The undersigned

SURNAME	TRENTINI
FIRST NAME	FILIPPO
PhD Registration Number	1287697

Thesis title:

Bayesian hierarchical Models for the integration of Genomic Platforms

PhD in	Statistics
Cycle	XXIII
Candidate's tutor	Muliere Pietro
Year of discussion	2012

DECLARES

Under his responsibility:

1) that, according to the President's decree of 28.12.2000, No. 445, mendacious declarations, falsifying records and the use of false records are punishable under the penal code and special laws, should any of these hypotheses prove true, all benefits included in this declaration and those of the temporary embargo are automatically forfeited from the beginning;

2) that the University has the obligation, according to art. 6, par. 11, Ministerial Decree of 30th April 1999 protocol no. 224/1999, to keep copy of the thesis on deposit at the Biblioteche Nazionali Centrali di Roma e Firenze, where consultation is permitted, unless there is a temporary embargo in order to protect the rights of external bodies and industrial/commercial exploitation of the thesis;

3) that the Servizio Biblioteca Bocconi will file the thesis in its 'Archivio istituzionale ad accesso aperto' and will permit on-line consultation of the complete text (except in cases of a temporary embargo);

4) that in order keep the thesis on file at Biblioteca Bocconi, the University requires that the thesis be delivered by the candidate to Società NORMADEC (acting on behalf of the University) by online procedure the

contents of which must be unalterable and that NORMADEC will indicate in each footnote the following information:

- thesis (*thesis* title) Bayesian hierarchical models for the integration of genomic platforms ;
- by (*candidate's surname and first name)* Filippo Trentini ;
- discussed at Università Commerciale Luigi Bocconi – Milano in (year of discussion) 2012 ;
- the thesis is protected by the regulations governing copyright (law of 22 April 1941, no. 633 and successive modifications). The exception is the right of Università Commerciale Luigi Bocconi to reproduce the same for research and teaching purposes, quoting the source;

5) that the copy of the thesis deposited with NORMADEC by online procedure is identical to those handed in/sent to the Examiners and to any other copy deposited in the University offices on paper or electronic copy and, as a consequence, the University is absolved from any responsibility regarding errors, inaccuracy or omissions in the contents of the thesis;

6) that the contents and organization of the thesis is an original work carried out by the undersigned and does not in any way compromise the rights of third parties (law of 22 April 1941, no. 633 and successive integrations and modifications), including those regarding security of personal details; therefore the University is in any case absolved from any responsibility whatsoever, civil, administrative or penal and shall be exempt from any requests or claims from third parties;

7) that the PhD thesis is not the result of work included in the regulations governing industrial property, it was not produced as part of projects financed by public or private bodies with restrictions on the diffusion of the results; it is not subject to patent or protection registrations, and therefore not subject to an embargo;

Date 31/01/2012

Signed (write first name and surname) FILIPPO TRENTINI

Bayesian hierarchical models for the integration of genomic platforms

Filippo TRENTINI

Department of Decision Sciences

Bocconi University - Milan

A thesis submitted for the degree of

PhilosophiæDoctor (PhD) in Statistics

January 31, 2012

To my Mentors, for their wisdom and their patience

To my family and my friends

Abstract

We consider modeling jointly microarray RNA expression and DNA copy number data.We propose separate Bayesian mixture models for the observed copy numbers and gene expression measurements that define latent Gaussian probit scores for the DNA and RNA, and integrate between the two platforms via a regression of the RNA probit scores on the DNA probit scores. Such a regression conveniently allow us to include additional sample specific covariates such as type of breast cancer and pathological complete response of patients in the study.

The two developed methods are aimed respectively to make inference on differential behaviour of genes in patients showing different subtypes of breast cancer and to predict the pathological complete response of patients borrowing strength across the genomic platforms.

Posterior inference is carried out via MCMC simulations. We demonstrate the proposed methodology using a novel data set consisting of 122 newly diagnosed breast cancer patients.

Contents

List of Figures

1

Introduction

1.1 Problem statement

The word cancer indicates a various group of diseases, which have in common an uncontrolled cell division leading to growth of abnormal tissue.
In general, cancer is a complex, hard to predict and very heterogeneous disease.
During the last decade, much research in the field of Biostatistics and Bioinformatics has focused on finding good indicators for cancer staging, which is the most important predictor of survival and is central in the development of good treatments.
These *biomarkers* (28), used together with traditional medical measures, could significantly improve patient care, if their meaning is correctly stated and assessed.

Cancer can cause multiple genomic alterations, including gene amplifications, altered gene expression, deletions and point mutations. These particular aberrations may be inherited or somatically acquired during progression from a normal cell to a cancerous one, and it turns out to be central, in Genetics studies, the understanding of how these changes drive cancer cell survival by perturbing the mechanism for cell cycle control, tumor vascularization, DNA repair, and metabolism.

In our work we focus on biomarkers that provide information about two specific genetic

aberrations occurring during cancer progression: gene expression microarray and array CGH data.
The first type of data suggests whether genes show over- or under- expression during cancer progression, while the second provides information about DNA copy amplifications or deletions in particular genomic locations, and they are specific for each type of cancer.

The microarray technology offers great potential to identify molecular signatures capable of differentiating cancer from normal tissues, predicting outcome, detecting preferred patterns and monitoring response to cancer treatment . Many methods have been designed for the analysis of the two types of data, but very few concentrated on the problem of the integration of such platforms.
We propose Bayesian Hierarchical models for the assessment of differential gene expression and for the prediction of clinical outcomes related to the patients, borrowing strength between the two different, but likely to be correlated, genomic platforms.

Our first goal was to develop a method to determine whether patients belonging to different subgroups of cancer, characterized by the presence or absence of specific receptors on tumor cells, show differential genetic behaviour. The second objective was to apply such method to predict clinical outcomes measured on the same set of patients, with specific priors allowing for model selection. The third and last aim was adapt the model to allow for the integration of multiple platforms.

1.2 Motivation

We address with this work to important problems which are frequently met in cancer treatment, such as *prediction of clinical outcomes* , *choice of treatment*, *assessment of different molecular profiles*.

A new strategy in targeted cancer therapeutics research is to develop effective biological agents that specifically target a group of disease markers, as opposed to the traditional symptom-drive therapeutic strategies. The succes of this new strategy depends on accu-

rately understanding the underlying biological process of the disease.
For example, it is now standard practice to treat breast cancer with different drugs based on the expression of biomarker genes, such as HER2 ans ER. Selecting genes that show differential expression, as we try and do with our model, is a first important step that can lead to a better understanding of genetic pathways that are responsible for cell cycles and apoptosis.
We thus strongly believe our model could help clinicians and biologists to develop targeted treatment strategies for cancer subtypes with different molecular profiles.

The motivating application of this Bayesian Mixture Model for the integration of data across different array platforms is to a data set, created by clinicians at MD Anderson Cancer Centre, consisting of 122 newly diagnosed breast cancer patients. The data set includes DNA copy number measurements using comparative genomic hybridization and mRNA expression measured using microarrays, which will be discussed later in details.

1.3 Related work

Many recent publications in the last decade address either the problem of detecting genetic aberrations in different types of cancer, working marginally with gene expression microarray and array CGH.
Statistical and computational models for integrating different types of data are becoming a popular topic in the recent literature, even though only a few integrate different types of array platforms, indicating the need of efforts to this problem.
For the specic problem of integrating arrayCGH and microarray data even fewer methods are available.

Pollack et al (33) was one of the earliest to investigate the direct association between the two data in breast cancer cell lines and tissue samples, and his approach is based mainly on descriptive statistics.
Van Wieringen and Van de Wiel (2009) (44), attempting to mitigate the high noise in

the raw expression measurements of the DNA and RNAs, proposed a sampling model for mRNA expression measurements based on estimated probabilities of copy number variations. They subsequently developed nonparametric adaptive tests to study whether the estimated copy number variations in the DNA level would induce differential gene expression at the RNA level.
More recently Choi et al (2010) (12) presented a double-layered mixture model (DLMM) (DLLM) that directly modeled segmental patterns in the copy number data to produce CNA proles, and simultaneously scored the association between copy number and gene expression data.

1.4 Contribution

Each type of cancer consists of many subtypes, characterized by major differences in their molecular regulation mechanism. The main goal of cancer research is eventually to improve firstly the diagnosis and then the treatment of cancer through more accurate disease classifications and patient stratification, which allows for the design of therapies which are targeted to specific cancer subtypes and potentially improve the effectiveness of existing regimens based on therapeutic response and adverse events.

Proposing data integration across different platforms, we expect to improve the statistical power of detecting differential behaviour among different disease subtypes and, subsequently, propose better targeted therapies or better predict clinical outcomes, related for example to the pathological response of patients to the treatment.

We specifically propose a fully Bayesian model for integrating aCGH data and gene expression microarray measurements, through which we convert the noisy raw intensity measurements of the DNAs and RNAs into probability of expression (30), which are subsequently modeled as latent parameters. Integration of the two platforms is realized by joint modeling the probabilities of expression through probit regressions. Hierarchical modeling is very convenient in order to extract relevant information from data which

contains a too large amount of additional noise to be processed directly and fully by medical experts.

2

Gene expression Microarrays and arrays CGH: description and existing models

It is broadly proved that the knowledge of human genome have and will change deeply the conception of medicine and biology.
After April 2003, when the International Sequencing Consortium announced that the Human Genome Project (49) had been completed and 99% of the human genetic code had been sequenced, people started talking about a post-genomic era. Efforts were no more aimed to sequence the human genome, but focused on harvesting the fruits hidden in the genomic *book*.
DNA, RNA and proteins are all intensively active, in a very complex and coordinated way, in mechanisms that are fundamental for the life of any living being, and before the beginning of this new era biologist, although conscious of this enormous complexity, did not have at their disposal appropriate technologies for such studies.
Molecular Biology was indeed based on "one experiment, one gene" criteria, a useful feature to clarify single biological processes, but completely insufficient to study the way a

whole organism responds to stimuli.
The advent of microarray technologies, developed and broadly studied in the last decade, tried to plug this gap and offered a promising tool, by providing a systematic way to survey variation in DNA and RNA.
We specifically dealed with two particular types of microarray: arrayCGH (aCGH) that measures DNA copy numbers and expression microarrays that measures RNA expression. They have been both broadly studied in the last years, though few models on the integration of the two platforms are available at the moment.
In this chapter of my thesis I would like to introduce separately these data and the basic notions of molecular biology we need to better understand the framework, describe the technology behind and mention at least one of the many methods which are available for the analysis of such data.

2.1 Basis of Molecular Biology

Many diseases involve problems at the cellular level and many begin with problems in the genetic code. This is the main thought that has driven my work and for a better understanding of data I dealt with it is important to give a brief introduction of cells and genetics.
The cell is the basic functional unit of every living being. It was discovered by Robert Hooke and is the functional unit of all known living organisms. It is the smallest unit of life and is classified as a living thing, and is often called the building block of life (1).
In the first part of 19^{th} century, Matthias Jakob Schleiden and Theodor Schwann developed the cell theory and stated that every organism is composed of one or more cells. Their theory stated that all cells come from preexisting cells, that vital functions of an organism occur within cells, and that all cells contain the hereditary information necessary for regulating cell functions and for transmitting information to the next generation of cells.

This theory was later confirmed and enhanced by an important discovery: all necessary information for a whole organism is present in the DNA of each individual cell in the organism itself.

Almost all of cells in our body contain, within each nucleus, the entire genome for that organism. This genome contains the organisms complete hereditary information in the form of deoxyribonucleic acid (DNA), that encodes a complete blueprint for all activities and structures within the organism.

In the human body, the genome consists of 23 pairs of chromosomes. One of each of this pair is inherited from the mother and the other from the father. Each chromosome is made of chains of DNA. DNA consists of two polymers (45) (large molecules of repeating subunits) made up of units called nucleotides, these molecules are wrapped around each other in a structure known as a double helix. Each nucleotide consists of a deoxyribose sugar, a phosphate group and one of the four nitrogen bases, guanine, adenine, thymine and cytosine. These bases, which are usually represented by their first letters, G, A, T and C, are where hereditary genetic information is actually encoded. It is worth noting that one of the two strands of the DNA double helix will suffice to describe this information; this is because of complementary base pairing, whereby an A on one strand always binds to a T on the other and a C always binds to a G (1).

Genes are essentially segments of the DNA structure described above. Loosely speaking, a gene is a section of DNA that defines a single trait by encoding a particular pattern, about 27,000 of which exist in humans; more technically, a gene is a locatable region of genomic sequence, corresponding to a unit of inheritance. The main purpose of genes is to act as a blueprint in the creation of proteins. Proteins are made of amino acids and are responsible for the structure and activity of an organism at a cellular level. They are created as follows; starting at the 5' end (the leading end) of a gene and proceeding to the 3' end (the tail end), the information contained in the gene is transcribed into a messenger ribonucleic acid (mRNA) strand. This process is performed by an enzyme called RNA polymerase, and it is important to highlight that the initial DNA sequence

containing a gene may also contain bits of sequence that will not be used: one feature of gene structure is that gene can have both "coding" regions, called exons, and "non coding" regions, called introns.
After transcription, processing inside the nucleus removes the non coding regions and "splices" the remaining parts together into mRNA, the final messenger RNA. This molecule leaves the nucleus of the cell where it is transcribed into a protein in a process called translation. This is performed by ribosomes, which read the code carried by mRNA molecules from the cell nucleus and create proteins combining any of the 20 amino acids in the body into complex polypeptide chains. These proteins are the building blocks of the organism and they carry out, or at least support the synthesis of DNA and RNA (1).
This whole process goes under the name of Central Dogma of Molecular Biology (15): DNA specifies RNA and RNA determines proteins.
The difference between how an organism is built and the way an organism appears resides in the two important concept of biology: the genotype and the phenotype.
The genotype (13) is related to what we have talked so far, that is the inherited instructions the organism carries within its genetic code. Not all the organisms with the same genotype look or act the same way because appearance and behaviour are modified by environmental and developmental conditions. Beyond that genomes belonging to the same species vary slightly from organism to organism in a phenomenon known as *genetic variation*.
On the other side, a phenotype is any *observable characteristic* or trait of an organism, such as its morphology, development, biochemical or physiological properties, behaviour and product of behaviour. Phenotypes result from the expression of an organism's genes as well as the influence of environmental factors and the interactions between the two.
Phenotypic variation, due to underlying heritable genetic variation, is a fundamental prerequisite for evolution by natural selection.
An amazing aspect of the human genome is that there is so little variation in the DNA sequence when the genome of one person is compared to that of another. Of the 3.2 bil-

lion bases, roughly 99.9% are the same between any two people. It is the variation in the remaining tiny fraction of the genome, 0.1%–roughly several million bases–that makes a person unique. This small amount of variation determines attributes such as how a person looks, or the diseases he or she develops.

Variation occurs whenever the order of the bases in a DNA sequence changes. Variations can involve only one base or many bases. If the two strands of a chromosome are thought of as nucleotides threaded on a string, then, for example, a string can break and the order of the beads can vary. One or more nucleotides may be changed, added, or removed. In chromosomes, these changes are called polymorphisms, insertions, and deletions. In addition to these changes, some DNA sequences called "repeats" like to insert extra copies of themselves several times. Chromosomes can also undergo more dramatic changes called translocations. These occur when an entire section of DNA on one chromosome switches places with a section on another. Not all variations in the genome's DNA sequences have an effect. Among the variations that do cause effects, some are more serious than others. The outcome depends on two factors: where in the genome the change occurs, (i.e., in a noncoding region, coding, or regulatory region) and the exact nature of the change.

No two humans are genetically identical. Even monozygotic twins, who develop from one zygote, have infrequent genetic differences due to mutations occurring during development and gene copy number variation has been observed.

Copy-number variations (35) (CNVs - a form of structural variation) are alterations of the DNA of a genome that results in the cell having an abnormal number of copies of one or more sections of the DNA. CNVs correspond to relatively large regions of the genome that have been deleted (fewer than the normal number) or duplicated (more than the normal number) on certain chromosomes. For example, the chromosome that normally has sections in order as A-B-C-D might instead have sections A-B-C-C-D (a duplication of "C") or A-B-D (a deletion of "C").

This variation accounts for roughly 12% of human genomic DNA and each variation may range from about one kilobase (1,000 nucleotide bases) to several megabases in size. CNVs

contrast with single-nucleotide polymorphisms (SNPs), which affect only one single nucleotide base.
The process of transcription of DNA into RNA and subsequent translation into proteins, referred to as gene expression, and copy number variation are strictly related to the genomic platforms I dealt with in my model.
These two processes can be discovered relatively by a well known multiplex technology, the DNA microarray, and by a cytogenetic technique known as array comparative genomic hybridization (aCGH). I will try and give a clearer idea on how these techniques work, and subsequently on how my data are structured.

2.2 Microarrays

By the Central Dogma of Molecular Biology we learned that if different genes are copied into RNA in different cells, that is they are differently expressed, different proteins will be produced and subsequently, since proteins are the chief actors within the cell, different types of cells will emerge.
The same gene may be highly expressed in one person, but show a very low expression in another and this difference may be due to a small mutation or due to different environmental exposure. For example, after an increase exposure to the sun, a person may begin to produce more dark pigment in his skin by expressing the melanin gene.
Moreover, despite the fact that every cell in our body contains the same DNA sequence, every gene is not expressed in every cell. Sticking to the example above, genes producing a protein that protects the skin from dangerous and harmful ultraviolet rays will only be expressed in skin cells. Gene expression microarrays measure mRNA expression, that is RNA copies a gene produces, and they represent powerful tools to study gene expression. When a gene is expressed it produces RNA which will help with the production of the final protein coded for by the gene itself and these arrays tell the scientists how much RNA a gene is making, if it is making any at all. They have been broadly used in the

past decade to try and find a link between a group of genes and a specific disease, as well as to develop targeted treatments.
The revolutionary aspect about microarrays is that they allow researches to measure the expression of every single gene in the whole human genome, so that even with a simplest comparison between the gene expression patterns of two different subgroups of individuals, scientists can quickly point out differences in the pattern.
Although it may sound rather simple, it has been a long way to figure out just what genes are doing. It wasn't until 1963 that people even knew genes expressed different amounts of RNA and until 1977 that clinicians had a practical tool to measure gene expression through a technique called "Northern Blot" (43), while the use of miniaturized microarrays for gene expression profiling was first reported in 1995 (36), and a complete eukaryotic genome (Saccharomyces cerevisiae) on a microarray was published in 1997 (27). Such techniques, allowing for the visualization of thousands of genes at a time, overcame one big disadvantage of blotting methods: they could only measure expression from one or a few genes at a time, through a slow and tedious process and clearly the study of the expression of the whole genome taking the one-gene-at-time approach would be like draining the ocean with a cocktail straw.

Most of the description that follows was taken from the website of Affymetrix (47), the company that manufactures microarrays I used for my analysis. Gene expression microarrays use the natural attraction between the DNA and RNA target molecules to determine expression level, that is how much RNA is being made, of a given gene. I have already talked about the four basis forming DNA and their natural binding; like DNA, RNA is composed of four basis, with uracil (U) instead of thymine (T), and the same pairing system. Unlike DNA, RNA appears in a single stranded form, and this allows it to bind easily to any other single stranded sequence, both it is DNA or RNA.
When a single strand of DNA, say ATCATG, matches a strand of RNA, say UAGUAC, the two strands are said to be complementary and will stick to each other. Though, even a single base that does not match its partner, could keep a single stranded sequence from

sticking to another.
Microarrays use this complementary bases attraction, known as hybridization, to allow researchers identify what RNA sequence are present in a sample to establish what genes are being expressed by that individual and how much they are being expressed.
We can focus on just one gene to illustrate how this technology works.
The first step is to build a DNA strand, a probe, onto a surface of a glass chip. Genes will be hundreds of thousands of bases long, while the probes are generally shorter by an order of magnitude. This is driven in part by the manufacturing process, as the cost of synthesis increases with the number of bases deposited. Thus, choosing probes to print requires a trade off between finding sequences that will be unique to the gene of interest and affordability. The final length decided on was 25 bases, and all Affymetrix probes are this length.
Scientists compare the 25 base probe sequence to the rest of the human genome sequence to make sure it does not match anywhere else, so that, when an RNA molecule binds to the probe, it will be clear that the gene was expressed, because the only possible match is from that gene and no other sequence in the genome would match. The short probe on the array measures expression of the complete gene by sampling for a small section of the gene.
Now that a probe is designed to measure expressed RNA, RNA is to be extracted from a biological sample, such as blood, tumor, or other body tissue. Once extracted, it is copied millions of times through a process known as PCR, in such a way that it is more easily detected on the array. While RNA is copied, molecules of a chemical called biotin are attached to each strand, and they act as molecular glue for fluorescent molecules that will later be washed over the array.
The entire prepared RNA sample is washed over the array to allow the hybridization to occur, thus somewhere the sequence of bases in the sample RNA matches that of a DNA probe, implying a perfect match and subsequently allowing the sample to stick to the probe.

