ABSTRACT

(Statistical Science)

BAYESIAN FUNCTION ESTIMATION USING OVERCOMPLETE DICTIONARIES WITH APPLICATION IN GENOMICS

by

Jen-hwa Chu

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Dr. Merlise A. Clyde, Supervisor

_____
Dr. Feng Liang, Supervisor

_____
Dr. Edwin S. Iversen

_____
Dr. Jeffrey R. Marks

_____

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University

2007

BAYESIAN FUNCTION ESTIMATION USING OVERCOMPLETE DICTIONARIES WITH APPLICATION IN GENOMICS

by

Jen-hwa Chu

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Dr. Merlise A. Clyde, Supervisor

_____
Dr. Feng Liang, Supervisor

_____
Dr. Edwin S. Iversen

_____
Dr. Jeffrey R. Marks

_____

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Statistical Science
in the Graduate School of
Duke University

2007

Abstract

In this dissertation we present a Bayesian approach for nonparametric function estimation based on continuous wavelet dictionaries, where the unknown function is modeled by a sum of wavelet functions at arbitrary locations and scales. By avoiding the dyadic constraints for orthonormal wavelet bases, the continuous wavelet dictionaries have greater flexibility to adapt to the structure of the data, and lead to sparser representations. The price for this flexibility is the computational challenge of searching efficiently over an infinite number of potential dictionary elements. We develop a reversible jump Markov Chain Monte Carlo algorithm which utilizes local features in the proposal distributions for the addition of new wavelet elements to improve mixing of the Markov chain. By utilizing continuous wavelets, we have the flexibility to handle data with non-equal spacing without resorting to interpolation or imputation of missing data.

In Chapter 1 we start with a review of wavelets and function estimation and provide an overview of array Comparative Genomic Hybridization (CGH) and gene expression data. Chapter 2 introduces the continuous wavelet dictionaries. We discuss the basic setting of the model and estimation. We present simulation results using standard wavelet test functions, which show that the new method leads to greater sparsity and improved mean square error over translational invariant wavelets, another overcomplete representation. We illustrates the method on non-equally spaced data, and show that the method compares favorably to methods using interpolation or imputation.

In Chapter 3 and 4 we present applications with array CGH, which is a technology used to detect DNA copy number alterations that could help identify the relevant genes for cancer development. This recent technology calls for new statistical meth-

ods for analyzing array CGH data. In Chapter 3 we present a hierarchical model to analyze multiple samples via a functional data analysis approach using the over-complete dictionaries. The hierarchical model is based on samples grouped according to the disease progression and survival status. The posterior probabilities of copy gain/loss are estimated for each gene at the group level. From that result, we can also classify new patients and identify the genes relevant to the group differences. We demonstrate the performance of our method using simulated and real data sets.

In Chapter 4 we extend our model to analyze gene expression and gene copy number alterations jointly. Both types of data have been linked to cancer development and progression and have been studied extensively to describe the pattern of expression levels and copy number changes in cancer. However, uncovering the genes related to cancer development is still a difficult task and few studies have combined analysis of both data types. Here we discuss our model and inference methods for joint analysis of these two genomic measurements. We present results from simulation studies and the breast cancer cell line data published by Hyman *et al.* (2002). We provide estimates for both gene expression levels and DNA copy numbers, along with the degree to which the two types of data are associated. We identify a subset of genes for which the expression levels are most likely attributable to gene copy number alterations across the samples, including some of the oncogenes that were previously associated with breast cancer and some new targets.

Acknowledgements

During the last five years in ISDS/DSS there are so many people who have helped me and encouraged me and I would like to thank all of them. First and foremost, my thanks go out to my advisors, Dr. Merlise Clyde and Dr. Feng Liang, for their support and encouragement throughout the whole process of my dissertation work. I have been fortunate to work with these two of the most brilliant and hard-working people in this field and I know I can always count on them to point me to the right direction when I hit a roadblock. I would also like to thank my committee members, Dr. Ed Iversen and Dr. Jeffrey Marks, for their comments on the genomic application work and CGH data analysis. I am grateful to Dr. David Banks, Dr. Bertrand Clarke and Dr. Marianna Pensky for introducing me to various aspects of data mining, high-dimensional data analysis and wavelets, providing a solid foundation for my thesis work.

I would like to thank all my friends here at Duke and Durham area. I could not have got through the toughest time without them. I also thank the wonderful ISDS/DSS staff, Pat, Kris, Anne, Nikki, Eric and Lance for their help throughout the years.

Finally, I would like to thank my parents, for their endless loving and support during the whole time of my study. Without their support the dissertation would not have been possible.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Massive high-dimensional data brought by new technology, such as microarray and gene expression data, generate an increasing demand for new data analysis tools. A fundamental statistical question there is how to retrieve or estimate a signal (or function) from the massive data. Many of these application problems can be thought of as a non-linear regression problem which is to estimate a function $f(x)$ conditional on data generated from f. For that purpose, non-parametric function estimation methods have been popular as they do not restrict f to have some predetermined form. In Bayesian non-parametric methods, the unknown function f is modeled with infinitely many parameters, which gives a wide range of possible functions.

Wavelets are an attractive option for non-parametric regression and have been of great interest for modeling features in many statistical application areas as they offer better localization and parsimony than other orthonormal bases. In this dissertation we propose a new class of non-parametric regression models based on continuous wavelets using an overcomplete wavelet dictionary, where the number of basis elements is greater than a regular orthonormal wavelet basis. In our model the unknown function f is modeled by a sum of an unknown number of wavelet functions at arbitrary locations and scales. We start by introducing the basics of wavelets and overcomplete dictionaries and discuss their advantages over other non-parametric methods.

The continuous wavelet dictionaries can be useful in many application areas. In this dissertation we focus on the application in genomic microarray data. The microarray data usually consist of measurements of thousands of genes or gene segments,

such as the numbers of DNA copies and the expression level. Nonparametric methods are frequently used in analysis of microarray data since their patterns usually do not have an obvious parametric form (Hsu *et al.*, 2005). We will review the basic of array CGH and expression data and motivate the use of wavelet-based methods in these application problems. We then provide an outline of the remaining chapters in the thesis.

1.1 Wavelets

Wavelet bases are generated by translation and dilation of a single wavelet function $\psi(x)$ called the mother wavelet. For example, the simplest wavelet basis is the Haar basis which is generated by the Haar function

$$\psi(x) = \begin{cases} -1/\sqrt{2} & \text{if} \quad -1 < x \leq 0 \\ 1/\sqrt{2} & \text{if} \quad 0 < x \leq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

The family of dyadic dilations and translations $\{\psi_{jk}\}_{j,k\in\mathbb{Z}}$ where $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$ forms an orthonormal basis for the space of L^2 functions, where j and k are indices for the scale and location, respectively. Some sample translations and dilations for the Haar wavelets family are graphed in Figure 1.1. Any L^2 function f can be represented as

$$f(x) = \sum_{j,k} \theta_{jk}\psi_{jk}(x),$$

where the coefficient θ_{jk} can be obtained by

$$\theta_{jk} = \int \psi_{jk}(u)f(u)du.$$

An advantage of wavelets is their flexibility. Since wavelets are localized in both time and frequency, they can be used to represent both smooth and locally bumpy

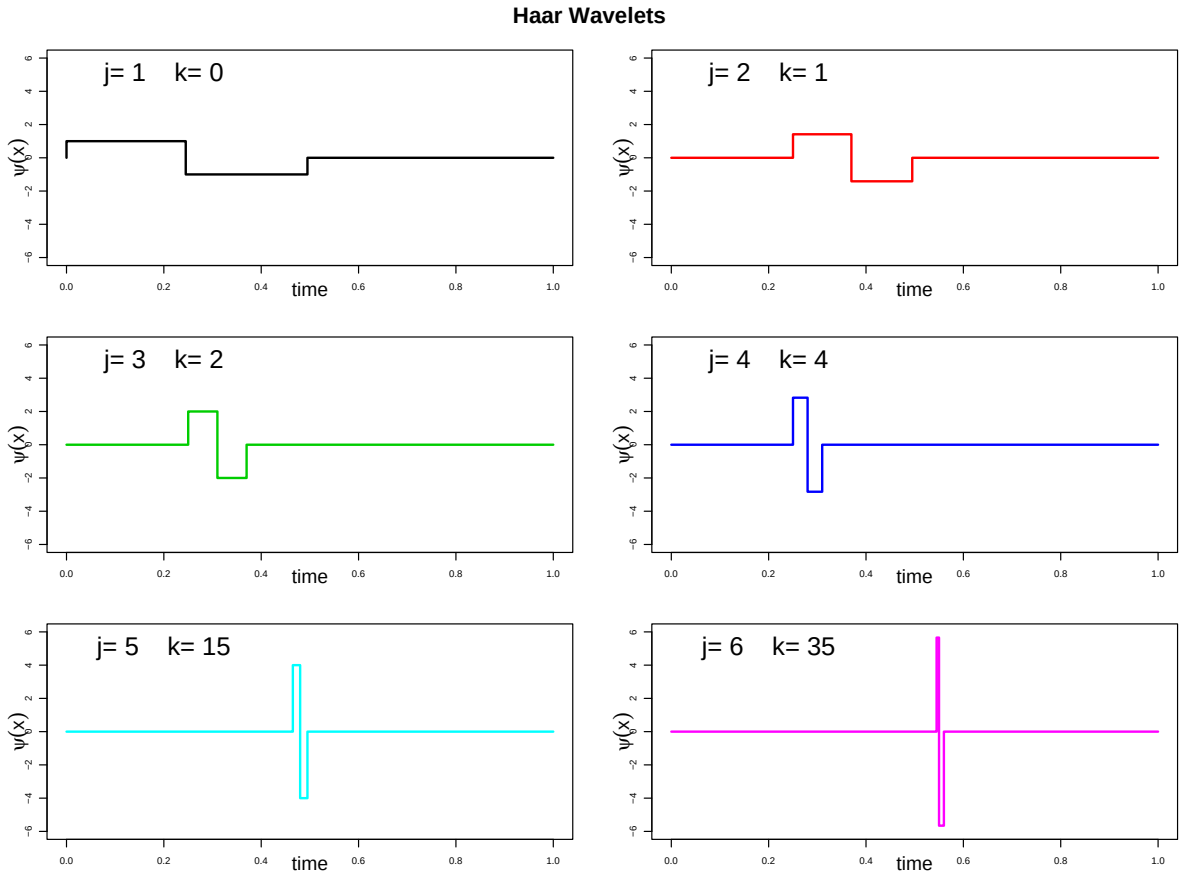

Figure 1.1: Some translations and dilations for the Haar wavelets family.

function with only a few coefficients. On the other hand, a Fourier basis is only localized in frequency and may not lead to a sparse representation for a wide range of functions. We may pick the wavelet family with suitable level of smoothness and regularity for a given application. For example, Daubechies' Compactly Supported Wavelets are among the most commonly used wavelet bases. From Figure 1.2 we can see the varying degree of smoothness of the wavelet functions in the Daubechies' family. The computation for wavelet estimation with regularly spaced data is extremely fast and easy with discrete wavelet transform (DWT) and the Cascade algorithm which takes advantage of the tree-like structure of the basis elements. As a result, wavelet bases have become very popular in signal processing and compression.

When the data are non-equally spaced, the DWT cannot be directly applied. We can treat the data as if they were equally spaced; or we can approximate (interpolate) the function at equally spaced dyadic points. In either case the structure of the data may be distorted. Extensions of the DWT for the non-equally spaced data have been proposed by Sardy *et al.* (1999) and Pensky and Vidakovic (1998). However, most computer packages for wavelets, such as wavethresh and waveslim in R, still cannot accommodate non-equally spaced data.

1.2 Overcomplete Dictionaries

In contrast to the orthonormal basis traditionally used for function estimation, overcomplete (or redundant) representations have been advocated due to their flexibility and adaptation. In an overcomplete dictionary, the number of basis vectors is greater than the dimensionality of the data. Suppose we have a vector $y \in \mathbb{R}^n$, and a collection of vectors $a_i \in \mathbb{R}^n, i = 1, ...m$, where $m > n$. The collection of a_i's is "more than a basis". Therefore we usually refer to it as an overcomplete "dictionary" rather than a basis, and the dictionary elements are referred to as "atoms". If we want to

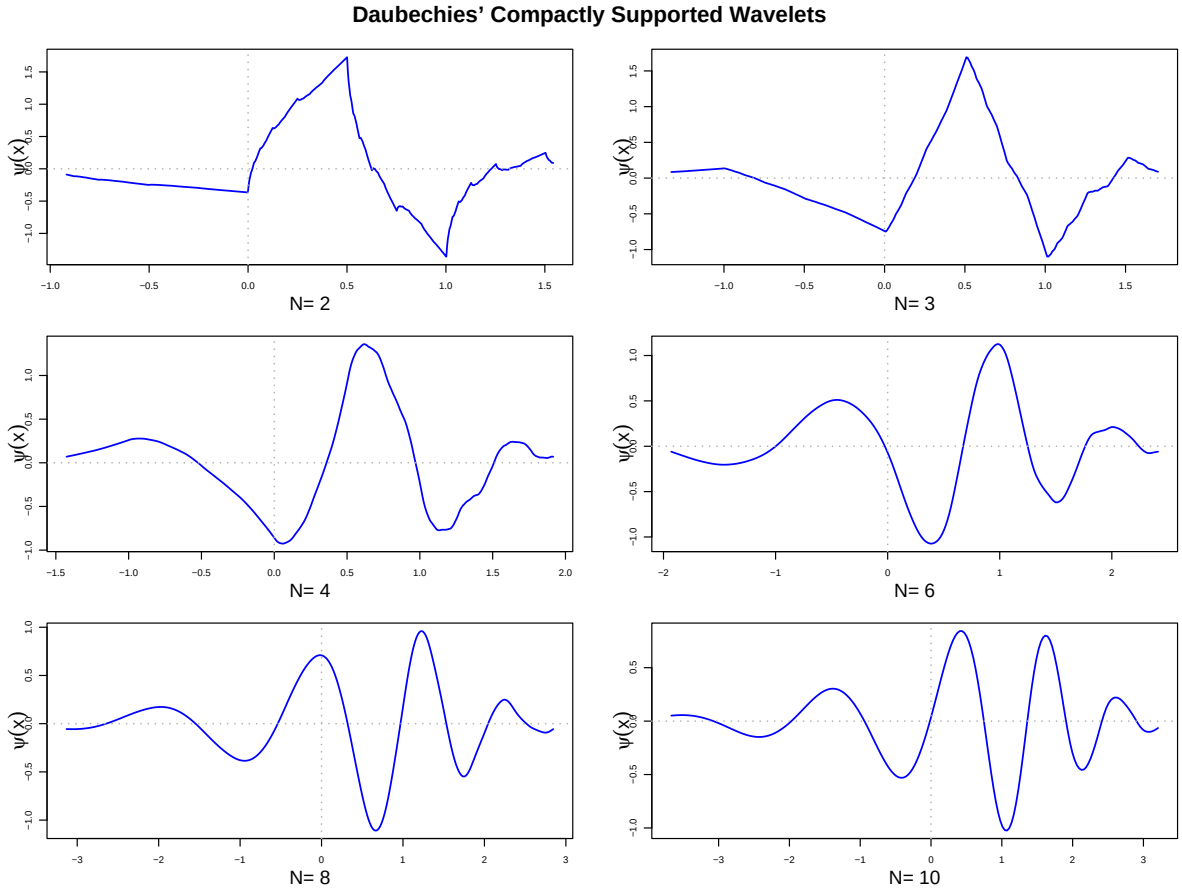

Figure 1.2: Graphs of wavelet functions from Daubechies' family. N here represents the number of vanishing moments of the wavelet. The larger N is, the smoother the wavelet function is.

represent our data y as a linear combination of the dictionary atoms, we will solve $\beta \in \mathbb{R}^m$ satisfying $y = \mathbf{A}\beta$, where $\mathbf{A} = \{a_1, ...a_m\}$. Since $m > n$, the solution is not unique.

