|\xi^{new}_{im})}{\pi(\xi^{old}_{im}|\xi^{old}_{i(m-1)}, \xi^{old}_{i(m+1)}, \boldsymbol{A})q(\xi^{new}_{im}|\xi^{old}_{im})}.$$

The ratio of interest can be evaluated as $\frac{\pi(\xi^{old}_{k(m+1)}|\xi^{new}_{km}, \boldsymbol{A})}{\pi(\xi^{old}_{k(m+1)}|\xi^{old}_{km}, \boldsymbol{A})}$, when $m \neq M$, and simply as 1 when $m = M$, by noting that $q(\xi^{new}_{im}|\xi^{old}_{im}) = \pi(\xi^{new}_{im}|\xi^{old}_{i(m-1)}, \boldsymbol{A})$, $\frac{\pi(\xi^{new}_{im}|\xi^{old}_{i(m-1)}, \xi^{old}_{i(m+1)}, \boldsymbol{A})}{\pi(\xi^{old}_{im}|\xi^{old}_{i(m-1)}, \xi^{old}_{i(m+1)}, \boldsymbol{A})} = \frac{\pi(\xi^{old}_{i(m+1)}|\xi^{new}_{im}, \boldsymbol{A})\pi(\xi^{new}_{im}|\xi^{old}_{i(m-1)}, \boldsymbol{A})}{\pi(\xi^{old}_{i(m+1)}|\xi^{old}_{im}, \boldsymbol{A})\pi(\xi^{old}_{im}|\xi^{old}_{i(m-1)}, \boldsymbol{A})}$, and considering that we update a single sample, sample k in our example.

Updating μ

Let $k = \{1, 2, 3, 4\}$ be the label for the four different states, δ_{0k} be the center of the truncated normal distributions in the prior specification of μ_k, n_k be the number of CGH in state k, $\bar{X}_k$ the mean of X's over those markers that are in state k and $\mathbf{I}_k$ denote the support of μ_k. Specifically

$$n_k = \sum_{m=1}^{M}\sum_{i=1}^{n}\mathbf{I}_{\{\xi_{im}=k\}}, \quad \bar{X}_k = \frac{1}{n_k}\sum_{m=1}^{M}\sum_{i=1}^{n}X_{im}\mathbf{I}_{\{\xi_{im}=k\}}.$$

The posterior probability for μ is obtained as:

$$\begin{aligned}
\pi(\mu_k|X, rest) &\propto \exp\{-\tfrac{1}{2\tau_k^2}(\mu_k-\delta_{0k})^2\}\exp\{-\tfrac{1}{2\sigma_k^2}\textstyle\sum_{i=1}^{n_k}(X_{ik}-\mu_k)^2\}\mathbf{I}_n \\
&= \exp\{-\tfrac{1}{2\tau_k^2}(\mu_k-\delta_{0k})^2\}\exp\{-\tfrac{1}{2\sigma_k^2}\textstyle\sum_{i=1}^{n_k}(X_{ik}-\bar{X}_k+\bar{X}_k-\mu_k)^2\}\mathbf{I}_k \\
&= \exp\{-\tfrac{1}{2\tau_k^2}(\mu_k-\delta_{0k})^2-\tfrac{1}{2\sigma_k^2}\,[\textstyle\sum_{i=1}^{n_k}(X_{ik}-\bar{X}_k)^2+\sum_{i=1}^{n_k}(\bar{X}_k-\mu_k)^2\}\mathbf{I}_k \\
&\propto \exp\{-\tfrac{1}{2\tau_k^2}(\mu_k-\delta_{0k})^2\}\exp\{-\tfrac{n_k}{2\sigma_k^2}\,(\bar{X}_k-\mu_k)^2\}\mathbf{I}_k \\
&= \exp\{-\tfrac{1}{2}[\mu_k^2\theta_k^2-2\mu_k(\tfrac{\delta_{0k}}{\tau_k^2}+\tfrac{\bar{X}_k}{(\frac{\sigma_k^2}{n_k})})+(\tfrac{\delta_{0k}^2}{\tau_k^2})+(\tfrac{\bar{X}_k^2}{(\frac{\sigma_k^2}{n_k})})]\}\mathbf{I}_k \\
&\propto \exp\{-\tfrac{1}{2}(\theta_k^2)[\mu-\eta_k]^2\}\mathbf{I}_n \rightarrow\sim N(\eta_k,(\theta_k^2)^{-1})\mathbf{I}_k
\end{aligned}$$