At this point RNA cannot be seen directly and it would be impossible to figure out how much RNA has stuck to the DNA probe on the array. To work around this problem, scientists use fluorescent molecules that sticks to the biotin, making the RNA glow in the dark. It is like pouring glitter over a paper I previously covered randomly with glue; after I shake it off, I would be able to see glitter sticked only to places where there was glue.
After all this a laser light is shone on the array, causing the stain to glow, and this scanning process produce an image where the intensities of different colours represent different level of gene expression. If a gene is highly expressed, many RNA molecules will stick to the probe, and the probe location will shine brightly when the laser hits in; viceversa whereas a gene was expressed at a lower level.
Precisely, the "relative expression level" is typically quoted as a number. In most cases, this number nominally corresponds to the log of the ratio of two intensity measurements, corresponding to the two dyes with which the two types of cells being compared have been respectively tagged (most commonly Cy5, red, and Cy3, green). Thus, the single number quoted is a derived value; it derives from the ratio of the two intensity values. And these intensity values are in turn derived from the image I was above talking about. For a better understanding of this converting process I recommend a technical report written by Dr. Keith Baggerly (48)
Our final data set was a matrix of intensity values for each gene and each individual in the sample we run the experiment on, appropriately pre-processed and normalised. Corresponding to the explosion of such technology, statistical methodology is being developed or adapted to extract meaningful information. A collection of existing statistical methods have been found useful, and new methods are emerging. The methods range from analysis of variance, mixed models, from multiple testing to cluster analysis, empirical Bayes and fully Bayesian methods, functional data analysis and networks.
Although it is not one of the latest, I will present here a method developed by Dr. Giovanni Parmigiani et al (30) which represents the starting point of our research.

2.2.1 Modeling gene expression: an example

The probability of expression (POE) model, as it is broadly known, is a 3-components mixture model on the data. With this fully Bayesian method they estimate allocation probabilities for each gene and condition to one of the 3 latent categories, providing a statistical definition of differential expression and defining in a more precise way a molecular profile for cancer subtypes.
The underlying assumption of their approach, and subsequently ours, is a broadly known and attested fact: gene expression can be usefully described as falling into categories of over- , under- or normal expression. These categories, which are latent, are defined by comparing the expression of a gene across samples and they can be interpreted as a gene being turned "on" or "off" compared with a normal expression. I would like to point out that in this specific case these latent categories correspond to aspects of physical reality, which could in principle be measured, and I would rather address to them as hidden states.
Hidden categories are particularly useful since they make the analysis robust to outlying observations of minor biological significance, but that can strongly help with clustering, they remove the normal component of variation that can be expected in the expression of a gene as a result of both technical and biological variability, they build an expression scoring that allows for better comparison and finally they allow for simple and interpretable definitions of molecular profiles, independent on the set of genes measured on the array.
Based on prior knowledge of gene expression distributions in microarray experiments, they use a mixture model to capture the gene expression distributions typically observed in microarray experiments. The mixture model permits correction for noise at both the sample- and gene-level. Specifically they use the following hierarchical mixture model (I kept the original notation) that borrows strength across genes to aid in estimation of

parameters, and tries to explain the variation of expression across tumors.

$$e_{gt} = \begin{cases} -1 & \text{if gene g has abnormally low expression in tumor t} \\ 0 & \text{if gene g has normal expression in tumor t} \\ 1 & \text{if gene g has abnormally high expression in tumor t,} \end{cases} \tag{2.1}$$

$$(a_{gt}|e_{gt} = e) \sim f_{e,g}(\cdot),\ e \in \{-1, 0, 1\}, \tag{2.2}$$

where a_{gt} represents the measured transcript abundance of gene g in tumor t ($g \in \{1, ..., G\}$ and $t \in \{1, ..., T\}$), and two important "characters" are what they call $\pi_g^- = P(e_{gt} = -1)$ and $\pi_g^+ = P(e_{gt} = 1)$, respectively the proportions of under - and over - expressed tumors in the population of unclassified tumors.

They assume that, for fixed π_g's and f's, e_{gt} are independent across genes and tumors and that, conditionally on e_{gt}, a_{gt} are independent.

As for the densities, $f_{0,g}(\cdot)$ describes the variation of expression for gene g in tumors that represent the modal expression for the cancer population of interest; variation that is due to both biological differences between tumors of the same subtypes, and measurements errors in the hybridization process. A priori information could help researcher to better identify the "normal" component. On the other side, densities $f_{-1,g}(\cdot)$ and $f_{1,g}(\cdot)$ describe the variation of expression for gene g in tumors that show under - or over - expression compared with the norm defined by $f_{0,g}(\cdot)$.

The posterior probabilities of differential expression are easily determined by Bayes's rule, as follows:

$$p_{gt}^+ = P(e_{gt} = 1|a_{gt}, \pi_g^+, \pi_g^-, f_{1,g}, f_{0,g}) = \frac{\pi_g^+ f_{1,g}(a_{gt})}{\pi_g^+ f_{1,g}(a_{gt}) + (1 - \pi_g^+ - \pi_g^-) f_{0,g}(a_{gt})} \tag{2.3}$$

if a_{gt} is in the support of $f_{1,g}$, and $p_{gt}^+ = 0$ otherwise.

$$p_{gt}^- = P(e_{gt} = -1|a_{gt}, \pi_g^+, \pi_g^-, f_{-1,g}, f_{0,g}) = \frac{\pi_g^- f_{-1,g}(a_{gt})}{\pi_g^- f_{-1,g}(a_{gt}) + (1 - \pi_g^+ - \pi_g^-) f_{0,g}(a_{gt})} \tag{2.4}$$

if a_{gt} is in the support of $f_{-1,g}$, and $p_{gt}^- = 0$ otherwise.

Thus they define the densities, sticking to some known and broadly used mixture models.

In particular, examples of these mixtures are discussed by Fraley and Raftery (1998 How many clusters?), who used such methods for finding outliers and sparse clusters.

$$\begin{aligned} f_{-1,g}(\cdot) &= U(-k_g^- + \alpha_t + \mu_g, \alpha_t + \mu_g) \\ f_{0,g}(\cdot) &= N(\alpha_t + \mu_g, \sigma_g) \\ f_{1,g}(\cdot) &= U(\alpha_t + \mu_g, \alpha_t + \mu_g + k_g^+) \end{aligned} \tag{2.5}$$

The systematic component $\alpha_t + \mu_g$, that we also used in our model, is both the centre of the distribution of the normal mRNA levels for gene g in tumor t and the threshold that separates under - and over - expression: μ_g is the effect of gene g on mean normal abundance, whereas α_t is a tumor-specific factor, providing for a normalization that takes into account only the common normal component and not the differentially expressed ones. (Colantuoni 2003 and Tseng 2001)

We leave out the detailed description of the other parameters since it will be precisely treated in the following chapter, where I explain my integration model.

Their Bayesian hierarchical model is completed by the prior assumption on all the set of parameters and a Metropolis-Hastings Markov chain Monte Carlo approach is used to obtain samples from the posterior distribution of the parameters, with a particular interest on the probabilities of the latent indicators being 0 or ± 1.

The main difference with our model, apart from the integration between the two platforms of data, which this work has no notion of, is the introduction of a regression on biological conditions, and subsequently further latent indicators in the model, that can be interpreted as indicators for differential expression. On the other side the POE approach could be described as a multi-step procedure: firstly, they get the probabilities of expression scores, which are related to the expression of a particular gene in a particular tumor; then they post-process these scores under different tumors, for example getting inference about differential expression, by simply comparing these scores. Incidentally, I could also make inference on scores for particular genes for particular tumors, though it is not the priority of my model.

Let's have a more precise look at how this post - processing procedure works.

Molecular profiles can be easily defined through the expression variable e_{gt}, restricting the number of genes from G to a smaller list that is considered interesting for their purposes. Otherwise the number of possible profiles based on the expression of all the genes in the data set would be too large and it would not allow reliable inference with data sets typically unbalanced between the large number of variables and the small amount of observations. If, for example, it is broadly known that gene A, B and C are interesting for the study, the 27 molecular profiles can be structured in the following way:
$\{(-1,-1,-1),(-1,-1,0),$
$(-1,0,0),...,(1,1,1)\}$.
However, these profiles are coarse classifications and they could be later refined, when scientists acquire additional evidence. Profiles defined in such a way are easy to be interpreted biologically, are independent from the array used to measure, as long as every array measures the same set of genes, and is independent from the algorithm that assigns tumors to profiles.
Although I will not go into detail for the post - processing procedure they use to cluster tumors according to molecular profile, I wanted to list the probabilities they use in this second step of their model.
Under the assumption of the genes being conditionally independent, they easily determine the joint probability of a tumor belonging to given profiles,

$$p(e_{1t},...,e_{Gt}|\theta) = \prod_g (p_{gt}^-)^{I_{\{e_{gt}=-1\}}} (p_{gt}^+)^{I_{\{e_{gt}=1\}}} (1-p_{gt}^- - p_{gt}^+)^{I_{\{e_{gt}=0\}}} \tag{2.6}$$

where I is a binary indicator and θ the set of all unknown parameters.
Another important measure, used to cluster similar molecular signatures, is the probability that two tumors show the same profile over a subset of genes G_0, that is, exploiting the conditional independence of e_{gt}'s,

$$p(t,t',G_0) = \prod_{g\in G_0} \{p_{gt}^- p_{gt'}^- + p_{gt}^+ p_{gt'}^+ + (1-p_{gt}^- - p_{gt}^+)(1-p_{gt'}^- - p_{gt'}^+)\} \tag{2.7}$$

On the other side, if we move our attention from tumor patterns across genes to gene patterns across tumors, rather interesting is the probability that a gene is normally expressed in all tumors, or is varying within the noise's range, and the expected number of tumors for which gene g is normally expressed. This probability is $p_g^0 = \prod_t (1 - p_{gt}^- - p_{gt}^+)$, while the number of such tumors $n_g = \sum_t (1 - p_{gt} - -p_{gt}+)$.
Clustering criteria in the post - processing stage, designed to find interesting subtypes, could use other two measures, easily computable with the model. The first one is the following,

$$p_{n^-,n+}(g) = p(e_{g1}, ..., egT \in E(n^-, n^+)|\theta) \tag{2.8}$$

the probability of a gene being under - expressed in n^- tumors and over -expressed in n^+; specifically, $E(n^-, n^+)$ is the set of T - dimensional vectors with values in $-1, 0, 1$ with n^- entries equal to -1 and n^+ equal to 1.
Lastly,

$$p(g, g') = \prod_t \{p_{gt}^- p_{g't}^- + p_{gt}^+ p_{g't}^+ + (1 - p_{gt}^- - p_{gt}^+)(1 - p_{g't}^- - p_{g't}^+)\} \tag{2.9}$$

the probability that genes g and g' express equally across all tumors. As we had previously stated, we will not be following this way of clustering subtypes of tumors according to their molecular profile, because our model focuses on claiming differential expression among fixed tumors's subtypes. Though, potentially, if we got rid of regressions on the clinical covariate indicating the subgroup our patients belong to, we could work with our method on a sample of unclassfied tumors, compute the same probabilities as the ones listed above, integrated with the information on copy number variation, and subsequently cluster tumors with similar procedures as the ones in Parmigiani et al's work. I want now to introduce the second technique that enabled our integration model to be designed: the array comparative genomic hybridization.

2.3 Array comparative genomic hybridization

Heterogeneity among human beings, induced by genetic modifications, is the basis for natural selection and evolution. One of this mutation is the so called copy number variation, I have introduced in the beginning of the chapter.
DNA sequencing helped scientists to find copy number variations in 5% of the human genome, but these are common mutations among all individuals. Since 2005 people have been noticing that some of these mutations were varying in the human population, meaning that they could be found in some individuals and not in others, and they were playing an important role in the progress of certain illnesses (21), (9). In 2008, it was estimated that approximately 0.4% of the genomes of unrelated people differ with respect to copy number (25).
Biologically, the choice of this data in my model is driven by the following fact: it has been largely proved that genetic diseases are caused by such mutations (31), although it is not yet clear which portion of the disease is caused by copy number variation. It is therefore very interesting to investigate the contribution of CNV to other common, complex diseases, such as cancer, and the impact of these mutations on gene expression.
Comparative Genomic Hybridization (CGH) has emerged as a dominant technique for this purpose (24), especially when combined with microarrays. The resulting arrayCGH techniques (32), (39), (40) use microarrays consisting of thousands or millions of genomic targets or "probes" that are spotted or printed on a glass surface. These probes usually span the whole genome with a resolution of the order ranging from 1MB (one million base pairs) for BAC (bacterial artificial chromosome) to 50-100 kb (kilo base pairs). A DNA *test* sample of interest is labeled with a dye (say Cy3) and then mixed with a diploid *reference* sample labeled with a different dye (say Cy5). The combined sample is then hybridized to the microarrays and intensities of both colors are measured through an imaging process. The quantity of interest is the $\log_2$ ratio of the two intensities for each probe. The collection of the intensity ratios then contain useful information about genome-wide

changes in copy numbers. Since in the reference the copy number of each DNA fragment is always two, the intensity ratio is determined by the copy number of the DNA in the test sample. If the copy number is also two, then the $\log_2$ intensity ratio equals zero. If there is a single copy loss in the test sample, the theoretical ratio is $\log_2 1/2 = -1$ assuming all the cells in the test sample lost one copy of the DNA fragment. If there is a single copy gain, the theoretical ratio is $\log_2 3/2 = 0.58$. Multiple copy gains are called *amplifications*, and the corresponding theoretical intensity ratios are $\log_2 4/2$, $\log_2 5/2$, etc. When both copies are lost, the theoretical ratio is $-\infty$ and a large negative value is usually observed in experiments.

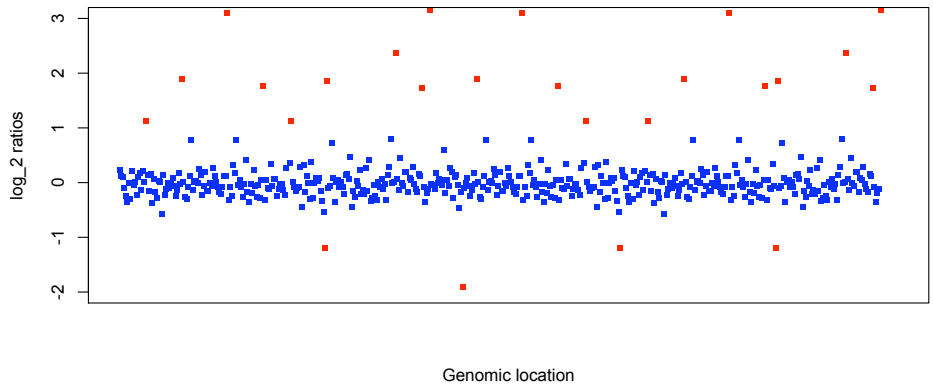

Figure 2.1: An illustration of arrayCGH sample data for one chromosome with a region of copy number losses and one with copy number gains. The X-axis is the chromosome location and the Y-axis is the log2 ratio copy number intensities. The blue squares correspond to log2 ratio concentrated on 0 while the red ones to log2 ratios either above 1 or below -1

Investigators have been interested in finding out

- which regions of the DNA have abnormal copy numbers;
- how many copies are lost or gained.

Regarding these aims, a number of methods have been proposed (Fridlyand et al., 2004; Guha et al., 2008; Hodgson et al., 2001; Olshen et al., 2004; Pollack et al., 2002). More recently, Baladandayuthapani et al. (2011) provided a thorough review of the arrayCGH technology and proposed bayesian mixture models, using functional data, that base statistical inference solely on posterior probabilities of CNA. One novel aspect of their approach is its ability to detect recurrent CNAs across multiple samples.

2.3.1 Modeling copy number variation: an example

I would like to review in this section one important model Dr. Guha et al developed in their paper of 2008, a starting point for our project of integrating the two different data. It represents one of the latest and most interesting *calling* methods for the analysis of aCGH data.

Calling methods, opposed to segmentation methods as the one described by Dr. Baladandayuthapani et al, model this data at probe level and call the unobserved state of each probe loss, gain or neutral.

Specifically, in their model, Guha et al. described a Bayesian Hidden Markov Model that takes into account dependence between neighboring probes by specifying the copy number states as the latent states in the HMM scheme.

Similarly to our approach, they did not restrict themselves to a segmentation method, but gave a specification of the likelihood which allows the use of decision rules, based on posterior probabilities, aiming to detect copy number aberrations.

Generally hidden Markov models assume the observed data, y_k come from a distribution depending on an unobserved variable, called hidden state, taking values in a subset of the integers. The hidden variable is normally indexed by either time or space and characterizes the *regime* in which the sampling process is at any point k of time or space. A further assumption of HMM's is a Markov dependence over the latent state variables and, hence,

$$p(s_k|s_1, s_2, ..., s_{k-1}) = p(s_k = j|s_{k-1} = i) = p_{ij}, \tag{2.10}$$

the process for s_k depends on the past realizations of $\mathbf{y}$ and $\mathbf{s}$ only through s_{k-1}, where λ_{ij} is a generic element of the transition matrix $\mathbf{\Lambda}=(p_{ij})$, with the vector of stationary probabilities $\boldsymbol{\pi}$ satisfying $\boldsymbol{\pi}$'$\mathbf{\Lambda}=\boldsymbol{\pi}$'.

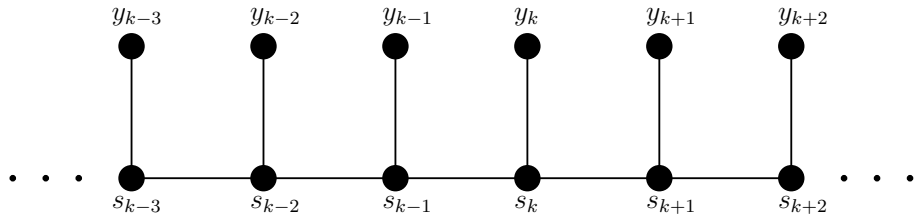

Figure 2.2: Graphical representation of HMM dependence. The distribution of y's conditionally to each state depends on all the other realizations only through the states to which it is connected by an edge.

There is only one element missing from the description of HMM's, that is the distributional assumption on y_k. Generally this could be expressed in the following way,

$$y_k|s_k \overset{\text{indep}}{\sim} \pi(\alpha_{s_k}, \beta s_k) \tag{2.11}$$

where α and β are the scale and shape parameters of a generic parametric distribution. In Guha et al.'s work $k = 1, , , n$, y_k denote the normalized log_2 ratios observed at probe k and s_k is a copy number state, associated with each probe, thus spatially indexed, that takes values in $\{1, 2, 3, 4\}$. These numbers respectively represent a single copy number loss, the copy neutral state, a single copy gain, and a generic amplification (i.e. multiple copies gain), at each mapped probe.

Denoted with μ_j the expected log_2 ratio of all probes for which $s_k = j$, the biological interpretation of copy number states allows us to postulate the ordering $\mu_1 < \mu_2 < \mu_3 < \mu_4$.

The conditional distribution of the normalized log_2 ratios is

$$y_k|s_k \overset{\text{indep}}{\sim} N(\mu_{s_k}, \sigma^2_{s_k}) \tag{2.12}$$

where $k = 1, ..., n$. Besides, the dependence of neighboring probes is modeled using a hidden Markov model (34), so that $\lambda_{ij} = p(s_j|s_i)$ are the elements of the transition matrix I mentioned above.

The elements of such matrix are assumed to be strictly positive, thus the process is an aperiodic and irreducible Markov process, with four positive recurrent states. The transition matrix Λ has a unique stationary distribution $\pi_\Lambda = (\pi_\Lambda(1), \pi_\Lambda(2), \pi_\Lambda(3), \pi_\Lambda(4))$, with $\pi_\Lambda(i)$ strictly positive for each i, and the copy number state of the first probe is also distributed as π_Λ.

All these assumptions lead to a unique joint likelihood of a given sequence $s_1, ..., s_n$ and the model is completed with the specification of the priors for the hyperparameters: the ixj elements of the transition matrix, means $\{\mu_1, ..., \mu_4\}$ and variances $\{\sigma_1^2, ..., \sigma_4^2\}$. Informative priors are chosen for biological reasons and posterior samples of the parameters are generated by a mix of Metropolis-Hastings algorithm, a stochastic version of the forward-backward algortihm and Gibbs sampling.