An overcomplete representation, however, has some advantages. The overcomplete dictionary can offer greater flexibility to match the structure of the data, since regular bases are not designed specifically for the data under consideration (Donoho and Elad, 2003). The overcomplete dictionary can also represent functions more parsimoniously (Chen *et al.*, 2001; Coifman *et al.*, 1992; Mallat, 1998; Wickerhauser, 1994). It has been shown that the overcomplete dictionaries outperform other orthonormal bases in application area such as biology (Olshausen and Field, 1997) and signal processing (Berg and Mikhael, 1999; DeBrunner *et al.*, 1997). Theoretically, overcomplete dictionaries can give better approximation (Candés and Donoho, 2004).

One important example of overcomplete dictionaries is continuous wavelets (Vidakovic, 1999, Sec. 3.1). The atoms $\psi_{a,b}(x)$, indexed by a scale parameter a and a location parameter b, are the dilations and translations of the mother wavelet $\psi(x)$

$$\psi_{a,b}(x) = \frac{1}{\sqrt{a}}\psi\left(\frac{x-a}{b}\right),$$

where $a \in \mathbb{R} \setminus \{0\}, b \in \mathbb{R}$ and the mother wavelet ψ satisfies the admissibility condition:

$$C_\psi = \int_{\mathbb{R}} \frac{|\Psi(\omega)|^2}{|\omega|} d\omega < \infty,$$

where $\Psi(\omega)$ here is the Fourier transformation of $\psi(x)$.

The continuous wavelet transform of an L_2 function f is defined as

$$\mathcal{CWT}_f(a,b) = \int f(x)\overline{\psi_{a,b}(x)}dx.$$

With the continuous wavelets transform the original function f can be represented as

$$f(x) = \frac{1}{C_\psi} \int \int \psi_{a,b}(x) \mathcal{CWT}_f(a,b) \frac{da}{a^2} db.$$

Overcomplete dictionaries can also be constructed by combining multiple orthonormal bases, such as Fourier, Gabor, or wavelets. Other example of overcomplete dictionaries include frames (Gröchenig, K., 2001; Wolfe *et al.*, 2004); stationary (non-decimated) wavelets (Nason and Silverman, 1995); and wavelet packets (Coifman and Meyer, 1990).

Since there is no unique solution to the representation problem, to find the "best" representation among all the overcomplete representations can be a challenging problem, as we have to search over a much larger space of possible representations. Algorithms for this type of search can be unstable if the data are slightly disturbed. The development of overcomplete dictionaries and efficient learning algorithms for them has been an active research area (Lewicki and Sejnowski, 1998).

1.3 Array CGH Data

In recent years researchers have found strong evidences that cancer is related to the abnormalities in the number of copies of genomic DNA (Pinkel and Albertson, 2005). Recent technological advancement has made it possible to measure the copy numbers of segments of DNAs. We will introduce Microarray Comparative Genomic Hybridization which provides a means to measure the DNA sequences copy number gains and losses and discuss the significance of DNA copy numbers in cancer research.

The Figure 1.3 demonstrates the Central Dogma of Molecular Biology. During a cell cycle a cell replicates its DNA by using its existing DNA as a template (Repli-

cation). During this process several possible mutations can occur. For example, a copy number gain or an amplification means that regions of DNA are duplicated or multiplied. On the other hand, a loss of DNA copy or a deletion is where regions are deleted or failed to replicate. Sometimes a segment of chromosome can be mixed up with other chromosomes, resulting in a translocation. These mutations are illustrated in Figure 1.4.

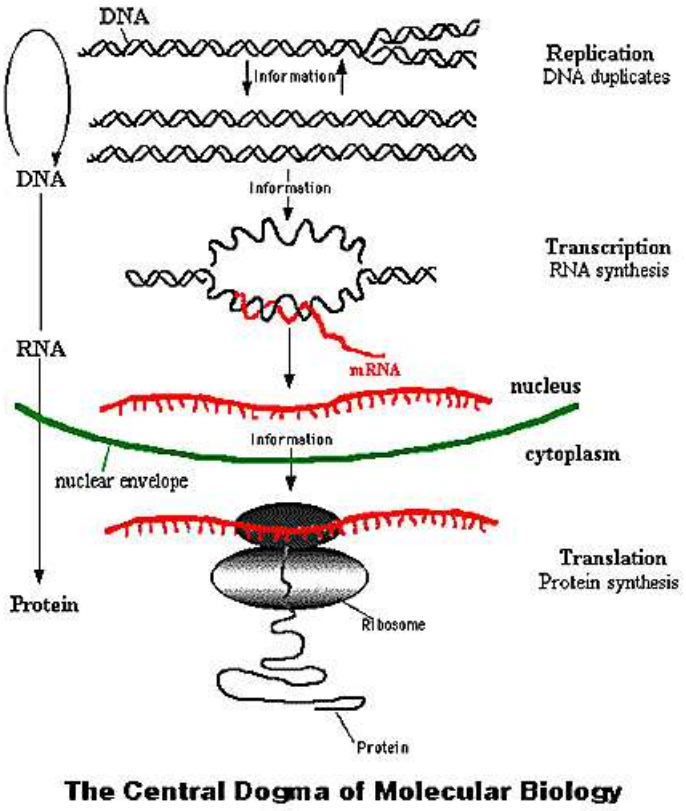

Figure 1.3: The Central Dogma of Molecular Biology. The graph illustrates the three stages of transcription of DNA to RNA to protein: Replication, Transcription and Translation. Source: www.accessexcellence.org

Normal cells have a checkpoint system that monitors and corrects this genomic instability. For example, mutations can be corrected or a cell can go through a programmed cell death so the mutation will not be passed on to other cells. However,

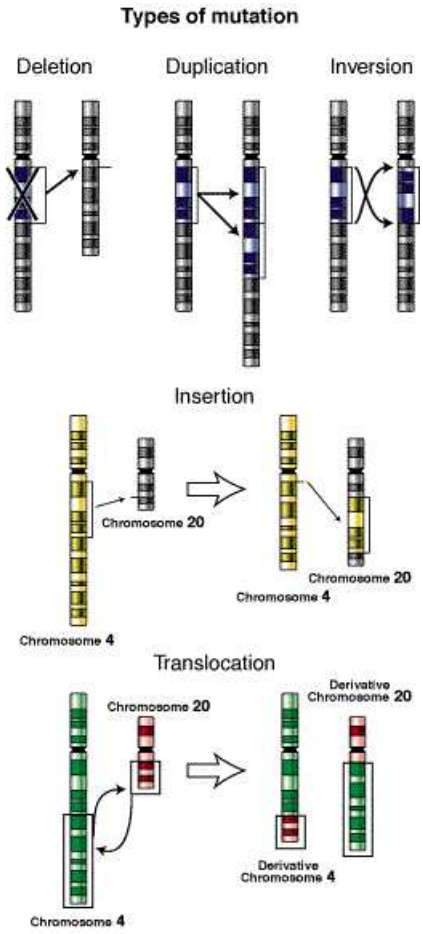

Figure 1.4: Some possible types of mutations that can occur during the cell cycle. In deletion, there is a loss of copy number for a chromosomal region. In duplication there is a gain in copy number. And in translocation there are both gains and losses for different chromosomal regions. Source: wikipedia.

if the system fails, the cell can survive the mutation and has the potential danger of growing uncontrollably. An amplified region may indicate an oncogene, a gene which stimulates cell growth and has become hyperactive. Conversely, a loss region may contain a tumor suppressor gene, which actively stops the tumor growth. These aberrations in copy numbers are frequently observed in tumor cells and it has been speculated that such genomic instability play a prominent role in cancer development. Therefore, it is important to understand these mutations.

Comparative Genomic Hybridization (CGH) is a recent technology advancement for detecting such alteration in DNA sequences. The samples from cancerous cells ("test" sample) and normal control cells ("reference" sample) are labeled with different fluorochromes, usually red and green so that they are easily distinguishable. The samples are then hybridized and superimposed on an array. The fluorescence intensity ratios of the test sample and the reference sample give a measurement of the relative copy number at each location on the chromosome. For example, if the tumor DNA was labeled green and the normal DNA was labeled red, when there is a deletion in copy number in the cancer DNA, the normal DNA will dominate and the hybridized sample will appear red. Conversely, when there is an amplification, the hybridized sample will appear green. When there is no gain or loss in the cancer genome, both DNA will be present equally and the resulting hybridization will be yellow. One advantage of the comparative hybridization procedure is the reduction in sources of variation, as whatever measurement error affecting both the test and the reference samples will likely be canceled (Pinkel and Albertson, 2005). The procedure of CGH is illustrated in Figure 1.5. Figure 1.6 shows the result from a CGH experiment. The x-axis is the chromosomes ordered from 1 to 22 and the sex chromosome. The y-axis is the $\log_2$ ratio of the test sample to the reference sample, centered by subtracting the median of all observations. We can clearly see the regions of copy

number aberrations in chromosomes 6,8,9,12 and 13. Note that in this experiment, the reference sample is an opposite sex, resulting in the copy number differences in the sex chromosomes.

With the invention of CGH technology the development of statistical methods of analyzing array CGH data has become an active research area. The main goal is to locate the regions of copy number changes accurately and to infer the number and significance of those changes (Willenbrock and Fridlyand, 2005). The step function is a good choice to model array CGH data. As we see from Figure 1.6, it is a common phenomenon that simultaneous gains or losses occur in multiple adjacent genes, sometimes spanning the whole chromosome (e.g., chromosome 13). Therefore, the step function naturally fits the block-like structure of the data (Hsu *et al.*, 2005). In later chapters we will introduce the overcomplete dictionary based on step functions, which is particularly suitable for detecting the breakpoints for DNA copy number aberrations.

1.4 Gene Expression Data

Gene expression is the process by which the genetic information is transferred from DNA to RNA and protein (transcription and translation, see Figure 1.3). There are various technologies to measure the abundance of the transcripts and their encoded proteins, or expression levels, including serial analysis of gene expression (SAGE) (Velulescu *et al.*, 1995), oligonucleotide arrays (Lockhart *et al.*, 1996), and cDNA microarrays (Schena *et al.*, 1995), all of which enable us to measure the expression levels of thousands of genes. In recent years the patterns of gene expression levels have been studied extensively for several complex diseases, such as cancer. For example, gene expression patterns have been used to classify cancers into biologically distinctive subcategories (Golub *et al.*, 1999; Perou *et al.*, 2000; Dhanasekaran *et al.*, 2001; Sørlie

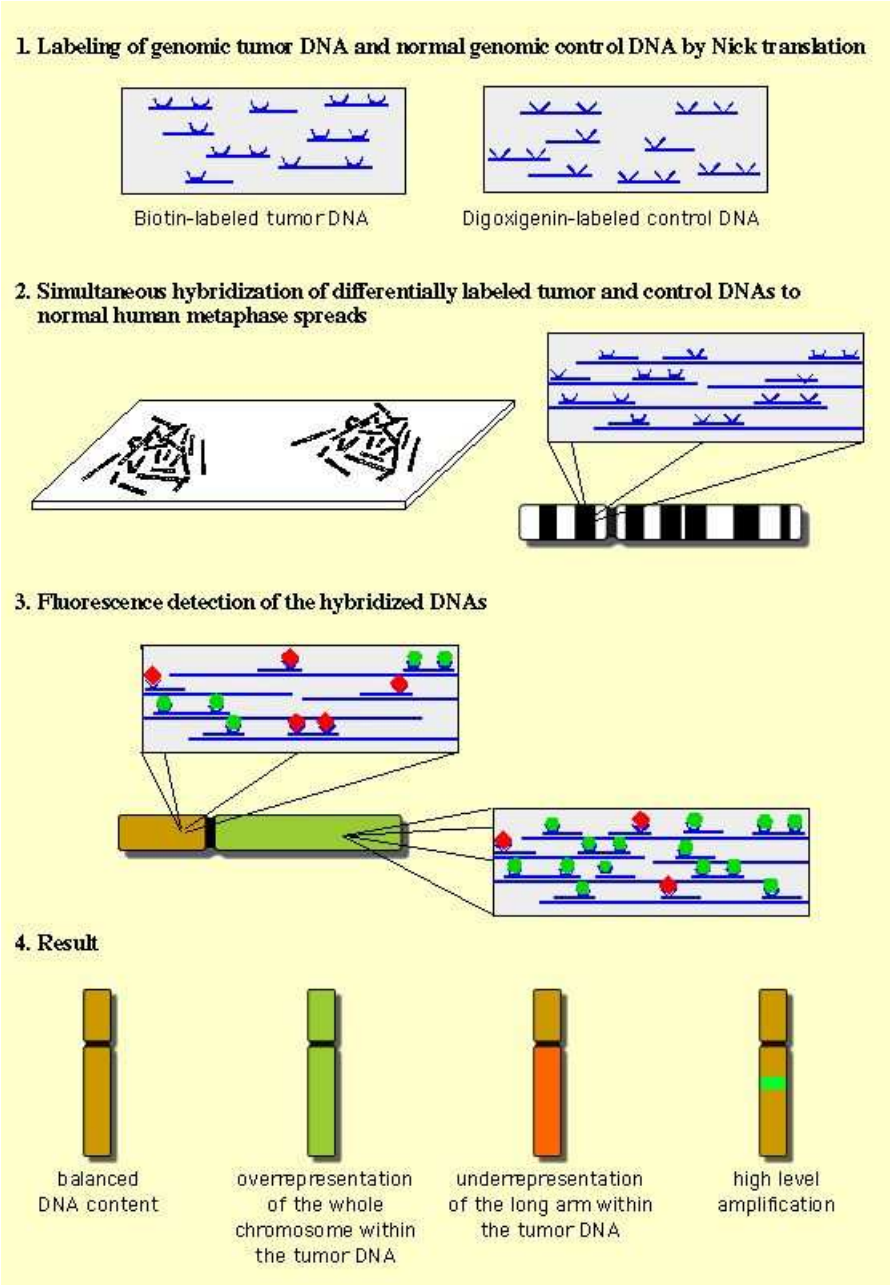

Figure 1.5: Comparative Genomic Hybridization. The genomic DNAs from the "test" sample and the "reference" sample are labeled with different fluorchromes, and then hybridized and superimposed. The resulting ratio of the fluorescence intensities of the two samples for a given location is a measurement of the ratio of copy numbers of the corresponding DNAs. Source: wikipedia.

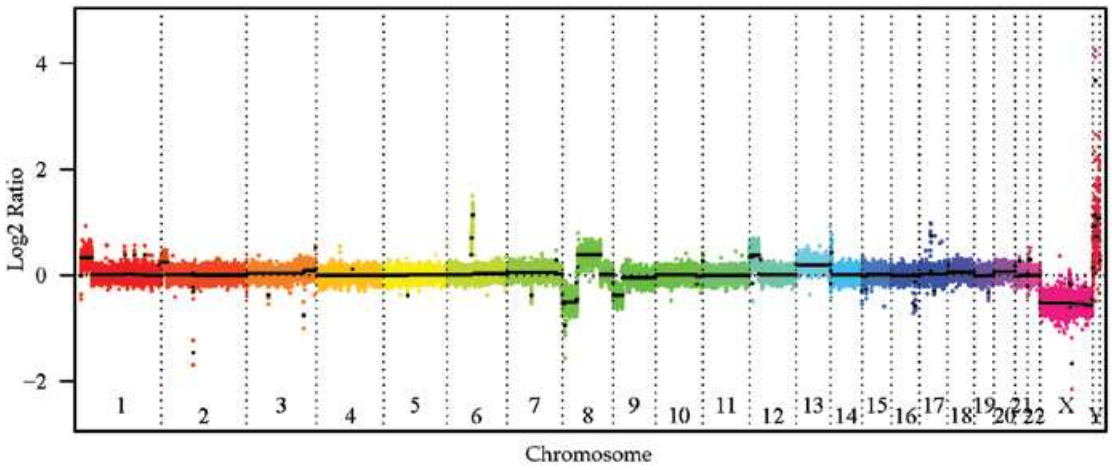

Figure 1.6: A plot of array CGH analysis. The x-axis is the chromosomes and the y-axis is the $\log_2$ ratio.

et al., 2001).