where $\theta_k = \tau_k^{-2} + n_k\sigma_k^{-2}$ and $\eta_k = \theta_k^{-2}(\delta_{0k}\tau_k^{-2} + \bar{X}_k n_k\sigma_k^{-2})$.

Updating σ^2

$$\begin{aligned}
\pi(\sigma_k^2|X, rest) &\propto \tfrac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)(\sigma_k^2)^{\alpha_j+1}}\exp\{-\tfrac{\beta_j}{\sigma_k^2}\}\tfrac{1}{(2\pi)^{\frac{n_k}{2}}(\sigma_k^2)^{\frac{n_k}{2}}}\exp\{-\tfrac{1}{2\sigma_k^2}V_k\} \\
&\propto \tfrac{1}{(\sigma_k^2)^{(\alpha_j+\frac{n_k}{2}+1)}}\exp\{-\tfrac{1}{\sigma_k^2}(\beta_j+\tfrac{V_k}{2})\} \rightarrow\sim IG(\alpha_j+\tfrac{n_k}{2},\beta_j+\tfrac{V_k}{2}),
\end{aligned}$$

where $V_k = \sum_{m=1}^{M}\sum_{i=1}^{n}(X_{im}-\mu_k)^2\mathbf{I}_{\{\xi_{im}=k\}}$.

Updating A

Let's focus on a single row of the transition matrix $\mathbf{A}$, then the distribution of the states arises from a multinomial distribution (except for the first element of each sample), and the prior distribution of any row of the matrix is $Dir(\phi, \phi, \phi, \phi)$:

$$\begin{cases} \pi(o_{h1}, o_{h2}, o_{h3}, o_{h4} | \mathrm{a}_{h1}, \mathrm{a}_{h2}, \mathrm{a}_{h3}, \mathrm{a}_{h4}) \sim Multinomial(\mathrm{a}_{h1}, \mathrm{a}_{h2}, \mathrm{a}_{h3}, \mathrm{a}_{h4}); \\ \pi(\mathrm{a}_{h1}, \mathrm{a}_{h2}, \mathrm{a}_{h3}, \mathrm{a}_{h4}) \sim Dir(\phi, \phi, \phi, \phi); \\ o_{hj} = \sum_{i=1}^{n} \sum_{m=1}^{M-1} \mathbf{I}_{\{\xi_{im}=h, \xi_{i(m+1)}=j\}}; \\ \sum_{h=1}^{4} o_{h1} + o_{h2} + o_{h3} + o_{h4} = n(M-1); \\ \mathrm{a}_{h1} + \mathrm{a}_{h2} + \mathrm{a}_{h3} + \mathrm{a}_{h4} = 1; \\ \pi_{\boldsymbol{A}}(\xi_{i1}). \end{cases}$$

We follow Guha et al. [2008] and generate a proposal $\boldsymbol{C}$ from the distribution $c_h|all \sim Dir(\phi + o_{h1}, \phi + o_{h2}, \phi + o_{h3}, \phi + o_{h4})$, ignoring the marginal distribution of state $\boldsymbol{\xi_1}$. We then accept the proposal with probability $\beta = min[1, \prod_{i=1}^{n} \frac{\pi_{\boldsymbol{C}}(\xi_{i1})}{\pi_{\boldsymbol{A}}(\xi_{i1})}]$.