Posterior inference and classification among different aCGH profiles focus on the marginal posterior probability $p(s_1, ..., s_n|y_1, ..., y_n)$. At each iteration of the Markov Chain Monte Carlo algorithm, the generated states could potentially be classified into profile, such as *focal aberrations* (19), amplifications or whole chromosomal changes, and the classification scheme results into a Bernoulli variable for each MCMC step and type of genomic aberration.

2.4 Integration

The main aim of our integration method for these two different measurements at DNA and RNA level, is the following: expression microarrays measure RNA expression which, by the central dogma of molecular biology, are resulted from the transcription of DNAs. Naturally, CNVs directly impact the intensities of RNA expression in that more copies of

the DNA should lead to higher levels of RNA expression. It is therefore of great interest to study the interaction between aCGH and microarray expression data.
Fewer methods are available for the integrated study of these two types of data.
Pollack et al was one of the earliest to investigate the direct association between the two data in breast cancer cell lines and tissue samples, and his approach is based mainly on descriptive statistics (33). Van Wieringen and Van de Wiel (44), attempting to mitigate the high noise in the raw expression measurements of the DNA and RNAs, proposed a sampling model for mRNA expression measurements based on estimated probabilities of copy number variations. They subsequently developed nonparametric adaptive tests to study whether the estimated copy number variations in the DNA level would induce differential gene expression at the RNA level. More recently Choi et al (12) presented a double-layered mixture model (DLMM) (*DLLM*) that directly modeled segmental patterns in the copy number data to produce CNA profiles, and simultaneously scored the association between copy number and gene expression data.
The DLMM assigned high scores to elevated or reduced expression measurement only if the expression changes are observed consistently across samples with copy number aberration. Similarly to the DLMM, we describe in this paper a Bayesian Mixture Model that converts the noisy raw intensity measurement of the DNA and RNAs into probability of expression, which are subsequently modeled as latent parameters. Thus the integration of the two platforms is realized by joint modeling the probabilities of expression through a probit regression where we include a clinical covariate leading to a complete different notion of differential *behaviour*. Differential behaviour in our case study does not refer to healthy vs sick, but it refers to differential behaviour of genes in subtypes of cancer, indicated by the clinical covariate I mentioned above.
Moreover, our aim is not only to evaluate the relative contribution of large genetic variants, as CNVs, to gene expression but also to be able to make inference using both differential expression of the genes and differential copy number variations of the same set of genes, all measured on a common sample of patients.

I would suggest to refer directly to their paper for the full analytical specification of the model.

3

A Bayesian framework for the integration of genomic platforms

Clinical studies for cancer patients often collect various types of genomic data, typically with the aim of systemically examine the origine and dynamics of the diseases. An important premise is that by integrating different types of genomics data, such as DNA copy number and RNA expression data, we will gain more knowledge about the underlying biological process. For example, a high or low correlation between a copy number aberration (CNA) for a gene marker and its abnormal RNA expression would indicate different disease mechanism and therefore different treatment selections.

We propose a Bayesian model-based framework to integrate different types of genomics data. The main features of the proposed framework are the following.

- We assume a mixture model ((30)) as sampling distribution for the observed expression data. The mixture implicitely defines a latent trinary indicator for differential expression status of each gene. By operating on the latent indicators, we probabilistically remove the high noise in the original data.

- We integrate diverse types of genomics data through a regression based on the latent variables in the mixture models. The regression formalizes the biologic belief that

CNA could impact gene expression. It also allows for easy incorporation of other covariates.

- The Bayesian model-based framework allows easy consideration of either data type alone or both. As in any Bayesian approach, posterior inference includes a probabilistic description of all uncertainties. This is important for coherent multiplicity control.

Integration models borrow information across multiple genomic platforms, using data that is measured on the same patients and genes. For illustration purpose, we consider two of the most commonly used genomic platforms: arrayCGH (aCGH) that measures DNA copy numbers and expression microarrays that measures RNA expression.

A new strategy in targeted cancer therapeutics research is to develop effective biological agents that specifically target a group of disease markers, as opposed to the traditional symptom-drive therapeutic strategies. The success of this new strategy depends on accurately understanding the underlying biological process of the disease.
Targeted therapy is being developed to differentiate prognosis for cancer subtypes. For example, treatments for breast cancer now depend on the expression of marker genes HER2 and ER.
The conception of our model started from clinical questions arising from such considerations and it is addressed specifically to breast cancer patients.

We analyze data from a breast cancer study that enrolled 122 patients from two disease subgroups, ER+/ HER2+, and triple negative (TN). Cancer cells for patients in the ER+ group tested positive for estrogen receptors, a protein related to hormone and regulation of gene expression. Patients are classified as HER2+ when the tumoral cells test positive for epidermal growth factor receptor 2. Finally TN patients lack three “receptors” in their cancer cells: ER, HER2, and progesterone receptors.
TN is chosen as baseline since it is clinically characterised as more aggressive and less responsive to standard treatment and associated poorer overall patient prognosis with

respect to the other two (16), (14).

Using the same samples from each of the 122 patients, we have performed experiments to obtain aCGH copy number data and RNA expression data for each sample. We used the Agilent Human 4×44K arrays for copy number variation and Affimetrix HG-U133A arrays for gene expression. After data pre-processing, we obtained log2 ratios between the sample copy number and the reference number for each of the 22,944 probes for aCGH, and expression quantifications for 11,306 genes for the microarray data. We then mapped 22,944 probes to the 11,306 genes and recorded the match between probe id's on the aCGH and the gene id's on the microarrays.

To summarize, on the microarray, we obtain 122 expression measurements for each gene; on the aCGH, we obtain 122 log2 ratios for each probe, with m_g probes corresponding to the same gene g, $g = 1, \ldots, 11,306$. Note that we did not combine the probe-level measurements on the aCGH into gene-level summaries as a convention. Typically researchers analyze probe-level intensties for CNA's due to the high correlation between adjacent probes (but not adjacent genes).

3.1 Probability Model

3.1.1 Sampling model for w and y

The proposed model integrates data across two different platforms: aCGH and gene expression microarray. The observed data are w_{bt} (the log2 ratio from aCGH) and y_{gt} (the expression level from microarray), for probe b belonging to gene g for sample t, $b = 1, ...B$, $g = 1, ..., G$, and $t = 1, ...T$. We write $\{b \in g\}$ for the set of aCGH probes that correspond to gene g. For each gene g and sample t, the observed continuos data are $\{(w_{bt})_{\{b \in g\}},\ y_{gt}\}$.. They are both assumed to be continuos since aCGH data, which would be intuitively considered discrete with point masses at $0, 1, \log_2 3, ...$, deviates from the theoretical values, due to the fact that not all the cells in a sample have the same copy

number aberration and due to other genetic contamination such as cross-hybridization. Below, we define two mixtures as sampling models for w and y.

The mixtures implicitely define two sets of latent indicators e^w_{bt} and e^y_{gt} that represent differential expression status of the DNA and RNA, respectively. We then integrate the two models by contructing a prior probit regression that links the latent variables from both platforms. A summary of the model is given in Figure 4.2. Below we will present the models in full detail.

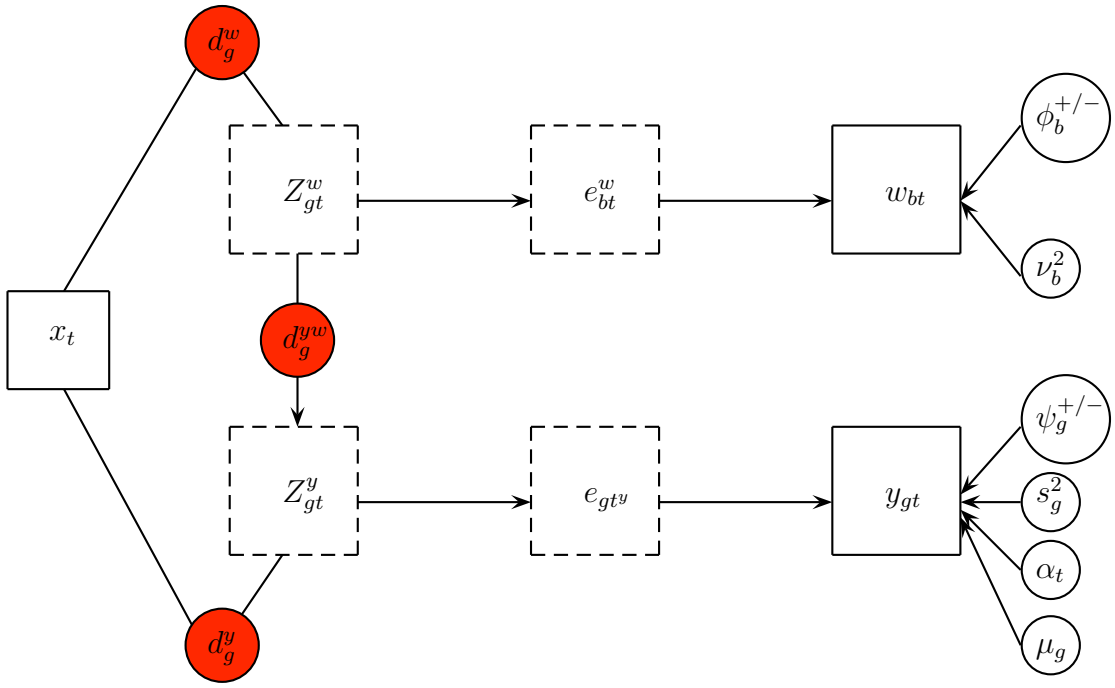

Figure 3.1: A graphical representation of the proposed Bayesian model.

3.1.2 Sampling model for w and y

We use a mixture model ((30)) to introduce a trinary latent indicator variables for the CNA state for each probe and the expression level state for each gene. Specificaly, let e^w_{bt} take values in the set $\{-1, 0, 1\}$, respectively corresponding to the copy-loss ($<$ 2 copy

number), copy-neutral (= 2 copy number), and copy-gain (> 2 copy number) states. Similarly, e^y_{gt} take values in the set $\{-1, 0, 1\}$, respectively corresponding to the under-, normal-, and over-expression states.

The latent indicators e^w_{bt} and e^y_{gt} define mixture models for copy number log2 ratios w_{bt} and for gene expression data y_{gt}:

$$f_w(w_{bt}|e^w_{bt}) =_d \begin{cases} U(-\phi^-_b, 0) & \text{if } e^w_{bt} = -1 \\ N(0, \nu^2_b) & \text{if } e^w_{bt} = 0 \\ U(0, \phi^+_b) & \text{if } e^w_{bt} = 1 \end{cases} \tag{3.1}$$

$$f_y(y_{gt} - \mu_g - \alpha_t|e^y_{gt}) =_d \begin{cases} U(-\psi^-_g) & \text{if } e^y_{gt} = -1 \\ N(0, s^2_g) & \text{if } e^y_{gt} = 0 \\ U(0, \psi^+_g) & \text{if } e^y_{gt} = 1 \end{cases} \tag{3.2}$$

In the integrated model, the mixture model for gene expression data y_{gt} includes a gene effect μ_g and a sample effect α_t. This is not the case in the mixture model for aCGH data w_{bt}.

The main reason is because w_{bt} is already a log ratio between the cancer sample copy number and the reference sample copy number and therefore the corresponding effects should have canceled out by taking the ratio.

The sampling model are indexed by ν^2_b and s^2_g representing normal ranges of variability in the observed measurements w_{bt} and y_{gt}. The parameters $\phi^{+/-}_b$ and $\psi^{+/-}_g$ define the tail overdispersion with respect to normality, associated with copy losses or gains for aCGH and under- or over-expression for microarrays.

3.1.3 Latent probit scores, regression and differential expression

The proposed model uses several latent indicators. I have already introduced in the previous section the indicators e's which are the latent indicators in the POE sampling models (3.1) and (3.2).

Latent trinary indicators d^w, d^y and d^{yw} will be used to model differential expression,

differential copy number variation and differential correlation across biological conditions of interest.

Anticipating the integration of both platforms in a regression model, we further introduce latent Gaussian variables z^w_{bt} and z^y_{gt} to define a probit scores for the trinary indicators e^w_{bt} and e^y_{gt}. The model here deviates from the POE model presented in the previous chapter, since prior probabilities for the latent trinary scores are implemented on a probit scale (11) to allow the integration of the two submodels by a simple normal regression at the level of these probit scores.

Specifically, define

$$e^w_{bt} = \begin{cases} -1 & \text{if } z^w_{bt} < -1 \\ 0 & \text{if } -1 \leq z^w_{bt} \leq 1 \\ 1 & \text{if } z^w_{bt} > 1 \end{cases} \quad \text{and} \quad e^y_{gt} = \begin{cases} -1 & \text{if } z^y_{gt} < -1 \\ 0 & \text{if } -1 \leq z^y_{gt} \leq 1 \\ 1 & \text{if } z^y_{gt} > 1 \end{cases} \tag{3.3}$$

Next we introduce the model element to represent differential CNA. We achieve this by introducing a prior for z^w_{bt} that allows for different CNAs across different conditions, in our case across different subtypes of breast cancer. We use two sets of indicators. Indicators $x_t \in \{0, 1\}$ describe the known biologic condition of sample t. We refer to $x_t = 0$ as the reference condition. Gene-specific latent indicators $d^w_g \in \{-1, 0, 1\}$ characterize differential CNA for probes in gene g, relative to $x_t = 0$. In the later data analysis we will use only two conditions. In general, for multiple conditions, x_t might be defined as a vector of binary indicators and d^w_g be a corresponding vector of indicators for differential CNA. For simplicity we continue the discussion with the single binary x_t. We assume

$$z^w_{bt} | z^w_b \sim N(z^w_b + x_t c_{d^w_g}, \sigma^2_a) \tag{3.4}$$

Here, z^w_b is a probe-specific mean that defines a baseline CNA status (e.g., for a reference subtype) and c_{-1}, c_0, c_1 are mean shifts under under-, regular and over-expression.

$$d^w_g = \begin{cases} -1 & p = p_1 \\ 0 & p = p_2 \\ 1 & p = p_3 \end{cases} \tag{3.5}$$

The integration of the two platforms is implemented as a regression with the probit scores. Setting up the regression first requires to match the unit of measurement for aCGH and RNA data. We define a gene-level score for the aCGH data, $z^w_{gt} = \frac{1}{m_g}\sum_{b\in g} z^w_{bt}$. Keeping in mind that there is a natural biological causal relationship between a copy number change to a DNA and differential gene expression for the corresponding RNAs, we write the integration as a regression of the RNA probit score z^y_{gt} on the CNA probit score z^w_{gt}. Finally the model includes a regression on the sample-specific condition x_t to represent differential gene expression. Similar to d^w_g we use a gene-specific trinary indicator $d^y_g \in \{-1, 0, 1\}$ for differential gene expression. A third set of indicators $d^{yw}_g \in \{-1, 0, 1\}$ describes correlation of CNV and RNA for gene g.

$$z^y_{gt}|z^w_{gt} \sim N(\alpha_g + x_t c_{d^y_g} + z^w_{gt}\lambda_{d^{yw}_g}, \tau_1^2), \tag{3.6}$$

where c_{-1}, c_0, c_1 are mean levels for under-, normal and over-expressed genes.

$$d^y_g = \begin{cases} -1 & p = p_1 \\ 0 & p = p_2 \\ 1 & p = p_3 \end{cases} \tag{3.7}$$

In summary we defined two trinary mixture probit models for latent scores z^w_{gt} and z^y_{gt} in the CNV and RNA submodel, respectively, and a regression of the RNA probit scores z^y_{gt} on the CNV probit scores. The mixtures are induced by latent indicators d^w_g and d^y_g of under-, normal and over-expression. A third set of indicators, d^{yw}_g, writes the regression as a mixture of three terms corresponding to negative, no and positive correlation of DNA and CNV scores.

3.1.4 Priors

3.1.4.1 CNA submodel

The prior on the reference scores z^w_b formalizes the dependence of adjacent probes. We assume Markov dependence. Assuming that the index b is ordered according to locus

proximity on the chromosome, the dependence across adjacent probes is described as

$$z_1 \sim N(0,1) \tag{3.8}$$

$$z_b^w | {z_{b-1}}^w, \beta_{b-1} \sim N(\beta_{b-1} z_{b-1}^w, \tau^2) \tag{3.9}$$

for $b \in \{2, ..., B\}$ The parameters $\beta = (\beta_1, ..., \beta_{b-1})$ are interpreted as partial correlation coefficients, defining the strength of dependence between $\log_2$ ratios associated with clones that are adjacent on the chromosome. For β_b we assume *a priori*

$$\beta_b \sim N(\sqrt{1-\tau^2}, \sigma^2)$$

for $b \in \{1, 2, ..., B-1\}$, with $\tau^2 < 1$, ensuring the the marginal variance of z_b's remains bounded.

We complete the model construction with conditionally conjugate priors for the remaining parameters. For the parameters of the sub-model related to CNV we assume $\nu_b^{-2} \sim G(a_\nu, b_\nu)$ for a normal range of CNA scores, $1/\phi_b^{+/-} \sim G(a_{\phi^{+/-}}, b_{\phi^{+/-}})$ for the range of low and high CNA scores, and $\sigma_a^{-2} \sim G(a_\sigma, b_\sigma)$ for the prior variance on probe and sample-specific specific scores z_{bt}^w. The prior on the level c_r, $r = -1, 0, 1$ is

$$c_{d_g^w} \sim \begin{cases} N(-k, \sigma_1^2) & \text{if } d_g^w = -1 \\ N(0, \sigma_2^2) & \text{if } d_g^w = 0 \\ N(k, \sigma_1^2) & \text{if } d_g^w = 1 \end{cases}$$

with σ_1 much larger than σ_2 and $k > 0$.

3.1.4.2 RNA submodel and integration model.

Finally we assume conditionally conjugate priors for the gene and slide specific effects in the RNA submodel. We use $\mu_g \sim N(\theta_\mu, \sigma_\mu^2)$, and $\alpha_t \sim N(0, \sigma_\alpha^2)$, subject to $\sum \alpha_t = 0$.

We assume $s_g^{-2} \sim G(a_s, b_s)$, for the range of variability in normal mRNA expression, and $1/\psi_g^{+/-} \sim G(a_{\psi^{+/-}}, b_{\psi^{+/-}})$, for the tail over-dispersion parameters.

For the gene-specific offset in the integration model we assume $\alpha_g \sim N(0,1)$. A normal prior is assumed for the shift c_1, c_0 and c_{-1} for over, normal and under-expression,

$$c_{d_g^y} \sim \begin{cases} N(-k, \sigma_1^2) & \text{if } d_g^y = -1 \\ N(0, \sigma_2^2) & \text{if } d_g^y = 0 \\ N(k, \sigma_1^2) & \text{if } d_g^y = 1 \end{cases}$$

A similar prior is used for the strength of correlation,

$$\lambda_{d_g^{yw}} \sim \begin{cases} N(-k, \sigma_1^2) & \text{if } d_g^{yw} = -1 \\ N(0, \sigma_2^2) & \text{if } d_g^{yw} = 0 \\ N(k, \sigma_1^2) & \text{if } d_g^{yw} = 1 \end{cases},$$

3.2 Simulation and Posterior Inference

3.2.1 Simulations

We carry out two simulation studies to validate inference under the model.

The first simulation investigates inference when the simulation truth is different from the assumed analysis model. We first generate two matrices of gene expression data (y_{gt}) and copy number variation (w_{bt}) measurements, respectively of dimensions $G \times T$ and $B \times T$, with $B = 2000$, $G = 1000$ (exactly two probes per gene) and $T = 50$. The clinical covariate x_t is set to be 1 for the first 10 patients (TN group) and 0 for the remaining 40 patients. Sample and gene effects α_t and μ_g for the sampling model are generated as $\alpha_t \sim N(0, \sigma_\alpha^2)$ subject to $\sum \alpha_t = 0$ and $\mu_g \sim N(\theta_\mu, \sigma_\mu^2)$. We set $e_{bt}^w = e_{gt}^y = 0$ for all genes, samples and probes and generate simualted gene expression values $y_{gt} \sim N(\alpha_t + \mu_g, s_g^2)$ and copy number variations $w_{bt} \sim N(0, \nu_b^2)$ for all samples, genes and probes, except for the last 50 genes for the first 10 samples, i.e., $t = 1, \ldots, 10$, $g = G - 49, \ldots, G$ and $b = B - 99, \ldots, B$. Those measurements were generated as $y_{gt} \sim U(-10, 0)$ and $w_{bt} \sim U(-2, 0)$, $t = 1, \ldots, 10$ and $g = G - 49, \ldots, G$, $b = B - 99, \ldots, B$.

The second simulation investigates inference under the assumed model, assuming the data arises from a similar mixture, allowing for some minor deviations from the assumed

analysis model. The aim is to verify that the simulation truth λ_g can indeed be recovered under reasonable assumption on sample size and parameters.