When the protein encoded by a gene is expressed in increased quantity, it is referred to as gene overexpression. Conversely, underexpression means a gene is expressed in decreased quantity. Overexpression and underexpression may result from the copy number change of the gene. An amplified region containing an oncogene may be overexpressed. On the other hand, if a loss region contains a tumor suppressor gene, that gene may be underexpressed. Identifying the genes that contribute to such genomic abnormalities can provide new insights in the mechanism behind cancer formation and development and help the researchers find therapies against those disease genes. There have been successful new treatments targeting amplified oncogenes such as ERBB2 and EGFR (Ross and Fletcher, 1999; Arteaga, 2001). However, the impact of DNA copy number changes on gene expression patterns remains unclear. A parallel study of expression levels and DNA copy number may help us understand the relationship between these two biological measurement and facilitate discovery of genes related to cancer progression.

1.5 Outline of the Dissertation

Chapter 2 of the thesis introduces the continuous wavelet dictionaries. We discuss the basic setting of the model, the methods for estimation and inference, and computational issues. We provide simulation results from standard test functions and the results from a real-life non-equally spaced data set.

In Chapter 3 and 4 we present applications of the model to problems in genomics. In Chapter 3 we present a model for array CGH (aCGH) data. The analysis of aCGH data requires functional estimation, so the overcomplete wavelet dictionaries naturally come into play. The CGH analysis also introduces other issues, such as summarizing information across multiple patients, patient classification and prediction, and comparisons between patients with different progression and survival status. We propose a hierarchical extension of the model from Chapter 2 and illustrate our methods with simulation studies and ovarian cancer data set from Duke Medical Center.

Extending from the model in Chapter 3, in Chapter 4 we consider a problem of joint analysis of DNA copy number data and gene expression data. Our model gathers information from two different biological measurements to detect the genes whose expression levels are associated with the copy number changes. We demonstrate the performance of our method using simulated data and the breast cancer cell line data from Hyman *et al.* (2002).

To conclude the dissertation, in Chapter 5 we present some ideas for future work, including computational improvement and model extensions.

Chapter 2

Continuous Wavelet Dictionaries

2.1 Introduction

Suppose we have observed data $\mathbf{Y} = \{Y_1, ..., Y_n\}$ at points $x_1, ..., x_n \in [0, 1]$ of some unknown function f measured with noise

$$Y_i = f(x_i) + \epsilon_i \quad \epsilon_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2).$$

A standard approach in nonparametric function estimation is to expand f with respect to an orthonormal basis, such as Fourier, Hermite or Legendre, and then to estimate the corresponding coefficients of the basis expansion. Wavelets, as a choice of orthonormal basis, are widely used in nonparametric function estimation and signal processing because they offer better localization and parsimony than other orthonormal bases (Mallat, 1989b; Donoho and Johnstone, 1998).

Given a wavelet function $\psi(x)$, let $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k), j, k \in \mathbb{Z}$, then the ψ_{jk}'s are an orthonormal basis for L_2 functions and any L_2 function can be represented as

$$f(x) = \sum_{j,k} \theta_{jk}\psi_{jk}(x).$$

For equally-spaced samples $x_1, ..., x_n$, we can obtain the coefficients θ_{jk} through filters by the cascade algorithm (Mallat, 1989a,b), which is fast as it takes advantage of the tree-like structure of the basis elements when j and k are integers.

However, orthonormal wavelets bases have constraints: the basis functions are subject to dyadic constraints on their location and scale. Some generalizations have been proposed, and they generally give much better results. For example,

the translation-invariant wavelet transform (Dutilleux, 1989; Nason and Silverman, 1995), which gives 2^J coefficients at each level ($2^{J-1} < n \leq 2^J$), is an example of an overcomplete representation, where the number of elements is greater than the dimensionality of the data. Other examples of overcomplete bases include Frames (Gröchenig, K., 2001; Wolfe *et al.*, 2004) and wavelet packets (Coifman and Meyer, 1990). However, with these bases the data are still required to be equally spaced and if not, some modification such as interpolations has to be done. The drawback of these modifications is that they will distort the structure of the data and complicate inference. Therefore it is appealing to generalize the wavelet basis functions to have more flexible locations and scales: not only can we apply it directly on the original data without any modification, but we will also have more flexibility to match the structure of the data.

In this chapter, we propose a Bayesian approach in function estimation using the continuous wavelet dictionary where the wavelet components have arbitrary locations and scales. It has been shown that in such an overcomplete setting, we may achieve greater sparsity and robustness against noise (Lewicki and Sejnowski, 1998; Donoho and Elad, 2003; Wolfe *et al.*, 2004; Donoho *et al.*, 2006). In practice, the expense for this setting is the computation in searching over the infinite model space. We will discuss a reversible jump Markov Chain Monte Carlo algorithm for inference on the parameters and strategies to achieve better convergence.

The remaining of this chapter is arranged as follows. In Section 2.2 we introduce the concept of the continuous wavelet dictionary. In Section 2.3 we discuss the prior specification. In Section 2.4 we describe the posterior inference by means of a reversible jump Markov Chain Monte Carlo sampling scheme. In Section 2.5 we present results from simulation studies and from a real example, which show that our new method leads to better performance in terms of sparsity and mean squared

error. Finally we present concluding remarks in Section 2.6.

2.2 The Model

Abramovich *et al.* (2000) introduced the stochastic expansions based on an overcomplete wavelet dictionary, where the function is modeled as a sum of wavelet components at arbitrary locations and scales, and the randomness of the locations, scales, and coefficients of the wavelet components are modeled by a marked Poison process. Our model is a modified version of their stochastic expansions model. Though their focus is not on Bayesian analysis, the stochastic expansion proposed in their paper suggests a prior choice for our model. We review the details of their setting in this section.

2.2.1 Stochastic Expansions

Suppose ϕ and ψ are the compact-supported scaling and wavelet functions that correspond to an r-regular multi-resolution analysis for some integer $r > 0$ (See Daubechies, 1992). Any function f may be decomposed as a sum of a coarse-scale function f_0 and a fine-scale function f_1. The function f_0 is given by

$$f_0(x) = \sum_{i=1}^{M} \eta_i \phi_{\lambda_i}(x), \tag{2.1}$$

where $\phi_{\lambda_i}(t) = a_i^{1/2}\phi(a_i(t - b_i))$ for some finite set of indices $\lambda_i = (a_i, b_i) \in (0, a_0) \times [0, 1]$, $i = 1, ...M$. The function f_1 is given by a stochastic expansion

$$f_1(x) = \sum_{\lambda \in S} \beta_\lambda \psi_\lambda(x), \tag{2.2}$$

where both λ and β_λ are random and ψ_λ is defined analogously as ϕ_λ but with $a \geq a_0$. Here the index $\lambda = (a, b)$ follows a Poisson process S on $\Lambda = [a_0, \infty) \times [0, 1]$

with intensity $\mu(\lambda)$ and given any realization of S, the wavelet coefficients β_λ are independent normal variables

$$\beta_\lambda \mid S \sim \mathsf{N}(0, \tau^2(\lambda)). \tag{2.3}$$

Both the variance $\tau^2(\lambda)$ and the intensity $\mu(\lambda)$ are assumed to depend on the scale a only

$$\tau_a^2 \propto a^{-\delta} \quad \text{and} \quad \mu_a \propto a^{-\zeta}, \tag{2.4}$$

where $\delta, \zeta \geq 0$, with $\delta + \zeta > 0$.

2.2.2 Continuous Wavelet Dictionary

Under the continuous wavelet dictionary (CWD) setting, we model the response variable Y as $N(f(x), \sigma^2)$ with

$$f(x) = f_0(x) + \sum_{k=1}^{K} \beta_{\lambda_k} \psi_{\lambda_k}(x), \tag{2.5}$$

where f_0 is a fixed scaling function as defined before. For example, we can take $f_0 = \bar{\mathbf{Y}}$, the sample mean, or the scaling function from the regular discrete wavelet transform (DWT). The unknown parameters here are the error variance σ^2, the number of wavelet elements K, and the corresponding location-scale index and coefficient for each wavelet component (β_λ, λ). Notice that in regular DWT where a and b have dyadic constraints and data are on an equally spaced grid, we can obtain the coefficients through filters without evaluating the wavelet function ψ directly. In CWD the basis elements do not have a tree-like structure needed for the cascade algorithm and in addition our data may not be equally spaced, therefore we will have to evaluate the wavelet function directly. Here we use the Daubechies-Lagarias local pyramid algorithm (Vidakovic, 1999, Sec. 3.5.4), which enables us to evaluate ϕ and ψ at an arbitrary point with preassigned precision.

In practice, wavelets are often used to represent functions from certain Besov spaces. Naturally one would ask under what kind of conditions, the random function f will still be in the same Besov space almost surely (a.s.). Note that the number of elements in (2.2) follows a Poisson distribution with intensity $\mu(\Lambda)$. When $\mu(\Lambda) < \infty$, we will have a finite number of elements (a.s.) and therefore f will belong to the same Besov space (a.s.) as the mother wavelet function ψ does for any reasonable choice of the probability distribution for β_λ. However, when $\mu(\Lambda) = \infty$ which is the case considered in Abramovich *et al.* (2000), extra conditions are needed for the random function f to be well-defined.

2.3 Prior Specification

2.3.1 Prior for $\lambda = (a, b)$

Following Abramovich *et al.* (2000), the prior for the scale parameter a takes this form

$$p(a) \propto a^{-\zeta}, \quad a_0 \leq a \leq a_1 \text{ and } \zeta > 0. \tag{2.6}$$

The hyperparameter ζ controls the relative number of the fine-scale wavelet component in the function. If ζ is large, we will have relatively few fine-scale (spiky) components in the function, and vice versa. The lower bound a_0 corresponds to the coarsest-scale component allowable in the function and the upper bound a_1 corresponds to the finest-scale. Theoretically, a can go up to infinity to span the whole space as in Abramovich *et al.* (2000). However in practice, allowing a to go up to infinity is not desirable, as when a increases we obtain spiky wavelet functions with very small support which have little or no effect on the likelihood. Therefore we set an upper bound for a so that the wavelet functions will have a large enough support. These bounds will depend on the data and the support of the wavelet ψ. For exam-

ple, suppose we have 1024 equally-spaced data points and use a mother wavelet with support of length 1. If we set $a > 1024$, the support of the wavelet function could fall entirely between two data points, and it will have no effect on the likelihood. As a result, the corresponding coefficient can not be estimated effectively and we could overfit the data.

The prior for the location parameter b takes the form

$$p(b) = \gamma \sum_{i=1}^{n} \frac{1}{n} \delta_{x_i}(b) + (1-\gamma), \quad 0 < \gamma < 1, \tag{2.7}$$

which is a mixture of point masses on all the data points and a uniform distribution on $[0, 1]$. This prior is a compromise of flexibility, which allows b to be at arbitrary positions, and efficiency, which focuses on the data points where the information is abundant. This mixture prior also enables us to search the dictionary elements more efficiently by using the information from residuals, which we will discuss in detail in the next section. Notice that when $\gamma = 1$ and $p(b)$ has support on data points only, we return to the non-decimated DWT setting, and when $\gamma = 0$, we have the continuous distribution from Abramovich *et al.* (2000).

2.3.2 Prior for K

For K, the number of wavelet components with non-zero coefficients, the stochastic expansion model in Abramovich *et al.* (2000) implies a Poisson prior distribution with mean

$$\mathbb{E}(K) = c_1 \int_{a_0}^{a_1} \int_{0}^{1} a^{-\zeta} db da,$$

where c_1 is some constant. Here we impose a negative binomial prior on K

$$p(K|r, q) = \binom{r+K-1}{K} q^r (1-q)^K.$$

The negative binomial distribution is a Gamma mixture of Poisson but is more flexible than the Poisson distribution, which has only one parameter that controls both the mean and the variance. We choose the hyperparameters r and q by specifying the probability of the null model $p(K = 0)$ and the range of K (for example, the 95% quantile of $p(K)$). We can easily solve these two equations to obtain the values of r and q.

Both the Poisson and negative binomial priors can be regarded as a limiting case for the mixture prior from Clyde *et al.* (1998) when the model space moves from being finite to being infinite. Recall that in the orthonormal wavelet model with N wavelet basis functions, the mixture prior implies a Binomial distribution (N, π) on the number of non-zero coefficients. When N goes to infinity as in the continuous wavelet dictionary model, if we let $\pi \propto 1/N$ and go to zero, we obtain the Poisson model with mean $\mu = \pi N$ for the number of non-zero coefficients. If we have a Gamma distribution on μ, we obtain the negative binomial model.

2.3.3 Prior for β_λ and σ^2

Given the location and scale of a wavelet function, we can set the prior distribution for the corresponding wavelet coefficient β_λ to be independent normal

$$p(\beta_\lambda \mid a) = \mathsf{N}(0, ca^{-\delta}), \tag{2.8}$$

where c is a tuning parameter specified by users and independent of a. A natural choice for c is to set $c = n$, the sample size, as in the unit-information prior (Kass and Wasserman, 1995). Note that β_λ being normally distributed is one of the condition for f_1 to be well-defined in the general setting in (2.2), where we can potentially have infinitely many elements in f_1. However, if we put an upper bound a_1 on a, with $\mu(\Lambda) < \infty$ we will have a finite number of elements almost surely, therefore the

normality of β is not necessary. Instead we can use a heavier-tailed prior for β, e.g. Laplace with a scale parameter that depends on a the same way as in (2.8)

$$p(\beta \mid a) = \frac{1}{2\sigma} \exp \frac{-|\beta|}{\sigma}, \quad \sigma^2 = ca^{-\delta}/2. \tag{2.9}$$

The heavy-tailed priors have been shown to have theoretical advantage over normal distribution, and may lead to greater sparsity and further reduction of the mean squared error (Johnstone and Silverman, 2004).

It may be appealing to use priors on β which take into account the dependence structure of the design matrix, such as g-priors (Zellner, 1986; Fernández *et al.*, 2001; Liang *et al.*, 2005). However it is known that Zellner's g-prior and its variations can not be directly applied on the regression coefficients in an over-complete setting (Liang *et al.*, 2005).

We set an non-informative prior for σ^2, $p(\sigma^2) \propto 1/\sigma^2$.

2.4 Posterior Inference

A big challenge here is how to search efficiently over a continuous model space. Since the dimensionality of the parameters may vary, we propose a reversible jump Markov Chain Monte Carlo (RJ-MCMC) techniques (Green, 1995) (see Appendix A). Our RJ-MCMC algorithm includes three types of movements: a birth step where we add a wavelet element, a death step where we delete a wavelet, and an update step where we move a wavelet element but leave the dimension K unchanged. A Metropolis-Hasting step is used to sample the parameters for wavelet elements $\{\beta, a, b\}$, because the corresponding full posterior does not have a close-form.

2.4.1 RJ-MCMC

For RJ-MCMC algorithms a good proposal distribution is necessary to speed up convergence. For example, proposing a "birth" of a new dictionary element from the prior on $(\beta, a, b|K+1)$ may simplify the calculation, but it often results in slow convergence since it does not necessarily lead to proposal values where the likelihood is high. Similarly, picking a component at random to remove may lead to frequent attempts to remove important wavelets. Here we discuss some of the proposal distributions we use, particularly the "birth" of the location parameter b, which is critical when we start the MCMC chain with few elements. The detail of the RJ-MCMC algorithm is in Appendix A, Sec 1.