C Rcpp

R is an open source programming language suited for statistical analysis. It uses a friendly command line interface and many packages are available directly on the CRAN web page[1] (most of them for statistical' s purposes). The main disadvantage of this programming language is its slowness, making it not suited for computationally intensive tasks. To address this problem, C, C++, and Fortran code can be linked and called at run time. Our MCMC is an example of computationally intensive task, therefore we used the two packages *Rcpp* and *RcppArmadillo*[2] to speed up our code. Essentially we use R as an interface but all the computations needed to obtain the MCMC chains are coded directly in C and, in particular, using the *Armadillo* library[3]. It is a linear algebra library that aims towards a good balance between speed and ease of use, with a syntax deliberately similar to Matlab. Algorithms 1 and 2 show the pseudo-code for the Metropolis step on $\boldsymbol{R}$. Inputs are: g the number of genes, m the number of CGH fragments, pR (i.e p_R see $\boldsymbol{R}$ update, section 4.3), *phi* the probability of A/D step, R vector form of the $\boldsymbol{R}$ matrix, *countnotwo* that contains for each m the number of samples not in state two.

[1] http://cran.r-project.org/
[2] http://dirk.eddelbuettel.com/code/rcpp.html
[3] http://arma.sourceforge.net/

Algorithm 1 ***R*** update - selection of genes to be changed

SET *sommo* to zero
SET nR to zero
while *sommo*$<g+1$ **do**
  PUT in *geom* a random number sampled from a geometric distribution with parameter pR
  ADD *geom* to *sommo*
  PUT *sommo* in *choiceg*$[nR]$
  INCREMENT *sommo*
  INCREMENT nR
end while
DECREASE nR
if $nR==0$ **then**
  PUT a random number sampled from a discrete uniform in the interval $[0, g-1]$ in *choiceg*$[0]$
  SET nR to one
end if

Algorithm 2 $\boldsymbol{R}$ update - reprise

for $0 \leq ggg < nR$ **do**
  PUT in *Rgene* positions of R that correspond to gene *ggg*
  SET *piRxi* to zero
  PUT in *ones* *Rgene*'s positions equal to one
  PUT in *counto* *ones*'s length
  PUT in *zeros* *Rgene*'s positions equal to zero
  PUT in *countz* *zeros*'s length
  CREATE vector *zerosok* of length *countz*
  SET all *zerosok*'s positions to zero
  SET *countzok* to zero
  for $0 \leq j < countz$ **do**
    if $countnotwo[zeros[\mathrm{j}]] \neq 0$ **then**
      PUT in $zerosok[countzok]$ $zeros[j]$
      INCREMENT *countzok*
    end if
  end for
  CREATE vector *posok* joining vectors *ones* and *zerosok*
  PUT in x a random number sampled in the interval $[0, 1]$
  if $countz = m$ **or** $countz = 0$ **or** $x < phi$ **or** $countzok = 0$ **then**
    PUT in *choice* one value in *posok* chosen at random
    CHANGE the value in $Rgene[choice]$
    ADD to *piRxi* $\log \frac{\pi(\boldsymbol{R}^{new}|\boldsymbol{\xi})}{\pi(\boldsymbol{R}^{old}|\boldsymbol{\xi})}$ (for details see Appendix B)
    ADD to *piRxi* $\log \frac{f(\boldsymbol{Y}_g|\boldsymbol{\xi}_R, \boldsymbol{R}^{new})}{f(\boldsymbol{Y}_g|\boldsymbol{\xi}_R, \boldsymbol{R}^{old})}$
  else
    PUT in *choice1* one value in *zerosok* chosen at random
    PUT in *choice2* one value in *ones* chosen at random
    SWAP the values in $Rgene[choice1]$ and $Rgene[choice2]$
    ADD to *piRxi* $\log \frac{\pi(\boldsymbol{R}^{new}|\boldsymbol{\xi})}{\pi(\boldsymbol{R}^{old}|\boldsymbol{\xi})}$ (for details see Appendix B)
  end if
  PUT in *random* a random number sampled in the interval $[0, 1]$
  if $random < \exp\{piRxi\}$ **then**
    return *Rgene*
  else
    return -1
  end if
end for

Bibliography

J.R. Pollack, T. Sørlie, C.M. Perou, C.A. Rees, S.S. Jeffrey, P.E. Lonning, R. Tibshirani, D. Botstein, A.L. Børresen, and P.O. Brown. 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*, 2002.