We set $\lambda_g = 2$ for $g = 1, \ldots, 50$ and $\lambda_g = 0$ for the remaining genes, $g = 51, \ldots, G$.

We start with the simulation of probe-specific average CNV scores

$$z_1 \sim N(0,1),\ z_b \sim N(\beta_{b-1} z_{b-1}, \frac{1}{4}),\ z_{bt} \sim N(z_b + x_t c_g)$$

for $b = 1, 2, \ldots, B$ and $t = 1, \ldots, 50$. The parameters for these distributions are generated as $\beta_b \sim N(\frac{3}{4}, \frac{1}{16})$ for $b = 1, \ldots, B-1$, $\sigma_a^{-2} \sim G(5,1)$, $c_g \sim N(1, \frac{1}{9})$ for the first 100 genes, $g = 1, \ldots, 100$, and $c_g \sim N(0, \frac{1}{400})$ for the remaining genes $g = 101, \ldots, 1000$.

To generate the simualted gene expression scores we first compute gene-specific $z^w_{gt} = \frac{\sum_{b \in g} z_{bt}}{m_g}$, for $m_g = 2$ probes per gene. We then use the data integration regression

$$z^y_{gt} \sim N(\alpha_g + x_t c_{d^y_g} + \lambda_g z^w_{gt}, 1)$$

to generate simulated probit scores for the RNA submodel. The paramters for this model are generated as $c_{d^y_g} \sim 0.3N(\frac{4}{5}, \frac{1}{100}) + 0.7N(0, \frac{1}{100})$ and $\alpha_g \sim N(0,1)$.

Once the latent scores are generated, we use the assumed sampling model to generate gene expression w_{gt} and CNV measurements y_{gt}. Hyperparameters for the sampling model are generated as $\phi_b^{+/-} = \pm 2$, $\nu_b^2 = \frac{1}{100}$, $b = 1, \ldots, B$ and $\psi_g^{+/-} = \pm 10$ and $s_g^2 = 1$ for $g = 1, \ldots, 1000$.

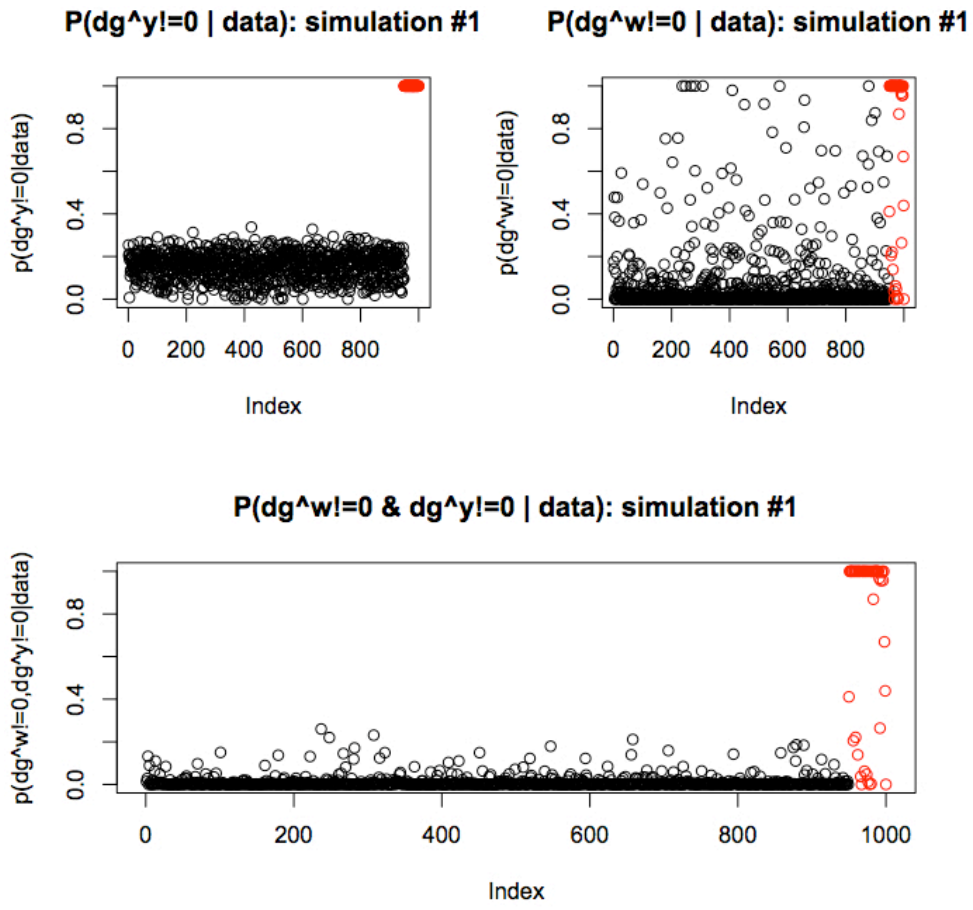

Figure 3.2: Posterior probabilities of differential gene expression, differential CNV and differential joint behaviour after simulation #1.

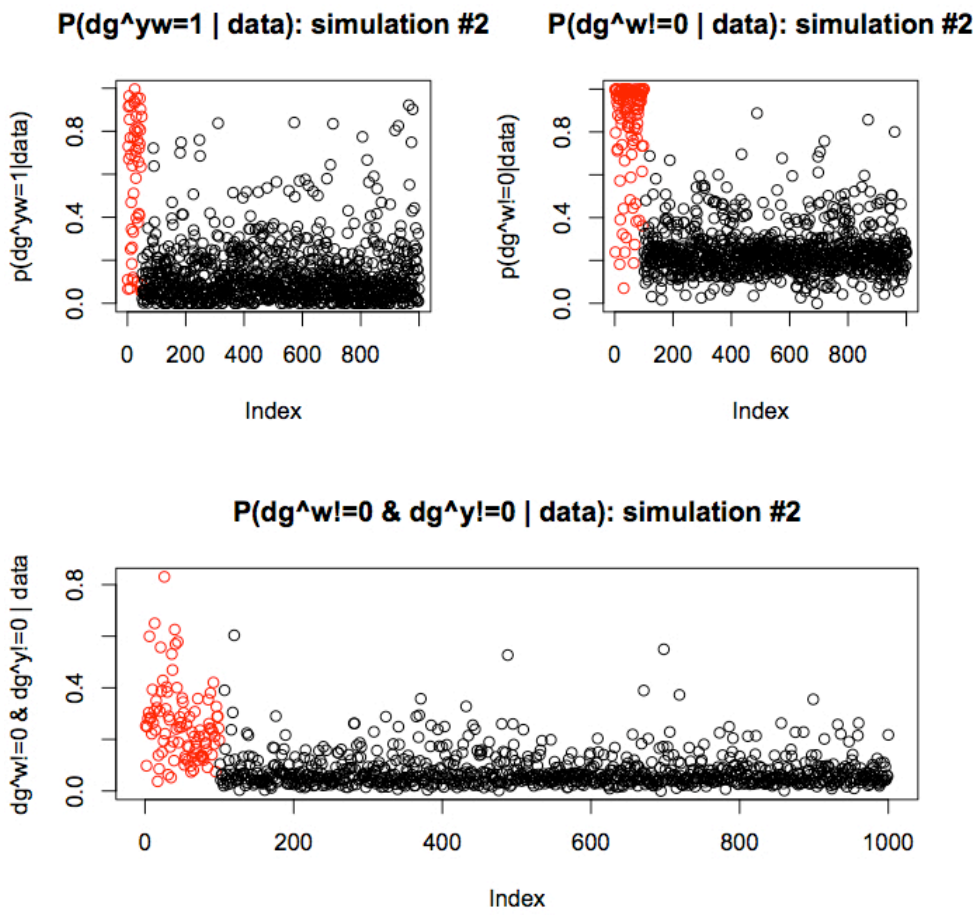

Figure 3.3: Posterior probabilities of positive interaction between the two platforms, differential CNV and differential joint behaviour after simulation #2.

As we can see from Figure 2, as we expected, posterior probabilities of differential gene expression, differential CNV and subsequently differential joint behaviour for the last 50 genes are among the highest.

Same thoughts for Figure 3, related to simulation 2, where we can focus on the first two plots and once more, as we expected, posterior probabilities of positive interaction between platforms for the first 50 genes and posterior probabilities of differential CNV for the first 100 genes are among the highest.

3.2.2 Posterior MCMC

Posterior inference in the proposed model is carried out using MCMC simulations; in particular, posterior simulations from the joint distribution of the parameters given the data is implemented by a Gibbs sampling scheme, iterating over all full conditional posterior distributions reported below. We start with a transition probability that jointly updates e^w_{bt} and z^w_{bt}. This is done by first sampling e^w_{bt} from its conditional posterior , conditional on the observed data and all currently imputed parameters except z^w_{bt}, and thus z^w_{bt} is sampled from the complete conditional posterior, now conditioning also on the already imputed e^w_{bt}.

$$e^w_{bt}|\{-z^w_{bt}\} \sim \begin{cases} -1 & p = p_- \\ 0 & p = p_0 \\ 1 & p = p_+ \end{cases} \tag{3.10}$$

where $\{-z^w_{bt}\} = \{z^w_{1t}, ..., z^w_{b-1,t}, z^w_{b+1,t}, ..., z^w_{Bt}\}$ and the probabilities are given by

$$p_- \propto \Phi(-1|\mu_1, \sigma_1^2)(\frac{1}{\phi_b^-})I(w_{bt} \in [-\phi_b^-, 0]) \tag{3.11}$$

$$p_0 \propto (\Phi(1|\mu_1, \sigma_1^2) - \Phi(-1|\mu_1, \sigma_1^2))\phi(w_{bt}|0, \nu_b^2) \tag{3.12}$$

$$p_+ \propto (1 - \Phi(-1|\mu_1, \sigma_1^2))(\frac{1}{\phi_b^+})I(w_{bt} \in [0, \phi_b^+]) \tag{3.13}$$

in turn properly normalized.

The conditional posterior probabilities for e^w_{bt} are proportional to the respective prior probabilities times the likelihood factor corresponding to w_{bt}, where the prior probabilities are the appropriate normal quantiles and the evaluation of the likelihood needs to include indicators for the finite support of the uniform kernels in (3.1).

z^w_{bt} is sampled form a truncated Normal distribution that guarantees the latent probit score to be consistent with the relative POE indicator of copy number aberration.

$$z^w_{bt}| \cdot \cdot \sim N(\mu_1, \sigma_1^2)[I(z^w_{bt} < -1)I(e^w_{bt} = -1) + \cdots] \tag{3.14}$$

where

$$\mu_1 = (\frac{\lambda_{d_g^{yw}}}{m_g}(z_{gt}^y - x_t c_{d_g^y}) - \frac{\lambda_g^2}{m_g^2}\sum_{-b} z_{bt}^w + \frac{1}{\sigma_a^2}(z_b^w + x_t c_{d_g^w})))/(\frac{\lambda_g^2}{m_g^2} + \frac{1}{\sigma_a^2}) \tag{3.15}$$

$$\sigma_1^2 = (\frac{\lambda_{d_g^{yw}}^2}{m_g^2} + \frac{1}{\sigma_a^2})^{-1} \tag{3.16}$$

and $m_g = \#\{b \in g\}$.

z_{gt}^w is simply defined to be the mean of the latent probit scores within the same gene.

$$z_{gt}^w = \frac{1}{m_g}\sum_{b\in g} z_{bt}^w \tag{3.17}$$

The parameters of the mixture in (3.1) are sampled from the updates Gamma distributions, since we assumed conjugate priors for them.

$$\phi_b^{-^{-1}}|\cdot\cdot\sim G(\alpha_{\phi^-} + m_b, \beta_{\phi^-})I(\phi_b^- > \max(0, -\min_t\{w_{bt} : e_{bt}^w = -1\})) \tag{3.18}$$

where $m_b = \#\{t : e_{bt}^w = -1\}$.

$$\nu_b^{-2}|\cdot\cdot\sim G(\alpha_\nu + \frac{n_b}{2}, (1/\beta_\nu + \frac{\sum_{t=1}^T w_{bt}^2}{2}I(e_{bt}^w = 0))^{-1}) \tag{3.19}$$

where $n_b = \#\{t : e_{bt}^w = 0\}$.

$$\phi_b^{+^{-1}}|\cdot\cdot\sim G(\alpha_{\phi^+} + r_b, \beta_{\phi^+})I(\phi_b^+ > \max(0, \max_t\{w_{bt} : e_{bt}^w = 1\})) \tag{3.20}$$

where $r_b = \#\{t : e_{bt}^w = 1\}$.

The conditional posterior probabilities for d_g^w, i.e. the trinary indicator for differential copy number aberration, are proportional to the respective prior probabilities times the likelihood factor corresponding to z_{bt}^w, where the prior probabilities are defined in such a way that p_2 is greater than $p_1 + p_3$, to express our prior belief that it is more likely for a gene to show similar copy numer aberrations in the two subgroups of breast cancer.

$$d_g^w \sim \begin{cases} -1 & p \propto p_1\phi(z_{bt}^w|z_b + c_{-1}x_t, \sigma_a^2) \\ 0 & p \propto p_2\phi(z_{bt}^w|z_b + c_0 x_t, \sigma_a^2) \\ 1 & p \propto p_3\phi(z_{bt}^w|z_b + c_1 x_t, \sigma_a^2) \end{cases} \tag{3.21}$$

While the posterior probabilities for the parameters in equation (3.4) and (3.9) are respectively the updated Normal and Gamma distributions obtained through the conjugate prior assumptions.

$$c_{d_g^w}|\cdot\cdot\sim\begin{cases} N(\frac{\frac{\sum_{t=1}^{T}\sum_{b\in g}x_t(z_{bt}^w-z_b^w)}{\sigma_a^2}-\frac{k}{\sigma_1}}{\frac{1}{\sigma_1^2}+\frac{\sum_{t=1}^{T}\sum_{b\in g}{x_t}^2}{\sigma_a^2}},\frac{1}{\frac{1}{\sigma_1^2}+\frac{\sum_{t=1}^{T}\sum_{b\in g}{x_t}^2}{\sigma_a^2}}) & \text{if } d_g^w=-1 \\ N(\frac{\frac{\sum_{t=1}^{T}\sum_{b\in g}x_t(z_{bt}^w-z_b^w)}{\sigma_a^2}}{\frac{1}{\sigma_2^2}+\frac{\sum_{t=1}^{T}\sum_{b\in g}{x_t}^2}{\sigma_a^2}},\frac{1}{\frac{1}{\sigma_2^2}+\frac{\sum_{t=1}^{T}\sum_{b\in g}{x_t}^2}{\sigma_a^2}}) & \text{if } d_g^w=0 \\ N(\frac{\frac{\sum_{t=1}^{T}\sum_{b\in g}x_t(z_{bt}^w-z_b^w)}{\sigma_a^2}+\frac{k}{\sigma_1}}{\frac{1}{\sigma_1^2}+\frac{\sum_{t=1}^{T}\sum_{b\in g}{x_t}^2}{\sigma_a^2}},\frac{1}{\frac{1}{\sigma_1^2}+\frac{\sum_{t=1}^{T}\sum_{b\in g}{x_t}^2}{\sigma_a^2}}) & \text{if } d_g^w=1 \end{cases} \tag{3.22}$$

$$\sigma_a^{-2}|\cdot\cdot\sim G(a_\sigma+\frac{BT}{2},(1/\beta_\sigma+\frac{\sum_{t=1}^{T}\sum_{b=1}^{B}(z_{bt}^w-z_b^w-x_tc_{d_g^w})^2}{2})^{-1}) \tag{3.23}$$

$$z_1^w|\cdot\cdot\sim N((\frac{\beta_1 z_2^w}{\tau^2}+\frac{\sum_{t=1}^{T}(z_{1t}-x_tc_{d_g^w})}{\sigma_a^2})/(\frac{\beta_1^2+\tau^2}{\tau^2}+\frac{T}{\sigma_a^2}),\frac{1}{\frac{\beta_1^2+\tau^2}{\tau^2}+\frac{T}{\sigma_a^2}}) \tag{3.24}$$

$$z_b^w|\cdot\cdot\sim N((\frac{\beta_{b-1}z_{b-1}^w+\beta_b z_{b+1}^w}{\tau^2}+\frac{\sum_{t=1}^{T}(z_{bt}-x_tc_{d_g^w})}{\sigma_a^2})/(\frac{\beta_b^2+1}{\tau^2}+\frac{T}{\sigma_a^2}),\frac{1}{\frac{\beta_b^2+1}{\tau^2}+\frac{T}{\sigma_a^2}}) \tag{3.25}$$

for $b=2,...,B-1$

$$z_B^w|\cdot\cdot\sim N((\frac{\beta_{B-1}z_{B-1}^w}{\tau^2}+\frac{\sum_{t=1}^{T}(z_{Bt}-x_tc_{d_g^w})}{\sigma_a^2})/(\frac{1}{\tau^2}+\frac{T}{\sigma_a^2}),\frac{1}{\frac{1}{\tau^2}+\frac{T}{\sigma_a^2}}) \tag{3.26}$$

$$\beta_b|\cdot\cdot\sim N(\frac{z_b^w z_{b+1}^w+\frac{\tau^2\sqrt{1-\tau^2}}{\sigma^2}}{{z_b^w}^2+\frac{\tau^2}{\sigma^2}},\frac{1}{\frac{{z_b^w}^2}{\tau^2}+\frac{1}{\sigma^2}}) \tag{3.27}$$

for $b=1,...,B-1$ Similarly to e_{bt}^w and z_{bt}^w, e_{gt}^y is sampled from its conditional posterior , conditional on the observed data and all currently imputed parameters except z_{gt}^y, and thus z_{gt}^y from the complete conditional posterior, now conditioning also on the already imputed e_{gt}^y.

$$e^y_{gt}|\{-z^y_{gt}\} \sim \begin{cases} -1 & p = p_- \\ 0 & p = p_0 \\ 1 & p = p_+ \end{cases} \tag{3.28}$$

where $\{-z^y_{gt}\} = \{z^y_{1t}, ..., z^y_{g-1,t}, z^y_{g+1,t}, ..., z^y_{Gt}\}$ and the probabilities are given by

$$p_- \propto \Phi(-1|\alpha_g + x_t c_{d^y_g} + z^w_{gt}\lambda_{d^{yw}_g}, 1)(\frac{1}{\psi^-_g})I((y_{gt} - \mu_g - \alpha_t) \in [-\psi^-_g, 0]) \tag{3.29}$$

$$p_0 \propto (\Phi(1|\alpha_g + x_t c_{d^y_g} + z^w_{gt}\lambda_{d^{yw}_g}, 1) - \Phi(-1|x_t c_{d^y_g} + z^w_{gt}\lambda_{d^{yw}_g}, 1))\phi(y_{gt} - \mu_g - \alpha_t|0, s^2_g) \tag{3.30}$$

$$p_+ \propto (1 - \Phi(1|\alpha_g + x_t c_{d^y_g} + z^w_{gt}\lambda_{d^{yw}_g}, 1))(\frac{1}{\psi^+_g})I((y_{gt} - \mu_g - \alpha_t) \in [0, \psi^+_g]) \tag{3.31}$$

in turn properly normalized.