Because of the local nature of wavelets, information in the residuals may aide in placing new wavelets. We choose a mixture proposal for the location parameter b of the new wavelet functions which is a mixture of point masses on the data points with weights that depend on the current residuals and uniform on $[0, 1]$. In particular, the proposal for the birth step is

$$q(b_{K+1}) = \gamma \sum_{i=1}^{n} \delta_{x_i}(b_{K+1}) v_i + (1-\gamma), \quad 0 < \gamma < 1, \tag{2.10}$$

where

$$v_i = \frac{|Y_i - \hat{f}(x_i)|}{\sum_{j=1}^{n} |Y_j - \hat{f}(x_j)|}$$

is proportional to the magnitude of the residual. Since the prior for b is also a mixture of point masses and uniform, it has density on the same measure as the proposal, which is a necessary condition for the transition kernel to be reversible. The proof of detailed balance condition and reversibility is in Appendix A, Sec 2.

The proposal for the death step is inversely proportional to the wavelet coefficient:

$$q(b_k \mid K) = \frac{1/|\beta_k|}{\sum_{i=1}^{K}(1/|\beta_i|)}, \tag{2.11}$$

so that small magnitude coefficients are more likely to be removed.

Finally, the proposal for the update step is

$$q(\tilde{b}_k \mid b_k) = \delta_{b_k}(\tilde{b}_k)u_k + \mathsf{N}(\tilde{b}_k; b_k, \sigma_b^2)(1 - u_k), \tag{2.12}$$

where

$$u_k = \begin{cases} 1 & \text{if } b_k \text{ is a data point} \\ 0 & \text{otherwise,} \end{cases}$$

which is a point mass at b_k if b_k is a data point and a random walk otherwise. Notice that this will cancel with the reverse proposal in either case, so the form (A.6) is still valid.

These proposal distributions can improve convergence in practice since a successful birth is more likely where the residual is large, and it makes more sense to kill a wavelet of which the coefficient is small since it will not change the likelihood dramatically. After T MCMC iterations post burn-in, each collection of the parameters $\{\boldsymbol{\beta}, \mathbf{a}, \mathbf{b}, K\}$ represents a sample from the posterior distribution , where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)^T$ and $\mathbf{a}$ and $\mathbf{b}$ are defined similarly. At each iteration we plug $\{\boldsymbol{\beta}, \mathbf{a}, \mathbf{b}, K\}$ back into equation (2.5), obtaining posterior samples $f^{(t)}(x_i), t = 1, .., T$ from $p(f \mid \mathbf{Y})$ which provide a full spectrum of description of the posterior distribution of f given the data $\mathbf{Y}$.

2.4.2 Estimation of f

In Bayesian model averaging (BMA), the posterior distribution of $f(x)$ given data $\mathbf{Y}$ is

$$\mathsf{P}(f(x) \mid \mathbf{Y}) = \sum_{j=1}^{J} \mathsf{P}(f(x) \mid M_j, \mathbf{Y}) \mathsf{P}(M_j \mid \mathbf{Y}), \tag{2.13}$$

and the mean estimate is given by

$$\mathbb{E}(f(x) \mid \mathbf{Y}) = \sum_{j=1}^{J} \mathbb{E}(f(x) \mid \mathbf{Y}, M_j) \mathsf{P}(M_j \mid \mathbf{Y}), \tag{2.14}$$

where $M_1, ...M_J$ are the models considered, and each model corresponds to a set of parameter $\{\boldsymbol{\beta}, \mathbf{a}, \mathbf{b}, K\}$. Since the number of models is infinite here, the exact posterior distribution cannot be obtained. However, we can use Markov chain Monte Carlo model composition (MC^3) (Madigan and York, 1995) to approximate (2.14). In particular, given T MCMC samples, we can take posterior mean as a point estimate for $f(x)$, that is,

$$\hat{f}_{BMA}(x) = \frac{1}{T} \sum_{t=1}^{T} \hat{f^{(t)}}(x) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}(\mathbf{Y} \mid x, \mathbf{a}^{(t)}, \mathbf{b}^{(t)}, \boldsymbol{\beta}^{(t)}, K^{(t)}), \tag{2.15}$$

where $\hat{f}^{(t)}$ is the model from the tth MCMC iteration. Standard Markov chain Monte Carlo results show that $\hat{f}_{BMA}$ converges to $\mathbb{E}(f \mid \mathbf{Y})$ (a.s.) as T goes to infinity (Smith and Roberts, 1993).

Alternatively we can provide point-wise posterior medians in a similar way.

2.4.3 Bayesian Credible Bands for f

We can construct point-wise credible bands for f

$$C_p(\alpha) = \{C : \mathsf{P}(f(x_i) \in C \mid \mathbf{Y}) = \alpha \quad \forall i\},$$

which covers $f(x_i)$ with $\alpha\%$ posterior probability at each data point x_i.

We can also construct simultaneous credible band to be the convex hull of all the posterior samples of f such that their square distances (at the data points) to the BMA estimate are below the $\alpha\%$ quantile. This idea is similar to the Baraud Confidence Set from Baraud (2004), where the confidence set of θ is taken as the ball $\mathcal{B}_n$ centered at some estimate $\hat{\theta}$.

For $\hat{f}^{(t)}$ from the MCMC iteration t, we take the Euclidean distance to the BMA estimate

$$D_t = \sum_{i=1}^{n} \{\hat{f}_{BMA}(x_i) - \hat{f}^{(t)}(x_i)\}^2,$$

then take the threshold D^*_α to be the $\alpha\%$ percentile of all D_t's and T^* to be the collection of indices t for which the distance D_t is below D^*_α

$$T^* = \{t : 1 \leq t \leq T, \quad D_t \leq D^*_\alpha\}.$$

Then the $\alpha\%$ simultaneous credible band can be defined as

$$C_s(\alpha) = [\min_{t \in T^*} f^{(t)}(x_i), \max_{t \in T^*} f^{(t)}(x_i)] \quad \forall i.$$

2.4.4 Model Selection

When the goal is model selection and a single model must be reported, we choose to report the model which is closest to the BMA estimate in the following sense:

$$f^* = \arg\min_{t \in \{1,\ldots,T\}} \sum_{i=1}^{n} \{\hat{f}_{BMA}(x_i) - \hat{f}^{(t)}(x_i)\}^2. \tag{2.16}$$

If β has a normal prior, we can reduce the Monte Carlo variation by replacing $\beta^{(t)}$ by its posterior mean when we calculate $\hat{f}^{(t)}$

$$\hat{\beta}^{(t)} = \mathbb{E}(\beta \mid \mathbf{Y}, \mathbf{a}^{(t)}, \mathbf{b}^{(t)}, K^{(t)}).$$

2.5 Examples

2.5.1 Simulation Studies

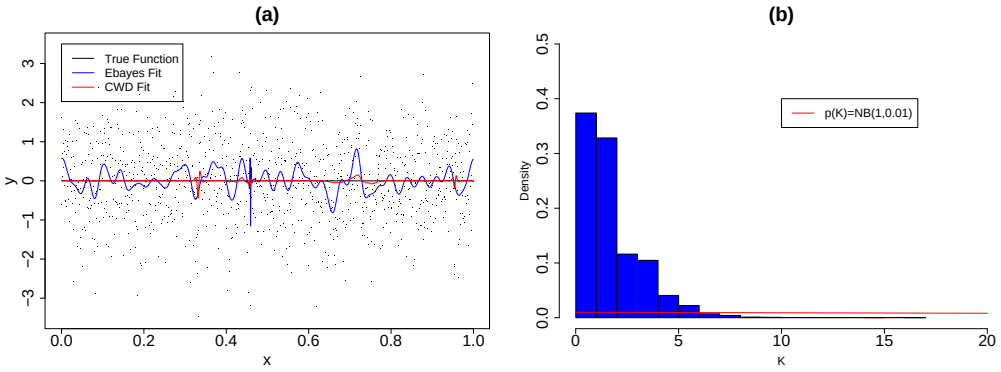

Figure 2.1: (a)The EBayes and CWD fit of the null function and (b) The posterior histogram for K overlaid with the prior NB(1,0.01)

As the stochastic representation allows extremely flexible representations, an initial concern is that the method may lead to over-fitting of the data. To test this, we applied the CWD method the null function $f(x) = 0$ observed with noise. The prior for the number of coefficients K is negative binomial with $r = 1$ and $q = 0.01$, which corresponds to 0.01 probability of the null model and 95% percentile at $K = 400$. The prior distribution function is relatively flat and covers a wide range of possible models (See Figure 2.1b). The results are shown in Figure 2.1. We can see from the posterior histogram that the null model (K=0) is the one with the highest posterior probability, while the empirical Bayes (EBayes) (Johnstone and Silverman, 2005) method, which always keeps the scaling function includes 1024 coefficients of the father wavelet.

We carried out a simulation study on four standard test functions from Donoho and Johnstone (1994): `bumps`, `blocks`, `doppler`, and `heavysine`. For each test function, 100 replications were generated with a fixed signal-to-noise ratio of 7. In each replicate, the function was simulated at 1024 equally spaced points x_i in $[0, 1]$.

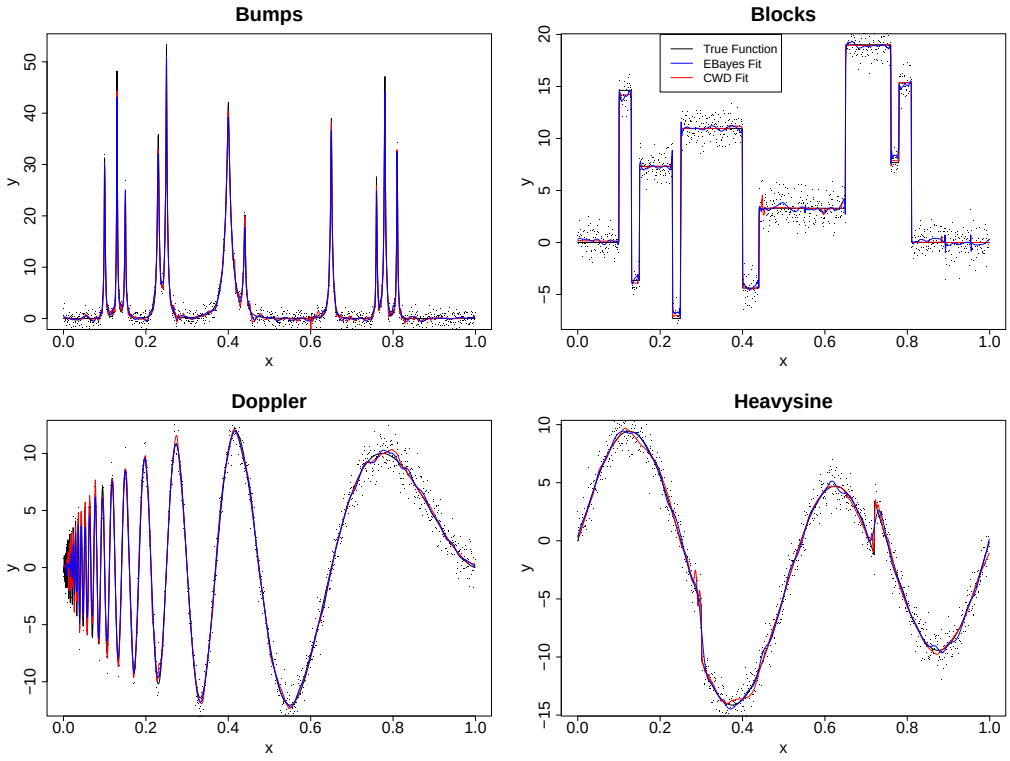

Figure 2.2: Four standard test functions from Donoho and Johnstone (1994). The data points and fitted function from Ebayes and CWD are from one replication for each function.

The hyperparameters for the variance and intensity are set to $\delta = 2$ and $\zeta = 1.5$. We tried several values and the result turned out not to be very sensitive to these hyperparameters. The default choice of wavelet in R (`la8`) was used in all functions except for `blocks`, where we used the father wavelet of Haar (step function). Unlike many other wavelet methods, we do not assume a boundary correction here, since some of the functions (e.g. `doppler`) are clearly not periodic. We set f_0 to be the constant function at the data mean $\bar{\mathbf{Y}}$, $c = n = 1024$, and the upper bound a_1 for a such that the wavelet elements cover at least 14-15 data points. Usual convergence diagnostic methods, such as Gelman and Rubin (1992) do not apply here since we are moving within an infinite model space and the parameters are not common to all models. Instead we look at K and the mean squared error, which have a coherent

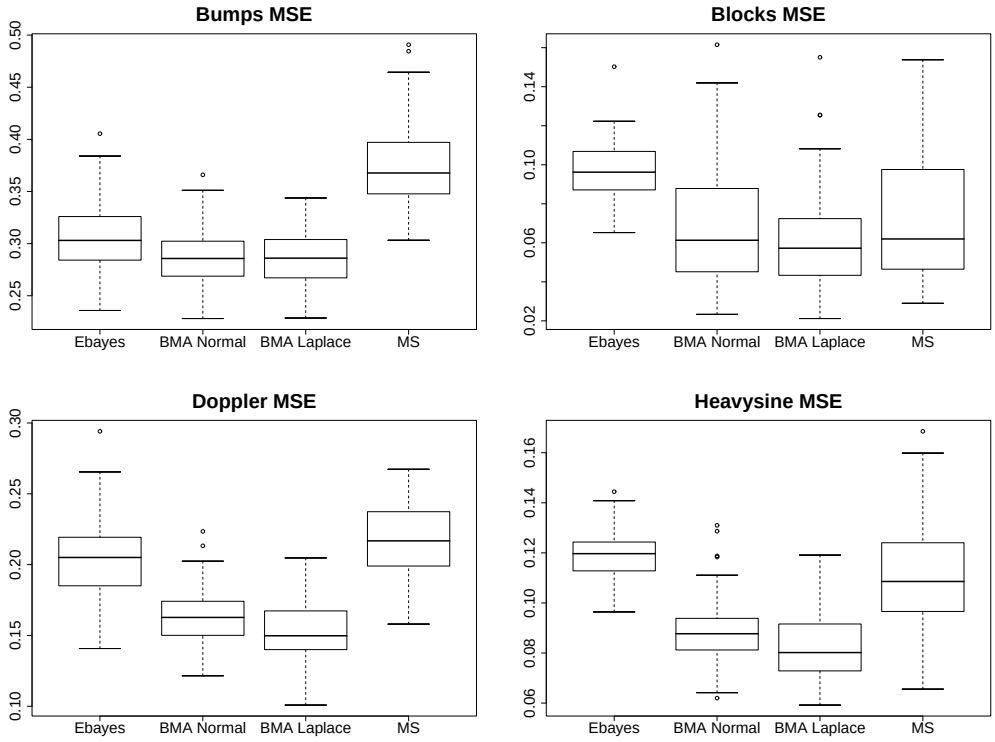

Figure 2.3: Box plot for mean squared error for four standard test functions using EBayes method (Johnstone and Silverman, 2005) and continuous wavelet dictionary (CWD) method with Bayesian model averaging (BMA) with normal and Laplace priors and model selection (MS) with normal prior

interpretation throughout the model space (Brooks and Giudici, 2000). The trace plots and the Gelman-Rubin shrink factor for K and mean squared error suggest that convergence usually occurs within 1 million MCMC iterations. The following results are based on 5 million iterations, which takes about 8-9 hours to run on 64-bit cluster computers.