W. Van Wieringen and M. A. Van de Wiel. Nonparametric testing for DNA copy number induced differential mRNA gene expression. *Biometrics*, 2009.

H. Choi, Z.S. Quin and D. Ghosh. A double-layered mixture model for the joint analysis of dna copy number and gene expression data. *Technical Report, Department of Statistics, Penn State University*, 2010.

S. Guha, Y. Li and D. Neuberg. Bayesian hidden Markov modelling of array cgh data. *JASA*, 2008.

K. Chin et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, 2006.

F.C. Stingo et al. A bayesian graphical modelling approach to microrna regulatory network inference. *Annals of Applied Statistics*, 2010.

L. Du, M. Chen, J. Lucas and L. Carlin. Sticky hidden Markov modelling of comparative genomic hybridization. *IEE transactions on signal processing*, 2010.

E. Fox, E.B. Sudderth, M.I. Jordan and A.S. Willsky. The sticky HDP-HMM: Bayesian nonparametric hidden Markov models with persistent states. *Technical Report, LIDS, MIT*, 2007.

M. Beal, Z. Ghahramani and C.E. Rasmussen. The infinite hidden Markov model . *Machine Learning, MIT Press*, 2002.

S. Zhong and J. Ghosh. A new formulation of coupled hidden Markov models. *Tech. Report, Dept. of Electrical and Computer Engineering, U. of Texas at Austin*, 2001.

S. Monni and M.G. Tadesse. A stochastic partitioning method to associate high-dimensional responses and covariates. *Bayesian Analysis*, 2009.

S. Richardson,L. Bottolo and J.S. Rosenthal. Bayesian models for sparse regression analysis of high dimensional data. *Oxford University Press*, 2010.

Z. Jia and S. Xu. Mapping Quantitative Trait Loci for expression abundance. *Genetics*, 2007.

R. Dahm. Friedrich Miescher and the discovery of DNA. *Developmental Biology*, 2005.

R. Dahm. A structure for Deoxyribose Nucleic Acid. *Nature*, 1953.

B. Albert, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter. Molecular biology of the cell,4th edition. *Garland Science*, 2002.

F.B. Churchill. William Johannsen and the genotype concept. *Journal of the History of Biology*, 1974.

E. Suárez, C.A. Sariol, A. Burguette and G. McLachlan. A tutorial in genetic epidemiology and some considerations in statistical modelling. *Puerto Rico Health Science Journal*, 2007.

M.M. Babu. Introduction to microarray data analysis. In *Computational Genomics: Theory and Application*, *Horizon Press*, 2004.

W.W. Lockwood, R. Chari, B. Chin and W.L. Lam. Recent advances in array comparative genomic hybridization technologies and their applications in human genetics. *European Journal of Human Genetics*, 2006.

V. Baladandayuthapani, Y. Ji, R. Talluri, L.E. Nieto-Barajas and J.S. Morris. Bayesian Random Segmentation Models to Identify Shared Copy Number Aberrations for Array CGH Data. *Journal of the American Statistical Association*, 2010.

C. Yau, O. Papaspiliopoulos, G.O. Roberts and C. Holmes. Bayesian Nonparametric Hidden Markov Models with the application to the analysis of copy-number-variation in mammalian genomes. *Journal of the Royal Statistical Society, Series B*, 2011.

P.J. Brown, M. Vannucci and T. Fearn. Baysian wavelength selection in multicomponent analysis. *Journal of Chemometrics*, 1998.