Again the conditional posterior probabilities for e^y_{gt} are proportional to the respective prior probabilities times the likelihood factor corresponding to y_{gt}, where the prior probabilities are the appropriate normal quantiles and the evaluation of the likelihood needs to include indicators for the finite support of the uniform kernels in (3.2).

z^y_{gt} is sampled form a truncated Normal distribution that guarantees the latent probit score to be consistent with the relative POE indicator of gene under-, normal- or over-expression.

$$z^y_{gt}| \cdot \cdot \sim N(\alpha_g + x_t c_{d^y_g} + z^w_{gt}\lambda_{d^{yw}_g}, 1)[I(z^y_{gt} < -1)I(e^y_{gt} = -1) + \cdots] \tag{3.32}$$

The posterior probabilities for the gene specific and sample specific effects are two updated Normal distributions, truncations due to the finite support of the uniform kernels in (3.2); moreover attention must be paid for the posterior probabilities of the α_t's since we are to take into account the condition $\sum_{t=1}^T \alpha_t = 0$.

$$\mu_g| \cdot \cdot \sim \begin{cases} N(\frac{\frac{\sum_{t:e^y_{gt}=0}(y_{gt}-\alpha_t)}{s^2_g} + \frac{\theta_\mu}{\sigma_\mu}}{N_g/s^2_g + 1/\sigma^2_\mu}, \frac{1}{\frac{N_g}{s^2_g} + \frac{1}{\sigma^2_\mu}}) \\ I((\max(\max_{t:e^y_{gt}=-1}(y_{gt} - \alpha_t), \max_{t:e^y_{gt}=1}(y_{gt} - \alpha_t - \psi^+_g)), \\ \quad \min(\min_{t:e^y_{gt}=-1}(y_{gt} - \alpha_t + \psi^-_g), \min_{t:e^y_{gt}=1}(y_{gt} - \alpha_t))) \end{cases} \tag{3.33}$$

where $N_g = \#\{t : e^y_{gt} = 0\}$. α_t's are sampled up to $T-1$, and subsequently α_T is defined as the opposite of the sum of the previous ones.

$$\alpha_t|\cdot\cdot\sim\begin{cases} N(\frac{(\sum_{g:e^y_{gt}=0}\frac{y_{gt}-\mu_g}{s^2_g}-\sum_{g:e^y_{gt}=0}\frac{y_{gT}-\mu_g+\sum_{-t}\alpha_{t'}}{s^2_g}-\sum_{-t}\alpha_{t'}}{\sum_{g:e^y_{gt}=0}\frac{1}{s^2_g}+\sum_{g:e^y_{gt}=0}\frac{1}{s^2_g}+\frac{2}{\sigma^2_\alpha}}, \frac{1}{\sum_{g:e^y_{gt}=0}\frac{1}{s^2_g}+\sum_{g:e^y_{gt}=0}\frac{1}{s^2_g}+\frac{2}{\sigma^2_\alpha}}) \\ I((\max(\max_{g:e^y_{gt}=-1}(y_{gt}-\mu_g), \max_{g:e^y_{gt}=1}(y_{gt}-\mu_g-\psi^+_g), \\ \max_{g:e^y_{gt}=-1}(-y_{gT}+\mu_g-\sum_{-t}\alpha_{t'}-\psi^-_g), \max_{g:e^y_{gt}=1}(-y_{gT}+\mu_g-\sum_{-t}\alpha_{t'})), \\ \min(\min_{g:e^y_{gt}=-1}(y_{gt}-\mu_g+\psi^-_g), \min_{g:e^y_{gt}=1}(y_{gt}-\mu_g), \\ \min_{g:e^y_{gt}=-1}(-y_{gT}+\mu_g-\sum_{-t}\alpha_{t'}), \min_{g:e^y_{gt}=1}(-y_{gT}+\mu_g-\sum_{-t}\alpha_{t'}+\psi^+_g))) \end{cases} \tag{3.34}$$

for $t = 1, ..., T-1$, and $N_t = \#\{g : e^y_{gt} = 0\}$.

$$\alpha_T = -\sum_{t=1}^{T-1}\alpha_t \tag{3.35}$$

The parameters in the mixture model (3.2) and the intercept in regression (3.6) are sampled respectively from the updated Gamma and Normal distributions.

$$\psi^{-^{-1}}_g|\cdot\cdot\sim G(\alpha_{\psi^-}+M_g,\beta_{\psi^-})I(\psi^-_g>\max(0,-\min_{t:e^y_{gt}=-1}(y_{gt}-\mu_g-\alpha_t))) \tag{3.36}$$

where $M_g = \#\{t : e^y_{gt} = -1\}$.

$$s^{-2}_g|\cdot\cdot\sim G(\alpha_s+\frac{N_g}{2},(1/\beta_s+\frac{\sum_{t:e^y_{gt}=0}(y_{gt}-\mu_g-\alpha_t)^2}{2})^{-1}) \tag{3.37}$$

$$\psi^{+^{-1}}_g|\cdot\cdot\sim G(\alpha_{\psi^+}+d^w_g,\beta_{\psi^-})I(\psi^+_g>\max(0,\max_{t:e^y_{gt}=1}(y_{gt}-\mu_g-\alpha_t))) \tag{3.38}$$

where $d^w_g = \#\{t : e^y_{gt} = 1\}$.

$$\alpha_g|\cdot\cdot\sim N(\frac{\sum_{t=1}^{T}(z^y_{gt}-x_tc_{d^y_g}-z^w_{gt}\lambda_{d^{yw}_g}}{T+1},\frac{1}{T+1}) \tag{3.39}$$

The conditional posterior probabilities for d^y_g and d^{yw}_g, i.e. the trinary indicator for differential gene expression and interaction between the two platforms, are proportional to

the respective prior probabilities times the likelihood factor corresponding to z^y_{gt}, where the prior probabilities are defined in such a way that p_2 is greater than $p_1 + p_3$, to express our prior belief that it is more likely for a gene to show similar gene expression in the two subgroups of breast cancer, and no interaction between the two platforms.

$$d^y_g \sim \begin{cases} -1 & p \propto p_1 \phi(z^y_{gt} | \alpha_g + c_{-1} x_t + \lambda_{d^{yw}_g} z^w_{gt}, 1) \\ 0 & p \propto p_1 \phi(z^y_{gt} | \alpha_g + c_0 x_t + \lambda_{d^{yw}_g} z^w_{gt}, 1) \\ 1 & p \propto p_1 \phi(z^y_{gt} | \alpha_g + c_1 x_t + \lambda_{d^{yw}_g} z^w_{gt}, 1) \end{cases} \tag{3.40}$$

$$d^{yw}_g \sim \begin{cases} -1 & p \propto p_1 \phi(z^y_{gt} | \alpha_g + c_{d^y_g} x_t + \lambda_{-1} z^w_{gt}, 1) \\ 0 & p \propto p_1 \phi(z^y_{gt} | \alpha_g + c_{d^y_g} x_t + \lambda_0 z^w_{gt}, 1) \\ 1 & p \propto p_1 \phi(z^y_{gt} | \alpha_g + c_{d^y_g} x_t + \lambda_1 z^w_{gt} 1) \end{cases} \tag{3.41}$$

Finally the posterior probabilities for the parameters in equation (3.6) are respectively the updated Normal and Gamma distributions obtained through the conjugate prior assumptions.

$$c_{d^y_g} | \cdot \cdot \sim \begin{cases} N(\frac{\sum_{t=1}^T x_{1t}(z^y_{gt} - \alpha_g - z^w_{gt}\lambda_{d^{yw}_g}) - \frac{k}{\sigma_1}}{\frac{1}{\sigma_1^2} + \sum_{t=1}^T {x_t}^2}, \frac{1}{\frac{1}{\sigma_1^2} + \sum_{t=1}^T {x_t}^2}) & \text{if } d^y_g = -1 \\ N(\frac{\sum_{t=1}^T x_t(z^y_{gt} - \alpha_g - z^w_{gt}\lambda_{d^{yw}_g})}{\frac{1}{\sigma_2^2} + \sum_{t=1}^T {x_t}^2}, \frac{1}{\frac{1}{\sigma_2^2} + \sum_{t=1}^T {x_t}^2}) & \text{if } d^y_g = 0 \\ N(\frac{\sum_{t=1}^T x_t(z^y_{gt} - \alpha_g - z^w_{gt}\lambda_{d^{yw}_g}) - \frac{k}{\sigma_1}}{\frac{1}{\sigma_1^2} + \sum_{t=1}^T {x_t}^2}, \frac{1}{\frac{1}{\sigma_1^2} + \sum_{t=1}^T {x_t}^2}) & \text{if } d^y_g = 1 \end{cases} \tag{3.42}$$

$$\lambda_{d^{yw}_g} | \cdot \cdot \sim \begin{cases} N(\frac{\sum_{t=1}^T z^w_{gt}(z^y_{gt} - \alpha_g - x_t c_{d^y_g}) - \frac{k}{\sigma_1^2}}{\frac{1}{\sigma_1^2} + \sum_{t=1}^T {z^w_{gt}}^2}, \frac{1}{\frac{1}{\sigma_1^2} + \sum_{t=1}^T {x_t}^2}) & \text{if } d^{yw}_g = -1 \\ N(\frac{\sum_{t=1}^T z^w_{gt}(z^y_{gt} - \alpha_g - x_t c_{d^y_g})}{\frac{1}{\sigma_2^2} + \sum_{t=1}^T {z^w_{gt}}^2}, \frac{1}{\frac{1}{\sigma_2^2} + \sum_{t=1}^T {x_{1t}}^2}) & \text{if } d^{yw}_g = -1 \\ N(\frac{\sum_{t=1}^T z^w_{gt}(z^y_{gt} - \alpha_g - x_t c_{d^y_g}) - \frac{k}{\sigma_1^2}}{\frac{1}{\sigma_1^2} + \sum_{t=1}^T {z^w_{gt}}^2}, \frac{1}{\frac{1}{\sigma_1^2} + \sum_{t=1}^T {x_t}^2}) & \text{if } d^{yw}_g = -1 \end{cases} \tag{3.43}$$

3.2.3 Multiplicities

The analysis involves massively many comparisons that lead to the selection of interesting genes based on differential gene expression, CNV and correlation of the two.
A useful generalization of frequentist type-I error rates to multiple hypothesis testing is the false discovery rate (FDR) introduced in Benjamini and Hochberg ((5)), and review in a Bayesian framework by Scott and Berger ((37), (38))).
Let d_i generically denote an indicator for the unknown truth of gene i being differentially expressed under two biologic conditions of interest. Depending on the question of interest d_i might refer to $e^w_{bt} \neq 0$, $e^y_{gt} \neq 0$, $d^{yw}_g \neq 0$, or any combination.

Let δ_i denote an indicator for the related decision, with $\delta_i = 1$ if we report $d_i = 1$. Let $D = \sum_{i=1}^{G} \delta_i$ denote the number of reported comparisons. The FDR is defined as $FDR = \sum_{i=1}^{G}(1 - d_i)\delta_i/D$, the fraction of false rejections, relative to the total number of rejections. The ratio involves the unknown parameters d_i as well as the data, indirectly through the decisions δ_i. As such it is neither Bayesian nor frequentist. Under a Bayesian perspective we compute the posterior expected FDR by averaging with respect to d_i, the only unknown quantity after conditioning on the data. Let $\bar{d}_i = P(d_i = 1 \mid data)$, then

$$F\bar{D}R = E(FDR \mid data) = \sum_{i=1}^{G}(1 - \bar{d}_i)\delta_i/D.$$

It can be shown ((29)) that under several loss functions that combine false positive and similarly defined false negative rates (or counts) the Bayes rule is of the form $\delta^*_i = I(\bar{d}_i > c)$, i.e., report all comparisons with posterior probability greater some cutoff c. We thus proceed in all in the following inference by setting a bound on FDR, and then finding a cutoff c that gives the desired posterior expected $F\bar{D}R$. An algorithm that allows us to compute FDR levels for number of discoveries, and therefore to select differentially expressed genes so that the FDR level is controlled at level α, is the following:

- Sort, from the lowest to the highest, the marginal posterior probabilities $\pi_i = (1-\bar{d}_i)$, to obtain $(\pi_{(1)}, ..., \pi_{(m)})$

- If $\pi_{(1)}/1 > \alpha$, then we claim there are no differentially expressed genes

- Otherwise if $(\pi_{(1)} + \pi_{(2)})/2 > \alpha$, then claim the gene corresponding to $\pi_{(1)}$ differentially expressed

- ...

- Continue until the first time $\sum_{i=1}^{G} \pi_{(i)}/G > \alpha$, and claim genes corresponding to $\pi_{(1)}, ...\pi_{(G-1)}$ differentially expressed

3.2.4 Posterior Inference on the breast cancer dataset

Our model at this point has the feature of selecting genes based on what we consider to be the posterior probability of differential expression, using either the marginal models on gene expression microarray data and aCGH data, or the integrating model.

Another important feature of this model is to make inference on the interaction between platforms, that would imply a positive interaction between aberrant copy numbers and over or under expression, and, as in Choi et al, inference on differential expression of genes conditional on aberrant copy number only.

The following graph represents the distribution of the posterior probabilities, obtained with both the marginal and the integrated models, we will use for inference later on.

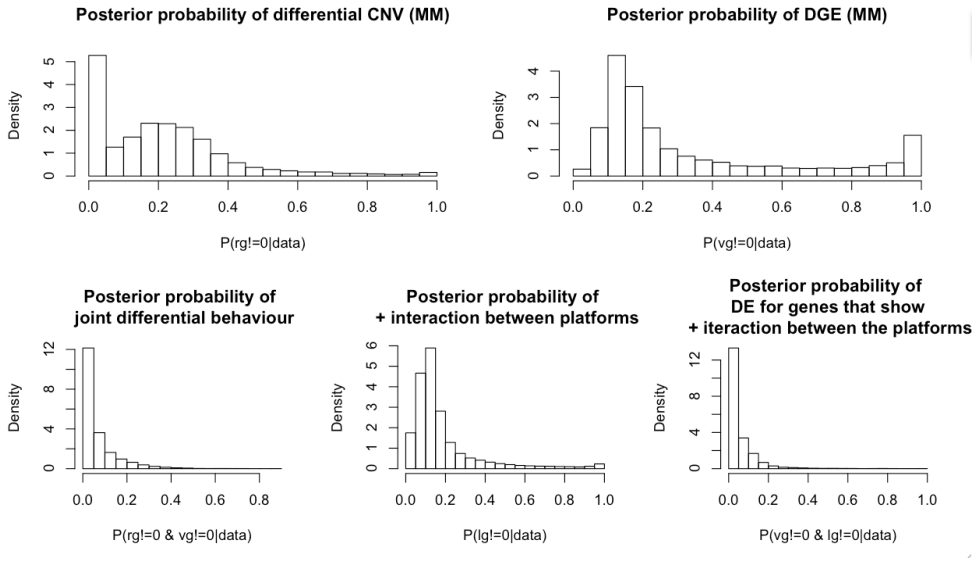

Figure 3.4: Posterior probabilities of differential gene expression and differential CNV (marginal models), and posterior probabilities of joint differential behaviour, positive interaction between platforms and differential expression of genes conditional on aberrant copy number only (joint model).

The purpose of selecting genes requires the definition of elements for hypothesis testing; for the selection of genes that show differential expression and CNV, through the marginal models, we define n_i to be respectively

$$n_i = P(d_i^y \neq 0|y_{it}) = P(d_i^y = -1|y_{it}) + P(d_i^y = 1|y_{it})$$
$$n_i = P(d_i^w \neq 0|w_{b(i)t}) = P(d_i^w = -1|w_{b(i)t}) + P(d_i^w = 1|w_{b(i)t})$$

while for the selection of genes that show joint differential behaviour, positive interaction between platforms and differential expression of genes conditional on aberrant copy number only.

$$n_i = P(d_i^w \neq 0, d_i^y \neq 0|w_{b(i)t}, y_{it}) = P(d_i^w = -1, d_i^w = -1|w_{b(i)t}, y_{it}) + P(d_i^w = 1, d_i^y = 1|w_{b(i)t}, y_{it})$$
$$n_i = P(d_i^{yw} = 1|w_{b(i)t}, y_{it})$$

$n_i = P(d_i^{yw} = 1, d_i^y \neq 0 | w_{b(i)t}, y_{it}) = P(d_i^{yw} = 1, d_i^y = -1 | w_{b(i)t}, y_{it}) + P(d_i^{yw} = 1, d_i^y = 1 | w_{b(i)t}, y_{it})$

where $t = 1, ...T$ and $b(i)$ indicates all the probes belonging to the gene i.

FDR levels were computed with the algorithm presented in the previous section for the distinct events of under expression and copy number deletion and over expression and copy number duplications. Therefore, in the n_i specified above, we introduced the distinction between d_i^y and d_i^w being either 1 or -1.

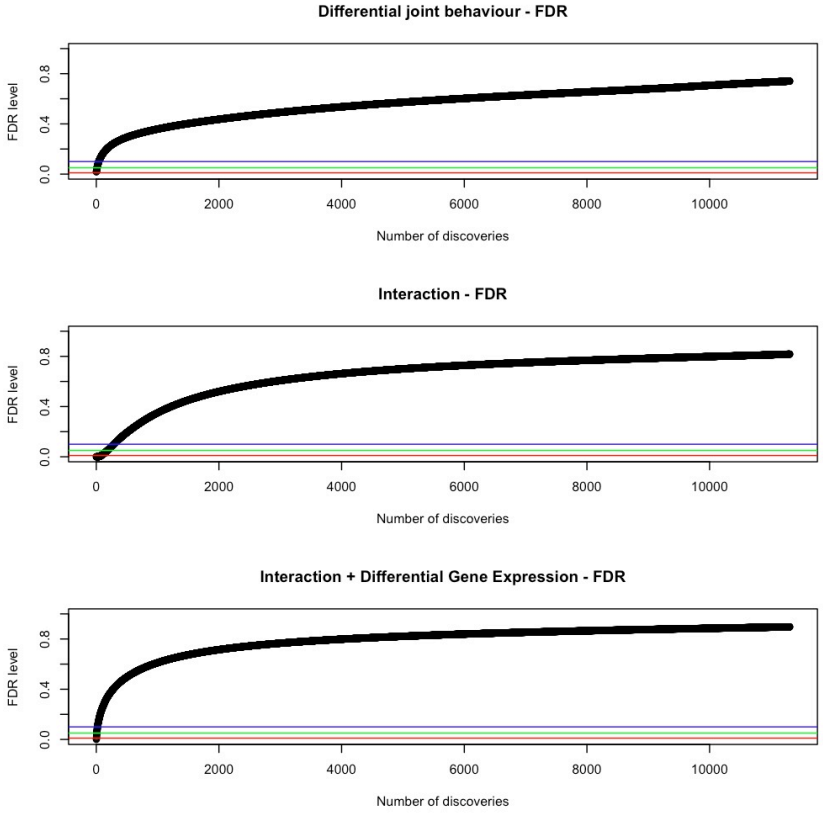

Figure 3.5: FDR levels for number of genes claimed to show jointly over expression and copy number duplications in TN group, and of those claimed to show jointly under expression and copy number deletions in TN group . The points below the red line are related to the selected list of genes that allow the percentage of false discoveries to be less than 1%,5% or 10%.

The lists of selected genes could be of greater interest for clinicians since they indicate which genes show differential expression and copy number variation between TN patients and patients who testes positively for ER or HER2 receptors. We therefore use both types of data to assess a differential behaviour of genes in different subgroups of breast cancer.