Figure 2.2 shows these test functions and the results from one replication for each function. We compared the CWD fits with the ones from EBayes based on mean-squared error

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{f}(x_i) - f(x_i)\}^2. \tag{2.17}$$

BMA estimates from both the normal and Laplace prior for β are presented here,

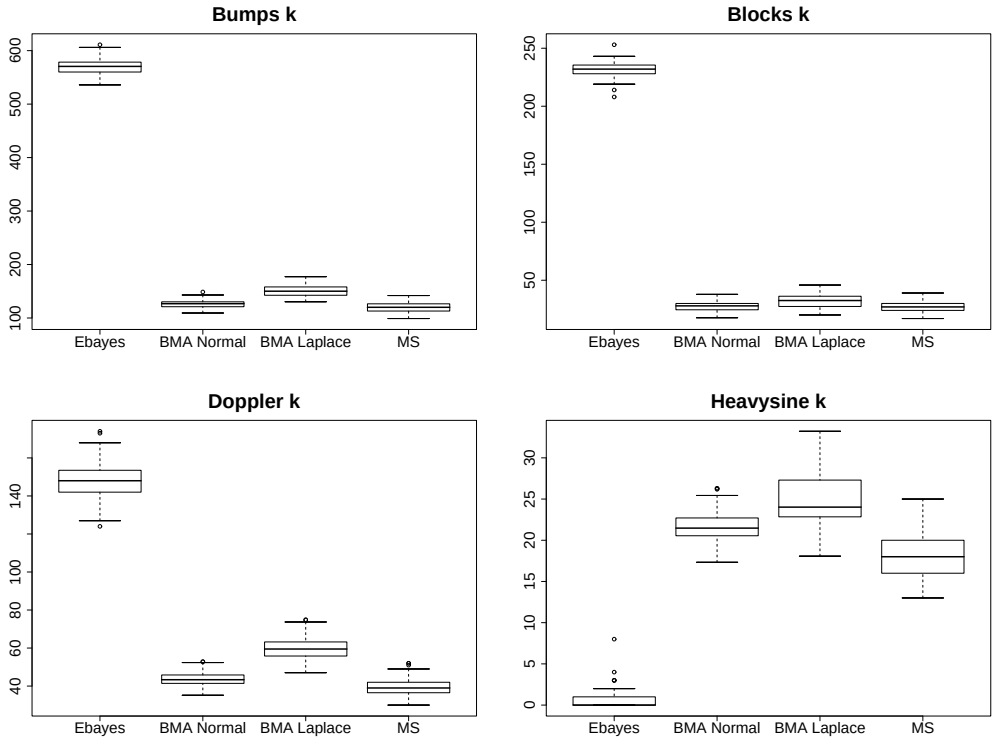

Figure 2.4: Box plot for the number non-zero coefficients for the four test functions using EBayes method (Johnstone and Silverman, 2005) and continuous wavelet dictionary (CWD) method with Bayesian model averaging (BMA) with normal and Laplace priors and model selection (MS) with normal prior

along with the model selection (MS) estimate from in (2.16) with normal prior, compared with Ebayes with Laplace prior. Figure 2.3 shows that the model average estimate in (2.15) has smaller MSE than EBayes for all four functions. Taking the heavy-tailed Laplace prior instead of normal does reduce the MSE slightly further except for `bumps`. If we take the MS estimate, then EBayes is doing better for `bumps` and `doppler`. However, if we compare the number of non-zero coefficients in $\hat{f}$ (See Figure 2.4) then CWD method clearly gives a much sparser representation than EBayes. Notice that the EBayes results for K do not include the 1024 coefficients from the scaling function, which are not shrunk.

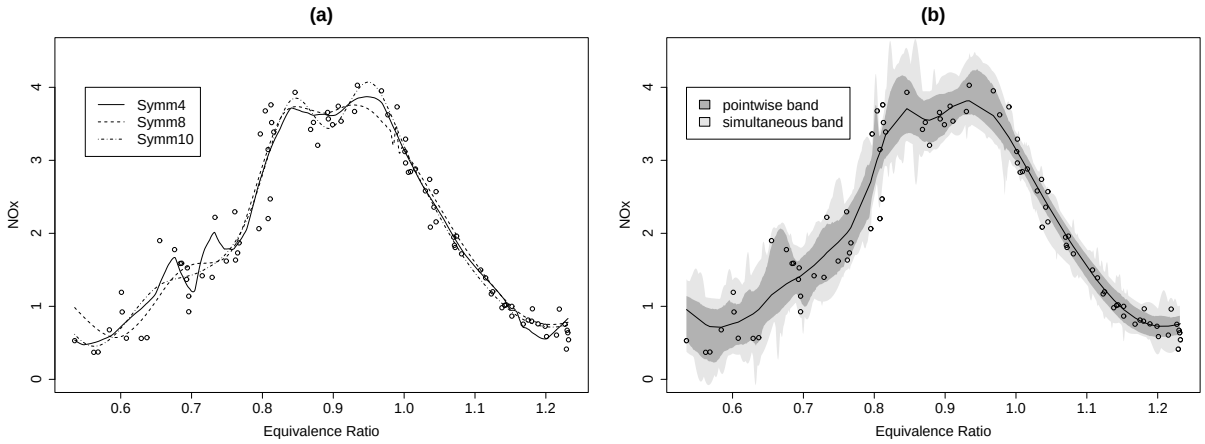

Figure 2.5: (a) The model selection CWD fits with normal prior for ethanol data from Brinkman (1981) and (b) The 95% simultaneous and pointwise credible bands with `symm8`

2.5.2 Application

One of the advantage of the CWD based method is that it can be applied directly to non-equally spaced data sets. To illustrate this point, we applied our method to a well-studied data set, `ethanol` data, from Brinkman (1981). This data set consists of $n = 88$ measurements from an experiment where ethanol was burned in a single cylinder engine. The concentration of the total amount of nitric oxide and nitrogen dioxide in the engine exhaust, normalized by the work done by the engine is related to the "equivalence ratio", a measure of the richness of the air ethanol mixture.

We applied our CWD method with 4, 8, and 10 vanishing moments of the least asymmetric Daubechies' wavelets (`symm4, symm8` and `symm10`). We use the same hyperparameters as in Section 2.5.1, except that the upper bound a_1 is lower since there are fewer data points. We report the model selection estimated curves defined in (2.16), in figure (2.5a) and the 95% pointwise and simultaneous credible bands with `symm8` in figure (2.5b).

This same data set was studied by Nason (2002) using the linear interpolation method. To compare with their result, we did a leave-one-out cross validation study

and calculated the cross validation score

$$\text{CV-score} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{f}^{-i}(x_i) - Y_i\}. \tag{2.18}$$

where $\hat{f}^{-i}$ is the estimated f from all the data except the ith point. With no attempts to optimize the hyperparameters, the CV score from CWD with `symm8` ranked 2nd out of the 60 combination reported in Nason (2002), and the estimated function looks very similar to their best combination. The credible bands here cover all but four data points. We can see that over the left region where there are fewer data points, there seems to be a lot more uncertainty, as the credible bands are wider and the estimates disagree (`symm4` gives two extra bumps). On the right hand side where all estimates seem to agree on the same downward slope, the credible bands are much narrower as we have more information here. We can see that CWD has managed to capture the main feature of the data without overfitting.

2.6 Conclusion

In this chapter we have introduced a Bayesian method for function estimation based on the stochastic expansion in a continuous wavelet dictionary. Despite the richness of the potential representations, RJ-MCMC algorithms are able to identify sparse representations. The simulation study shows that the new method leads to greater sparsity and improved mean squared error performance than the current wavelet-based methods. Because the models do not require the data to be equally spaced, this will permit wavelet methods to be used in a greater variety of applications. We have also introduced a new approach for constructing simultaneous credible bands in overcomplete settings.

Chapter 3

On Detecting Gene Copy Number Changes and Patient Classification

3.1 Introduction

Changes in the number of genomic DNA copy numbers have been associated with cancer (Lockwood *et al.*, 2006). As a result, detection of DNA copy changes is important for identifying relevant genes for cancer development and patient classification. Array comparative genomic hybridization (CGH) is a current technology used to detect DNA copy number alterations. The test and reference DNA samples are labeled with different colors then hybridized and superimposed on an array. Then the log ratios of the intensity of the test and reference samples are obtained. Those ratios are measurements of DNA copy number changes.

There have been many studies of analysis of array CGH data to identify the genes and contiguous chromosomal regions with copy number changes. There are two main estimation problems with CGH data. First is to locate the copy number transition or breakpoints, which is commonly referred to as segmentation (Willenbrock and Fridlyand, 2005). The other problem is to infer the number and statistical significance of the alterations. Numerous methods and algorithms have been proposed for these problems. Olshen *et al.* (2005) proposed a non-parametric change point method, Circular Binary Segmentation (DNAcopy). Hupe *et al.* (2004) used a Gaussian model-based approach (GLAD). CLAC by Wang *et al.* (2005) involved building a hierarchical clustering trees along each chromosome. Picard *et al.* (2005) used a penalized-likelihood criterion to estimate breakpoints. Myers *et al.* (2004) proposed

an EM algorithm-based method and Hsu *et al.* (2005) used wavelets for smoothing and denoising. Some Bayesian methods also have been proposed, including CGH-MIX (Broët and Richardson, 2006), a Hidden-Markov-Model (HMM) based method, and RJaCGH (Rueda and Daz-Uriarte, 2006), which used reversible jump MCMC to explore an unknown number of hidden states. See Lai *et al.* (2005) and Willenbrock and Fridlyand (2005) for comparative studies and summarization for some of these methods.

Lai *et al.* (2005) concluded that when the segments of aberration are short, it is hard to detect their existence, especially when the signal-to-noise ratio (SNR) is low. Therefore, when we have multiple samples, it is important to borrow strength from the information across the samples to assess the significance of the segments. Moreover, we are usually interested in the area where the aberrations occur more frequently among the patients, as these are more likely candidates for future research. The only method that analyzes multiple samples at the same time and gives inference across samples is cghMCR (Aguirre *et al.*, 2004), in which they define minimal common regions (MCR) as contiguous regions of gains/losses with a certain recurrence rate across the samples.

Here we propose a Bayesian hierarchical model on the samples from the same cancer group, which allows us to find the common structure of patients with tumor growth and/or survival status. We combine the segmentation problem and the inference problem and obtain the segments and quantitative measure of gain/loss automatically in one step. There are several other improvements over the currently available methods. Our model incorporates the physical distance of the probes, which is useful information that other methods often ignore by assuming equally-spaced probes. In theory, if the two probes indicating the same direction of change are very far apart, it will be less likely that they refer to the same alteration than if they

were closer (Lai *et al.*, 2005). Also, we do not assume a discrete state-space for the copy numbers as in many HMM based methods, because when we compare different samples, it is reasonable to assume that the same log ratio levels do not always correspond to the same underlying copy number, as each individual sample contains a different proportion of tumor cells. The model is highly flexible and operates on individual chromosomes as well as a genome-wide level, and forces a standard measurement error for all chromosomes and patients. The Bayesian method provides the posterior probability of gain/loss and also a measure of uncertainty in the level of gain/loss. Finally, we can summarize the overall results from different cancer groups and use it for patient classification, and we can also identify the genes and regions most relevant to the different cancer development and survival status.

The remaining of this chapter is arranged as follows. First we give an overview of the statistical model in Section 3.2. In Section 3.3 we go through the methods for estimation, inferences and prediction. In Section 3.4 we present results from simulated and real data. Finally we present concluding remarks discussion of related issues in Section 3.5.

3.2 Model

Let n be the number of probes, and $x_1, ...x_n$ be the physical location of the probes. The locations are rescaled to $[0, 1]$. Suppose we have L groups of patients indexed by $l = \{1, 2, ...L\}$, and in each group we have J_l patients indexed by $j = 1, ...J_l$. Let Y_{ijl} be the log ratio of fluorescence intensities between tumor and reference samples of probe i of patient j in group l. Since the main goal is to find the copy number changes that span over consecutive probes, we model Y by a step function

$$\mathbb{E}(Y_{ijl}) = f_{jl}(x_i) = \sum_{k=1}^{K_l} \beta_{jkl} h_{kl}(x_i), \quad h_{kl}(x) = I(a_{kl} \leq x < b_{kl}) \tag{3.1}$$

$$Y_{ijl} = f_{jl}(x_i) + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2).$$

We assume that the patients in the same groups share the same segmentations. Here h_{kl} is the element which represents the $k\underline{\text{th}}$ segment where a copy number alteration occurs in the $l\underline{\text{th}}$ group, spanning from a_{kl} to b_{kl}. This setting is related to the continuous wavelet dictionary in Chapter 2 with Haar wavelets, but the dictionary elements are parametrized differently with the end points of the segments, instead of the scale and location. Here $\mathbf{a}_l = \{a_{1l}, ...a_{K_l l}\}, \mathbf{b}_l = \{b_{1l}, ...b_{K_l l}\}$ are the end points for the segments for group l. We set $0 \leq a_{1l}$, $b_{lK_l} \leq 1$, $a_{lk} < b_{lk}$ for all k and $b_{lk} \leq a_{l(k+1)}$ for $k = 1, ..K_l - 1$, so no segment will overlap. The part not covered by any segment will have a zero mean, which means the copy number there is unchanged. These segments span the space of the step functions, as in the continuous wavelet case with Haar wavelets. Since any L_2 function can be approximated by a step function, these collections of segments can represent a rich enough class of functions for our purpose.

We assume a uniform prior for $\mathbf{a}_l$ and $\mathbf{b}_l$ on the set of data points $x_1, ..., x_n$, and a negative binomial prior on the number of segments K_l. The uniform prior combined with a restriction on no overlapping segments induces a prior on segment length $b_{kl} - a_{kl}$, which favor short segments over long ones. The prior suggests that *a priori* it is more likely to find segments of gain/loss where the data points are more concentrated and more information is available. Depending on the size of the data and other information, we may have a prior expectation on the distribution of the number of segments, which can be used to set the hyperparameter of the negative binomial prior. Without prior information, we have found that a flat prior such as

$NB(1, 0.01)$ often works as well.

The coefficient β_{jkl} of the segment k for patient j in group l indicates the degree to which the copy number changes at that segment. The hierarchical nature of the model comes from the assumption that *a priori* the individual coefficients β_{jkl} have a normal distribution centered at some common mean μ_{kl}, which in turn have a normal distribution common to all the groups

$$\beta_{jkl} \mid \mu_{kl}, \sigma_l^2 \sim \mathsf{N}(\mu_{kl}, \sigma_l^2)$$

$$\mu_{kl} \mid a_{kl}, b_{kl} \sim \mathsf{N}(0, c(b_{kl} - a_{kl})^{\delta}),$$

which means the individual copy number changes within the same group are all correlated and centered at zero. The prior for μ_{kl} is taken from the continuous wavelet dictionary from Abramovich *et al.* (2000) and Chapter 2, where the variance depends on the segment length. Here c is a tuning parameter which does not depend on $\mathbf{a}$ and $\mathbf{b}$. We can see the shrinkage effect of the prior from the posterior of $\boldsymbol{\mu}$ in (B.1, See Appendix B). If we set $\delta > 0$, then the longer the segment, the less it will be shrunk toward zero. Therefore, this prior has the effect of keeping the magnitudes of long segments closer to the sample mean while shrinking the short ones which might have been outliers. To see the biological explanation behind this setting, suppose we have 2 segments A and B from Figure 3.1. The data plotted here are the sample mean for all patients in a group, and they are the same for the two segments, except that segment B is longer and more spread out. The prior will shrink μ_{kl} for A more toward zero, and will more likely find A insignificant. This is consistent with the biological background because the probes in the left hand side of the dash line are more likely to belong to the same segment due to their proximity, and segment A is

more likely just outliers.