P.J. Brown, M. Vannucci and T. Fearn. Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, series B*, 1998.

P.J. Brown, M. Vannucci and T. Fearn. Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society, series B*, 2002.

E.I. George and R.E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, 1997.

A.P. Dawid. Some Matrix-Variate Distribution Theory:Notationl Considerations and a Bayesian Application. *Biometrika*, 1981.

J.M. Kidd, G.M. Cooper, W.F. Donahue et al. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 2008.

G.R. Bignell, J. Huang, J. Greshock et al. High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Research*, 2004.

X. Zhao, C. Li, J.G. Paez et al. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Research*, 2004.

A.J. Iafrate, L. Feuk, M.N. Rivera M.L. Listewnik, P.K. Donahoe, S.W. Scherer and C. Lee. Detection of large-scale variation in the human genome. *Nature genetics*, 2004.

D.P. Locke, R. Segraves, L. Carbone N. Archidiacono, D.G. Albertson, D. Pinkel and E.E. Eichler. Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome research*, 2003.

A. Fortna, Y. Kim, E. MacLaren et al. Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biology*, 2004.

G. Parmigiani, E.S. Garret, R. Anbazhagan and E. Gabrielson. A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society, series B*, 2002.

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*, 2004.

P. Broët and S. Richardson. Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. *Bioinformatics*, 2006.

J. Fridlyand, A.. Snijders, D.G. Pinkel, D.G. Albertson and A.N. Jain. Application of Hidden Markov Models to the analysis of the array CGH data. *Journal of Multivariate analysis*, 2004.

G. Hodgson, J.H. Hager, S. Volik, S. Hariono, M. Wernick, D. Moore, N. Nowak, D.G. Albertson, D. Pinkel, C. Collins, D. Hanahan and J.W. Gray. Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nature genetics*, 2001.

G. Sen and J.W. Srivastava. On tests for detecting a change in mean. *Annals of Statistics*, 1975.

Jason Eisner. An interactive spreadsheet for teaching the forward-backward algorithm. *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching NLP and CL*, 2002.

E.I. George and R.E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 1993.

N. Sha, M. Vannucci, M. Tadesse, P. Brown, I. Dragoni, N. Davies, T. Roberts, A. Contestabile, N. Salmon, C. Buckley and F. Falciani. Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics*, 2004.

M. Smith and R. Kohn. Non Parametric Regression using Bayesian Variable Selection. *Journal of Econometrics*, 1996.

T.J. Mitchell and J.J. Beauchamp. Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 1988.

L. Bottolo and S. Richardson. Evolutionary Stochastic Search for Bayesian Model Exploration. *Bayesian Analysis*, 2010.

D. Madigan and J. York. Bayesian graphical models for discrete data. *International statistical review*, 1995.

J.G. Scott and J.O. Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, 2010.

C.C. Holmes and L. Held. Bayesian Auxiliary Variable Models for Binary and Multinomial Regression. *Bayesian Analysis*, 2006.

J.H. Albert and S. Chib. Bayesian analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 1993.

M.P. Scott-Boyer, G.C. Imhoolte, A. Tayeb, A. Labbe, C.F. Deschepper and R. Gottardo. An Integrated Hierarchical Bayesian Model for Multivariate eQTL Mapping. *Statistical Applications in Genetics and Molecular Biology*, 2012.

J. Yin and H. Li. A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics*, 2011.

T.T. Cai, H. Li, W. Liu and J. Xie. Covariate Adjusted Precision Matrix Estimation with an Application in Genetical Genomics. *Biometrika*, 2011.

B. Jones, C. Carvalho, A. Dobra, C. Carter and M. West. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 2005.

K. Wang, Z. Chen, M.G. Tadesse, J. Glessner, S.F.A. Grant, H. Hakonarson, M. Bucan and M. Li. Modeling genetic inheritance of copy number variations. *Nucledi Acids Research*, 2008.