The lists of genes selected for respectively their joint differential behaviour (one for over expression and copy number duplications, the other for under expression and copy number deletions), the positive interaction between the two platforms and the differential expression, respectively over and under expression) of genes conditional on aberrant copy number only are shown in the following graph.

```
 [1] MCM4     BTG3     STIL     POLR1E   CHODL    HMGN4    USP25    SH2D2A   PDSS1
[10] C1orf38  TIMM44   NFKBIE   NDRG1    TPX2     PVR      SLC2A3   PRKD3    INSL4
[19] ICAM1    FAM107A  HLA-E    MOBKL2B  C11orf75 RECK     C1S      FXYD5    HLA-DRA
[28] ST8SIA1  GAS1     TRIT1    CDH3     VIM      HEPH     AK2      STK38    C1GALT1
[37] IRAK1    CTPS     SLAMF7   CCDC86   CHD1L    SHCBP1   CD14     KARS     C10orf10
[46] FOLH1    GTPBP2   C12orf32 PEPD     MYC      PLCG2    E2F3
```

Figure 3.6: Lists of selected genes showing over expression and copy number duplications in TN group

```
 [1] IGF1R    TFF1     FAM130A1 MYO5C    TMEM62   CA12     PRKAG1   IRS1     AMIGO2
[10] CYB5R1   SLC16A6  STC2     RND1     FUT8     PANK3    MON2     UQCRQ    CUEDC1
[19] BCAS4    SLC22A5  TSFM     TOM1L1   DUSP10   ZNF281   ORMDL2   PURA     TEGT
[28] HDAC11   RTF1     C6orf211 IFT140   PEX11A   TBC1D9   BCAS1    KCTD3    ACSS3
[37] H2AFY    MSX2
```

Figure 3.7: Lists of selected genes showing under expression and copy number deletions in TN group

[1]	ALG8	APPBP2	ASH2L	CLTC	CSNK1G3	DDHD2	ERBB2	ERLIN2
[9]	KCTD3	LLGL2	LRRC59	NME1	PPME1	PSMD3	STARD3	TLK2
[17]	TMEM49	ZNF16	AMZ2	COIL	DDX28	ARNT	BPTF	C1orf27
[25]	DPY19L4	FGFR1	HISPPD2A	LMNA	LSM1	PERLD1	PPM1D	RAB2A
[33]	TPR	ASPH	GRB7	NDUFC2	S100A13	WHSC1L1	FEM1C	LASP1
[41]	UTP18	C11orf67	MED13	KIF1B	MTMR4	ORAI3	CCT2	EIF4EBP1
[49]	RPS6KB1	EXT1	LYK5	SLC35B1	WIPF2	ZC3H11A	CCNE2	PHB
[57]	RBM35A	SMARCD2	PSMB3	REXO2	WDR68	CROP	HEATR6	PSMD12
[65]	BAG4	CCNH	FADD	HN1	NCBP1	RHOT1	RNFT1	USP12
[73]	MTDH	TMEM97	YTHDF3	AZIN1	PPFIA1	COG7	HSF1	PROSC
[81]	TMEM104	PIP4K2B	RPL23	SCNM1	NCOA2	TEX2	TTC35	CLNS1A
[89]	TIPRL	COPS2	YY1AP1	CASC3	TUBD1	TH1L	PFDN2	RAD52
[97]	GGPS1	MAP3K3	MED1	HHAT	INTS8	MCOLN3	ATP5L	HOXC13
[105]	RAB11FIP1	TFAP2C	ZNF217	PSMC5	EIF2B5	GYG1	IQWD1	RNF43
[113]	RNPEP	TMEM33	TRIM37	UBR5	FZD6	MTERFD1	UBE2Z	GNAI3
[121]	PHF10	TCEA1	C8orf33	GSTT1	MRPS27	PHYH	PMF1	SNF8
[129]	CACNG4	TICAM2	UBE2Q1	EIF2S1	SRF	FPGT	GPR137B	NAE1
[137]	SF3B4	ZBTB5	ACOT2	MYST2	PRKAR1A	UPF2	LACTB2	RGS11
[145]	ZMYND11	PCGF2	ZNF146	CCT3	ENY2	MED13L	ATG5	COMMD10
[153]	MTCH1	PIGT	RAD51C	UBE2W	CD46	SEPHS1	G6PC3	BRD7
[161]	SUPT4H1	RAD54B	CCDC47	GPRC5C	TAF9	ATP6V1C1	CASP1	UBN1
[169]	ARL6IP1	MYO7A	ADSL	ADIPOR1	TERF1	ADSS	CHD1	SPOP
[177]	ZGPAT	RNF19A	C17orf71	ATF6	RCOR3	GPATCH2	SPAG9	IFT20
[185]	LANCL2	CALCOCO2	C13orf27	C15orf29	FNDC3A	RAB22A	STX16	DUSP12
[193]	IL6ST	ARMC1	MAP4K1	UBTD1	POLG2	FTSJ3	HIRIP3	RAPGEFL1
[201]	SNRPE	SPATA5L1	RRP15	ACAD8	SNX11	ZNF688	RNASEL	DHTKD1
[209]	ST3GAL1	RAB26	APH1A	MCM4	JARID1A	SHARPIN	MRPL13	SPG11
[217]	CDC23	PHF20L1	FLRT3	MRPL24	HMBS	ADAM9	SUMO2	GPR172A
[225]	CLK2	ATP9A	PLOD3	CETN3	UQCRB	DHX40	BAT2D1	NMT1
[233]	OSBPL2	GRHL2	MTAP					

Figure 3.8: Lists of selected genes showing positive interaction between platforms

[1]	REXO2	TBC1D4	BRD7	RPS20	RDX	MCM4	JRKL	RPL38	E2F3
[10]	MTAP	BTG3	ASB7	C13orf27	RHBDF2	CASP1	ANP32B	MPZL2	EGFR
[19]	KIF2C	CD3G	NDRG1	NOC3L	TFAP2C	AK2	BTBD3	RIPK2	ANKRD27
[28]	TMEM38B	RFWD3	FOXM1	C11orf75	NAT10	S100A10	B4GALT5	SHFM1	NUP153
[37]	HMBS	ANP32E	RASGRP1						

Figure 3.9: Lists of selected genes, respectively showing over expression conditional on aberrant copy number only

[1]	KCTD3	S100A13	HHAT	GGPS1	ACOT2	GPRC5C	G6PC3	RAB26
[9]	HISPPD2A	GPR137B	ORAI3	FPGT	LRRC6	RNPEP	UBE2W	COQ7
[17]	ATP6V0C	PEX7	VEZF1	CCDC56	CYB5R1	SELENBP1	RNF43	COG2
[25]	RGS11	MRPL24	TMEM62	KIAA0196	ZC3H3	AP1M2	SPG11	EMP2
[33]	AZGP1	SPATA20	C1orf25	SUPT4H1	PBX1	IQWD1	C1orf66	HOXC13
[41]	PJA2	TMCO1	GLUL	DHX40	SYNJ2BP	GSTT1	ADIPOR2	SOX13
[49]	FLJ11506	CD46	LGALS8	BCL9	C15orf24	RAB11FIP1	IL6ST	ZMYND11
[57]	CD2BP2	ANKRA2	MED13L	PRPF6	SRP14	FADD	PEX19	SLC4A2
[65]	UBR5	MAPBPIP	MCCC2	SCCPDH	SLC9A3R1	ALG8	FAM63A	PSEN2
[73]	C1orf218	ST3GAL1	URG4	TFAP2A	LASS2	TOM1L1	MYO5C	ADIPOR1
[81]	C19orf21	CCND1	HOXB7	STAU2	RNF14	PRODH	UQCC	AMZ2
[89]	TMEM121	CAB39L	CACNG4					

Figure 3.10: Lists of selected genes, respectively showing under expression conditional on aberrant copy number only

It is moreover to be specified that our joint model selects differently from the marginal models and this can be seen from the following plot. If the list of selected genes coincided with the intersection of the lists obtained through the marginal models there would be no point in running the integration model.

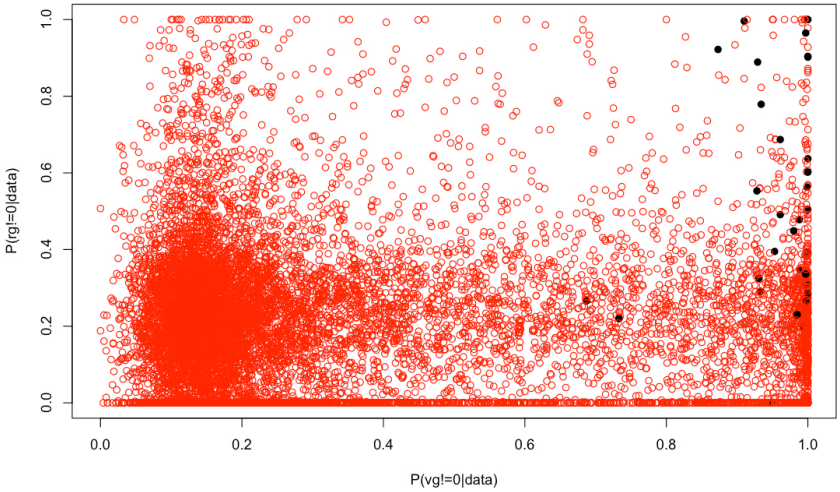

Figure 3.11: Plot of posterior probabilities of differential gene expression and differential CNV, with black dots indicating genes claimed to show joint differential behaviour by the integrated model

As we can clearly see, our model turns out to be different, borrowing strength from the two platforms, from the intersection of the two marginal models, although it still selects, coherently, mostly genes in the upper right corner. A simple model checking was achieved plotting posterior probabilities of differential gene expression and difference in means of the gene expression measurements for TN and non TN group. Same was done for posterior probabilities of positive interaction between platforms and sample correlations. The plot is exactly what we were expecting, a perfect description of the data is provided for by posterior probabilities obtained by our model.

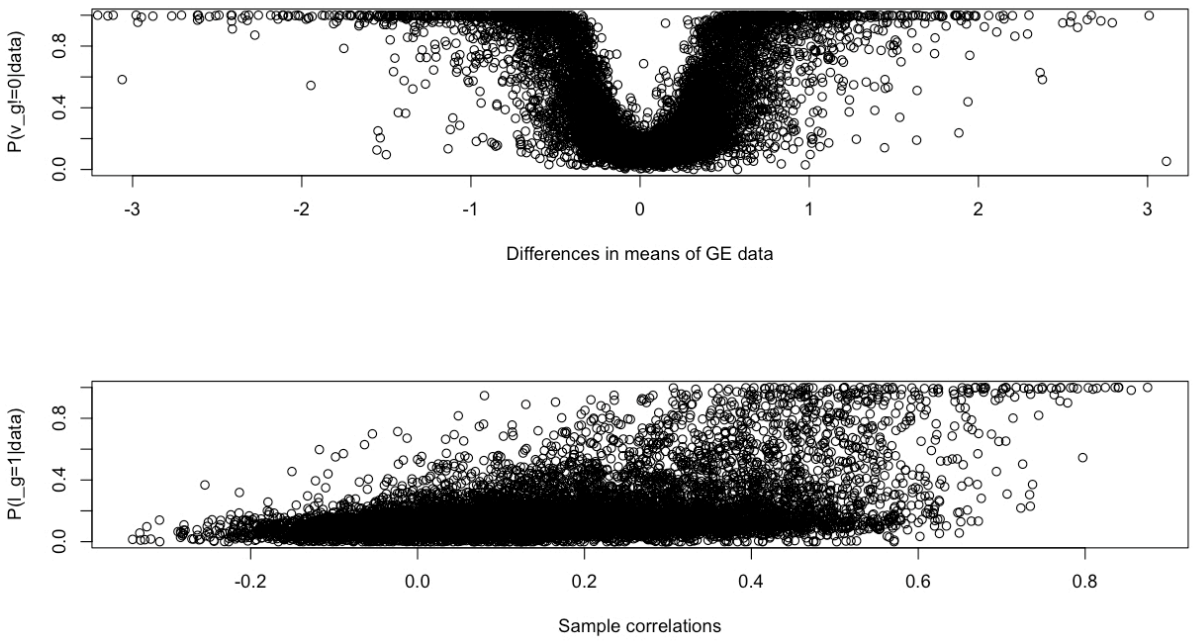

Figure 3.12: Comparison between sample measures and posterior probabilities

3.3 Conclusion

The four lists of genes we selected were analysed using Pathway Analysis GeneGo (50) in order to show biological meanings and networks.

The list of jointly over expressed genes has one big network centered on MYC, the list of over epxressed genes conditional on copy number aberrations only has one small network centred on Caspase-1, while the genes that showed positive interaction between the two platforms seem to be well connected.
When we excluded MYC from the first list, there was no big network and that indicates MYC might be a key gene in order to build a big network.
The next three plots show the three pathways we cited above.

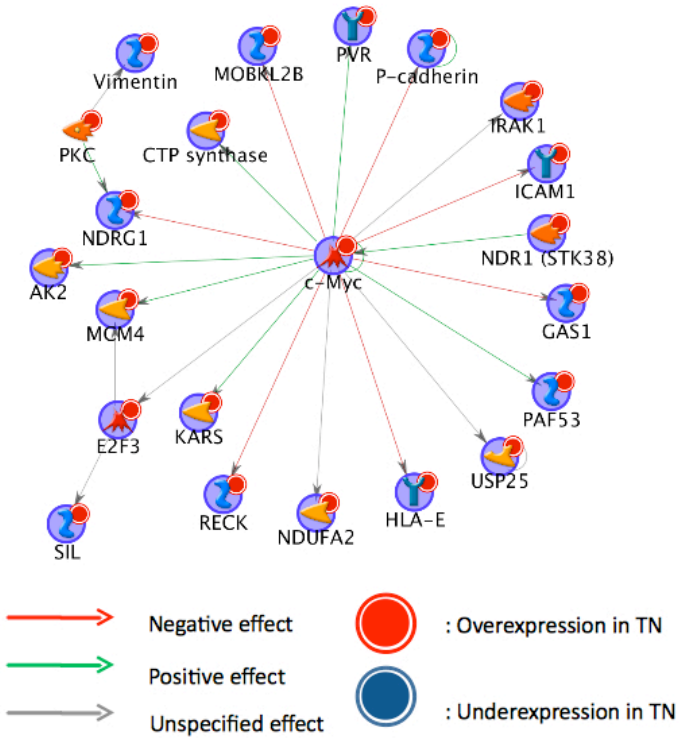

Figure 3.13: Network for the list of jointly over expressed genes

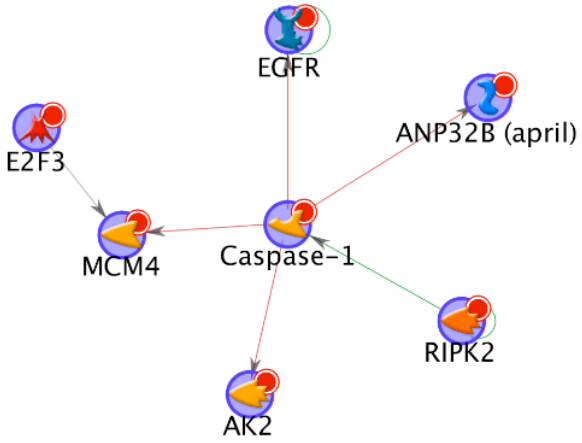

Figure 3.14: Network for the list of over expressed genes conditional on copy number aberrations only

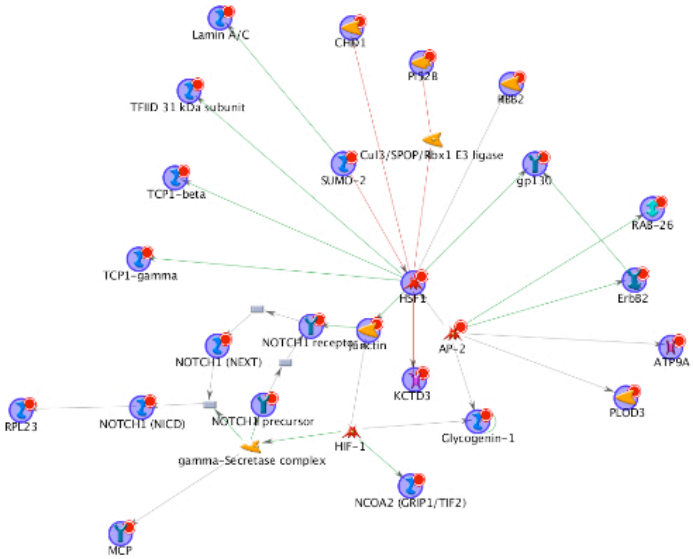

Figure 3.15: Network for the list of genes that show positive interaction between the two platforms

This result is promising since MYC is a key regulator of cell growth, proliferation, metabolism, differentiation, and apoptosis and MYC deregulation contributes to breast cancer development and progression and is associated with poor outcomes.
Multiple mechanisms are involved in MYC deregulation in breast cancer, including gene amplification, transcriptional regulation, and mRNA and protein stabilization, which correlate with loss of tumor suppressors and activation of oncogenic pathways. The heterogeneity in breast cancer is increasingly recognized. Breast cancer has been classified into 5 or more subtypes based on gene expression profiles, and each subtype has distinct biological features and clinical outcomes. Among these subtypes, basal-like tumor is associated with a poor prognosis and has a lack of therapeutic targets (46).
MYC is overexpressed in the basal-like subtype and may serve as a target for this aggressive subtype of breast cancer. Tumor suppressor BRCA1 inhibits MYC's transcriptional and transforming activity. Loss of BRCA1 with MYC overexpression leads to the development of breast cancerespecially, basal-like breast cancer. As a downstream effector of estrogen receptor and epidermal growth factor receptor family pathways, MYC may contribute to resistance to adjuvant therapy. Targeting MYC-regulated pathways in combination with inhibitors of other oncogenic pathways may provide a promising therapeutic strategy for breast cancer, the basal-like subtype in particular (46).

4

Predicting clinical outcomes for breast cancer patients using integrated platforms

The idea of this chapter raises from the question of whether or not we could use the same latent structure underneath gene expression and copy number variation data to make inference on a clinical outcome of new patients in the study.
The chosen approach is to state a model for y_{gt} and w_{bt}, $p(w_{bt}, y_{gt}|\theta)$, and to assume a Bernoulli distribution for u_t. This leads us to the sought model $p(u_t|w_{bt}, y_{gt})$ and posterior probabilities of u_t being 1 give us a measure for the prediction of the outcome of the new patient.
The advantages of our model with respect to, for example, a simple logistic regression $p(u_t|y_{gt}, w_{bt})$ are mainly two:

- the noise reduction achieved through the assumption of a latent structure underneath our data, i.e. the latent POE scores for gene expression and copy number variation;

- it allows for a natural variable selection within the model itself; indicators d_g^{wu} and d_g^{yu} in equation (4.4) and (4.5) (with Bernoulli priors with probability $1-p$ and p very close to 1) allow for a reduction of the number of covariates (genes) and avoid the problem of overestimation.

In summary, as a new patient comes into a study and we have measurements of his gene expression and copy number variation, we run the model $p(w_{bt}, y_{gt}|\theta)$ and assume for his clinical outcome u_t a Bernoulli distribution with probability π. Through MCMC methods we obtain updated posterior probabilities of u_t being 1 that give us a measure for the prediction of his outcome.
In this particular case the outcome refers to the pathological complete response to the treatment of patients in a breast cancer study.
This response was defined as a complete disappearance of the tumor with no more than a few scattered tumor cells detected by the pathologist in the resection specimen (6).

4.1 Probability Model

4.1.1 Sampling model for w and y

As before we use a mixture model (30) to introduce a trinary latent indicator variables for the CNA state for each probe and the expression level state for each gene.
The latent indicators e_{bt}^w and e_{gt}^y define mixture models for copy number log2 ratios w_{bt} and for gene expression data y_{gt}:

$$f_w(w_{bt}|e_{bt}^w) =_d \begin{cases} U(-\phi_b^-, 0) & \text{if } e_{bt}^w = -1 \\ N(0, \nu_b^2) & \text{if } e_{bt}^w = 0 \\ U(0, \phi_b^+) & \text{if } e_{bt}^w = 1 \end{cases} \tag{4.1}$$

$$f_y(y_{gt} - \mu_g - \alpha_t | e^y_{gt}) =_d \begin{cases} U(-\psi_g^-) & \text{if } e^y_{gt} = -1 \\ N(0, s_g^2) & \text{if } e^y_{gt} = 0 \\ U(0, \psi_g^+) & \text{if } e^y_{gt} = 1 \end{cases} \tag{4.2}$$

with the same parameters' specifications as in (3.1.2).

4.1.2 Latent probit scores, regression and probabilistic assumption on the outcome for new patients

Anticipating the integration of both platforms in a regression model, we further introduce latent Gaussian variables z^w_{bt} and z^y_{gt} to define a probit scores for the trinary indicators e^w_{bt} and e^y_{gt}.

As specified in the previous chapter, the model here deviates from the POE model presented in the previous chapter, since prior probabilities for the latent trinary scores are implemented on a probit scale (11) to allow the integration of the two submodels by a simple normal regression at the level of these probit scores, and borrow strength across the genomic platforms to achieve better predictive performances of the clinical outcome. Specifically, define

$$e^w_{bt} = \begin{cases} -1 & \text{if } z^w_{bt} < -1 \\ 0 & \text{if } -1 \leq z^w_{bt} \leq 1 \\ 1 & \text{if } z^w_{bt} > 1 \end{cases} \quad \text{and} \quad e^y_{gt} = \begin{cases} -1 & \text{if } z^y_{gt} < -1 \\ 0 & \text{if } -1 \leq z^y_{gt} \leq 1 \\ 1 & \text{if } z^y_{gt} > 1 \end{cases} \tag{4.3}$$

The next two equations embody our assumption that positive or negative clinical response of patients could be related to differential behaviour of a small subgroups of the 11.306 genes, i.e. copy number variation and gene expression. We assume

$$z^w_{bt} | z^w_b \sim N(z^w_b + d^{wu}_g p_g u_t, \sigma^2_a) \tag{4.4}$$

Here, z^w_b is a probe-specific mean that defines a baseline CNA status (e.g., for a reference subtype), u_t is the clinical outcome mentioned above measured on the 122 patients and d^{wu}_g

is a binary indicator introduced for controlling the number of covariate in the regression. The integration of the two platforms is implemented as a regression with the probit scores. Setting up the regression first requires to match the unit of measurement for aCGH and RNA data. We define a gene-level score for the aCGH data,
$z^w_{gt} = \frac{1}{m_g} \sum_{b \in g} z^w_{bt}$.

$$z^y_{gt} | z^w_{gt} \sim N(\alpha_g + d^{yu}_g q_g u_t + z^w_{gt} \lambda_{l_g}, 1), \tag{4.5}$$

where λ_{l_g} characterizes the relationship between the two platforms, l_g is a trinary indicator for positive, negative or null interaction between the two platforms of data, d^{yu}_g is a binary indicator introduced for controlling the number of covariate in the regression and u_t is the same variable as above.

As new patients $t+1, ..., t+n$ come into the study, and supposedly they do not have an information on pathological complete response, an assumption on their outcome is made, as follows:

$$u_{t+i} \sim_{iid} \text{Bernoulli}(\pi),\ i = 1, ..., n. \tag{4.6}$$

so that, as mentioned in the beginning of the chapter, we can learn about u_t through the above prior and $p(w_{bt}, y_{gt} | u_t, \theta)$, using Bayes formula and MCMC methods.