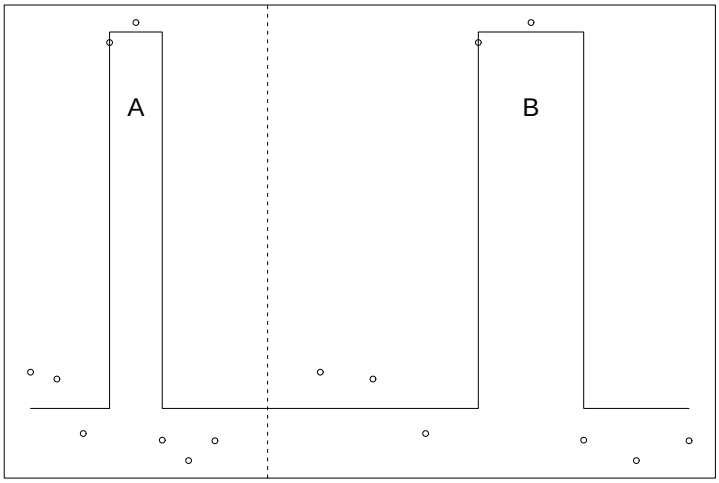

Figure 3.1: Two sample segments. The data on the left and right hand side of the dash line are identical except for the distance between probes.

We also have non-informative priors for σ^2 and σ_l^2 (Hill, 1965; Hobert and Casella, 1996; Sun *et al.*, 2001). It is straight forward to show the propriety of the joint posterior distribution in our setting

$$p(\sigma^2) \propto \sigma^{-2}, \quad p(\sigma_l^2) \propto (\sigma_l^{-2})^{1/2}.$$

The likelihood of the model, $L(K, \mathbf{a}, \mathbf{b}, \boldsymbol{\beta}, \boldsymbol{\mu}, \sigma_l^2, \sigma^2 \mid \mathbf{Y})$ is just a multivariate normal density with mean f and variance $\sigma^2 I$. Therefore the joint posterior distribution is proportional to

$$L(\mathbf{Y}; K, \mathbf{a}, \mathbf{b}, \boldsymbol{\beta}, \boldsymbol{\mu}, \sigma_l^2, \sigma^2) p(\mathbf{a}, \mathbf{b}) p(\boldsymbol{\beta} | \boldsymbol{\mu}, \sigma_l^2) p(\boldsymbol{\mu} | \mathbf{a}, \mathbf{b}) p(\sigma_l^2) p(\sigma^2).$$

3.3 Methods

Since the model has a varying number of segments, we use a Reversible Jump Markov Chain Monte Carlo (RJ-MCMC) algorithm (Green, 1995), which allows us to jump

between models with different dimensions. The RJ-MCMC here involves four kinds of steps: birth, death, split and merge. In birth/death step, we propose to add/delete a segment of copy gain/loss. In split step, we propose to split one segment into two adjacent ones, and in merge move, we propose to merge two adjacent segments into one. Though those are standard moves, one would have to make clever proposals or other changes to improve the acceptance rate, which is often difficult in Reversible Jump. Most of the innovative ideas are slight variations of Chapter 2, however, with different parametrization and conjugate structure of the model we are able to integrate out some of the parameters and preform block Gibbs sampling, which speeds up the algorithm and improves the mixing from our previous work. The details of the RJ-MCMC is in Appendix B.

After we run the algorithm for a large number of times and discard the first iterations, we will have samples from the joint distribution from which we can make inference. For example, to get the posterior probability of a copy number change common to a group, let g_{il} be the underlying mean copy number of probe i for group l:

$$g_{il} = \begin{cases} \mu_{kl} & \text{if} \quad a_{kl} \leq x_i < b_{kl} \quad \text{for some k} \\ 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

Let s_{il} be the underlying state of probe i in group l, which can be gain(G), loss(L) or unchanged(U), then the posterior probability

$$p(s_{il} = G \mid \mathbf{Y}) = p(g_{il} > 0|\mathbf{Y})$$

$$p(s_{il} = L \mid \mathbf{Y}) = p(g_{il} < 0|\mathbf{Y})$$

$$p(s_{il} = U \mid \mathbf{Y}) = p(g_{il} = 0|\mathbf{Y}),$$

which can be estimated by the ergodic average over all models from the MCMC run, for example,

$$\hat{p}(s_{il} = G \mid \mathbf{Y}) = \sum_{t=1}^{T} \frac{1}{T} I(g_{il}^t > 0), \tag{3.3}$$

where T denotes the total number of MCMC draws after burn-in and g_{il}^t denotes the estimation for g_{il} from the tth draw from the posterior. The probability for loss and unchanged can be estimated similarly. Notice that there is a positive probability that $g_{il} = 0$, and no threshold needs to be specified as it has been applied implicitly. Though the segmentation/classification is done at group level, not the individual level, we can still get the estimation of each individual patient by looking at the posterior draw of $\boldsymbol{\beta}$'s.

In addition, given new data $\mathbf{Y}^* = \{y_1^*, ...y_n^*\}$, we can calculate the posterior probability that $\mathbf{Y}^*$ belongs to group l:

$$p(\mathbf{Y}^* \in l \mid \mathbf{Y}, \mathbf{Y}^*) = \frac{\pi(\mathbf{Y}^* \in l) f(\mathbf{Y}^* \mid \mathbf{Y}^* \in l, \mathbf{Y})}{\sum_s \pi(\mathbf{Y}^* \in s) f(\mathbf{Y}^* \mid \mathbf{Y}^* \in s, \mathbf{Y})}, \tag{3.4}$$

where $\pi(\mathbf{Y}^* \in l)$ is the prior probability for group l and $\theta_l = \{\mathbf{a}_l, \mathbf{b}_l, \boldsymbol{\mu}_l, \sigma_l^2\}$ denotes the parameter particular for group l. The posterior distribution of the new data $\mathbf{Y}^*$, given that $\mathbf{Y}^*$ belongs to group l, is

$$\begin{aligned}
& f(\mathbf{Y}^* \mid \mathbf{Y}^* \in l, \mathbf{Y}, \theta_l) \qquad (3.5)\\
= \;& f(\mathbf{Y}_0^* \mid \mathbf{a}, \mathbf{b}) f(\boldsymbol{\mu} \mid \mathbf{a}, \mathbf{b}) \prod_k \int f(\mathbf{Y}_k^* \mid \beta_{kl}^*, \sigma^2, a_{kl}, b_{kl}) f(\beta_{kl}^*, \mid \mu_{kl}, a_{kl}, b_{kl}, \sigma_l^2) d\beta_{kl}^* \\
= \;& f(\mathbf{Y}_0^* \mid \mathbf{a}, \mathbf{b}) f(\boldsymbol{\mu} \mid \mathbf{a}, \mathbf{b}) \\
\times \;& \prod_k \left\{ \mathsf{N}\left(\bar{\mathbf{Y}}_k^*; \mu_{kl}, \frac{\sigma^2}{n_{kl}} + \sigma_l^2 \right) (2\pi\sigma^2)^{-\frac{n_{kl}-1}{2}} n_{kl}^{-1/2} \exp - \frac{\sum_i (y_{ik}^* - \bar{\mathbf{Y}}_k^*)^2}{2\sigma^2} \right\},
\end{aligned}$$

where $\mathbf{Y}_k^* = \{y_i^* \in \mathbf{Y}^*; a_{kl} \leq x_i < b_{kl}\} = \{y_{ik}^*, i = 1, ...n_{kl}\}$, and $\mathbf{Y}_0^* = \{y_i^*; \sum_k I(a_{kl} \leq x_i < b_{kl}) = 0\}$ is the part not "covered" by any basis elements.

Notice that the posterior probability of the new data can be factored into k elements and the non-aberration areas. Ideally when we compare the groups we would like to isolate the effect of each probe on the group differences. Though the posterior cannot be factored into the multiple of individual probes, we can define the contribution of an individual probe as follows:

$$\begin{aligned} & \tilde{p}(y_i^* \mid \mathbf{Y}^* \in l, \theta_l) \qquad (3.6) \\ = & \begin{cases} \mathsf{N}\left(\bar{\mathbf{Y}}_k^*; \mu_{kl}, \frac{\sigma^2}{n_{kl}} + \sigma_l^2\right)^{\frac{1}{n_{kl}}} & \\ \times \mathsf{N}\left(\mu_{kl}; 0, c(b_{kl} - a_{kl})^{\delta}\right)^{\frac{1}{n_{kl}}} \mathsf{N}\left(y_i^*; \bar{\mathbf{Y}}_k^*, \sigma^2\right) & \text{if} \quad \exists k \quad a_{kl} \leq x_i < b_{kl} \\ \mathsf{N}\left(y_i^*; 0, \sigma^2\right) & \text{otherwise,} \end{cases} \end{aligned}$$

and express (3.5) as the product of $\tilde{p}(y_i^* \mid \mathbf{Y}^* \in l, \theta_l)$'s. Let $\theta_l^1, ..., \theta_l^T$ denote the T draws from the posterior distribution, then we can calculate the ratio of predictive density with the same form as the Bayes factor:

$$\begin{aligned} \hat{B}_{01}(\mathbf{Y}^*) &= \frac{f(\mathbf{Y}^* \mid \theta_0^t)}{f(\mathbf{Y}^* \mid \theta_1^t)} = \frac{\sum_{t=1}^T \prod_{i=1}^n \tilde{p}(y_i^* \mid \mathbf{Y}^* \in 0, \theta_0^t)}{\sum_{t=1}^T \prod_{i=1}^n \tilde{p}(y_i^* \mid \mathbf{Y}^* \in 1, \theta_1^t)} \qquad (3.7) \\ &= \frac{\sum_{t=1}^T \exp \sum_{i=1}^n \log(\tilde{p}(y_i^* \mid \mathbf{Y}^* \in 0, \theta_0^t))}{\sum_{t=1}^T \exp \sum_{i=1}^n \log(\tilde{p}(y_i^* \mid \mathbf{Y}^* \in 1, \theta_1^t))}. \end{aligned}$$

Therefore, we can break down the predictive ratio into individual probes and see which ones differ significantly from 1 to identify the probes that are relevant to the group features. Note that this factorization only works for an individual model drawn from the posterior, but not the ergodic average. However, looking at the marginal contribution at each probe for all individual model compared to the baseline model should give us an idea where these two groups differ.

3.4 Results

3.4.1 Ovarian cancer data

As an illustration, we applied our method on the ovarian cancer data set from Duke University Medical Center. Sixty-nine patients are sorted into 4 different cancer groups (Borderline, Early stage, Late stage/Short term survivor and Late stage/Long term survivor). Each sample (patient) has 2,016 observations on the genomic sequence, with about 5% of the data missing.

Figure 3.2 shows the estimate at the group level g_{il} (solid line) with the group sample mean (points). The colored areas indicate the pointwise and simultaneous Bayesian credible band. Notice that in some segments it looks like the sample mean differs significantly from zero, while we find it insignificant. For example, the first part of chromosome 3 in the early stage group. This is due to a few outliers in that group driving the sample mean from zero. Since our goal is to find the area where the copy number changes are common to the patients, we tend to not declare a segment gain or loss when the gain or loss is only present in a few samples. If we plot the fitted values for individual samples (Figure 3.3), we can see that overall we capture the overall structure of the data well.

The estimated posterior probability of copy gain or loss is shown in Figure 3.4. Samples from the borderline group look like pure noise and do not suggest any copy number changes anywhere. But for the other groups, we can clearly see the area where a copy number alteration is most likely. We compare our results with the minimum common region found by cghMCR package (Zhang and Feng, 2006). First we try the default threshold value in (Aguirre *et al.*, 2004), and no MCR are found. When we lower the thresholds we get somewhat comparable results, which are shown in the shade areas in Figure 3.4. We can see that MCRs mostly agree with the posterior

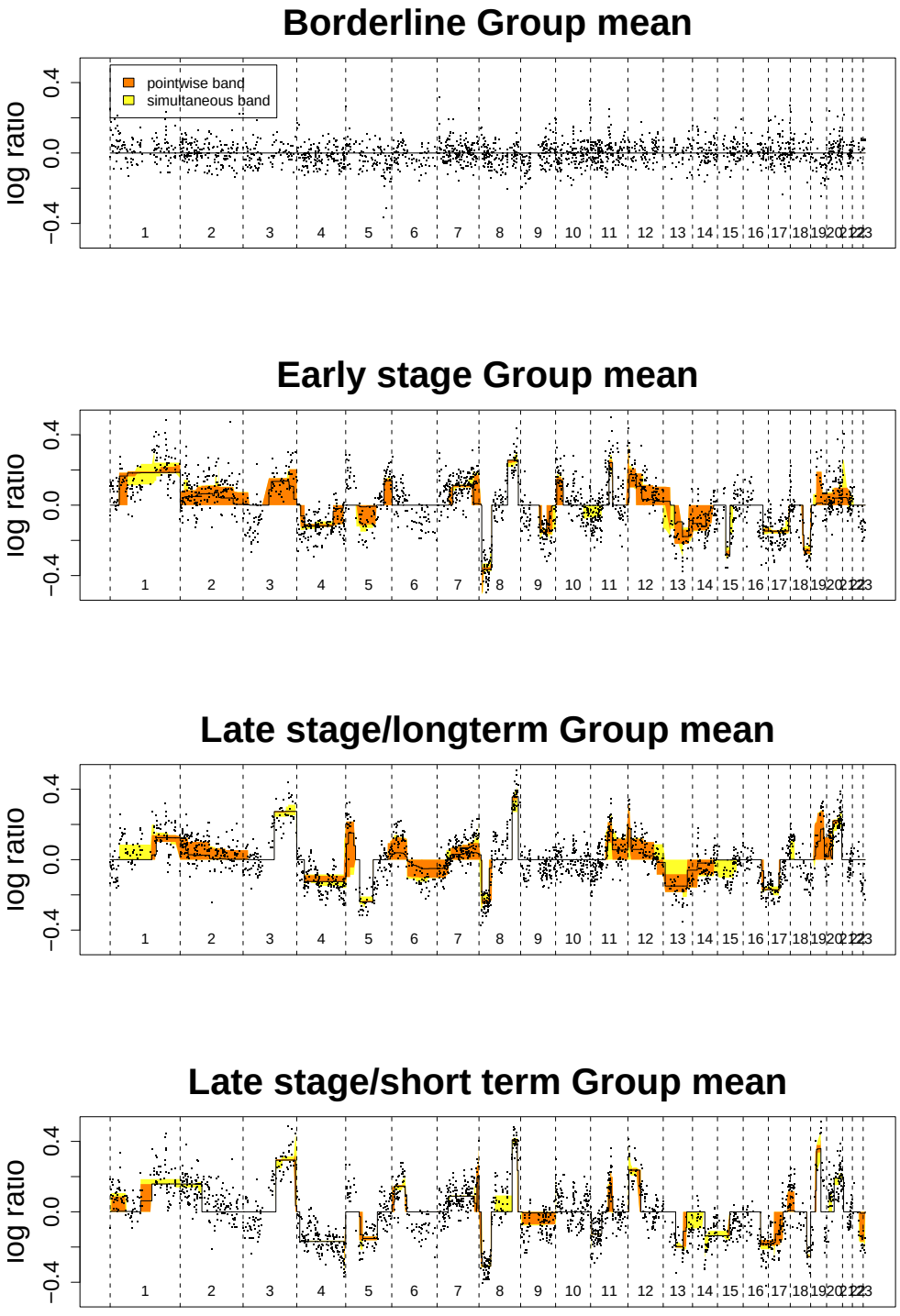

Figure 3.2: The estimate of the copy number change at the group level g_{il} from the ovarian cancer data (the solid line) and the group sample mean (points) with Bayesian credible bands

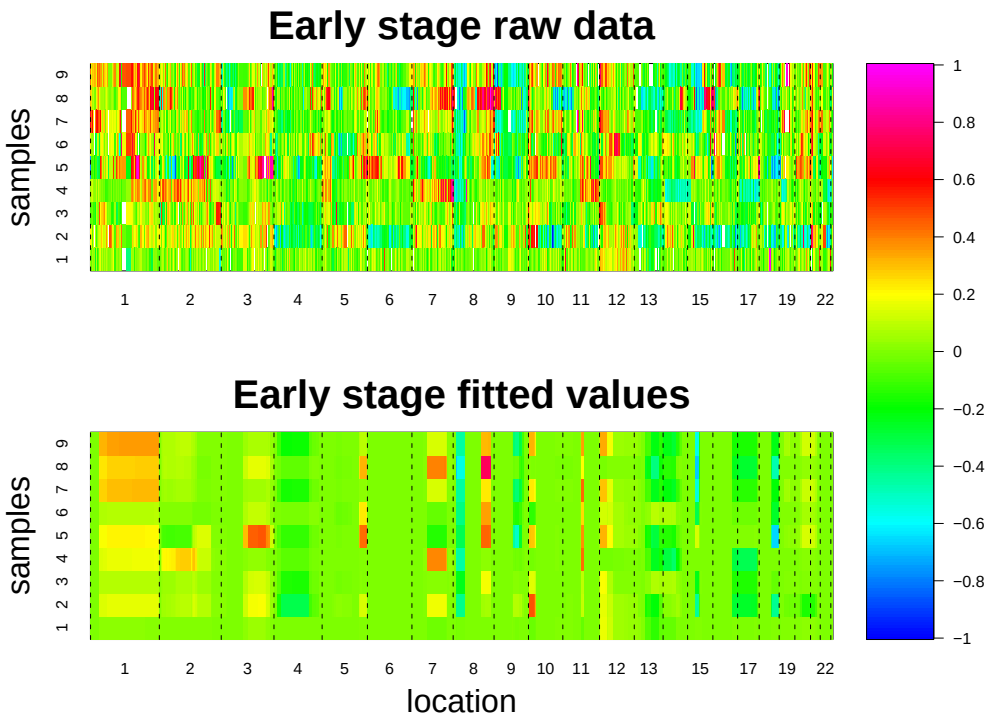

Figure 3.3: Heatmap for individual samples in early stage group

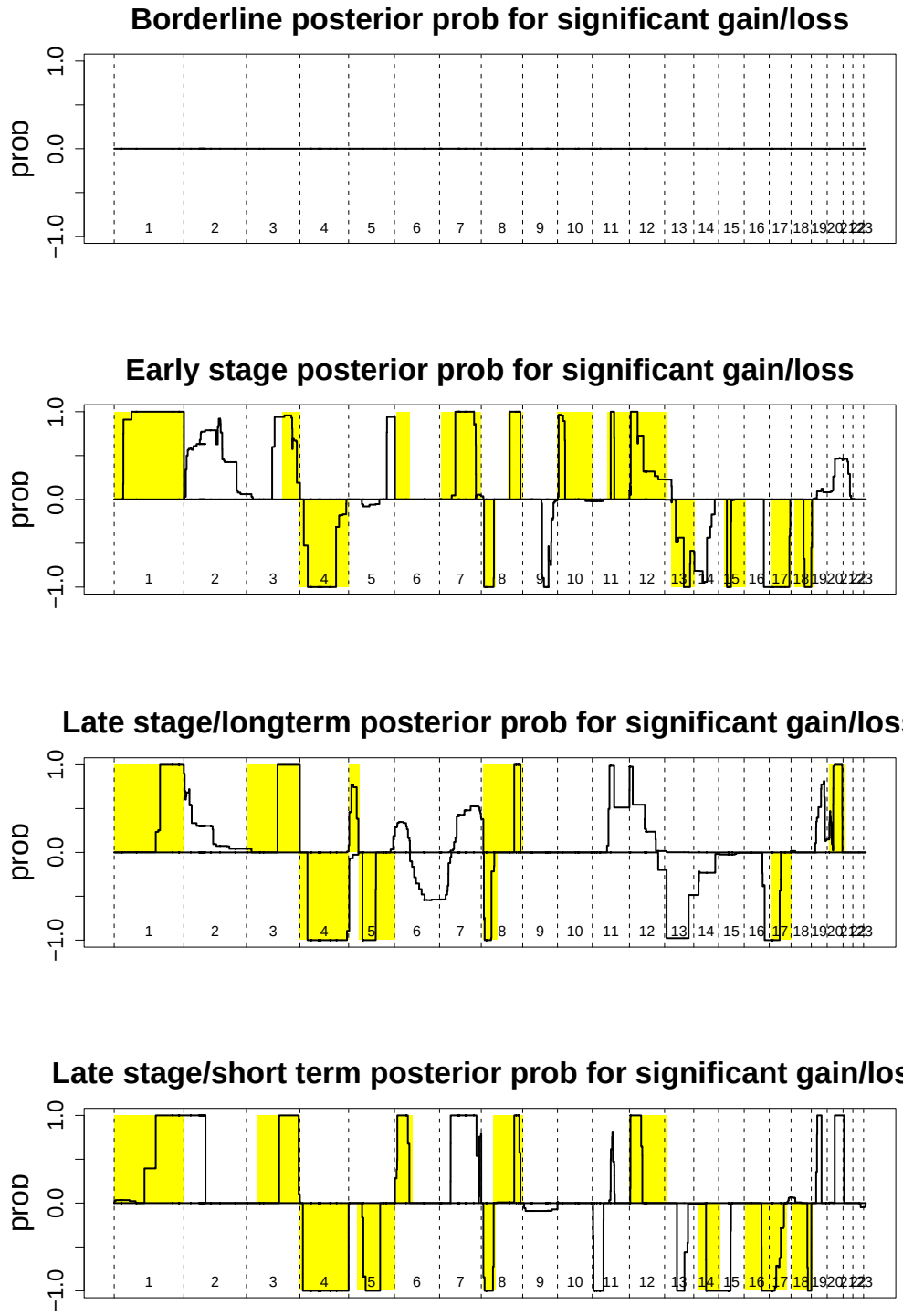

Figure 3.4: The posterior probability of gain/loss at the group level. The area above zero indicates the probability of copy gain and the area below zero indicates the probability of copy loss. The yellow area is the minimum common region found by cghMCR.

probability found by our method, though there are some obvious discrepancies. For example, In chromosome 8 of the late stage/longterm group there is a region declared as both gain and loss by cghMCR, but looking at the sample mean it is obviously a loss region. Also in chromosome 7 of the late stage/short term group, we declare it a gain region with very high posterior probability, but looking at the sample mean you can see that though they are all positive, the magnitude is smaller than other significant regions, therefore it was not called a gain region by cghMCR.

3.4.2 Simulated Data

To make a full comparison of our method and cghMCR, we test both methods on a simulated data set. We generate 10 genomic sequences from the posterior draws of the ovarian data, to create a realistic $\log_2$-ratio-copy number profile. We try to choose the samples with different numbers of segments to get a variety of mean function while keep the setting more realistic. The number of patients from each group are kept the same as the real data (9 early stages, 25 late stage/longterm, 30 late stage/short term), except that we can discard the 5 borderline case since the posteriors are all null functions.

The simulated data are generated from the following equations:

$$\hat{f}_{jl}(x_i) = \sum_{k=1}^{K_l} \hat{\beta}_{jkl} h_{kl}(x_i), \quad h_{kl}(x) = I(a_{kl} \leq x < b_{kl}) \tag{3.8}$$

$$\hat{Y}_{ijl} = \hat{f}_{jl}(x_i) + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2).$$

Note that the simulated data only depend on the posterior draws for $\boldsymbol{\beta}, \mathbf{a}, \mathbf{b}$ and σ^2, not on $\boldsymbol{\mu}$ and σ_l^2.

We run cghMCR on the simulated data with 3 set of parameters, the default value in (Aguirre *et al.*, 2004) and two other ones with lower thresholds. For our method,

we set the classification rule to be the highest probability model:

$$s^*_{il} = \arg\max_S p(s_{il} = S|\mathbf{Y}), \quad S = \{G, U, L\}, \tag{3.9}$$

where we put each probe in the category with the highest posterior probability. To see if the decision rule affects our performance, we also try two other decision rules with lower threshold for gain or loss:

$$s^*_{il} = \begin{cases} G & \text{if} \quad p(s_{il} = G|\mathbf{Y}) \geq p^* \\ L & \text{if} \quad p(s_{il} = L|\mathbf{Y}) \geq p^* \\ U & \text{otherwise.} \end{cases} \tag{3.10}$$

with $p^* = 0.25, 0.1$, and we declare a copy number gain or loss when the posterior probability exceeds that threshold $p^* = 0.25$.

The overall results can be further summarized by the following summary statistics:

Correct Classification Rate (CCR) The proportion of probes assigned to the correct class.

False Discovery Rate (FDR) The proportion of unchanged probes assigned to class Gain or Loss.

True Positive Rate (TPR) The proportion of probes predicted to be gain or loss of all the true gains/losses.

True Negative Rate (TNR) The proportion of probes predicted to be unchanged of all the true unchanged probes.

Table 3.1: Summary statistics from simulation study

	Hierarchical model			cghMCR		
Thresholds	High	Med	Low	High	Med	Low
CCR	0.8465	0.9185	0.9120	0.5789	0.5888	0.7901
FDR	0.0022	0.0213	0.0966	0.0000	0.0011	0.0203
TPR	0.6417	0.8269	0.9126	0.0129	0.0363	0.5189
TNR	0.9989	0.9866	0.9273	1.0000	0.9999	0.9920

We see that cghMCR achieves great FDR and true negative rate but is not a great method overall since it predicts unchanged most of the time and the overall correction rate is lower. If we look at the correct classification rate, which is a good indication of the overall performance of the method, our method did better than cghMCR for all combination of thresholds tested.

Note that the simulated data has the same noise level as the real data and the resulting signal-to noise ratio is less than one. We perform the same analysis with simulated data with lower noise (signal-to-noise ratio=3):

Table 3.2: Summary statistics from simulation study, low noise

	Hierarchical model			cghMCR		
Thresholds	High	Med	Low	High	Med	Low
CCR	0.9886	0.9895	0.9886	0.6112	0.7718	0.9906
FDR	0.0059	0.0071	0.0104	0.0039	0.0049	0.0088
TPR	0.9791	0.9824	0.9836	0.0892	0.4676	0.9868
TNR	0.9957	0.9948	0.9923	0.9997	0.9982	0.9935

In cghMCR with lower thresholds, where the thresholds are set at the 67% and 33% percentiles of the data and the recurrence rate at 50%, the results are comparable to our methods. But we can see that the performance of cghMCR depends on the threshold values, while our method gives consistently good performance for all three decision rules we tried.

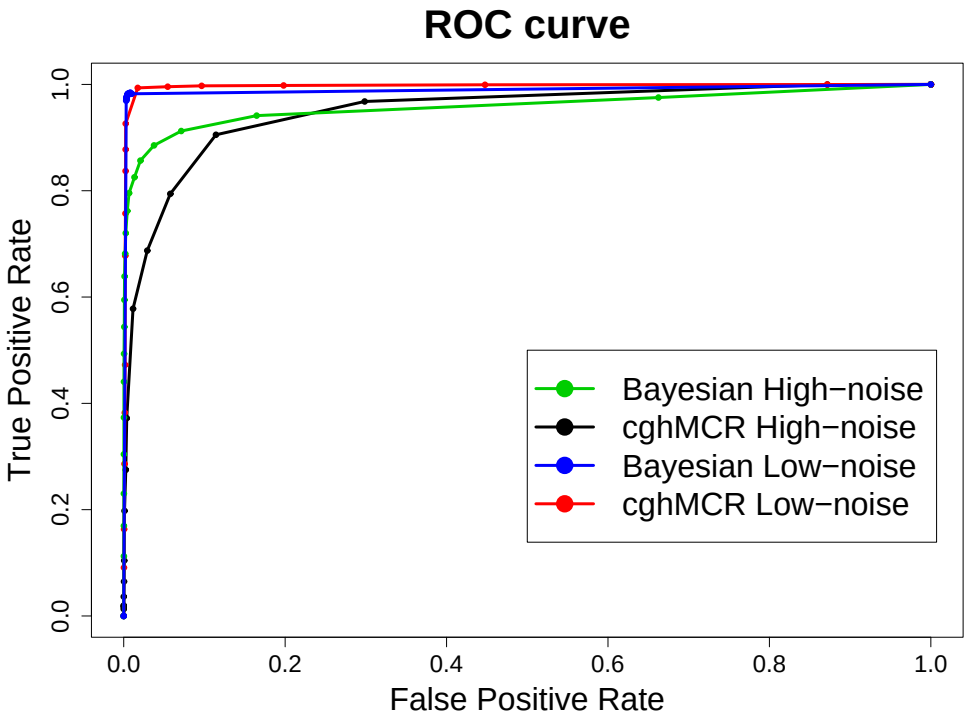

Figure 3.5: Receiver operating characteristic (ROC) curves for our methods and cghMCR measured at two different signal-to-noise ratios. The curves were generated by measuring the true and false positive rates on simulated data at 21 threshold values for each method.

We calculate the ROC profiles for each method and each noise level, and results are shown in Figure 3.5. We run both methods with a sequence of threshold values, and each threshold give a TPR and TNR. The false positive rate (FPR) is defined as 1-TNR, and each pair of TPR and FPR is represented by a point in the ROC curve, with the upper left corner representing the optimal result. We can see that in the high noise scenario, which is more realistic, our method is doing better than cghMCR. However in the low noise scenario, both methods seem to be doing equally well.

3.4.3 Classification

To demonstrate the new patient classification, we run out-of-sample prediction on another simulated data set. The setting is similar to Lai *et al.* (2005), with various aberration segment widths. The simulated data consist of 4 groups, each with 10 patients. The first group has three "gain" segments with widths of 5,10 and 20 probes. And the remaining three groups have only two of the corresponding segments while missing one of them. Gaussian noise was added to make the SNR around 2. Figure 3.6 shows the features of each group and their difference.

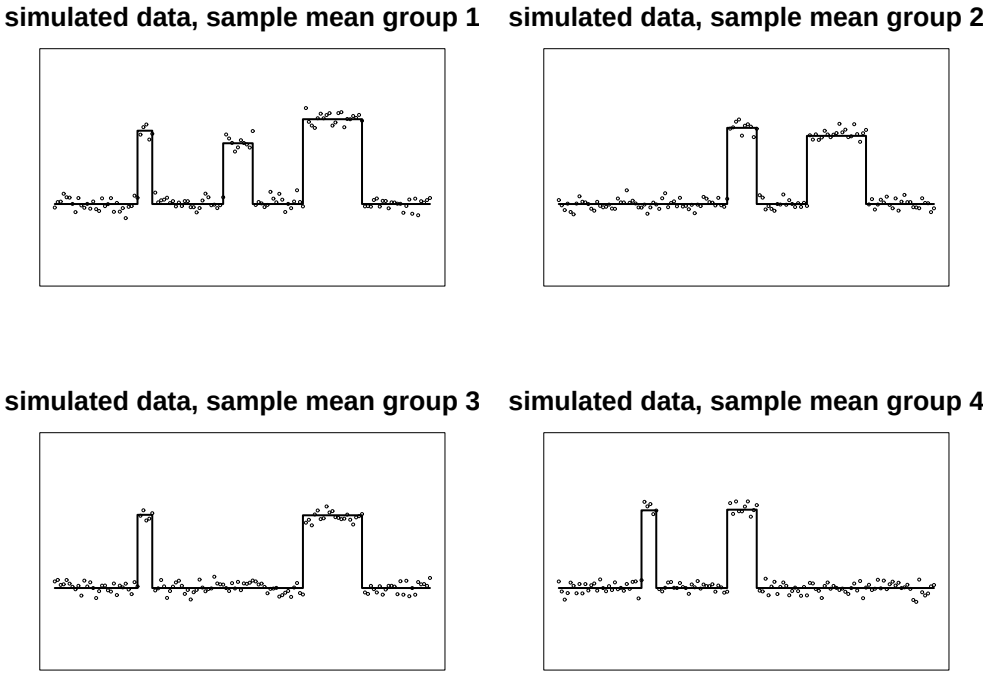

Figure 3.6: Sample mean of the simulated data for each group. The solid lines are the mean underlying copy numbers

The predictive probabilities of each group in (3.4) were estimated by fitting the model after taking out 20% of the patients in each group. Given a uniform prior on the groups and taking the group with the highest posterior probability as the

prediction, we obtain the confusion matrix:

Table 3.3: Confusion matrix for cross validation of the simulated data. The numbers in parenthesis show the average posterior probability for each group.