A graphical representation of the above model is the following

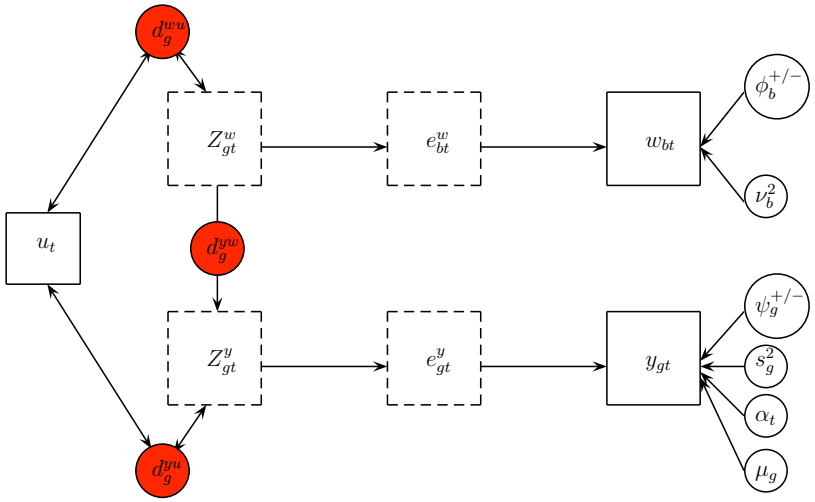

Figure 4.1: Graphical representation of the prediction model

4.1.3 Priors

As in the previous chapter the prior on the reference scores z_b^w formalizes the dependence of adjacent probes. We assume Markov dependence.
Assuming that the index b is ordered according to locus proximity on the chromosome, the dependence across adjacent probes is described as

$$z_1 \sim N(0,1) \tag{4.7}$$

$$z_b^w | {z_{b-1}}^w, \beta_{b-1} \sim N(\beta_{b-1} z_{b-1}^w, \tau^2) \tag{4.8}$$

for $b \in \{2, ..., B\}$ The parameters $\beta = (\beta_1, ..., \beta_{b-1})$ are interpreted as partial correlation coefficients, defining the strength of dependence between $\log_2$ ratios associated with clones that are adjacent on the chromosome. For β_b we assume *a priori*

$$\beta_b \sim N(\sqrt{1-\tau^2}, \sigma^2)$$

for $b \in \{1, 2, ..., B-1\}$, with $\tau^2 < 1$, ensuring the the marginal variance of z_b's remains bounded.
We complete the model construction with conditionally conjugate priors for the remaining parameters.
Priors are defined as in section 3.1.4, with the only exception of parameters in equations (4.4) and (4.5) p_g and q_g, and the binary indicators d_g^{wu} and d_g^{yu}. For both the first parameters an informative prior is assumed

$$p_g \sim N(\hat{p_g}, \sigma(\hat{p_g})^2)$$

$$q_g \sim N(\hat{q_g}, \sigma(\hat{q_g})^2)$$

with $\hat{p_g}$, $\hat{q_g}$ and the variances estimated using the data. While for the two indicators

$$d_g^{wu} \sim Bernoulli(1-p)$$

$$d_g^{yu} \sim Bernoulli(1-p)$$

with p very close to 1 to allow for the selection of a very small subgroups of genes as covariates in the two regressions.

4.2 Variable Selection

Both the models in chapter 3 and the above one introduce variable selection through the prior specifications on parameters c_{d_g}'s, λ_{d_g}, in chapter 3 and d_g^{wu}, d_g^{yu} in the current one. In the first model, to allow for variable selection, we assumed a mixture of three normals on c_{d_g}'s and λ_{d_g}, one centered in zero with very small variance and two vague normals to account for differential behaviour, while in the current one we used binary indicators multiplying parameters p_g and q_g with Bernoulli prior with probability $1-p$ very close to zero.

Both these choices were taken in order to allow for selection of small subsets of "interesting" genes without the convergence of the Gibbs sampler becoming too slow, but a further discussion on that is needed, at least to account for better and newer approaches to this problem.

While the one in the previous chapter may be seen as an "a posteriori" variable selection, achieved through the mixture prior on c_{d_g}'s and λ_{d_g}, in the current model, for a given response variable u_t, in our particular case pCR of patients in a breast cancer study, and a set of potential predictors, i.e. the pairs (w_{bt}, y_{gt}), the variable selection problem, or subset selection problem, name that fits better with our aim of selecting a subset of "interesting" genes, can be seen as a special case of a model selection problem.

The setup could be summarized as follows.

Letting γ index the subset of $\{1, ..., G\}$, each submodel is of the form

$$M_\gamma : u_t = \beta_0 + \beta_\gamma^{\mathbf{w}} \underline{\mathbf{z}}_{\gamma\mathbf{t}}^{\mathbf{w}} + \beta_\gamma^{\mathbf{y}} \underline{\mathbf{z}}_{\gamma\mathbf{t}}^{\mathbf{y}} + \epsilon_t \tag{4.9}$$

where $\underline{z}^w_{\gamma t}$ and $\underline{z}^y_{\gamma t}$ are the column vectors corresponding to the γ^{th} subset of covariates, β^w_γ and β^y_γ the row vectors of regression coefficients, and $\epsilon_t \sim N(0, \sigma^2)$.
The problem is that of choosing the "best" predictor subset $(\underline{\mathbf{z}}^{\mathbf{w}}_{\gamma\mathbf{t}}, \underline{\mathbf{z}}^{\mathbf{y}}_{\gamma\mathbf{t}})$ for u_t.
We dealt with such variable selection problem through shrinkage, in particular by multiplying parameters in the regression with binary indicators $d^u \sim Bernoulli(1-p)$ and p very close to 1.
Respectively, the mixture priors in section 3.1.4 for $c_{d^w_g}$, $c_{d^y_g}$ and $\lambda_{d^{yw}_g}$, and the prior on the binary indicator mentioned above, allowed us to select subsets of genes showing gene differential expression conditional on copy number aberration only, differential joint behavior and positive interaction between platforms, and, in the second scenario, implied a within-the-model variable selection to overcome the issue in overestimating u_t, due to the large number of genes (variables) in the study.
As i mentioned above, this choice was justified by an efficiency of the model along with a pretty fast convergence of the MCMC sampler.
Other more robust and efficient methods related to this framework were discussed recently in literature and our aim is that of comparing these with our own choice and checking for feasibility, since we are to take into account the large number of parameters in the models.

4.2.1 Non-local prior

Johnson and Rossell in their 2010 work (23) define non-local prior densities within the framework of Bayesian hypothesis testing.
They start considering the problem

$$\begin{aligned} H_0 &: \theta \sim \pi_0(\theta) \\ H_1 &: \theta \sim \pi_1(\theta) \end{aligned} \tag{4.10}$$

that requires $\pi_0(\theta)$ and $\pi_0(\theta)$ to be 0 on Θ_1 and Θ_0 respectively. Their focus is on the violation of such form, since it often happens in this context that $\pi_1(\theta)$ is positive on Θ_0.

They refer to such priors as "local alternative prior densities", which do not reflect a theory that is fundamentally different from the null hypothesis, and propose two classes of non-local prior densities that overcome this issue, but require the specification of a scale parameter that implicitly defines the total deviation from the null hypothesis (23).
Our prior specification for $c'_{d_g}s$ and λ_{d_g} in 3.1.4 could be pictured as a Bayesian hypothesis test of this kind, with the two local alternative prior densities defined as vague Gaussian distributions centered in the scale parameter referring to the density in H_1 (in our specific case roughly ± 1).
The plausible modification to our prior specifications is thus a mixture of the same "Gaussian spike" distribution centered in 0, and two inverse moment prior densities defined according to

$$\pi_I(\theta) = \frac{\kappa\tau^{\nu/2}}{\Gamma(\nu/2\kappa)}\theta^{-(\nu+1)}\exp\left[-\left(\frac{\theta^2}{\tau}\right)^{-\kappa}\right] \tag{4.11}$$

for $\kappa, \nu, \tau > 0$.
Their functional form is related to inverse gamma density functions, so that their behaviour near 0 is similar to that of an inverse gamma density near 0, which suits perfectly for our case.

4.2.2 Horseshoe prior

The alternative for the prior specifications on p_g and q_g, called horseshoe prior, was described by Carvalho, Polson and Scott (10).
Their new approach to sparsity was proven to be robust, and closely related to the answer obtained by Bayesian model averaging under a point-mass mixture prior in sparse situations.
Suppose we observe $y|\theta \sim N(\theta, \sigma^2)$m with θ assumed to be sparse, they proposed the following model for estimation and prediction:

$$\theta_i|\lambda_i \sim N(0, \lambda_i^2),\ \lambda_i|\tau \sim C^+(0, \tau),\ \tau|\sigma \sim C^+(0, \sigma) \tag{4.12}$$

where $C^+(0, a)$ is a standard half-Cauchy distribution on the positive reals with scale parameter a. Additionally a Jeffrey's prior is assumed for the variance, $p(\sigma^2) \propto 1/\sigma^2$. θ is estimated using the posterior mean under the model, which is called the horseshoe prior, since, for fixed values $\sigma^2 = \tau^2 = 1$,

$$\mathbb{E}(\theta_i|y) = \int_0^1 (1 - \kappa_i) y_i p(\kappa_i|y_i) d\kappa_i = [1 - \mathbb{E}(\kappa_i|y)] y_i \tag{4.13}$$

where $\kappa_i = \frac{1}{1+\lambda_i^2}$ and the expected value in the right side of the equation can be interpreted as the amount of shrinkage toward 0, a posteriori. The half-Cauchy prior on λ_i implies a horseshoe-shaped $Be(\frac{1}{2}, \frac{1}{2})$ prior for the shrinkage coefficient κ_i. The left side of the horseshoe, $\kappa_i \approx 0$, yields virtually no shrinkage and describes signal, while the right side, $\kappa_i \approx 1$ yields nearly total shrinkage and describes noise.

They also proved tight bounds for the univariate density $p(\theta)$, although the analytic form could not be written; in particular $\lim_{\theta \to 0} p(\theta) = \infty$ and, for $\theta \neq 0$,

$$\frac{K}{2} \log(1 + \frac{4}{\theta^2}) < p(\theta) < K \log(1 + \frac{2}{\theta^2}) \tag{4.14}$$

where $K = 1/\sqrt{2\pi^3}$

4.3 Posterior MCMC

The model is run for the same samples from each of the 122 breast cancer patients for which we obtained aCGH copy number data and RNA expression data. We used the Agilent Human 4 × 44K arrays for copy number and Affimetrix HG-U133A arrays for gene expression. After appropriate data processing, we obtained log2 ratios between the sample copy number and the reference number for each of the 22,944 probes for aCGH, and expression quantifications for 11,306 genes for microarray. We then mapped 22,944 probes to the 11,306 genes, which gave us a matching between the probe ids on the aCGH and the gene ids on the microarrays.

On the same 122 patients we have the measure of their pathological complete response

after six months of treatments as I mentioned above; we thus used the data on 100 patients to obtain samples from the posterior distribution of the parameters and the other 22 to check for prediction performances. Posterior inference is carried out using a complete Gibbs sampling scheme that differs from the one presented in 3.2.2 in the following full conditionals.

Again we start with a transition probability that jointly updates e^w_{bt} and z^w_{bt}. This is done by first sampling e^w_{bt} from its conditional posterior , conditional on the observed data and all currently imputed parameters except z^w_{bt}, and thus z^w_{bt} is sampled from the complete conditional posterior, now conditioning also on the already imputed e^w_{bt}.

$$e^w_{bt}|\{-z^w_{bt}\} \sim \begin{cases} -1 & p = p_- \\ 0 & p = p_0 \\ 1 & p = p_+ \end{cases} \tag{4.15}$$

where $\{-z^w_{bt}\} = \{z^w_{1t}, ..., z^w_{b-1,t}, z^w_{b+1,t}, ..., z^w_{Bt}\}$ and the probabilities are given by

$$p_- \propto \Phi(-1|\mu_1, \sigma_1^2)(\frac{1}{\phi_b^-})I(w_{bt} \in [-\phi_b^-, 0]) \tag{4.16}$$

$$p_0 \propto (\Phi(1|\mu_1, \sigma_1^2) - \Phi(-1|\mu_1, \sigma_1^2))\phi(w_{bt}|0, \nu_b^2) \tag{4.17}$$

$$p_+ \propto (1 - \Phi(-1|\mu_1, \sigma_1^2))(\frac{1}{\phi_b^+})I(w_{bt} \in [0, \phi_b^+]) \tag{4.18}$$

in turn properly normalized.

The conditional posterior probabilities for e^w_{bt} are proportional to the respective prior probabilities times the likelihood factor corresponding to w_{bt}, where the prior probabilities are the appropriate normal quantiles and the evaluation of the likelihood needs to include indicators for the finite support of the uniform kernels in (3.1).

z^w_{bt} is sampled form a truncated Normal distribution that guarantees the latent probit score to be consistent with the relative POE indicator of copy number aberration.

$$z^w_{bt}| \cdot \cdot \sim N(\mu_1, \sigma_1^2)[I(z^w_{bt} < -1)I(e^w_{bt} = -1) + \cdots] \tag{4.19}$$

where

$$\mu_1 = (\frac{\lambda_{l_g}}{m_{bg}}(z^y_{gt} - \alpha_g - d^{yu}_g q_g u_t) - \frac{\lambda^2_g}{m^2_{bg}} \sum_{-b} z^w_{bt} + \frac{1}{\sigma^2_a}(z^w_b + d^{wu}_g p_g u_t))/(\frac{\lambda^2_g}{m^2_{bg}} + \frac{1}{\sigma^2_a}) \quad (4.20)$$

$$\sigma^2_1 = (\frac{\lambda^2_{l_g}}{m^2_{bg}} + \frac{1}{\sigma^2_a})^{-1} \quad (4.21)$$

and $m_{bg} = \#\{b \in g\}$.

The conditional posterior probabilities for e^w_{bt} and z^y_{gt} are computed as above

$$e^y_{gt}|\{-z^y_{gt}\} \sim \begin{cases} -1 & p = p_- \\ 0 & p = p_0 \\ 1 & p = p_+ \end{cases} \quad (4.22)$$

where the probabilities are given by

$$p_- \propto \Phi(-1|\alpha_g + d^{yu}_g q_g u_t + z^w_{gt}\lambda_{l_g}, 1)(\frac{1}{\psi^-_g})I((y_{gt} - \mu_g - \alpha_t) \in [-\psi^-_g, 0]) \quad (4.23)$$

$$p_0 \propto (\Phi(1|\alpha_g + d^{yu}_g q_g u_t + z^w_{gt}\lambda_{l_g}, 1) - \Phi(-1|d^{yu}_g q_g u_t + z^w_{gt}\lambda_{l_g}, 1))\phi(y_{gt} - \mu_g - \alpha_t|0, s^2_g) \quad (4.24)$$

$$p_+ \propto (1 - \Phi(1|\alpha_g + d^{yu}_g q_g u_t + z^w_{gt}\lambda_{l_g}, 1))(\frac{1}{\psi^+_g})I((y_{gt} - \mu_g - \alpha_t) \in [0, \psi^+_g]) \quad (4.25)$$

in turn properly normalized.

$$z^y_{gt}| \cdot \cdot \sim N(\alpha_g + d^{yu}_g q_g u_t + z^w_{gt}\lambda_{l_g}, 1)[I(z^y_{gt} < -1)I(e^y_{gt} = -1) + \cdots] \quad (4.26)$$

The conditional posterior probabilities for d^{wu}_g and d^{yu}_g are proportional to the respective prior probabilities times the likelihood factor corresponding to z^w_{bt}, where the prior probabilities are defined in such a way that p is very close to 1, to express our prior belief that the prediction of u_t for new patients in the study must be depending on a few list of genes .

$$d^{wu}_g \sim \begin{cases} 0 & \text{p} \propto p(G - \sum_{-g} d^{wu}_g) \prod_{t=1}^T \prod_{b \in g} \phi(z^w_{bt}|z_b, \sigma^2_a) \\ 1 & \text{p} \propto (1-p)(1 + \sum_{-g} d^{wu}_g) \prod_{t=1}^T \prod_{b \in g} \phi(z^w_{bt}|z_b + p_g u_t, \sigma^2_a) \end{cases} \quad (4.27)$$

$$d_g^{yu} \sim \begin{cases} 0 & \text{p} \propto p(G - \sum_{-g} d_g^{yu}) \prod_{t=1}^{T} \phi(z_{gt}^y | \alpha_g + \lambda_{l_g} z_{gt}^w, 1) \\ 1 & \text{p} \propto (1-p)(1 + \sum_{-g} d_g^{yu}) \prod_{t=1} \phi(z_{gt}^y | \alpha_g + q_g u_t + \lambda_{l_g} z_{gt}^w, 1) \end{cases} \tag{4.28}$$

While the posterior probabilities for the parameters p_g and q_g are the updated Gaussian distributions obtained through the conjugate prior assumptions.

$$p_g \sim \begin{cases} N(\hat{p_g}, \sigma(\hat{p_g})^2) & \text{if } d_g^{wu} = 0 \\ N(\frac{\frac{\sum_{t=1}^{T} \sum_{b \in g} u_t (z_{bt}^w - z_b^w)}{\sigma_a^2} - \frac{\hat{p_g}}{\sigma(\hat{p_g})^2}}{\frac{1}{\sigma(\hat{p_g})^2} + \frac{\sum_{t=1}^{T} \sum_{b \in g} u_t^2}{\sigma_a^2}}, \frac{1}{\frac{1}{\sigma(\hat{p_g})^2} + \frac{\sum_{t=1}^{T} \sum_{b \in g} u_t^2}{\sigma_a^2}}) & \text{if } d_g^{wu} = 1 \end{cases} \tag{4.29}$$

$$q_g \sim \begin{cases} N(\hat{q_g}, \sigma(\hat{q_g})^2) & \text{if } d_g^{yu} = 0 \\ N(\frac{\sum_{t=1}^{T} u_t (z_{gt}^y - \lambda_{l_g} z_{gt}^w) - \frac{\hat{q_g}}{\sigma(\hat{q_g})^2}}{\frac{1}{\sigma(\hat{q_g})^2} + \sum_{t=1}^{T} u_t^2}, \frac{1}{\frac{1}{\sigma(\hat{p_g})^2} + \sum_{t=1}^{T} u_t^2}) & \text{if } d_g^{yu} = 1 \end{cases} \tag{4.30}$$

Finally, the clue difference with the model in chapter 3, is the conditional posterior distribution of u_t for new patients coming into the study, that we shall use to predict the clinical outcome of such patients.

It is an updated Bernoulli with probabilities that are proportional to the respective prior probabilities of u_t times the likelihood corresponding to the latent factors z_{bt}^w and z_{gt}^y

$$u_{t+i} \sim \begin{cases} 0 & \text{p} \propto (1-\pi) \prod_{g=1}^{G} \prod_{b \in g} \phi(z_{bt}^w | z_b^w, \sigma_a^2) \prod_{g=1}^{G} \phi(z_{gt}^y | \alpha_g + \lambda_{l_g} z_{gt}^w, 1) \\ 1 & \text{p} \propto \pi \prod_{g=1}^{G} \prod_{b \in g} \phi(z_{bt}^w | z_b^w + d_g^{wu} p_g, \sigma_a^2) \prod_{g=1}^{G} \phi(z_{gt}^y | \alpha_g + d_g^{yu} q_g + \lambda_{l_g} z_{gt}^w, 1) \end{cases} \tag{4.31}$$

with $i = 1, ..., n$.

4.4 Posterior inference on the breast cancer dataset

4.4.1 Sensitivity, specificity and the ROC curve

Sensitivity and specificity have their origins in screening tests for diseases.

When a single test is performed, the person may in fact have the disease or the person

Test result / **True status**	Positive	Negative
Response (+)	a	b
No response (-)	c	d

Table 4.1: False positive and negative counts

may be disease free; in our particular case that is represented by patients who showed pathological complete response and patients who did not. The test result may be positive, indicating the *"complete disappearance of the tumor with no more than a few scattered tumor cells detected by the pathologist in the resection specimen"*, or the test result may be negative, indicating the absence of this response to the treatment.
Table 4.1 displays test results in the columns and true status of the person being tested in the rows. We defined sensitivity as the probability that the test says a patient had a pathological complete response when in fact he had. This is $P(T^+|S^+) = \frac{a}{a+b}$. It is a measure of how likely it is for the test to detect the pathological complete response of a patient who showed the disappearence of the tumor after 6 months of treatment.
On the other side, the specificity is the probability that the test says a person does not show response to the treatment when in fact he did not respond to such treatment. This is $P(T^-|S^-) = \frac{a}{c+d}$.
Ideally, a test should have high sensitivity and high specificity. Sometimes there are trade-offs in terms of sensitivity and specificity. For example, we can make a test have very high sensitivity, but this sometimes results in low specificity. Generally we are able to keep both sensitivity and specificity high in screening tests, but we still get false positives and false negatives.

A receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot of the sensitivity, or true positive rate, vs. false positive rate (1 - specificity or 1 - true negative rate), for a binary classifier system as its discrimination threshold is varied. The

ROC can also be represented equivalently by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate)(7).
To draw an ROC curve, only the true positive rate (TPR) and false positive rate (FPR) are needed. TPR determines a classifier or a diagnostic test performance on classifying positive instances correctly among all positive samples available during the test. FPR, on the other hand, defines how many incorrect positive results occur among all negative samples available during the test.
A ROC space is defined by FPR and TPR as x and y axes respectively, which depicts relative trade-offs between true positive (benefits) and false positive (costs). Since TPR is equivalent with sensitivity and FPR is equal to 1 - specificity, the ROC graph is sometimes called the sensitivity vs (1 - specificity) plot. Each prediction result or one instance of a confusion matrix represents one point in the ROC space.
The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). The (0,1) point is also called a perfect classification. A completely random guess would give a point along a diagonal line (the so-called line of no-discrimination) from the left bottom to the top right corners. An intuitive example of random guessing is a decision by flipping coins (head or tail).
The diagonal divides the ROC space. Points above the diagonal represent good classification results, points below the line poor results. Note that the output of a poor predictor could simply be inverted to obtain points above the line.
The area measures discrimination, that is, the ability of the test to correctly classify those with and without response to the treatment. Consider the situation in which patients are already correctly classified into two groups. You randomly pick on from the group who showed pCR and one from the group who does not and do the test on both. The patient with the more abnormal test result should be the one from the pCR group. The area under the curve is the percentage of randomly drawn pairs for which this is true (that is,

the test correctly classifies the two patients in the random pair).
Summarizing the ROC curve indicates several things:

- it shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity);
- the closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test;
- the closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test;
- the area under the curve is a measure of text accuracy.

4.4.2 Predictive performances

The following is the plot of the empirical ROC curve

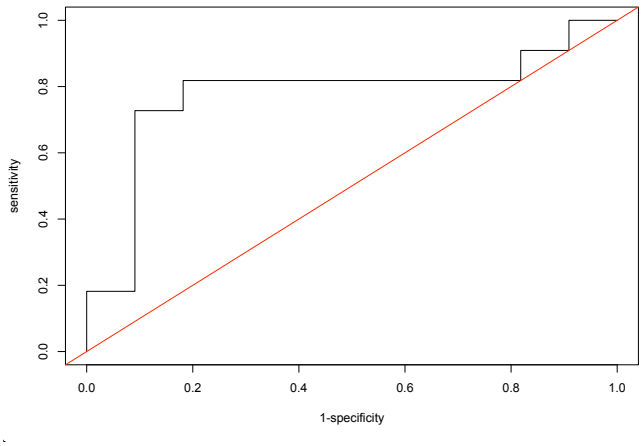

Figure 4.2: ROC curve.

obtained with some simulated counts of false positive and negative.
The approach we used in the estimation of the ROC curve was LLoyd and Yong's one

(26), who is proved to be performing better than the empirical estimation. They proposed to estimate this curve from kernel smoothing of the distribution functions of the diagnostic measurement underlying the binary decision rule, i.e. the conditional posterior probabilities of positive pathological complete response, and showed the significant accuracy achieved by this method for realistic sample size compared with the empirical estimation.

The tests we performed were done for a sample of 22 patients for which we had previously measured their pCR.

Tests are based on the posterior probabilities of the clinical outcome being 1,

$P(u_t = 1|data)$, obtained running the Gibbs Sampler for 30.000 iterations.

Although we have not found any existing model on prediction using integrated platforms to compare our perfomances with, the area under the curve is slightly above .9, which represents an excellent predictive performance.

Moreover, the flexibility of the model allowed us to make a comparison between the predictive performances using either gene expression only (which is broadly found in recent literature) or the integrated platforms of gene expression and copy number variation data. The results are shown in the figure below and confirm our choice of borrowing information between the two genomic platforms.

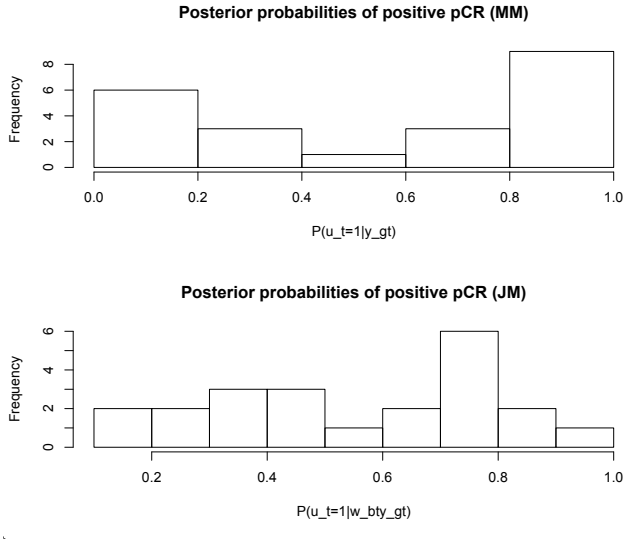

Figure 4.3: Histograms of the posterior probability of positive pCR in the marginal and joint models

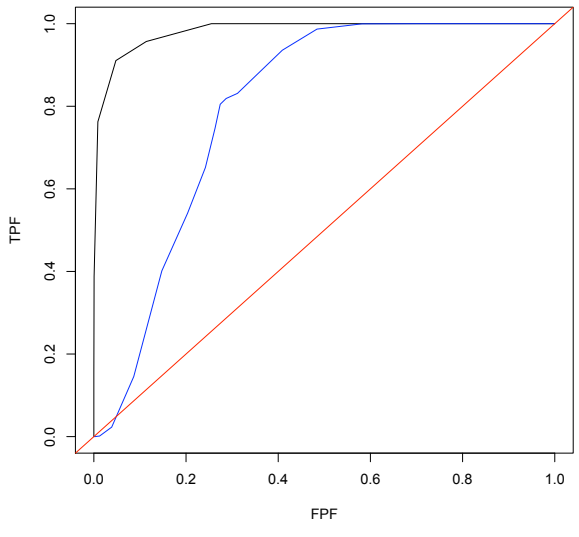

Figure 4.4: Comparison between the ROC curves obtained using the marginal (black line) and the integrated models (blue line)

Therefore, we showed the integrated model seems to predict perfectly the pathological complete response of patients, much better that the marginal one, although patients within our study come from different biological conditions.
It could be proven on a different data set, with all patients showing the biological condition of interest, that predictive performances of our method are even more precise.

4.5 Conclusion

The idea of this second model has raised after some data on the pathological complete response of the patients in the study became available. It could be applied to any clinical outcome of such patients and be of central importance in problems such as the randomization of patients in a clinical trial.
The approach is that of keeping the latent structures we defined in the model for assessing differential expression, i.e. the POE models for gene expression and copy number variation data, respectively w_{bt} and y_{gt}, and defining the latent probit scores through regression on the clinical outcome of patients, u_t, instead of on their biological condition, such as the subgroup of breast cancer showed by their cancerous cells.
Although it seems rather counter intuitive to define a model for $p(w, y|u)$ instead of a simple logistic regression of u_t on the data, the assumption of a prior Bernoulli distribution on the pathological complete response of new patients coming into the study, and the exploitation of the Bayesian framework, lead us to the searched model $p(u|w, y)$. It moreover allowed us for a massive reduction of the noise in the data and a within selection of a small subset of genes, among the 11,306, contributing to the prediction of the chosen outcome.
It seems to be quite a new approach in recent literature, also with respect to the previous model, since we are lately assisting to a growth in the number of models dealing with the integration of genomic platforms for the assessment of differential expression, and yet

very few of such models, if any, allow for prediction of a clinical outcome.
Results shown in the previous section seem rather promising, showing the better performances obtained using the integrated data with respect to gene expression data marginally. For possible drawbacks and modification of the assumptions of the model one may refer to the discussion chapter.

5

Discussion

At the beginning of this work I was facing the problem of analyzing two different types of measurements on a sample of 122 breast cancer patients. I was given information about the expression of their genes and for the same genes a data on their copy number variations.
On the same sample we had information about the subgroup of breast cancer the patients belong to, either triple negative or ER+/HER2+ and the pathological response of the patients after six months of treatment.
The biological questions we were asked by clinicians working directly with the patients were the following:

- Is there any significant genetic difference between patients belonging to triple negative subgroup and other patients?
- Are there specific genes which "behave" differently in different subgroups?
- Is it possible to detect which genes among many show this differential behaviour, so that targeted therapies can be developed?
- How can the interaction between DNA and RNA stated by the central dogma of molecular biology can be quantified?

- With a further information on a particular clinical outcome for the patients, how can we detect a list of genes which distinguish a positive response from a negative one?

- Is is possible to predict such outcome, and doed the use of integrated platforms lead to better predition performances?

We thus tried and developed a flexible model that would take into account all these questions, would able to answer all of them and would work for our particular data set as well as for any other.
We studied existing literature and took from it all the broadly approved statistical assumptions on these types of data, noticing that really few works were dealing with the integration of the two platforms (gene expression microarrays and arrays CGH).
Finally we structured two different Bayesian hierarchical models, the first of which allowed us extend the concept of differential gene expression and relate it to copy number variations of the same genes and quantified the interaction between variations in the copy number and the resulting expressions of genes.
As for the second model, it proved to be able to integrate different genomic platforms and predict the pathological complete response of patients in the breast cancer study, with much better predictive performances that were achieved using marginal data.
Clinicians were given the list of genes that showed particular behaviour in TN subgroup with respect to others and suggested interesting existing molecular pathways using those genes.
These pathways moreover showed the centrality of MYC, a gene which has been widely studied lately in problems related with the poor prognosis of triple negative breast cancer, and brought to light other possible ways of developing targeted treatments.
There are few possible weaknesses of the procedure, mainly related to the prior specification for parameters directly related to differential expression and prediction. We were dealing with highly parametrized models and few observations data sets, reason why we

chose some easier shortcuts in order to achieve faster MCMC convergency.
It is also to be pointed out that the model was run for 30000 iterations and it was, at any iteration, sampling millions of parameters; the application of our model to the data sets regarding the breast cancer study was a main hump we were able to get over with the use of C language instead of statistical softwares such as R which were not computationally adequate for our purpose.
Some interesting modifications that could be applied to our prior specifications were mentioned in Chapter 4 and are now to be implemented, since we found in literature new and more efficient approaches to the issue of sparsity.
Also, it was very hard to compare our models' performances with other methods, either due to the lack of codes or to the scarcity of works on the specific topic of integrated platforms prediction, and yet we reckon very important to check whether there was an effective improving.
We then discussed in the beginning the starting point of all our work, i.e. the central dogma of molecular biology, which leads to another important research question that could be addressed to us and thus developed starting from these models: would it be possible to take into account three or more platforms, one of which could be related to the third level of such dogma, i.e. proteins? This problem could require massive modifications of the assumptions we made throughout this thesis, and yet be a completely new field to pay much attention to.
Finally, all this project was focused on a specific data set, with rather particular features. The natural hierarchical structure and correlation between DNA and RNA makes very hard to think of the application of our methodology to different problems, though an interesting path to follow could be that of demographical sciences, where this hierarchical structure could be found for example in data at country level and regional level.

Bibliography

[1] Albert, B., Johnson, A., Lewis, J. , Raff, M., Roberts, K., Walter, P. (2002), *Molecular biology of the cell*, 4^{th} edition, published by Garland Science 8, 9, 10

[2] Andre, F., Job, B., Tordai, A., Michiels, S., Liedtke, C., Richon, C., Yan, K., Wang, B., Vassal, G., Delaloge, S., Hortobagyi, G. N., Symmans, W. F., Lazar, V. and Pusztai, L., (2009) *Molecular characterization of breast cancer with high-resolution oligonucleotide comparative genomic hybridization*, Clinical Cancer Research **15**(2) pp. 441–451.

[3] Baggerly, K. A., Coombes K. R., Morris, J. S., *An introduction to High - Throughput Bioinformatics Data*, in "Bayesian Inference in Gene Expression and Proteomics". Ed(s) Do, K. A., Müller, P., Vannucci, M. New York: Cambridge University Press pp 1–33.

[4] Baladandayuthapani, V., Ji, Y., Talluri, R., Nieto-Barajas, L. E., Morris, J. S., (2010) *Bayesian Random Segmentation Models to Identify Shared Copy Number Aberrations for Array CGH Data*, JASA In Press.

[5] Benjamini, Y., Hochberg, Y., (1995) *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society B **57**, pp 289–300. 47

[6] Bonnefoi, H. (2007) *Validation of gene signatures that predict the response of breast*

cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial., Lancet Oncology **8**(12) pp. 1071–1078. 60

[7] Bradley, A. P. (1997) *The use of the area under the curve in the evaluation of machine learning algorithms*, Pattern Recognition - Elsevier. 72

[8] Brown, P., Vannucci, M., (1998) *Multivariate Bayesian variable selection and prediction*, Journal of the Royal Statistical Society B **60**(3), pp 627–641.

[9] Cappuzzo F., Hirsch, F.R. et al. (2005) *Epidermal growth factor receptor gene and protein and gefitinib sensitivity in non-small-cell lung cancer*, Journal of the National Cancer Institute **97**(9): pp 643–655 21

[10] Carvalho C.M., Polson N.G., Scott J.G. (2010) *The horseshoe estimator for sparse signal*, Biometrika, pp 1–16 66

[11] Chib, S., Greenberg, E., (1998) *Analysis of multivariate probit models*, Biometrika **85**(2), pp 347–361 34, 61

[12] Choi, H., Qin, Z. S., Ghosh, D., (2010) *A double-layered mixture model for the joint analysis of DNA copy number and gene expression data*, Journal of Computational Biology **17**(2), pp 121–137 4, 26

[13] Churchill, F.B. (1974) *William Johannsen and the genotype concept*, Journal of the History of Biology **7** pp 5–30. 10

[14] Chustecka, Z. (2007). *Survival Disadvantage Seen for Triple-Negative Breast Cancer*, Medscape Medical News 31

[15] Crick, F. (1970) *Central dogma of molecular biology*, Nature **227** (5258): pp 561–3. 10

[16] Dent, R., Trudeau, M., Pritchard, K. I., Hanna, W.M., Kahn, H.K. et al (2007). *Triple-Negative Breast Cancer: Clinical Features and Patterns of Recurrence*, Clinical Cancer Research (American Association for Cancer Research) **13**, pp 442–4434 31

[17] Do, K. A., Müller, P., Tang, F., (2005) *A Bayesian Mixture Model for differential gene expression*, Journal of the Royal Statistical Society C **54**(3), pp 627–644.

[18] Efron, B., Tibshirani, R., (2006) *On testing the significance of sets of genes*, The Annals of Applied Statistics **1**, pp 101–129

[19] Fridlyand, J., Snijders, A., Pinkel, D.G., Jain, A. N., (2004) *Application of Hidden Markov Models to the analysis of the array CGH data*, Journal of Multivariate Analysis **90**, pp 132–153 25

[20] evaert O., Smet F. D., Timmerman D., Moreau Y. and Moor B. D. (2006), *Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks* , Bioinformatics **22**, pp 184190.

[21] Gonzalez, E. et al. (2005) *The Influence of CCL3L1 Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility* Science **307** (5714): 14341440 21

[22] Guha, S., Li, Y., Neuberg, D., (2008) *Bayesian Hidden Markov modeling of array CGH data*, JASA **103**, pp 485–497

[23] Johnson V.E., Rossell D. (2010) *On the use of non-local prior densities in Bayesian hypothesis tests*, JRSSB **72**, pp 143–170 65, 66

[24] Kallioniemi, A.,Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J., Waldman, F., Pinkel, D. (1992) *Comparative genomic hybridization for molecular cytogenetic analysis of solid tumor*, Science **258**, pp 818–821 21

[25] Kidd, J.M., Cooper, G.M., Donahue, W.F. et al. (2008) *Mapping and sequencing of structural variation from eight human genomes*, Nature **453**(7191): pp 56–64 21

[26] Lloyd, C.J., Yong, Z. (1999) *Kernel estimators of the ROC curve are better than empirical* Statistics & Probability Letters **44**: pp 221–228 74

[27] Lashkari, D.A., DeRisi, J.L., McCusker, J.H., Namath, A.F., Gentile, C., Hwang, S.Y., Brown, P.O., Davis, R.W. (1997) *Yeast microarrays for genome wide parallel genetic and gene expression analysis* Proc Natl Acad Sci USA **94**(24): pp 13057–13062 13

[28] Ludwig, J.A., Weinstein, J.N. (2005) *Biomarkers in cancer staging, prognosis and treatment selection*, Nature **5**, pp 845 – 856 1

[29] M*ü*ller, P., Parmigiani, G., Rice, K. (2007) *FDR and Bayesian multiple comparison*, in "Bayesian Statistics". Ed(s) Bernardo, J. M., Bayarri, S., Berger, J. O., Dawid, A. P., Heckerman D., Smith, A. F. M., West, M. Oxford University Press. 47

[30] Parmigiani, G., Garrett, E. S., Anbazhagan, R., Gabrielson, E., (2002) *A statistical framework for expression-based molecular classification in cancer*, Journal of Royal Statistical Society B **64**, pp 717–736 4, 15, 29, 32, 60

[31] Pinkel, D., Albertson, D. G., (2005) *Array comparative genomic hybridization and its applications in cancer*, Nature Genetics **23**, pp 41–46 21

[32] Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W., Chen, C., Zhai, Y., Dairkee, S., Ljung, B., Gray, J. W., Albertson, D. G., (1998) *High resolution analysis of DNA copy number variation using comparative genomic hyridization to microarrays*, Nature Genetics **20**, pp 207–211 21

[33] Pollack, J., Sorlie, T., Perou, C., Ress, C., Jeffrey, S., Lonning, P. R. T., Botstein, D., Borresendale, A. L., Brown, P., (2002) *Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumor*, proceedinfs of the National Academy of Sciences **99**, pp 12963–12968 3, 26

[34] Rabiner, L.W. (1989) *A tutorial on Hidden Markov Models and selected applications in speech recognition*, Proceedings of the IEEE **77**(2): pp 257–286 25

[35] Redon, R. et al (2006) *Global variation in copy number in the human genome*, Nature **444**, pp 444–454 11

[36] Schena, M., Shalon, D., Davis, R.W., Brown, P.O. (1995) *Quantitative monitoring of gene expression patterns with a complementary DNA microarray* Science **270** (5235): pp 467–470 13

[37] Scott, J. G., Berger, J. O., (2006) *An exploration of aspects of Bayesian multiple testing*, JSPI **136**, pp 2144–2162 47

[38] Scott, J. G., Berger, J. O., (2010) *Bayes and empirical-Bayes multiplicity adjustment in the variable selection problem*, The Annals of Statistics **38**(5), pp 2587–2619 47

[39] Snijders, A., Nowak, N., Segraves, R., Blackwood, S., Brwon, N., Conroy, J., Hamilton, G., Hindle, A., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., yue, J., Gray, J. W., Jain, A., Pinkel, D., Albertson, D. G., (1998) *Assembly of microarrays for genome-wide measurement of DNA copy number*, Nature Genetics **29**, pp 263–264 21

[40] Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Döhner, H., Cremer, T., Lichter, P., (1997) *Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances*, Genes Chromosomes Cancer **20**(4), pp 399–407 21

[41] Storey, J. D., (2002) *A direct approach to false discovery rate*, Journal of Royal Statistical Society B **64**(3), pp 479–498

[42] Storey, J. D., (2003) *The positive flase discovery rate: a Bayesian interpretation and the q-value*, The Annals of Statistics **31**(6), pp 2013–2035

[43] Taniguchi, M., Miura, K., Iwao, H., Yamanaka, S. (2001) *Quantitative Assessment of DNA Microarrays Comparison with Northern Blot Analysis*, Genomics **71** pp 34–39. 13

[44] Van Wieringen, W., Wiel, M. A., (2090) *Nonparametric testing for DNA copy number induced differential mRNA gene expression*, Biometrics **65**(1), pp 19–29 3, 26

[45] Watson, J. D., Crick, F.H. (1953) *Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid*, Nature **171** (4356): pp 737–738 9

[46] Xu, J., Chen, Y., Olopade, O.I. (2010) *MYC and breast cancer*, Genes & cancer **1** (6): pp 629–640 57

[47] *http://www.affymetrix.com* 13

[48] *http://bioinformatics.mdanderson.org/tut-image.html* 15

[49] *http://www.ornl.gov* 7

[50] *http://www.genego.com* 54