		Predicted Group			
		1	2	3	4
True Group	1	8 (0.8000)	2 (0.2000)	0 (0.0000)	0 (0.0000)
	2	2 (0.2761)	8 (0.7239)	0 (0.0000)	0 (0.0000)
	3	2 (0.2553)	0 (0.0000)	8 (0.7446)	0 (0.0000)
	4	4 (0.3791)	0 (0.0000)	0 (0.0000)	6 (0.6209)

This gives us a 75% correction rate. Sometimes group 2,3,4 patients are misclassified as group 1 because the fitted model for group 1 failed to pick up the corresponding signal, due to high noise. As Lai *et al.* (2005) showed, the segmentation based methods have not done well under high-noise scenarios, especially when the segments are short. Given the difficulty of detecting these small segments with noisy data, our method is preforming reasonably well.

We can also identify the probes most relevant to the group difference by plotting the marginal contribution at each probe. Figure 3.7 shows the log ratios for all posterior draws, and one can clearly see the region where the log ratios are below zero. Compared to Figure 3.6, those regions correspond to the part where the two groups differ. We also see a larger variance in all the aberration segments because there is another layer of uncertainty that is the within group variance.

We return to the ovarian cancer data and run the same out-of-sample prediction. As we can see from Figure 3.2, all the groups except the borderline group have very

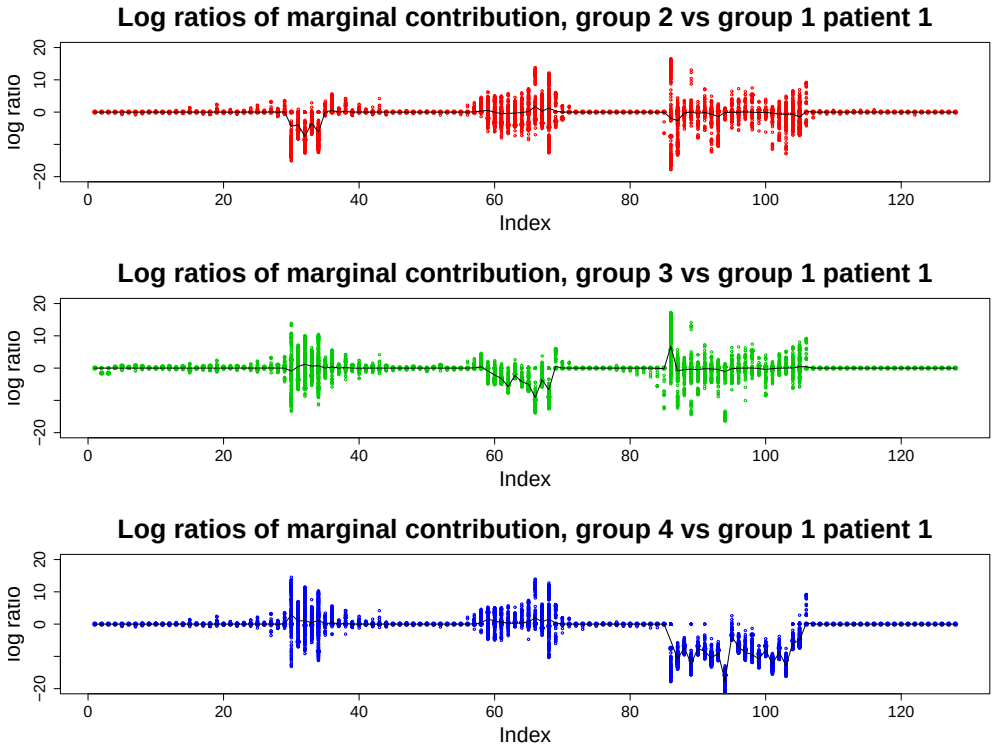

Figure 3.7: The log ratios of marginal contribution at each probe for one simulated sample in group 1. Group 1 is the baseline model. The solid line is the average of all posterior draws, which should not be taken as the model averaging estimate. 500 thinned samples from MCMC draw are plotted at each location.

similar features. As a result, we could not separate those groups very well. However, the marginal contribution can still show subtle differences. For example, if we take the early stage group as the baseline, and plot the log ratio at each position for patient 7, from Figure 3.2 we can see a gain region present in chromosome 19 for both the late stage groups, but not in the early stage group. This is clearly seen in Figure 3.8. As we can see in Figure 3.3, this particular patient does have a gain segment in chromosome 19, which is a feature in the late stage groups, while most in the early stage group do not. Not surprisingly, she was (mistakenly) classified as late stage.

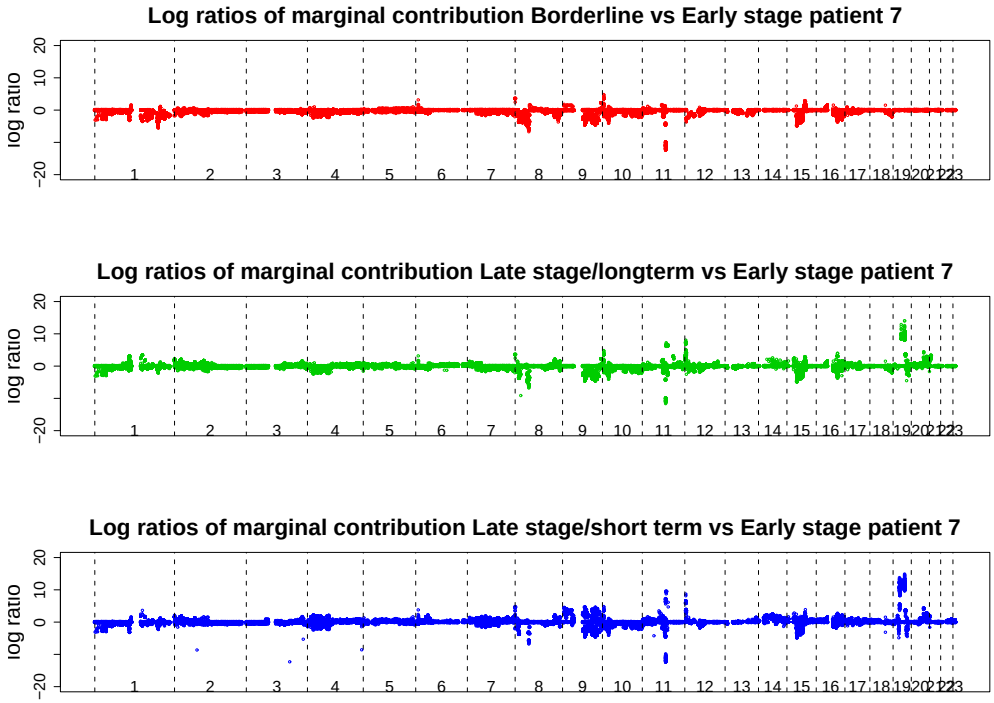

Figure 3.8: The log ratios of marginal contribution at each probe for one patient in the early stage group. Early stage is the baseline model. 100 thinned samples from MCMC run are plotted at each location.

3.5 Discussion

We propose a procedure based on a hierarchical model to detect the region where the gene copy number changes. We apply the method to the ovarian cancer data and the simulated data and the results show that our method can capture the features shared among the majority of the patients in a group. We find our model-based approach works better than cghMCR, the only current method that deals with multiple samples. In addition, the features found in the groups can again be used for new patient classification, and we can also identify the genes relevant to the group differences, which can be useful in both the basic research and clinical applications.

One of the main issues of the model-based approach is if our model could ad-

equately describe the data structure. Although our model assumes the same segmentation for all the patients in the same group, which is a strong assumption, the hierarchical nature of the model accounts for the heterogeneity within the group. For example, looking at the early stage plot in Figure 3.2, there is bump in chromosome 7. While the upper panel of the heatmap (Figure 3.3) shows that such bump is not common among all the patients, the hierarchical model will shrink the individual coefficients near zero for those who do not share that gain region, as we can see from patient 5 and 6 in the lower panel in Figure 3.3. When we look at the fitted values for individual patients the model does give a good fit overall, even though it can keep a very small non-zero coefficient for someone when it's supposed to be zero.

In conclusion, the complexity of CGH data calls for a model-based approach to analyze multiple samples jointly to get more accurate information. Our model offers an efficient way to find the common feature among the samples.

Chapter 4

Joint Analysis of Gene Expression and DNA Copy Number Data

4.1 Introduction

It has been suggested that the changes in number of genomic DNA copy numbers play a prominent role in cancer activity (Lockwood *et al.*, 2006). Detection of DNA copy changes has been simplified due to array comparative genomic hybridization (CGH) technology, which measures DNA copy numbers indirectly. In recent years a great amount of research has been conducted in this area, in both applied and methodological work (Willenbrock and Fridlyand, 2005; Olshen *et al.*, 2005; Hupe *et al.*, 2004; Wang *et al.*, 2005; Myers *et al.*, 2004; Broët and Richardson, 2006; Lai *et al.*, 2005). Cancer development can also appear in the form of gene over or under expression (Lockhart and Winzeler, 2000; Perou *et al.*, 1999, 2000; Sørlie *et al.*, 2001, 2003), which may result from copy number changes. When copy numbers increase, the extra RNA transcripts and their encoded proteins may lead to gene overexpression, and vice versa. However, the gene expression patterns are noisy and it is still unclear to what extent copy numbers affects expression level, especially on the individual level. Combining information from DNA copy numbers and expression levels may reduce the source of variation and lead to more accurate and reliable results.

To date, only a few studies have analyzed both DNA copy number and gene expression level simultaneously. Hyman *et al.* (2002) studied 14 breast cancer cell lines and concluded that 44% of the highly amplified genes showed overexpression. Pollack *et al.* (2002) analyzed 37 breast cancer tumor samples and found that a 2-fold

change in DNA copy number was associated with 1.4-1.5 fold changes in expression level. They also found that 7-12 percent of the variation in expression level may be attributed to copy number variation. For other cancer types, Platzer *et al.* (2002) found only 3% of the genes showed overexpression among all the amplified genes in a colon cancer study. Therefore, the effect of copy number alteration on the expression level may be cancer specific. Those studies are either exploratory or based on a hypothesis-testing framework and are mostly done on a gene-by-gene scale. More recently, Berger *et al.* (2006) proposed a generalized singular value decomposition (GSVD) algorithm that iteratively selects the genes with highly similar patterns of variation for both data sets. Chin *et al.* (2006) did a more comprehensive analysis of gene expression and CGH data for breast cancer and identified 66 genes with highly amplified copy numbers that are associated with expression levels. Nine of them are considered druggable. Their results also suggested some strong association of DNA copy numbers with outcome and other clinical variables, such as stage and survival time.

In this chapter we propose a Bayesian hierarchical model to analyze gene expression and copy number data jointly. We extend the hierarchical model for CGH data from Chapter 3 to incorporate gene expression data. Our main goal here is to identify the genes in which pattern of expression is related to DNA copy number, as these are likely genes related to cancer progression. We can also detect the copy number change regions and their significance. The level of association between gene expression level and DNA copy number is defined by a linear regression coefficient θ and inferences about θ are based on posterior probabilities. Our method can easily accommodate missing and misaligned data, which is an improvement over GSVD.

The remaining chapter is arranged as follows. In Section 4.2 we present the hierarchical model. In Section 4.3 we outline our methods of estimation and inference.

In Section 4.4, we present the result using simulated data and the breast cancer cell line cDNA data from Hyman *et al.* (2002). We conclude in Section 4.5 with a discussion of the extensions of our model.

4.2 Approach

Suppose we have J patients indexed by $j = 1, ...J$ and n probes located on the genomic sequence with their physical locations labeled as $x_1, ...x_n$. Let c_{ij} be the observation of the CGH data of probe i of patient j, which is the log ratio of fluorescence intensities between tumor and reference samples. As in Chapter 3, we model c_{ij} by a step function of the form

$$c_{ij} = f_j(x_i) + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} \mathsf{N}(0, \sigma^2), \tag{4.1}$$

$$f_j(x_i) = \sum_{k=1}^{K} \beta_{jk} h_k(x_i), \quad h_k(x) = I(a_k \leq x < b_k). \tag{4.2}$$

Most model-based methods take similar approaches (Lai *et al.*, 2005). Here h_k is the kth segment where a copy number alteration occurs, spanning from a_k to b_k. In this CGH segmentation model, all the patients share the same segmentation, but may have different coefficients β_{jk}. All segments are mutually exclusive and do not overlap. We use $\mathbf{a} = \{a_1, ...a_K\}, \mathbf{b} = \{b_1, ...b_K\}$ to denote the beginning and end points for the segments, respectively.

In the Bayesian framework, we have to specify prior distribution for all unknown parameters. The prior settings for the parameters related to the copy numbers are similar to Chapter 3. We have uniform prior distribution for $\mathbf{a}$ and $\mathbf{b}$ on the set of data points $x_1, ..., x_n$. We take a negative binomial prior distribution $NB(r, q)$ on the number of segments K. We can choose the hyperparameter r and q based on the size of data and other information about a reasonable range of K. From our

experience a flat prior such as $NB(1, 0.01)$ can often work as well in the absence of prior information. Note that although these segment can be as small as one probe, in practice we cannot detect those very small regions with much confidence.

The prior distributions for the coefficients β_{jk}'s and σ^2 are given below

$$\beta_{jk}|\mu_k, \sigma_l^2 \sim \mathsf{N}(\mu_k, \sigma_l^2),$$

$$\mu_k \mid a_k, b_k \sim \mathsf{N}(0, c(b_k - a_k)^\delta),$$

$$p(\sigma_l^2) \propto (\sigma_l^{-2})^{1/2}, \quad p(\sigma^2) \propto \sigma^{-2}.$$

The β_{jk}'s are centered around a common mean μ_k, which is centered at 0. The hierarchical structure allows us to extract common features across patients, while it also accounts for heterogeneity among patients. The hyperparameters c and δ control the level of shrinkage in relation to the segment length. A reasonable choice for c is to set $c = J$, the sample size, as in the unit-information prior (Kass and Wasserman, 1995). For δ, generally we choose $\delta > 0$ so the short segments will be shrunk more toward zero as they will more likely be outliers instead of real copy number aberrations. The non-informative priors given for the variance parameters, σ^2 and σ_l^2, are improper, but it can be easily shown to have proper posterior distributions (Hill, 1965; Hobert and Casella, 1996; Sun *et al.*, 2001).

Our prior specification also implies a model for the CGH data at the population level,

$$f(x_i) = \sum_{k=1}^{K} \mu_k h_k(x_i). \tag{4.3}$$

We will refer to the function above as the mean function of f_j.

Let d_{gj} denote the probe measurement on log scale for patient j at location x_g. We model the gene expression data as the following

$$d_{gj} = \alpha_g + \theta_g f_j(x_g) + \epsilon_{gj}, \quad \epsilon_{gj} \overset{iid}{\sim} \mathsf{N}(0, \sigma_d^2). \tag{4.4}$$

Here θ_g represents the level at which the expression associates with the latent gene copy number. There is no association when x_g is not covered by any gain/loss segment, i.e., $f_j(x_g) = 0$, so in that case, we set $\theta_g = 0$. Since previous studies suggest that associations between gene expression levels and DNA copy numbers are not likely to be present in all genes, we expect only a portion of the θ's would be significantly different from zero. Therefore we propose a mixture prior for θ_g with a point mass at zero with probability π and a normal component centered at some common mean ξ which has a normal prior centered at zero, namely,

$$\begin{aligned} p(\theta_g) &= \begin{cases} \pi\delta_0(\theta_g) + (1-\pi)\mathsf{N}(\theta_g; \xi, \tau^2) & \text{if } x_g \in [a_k, b_k) \\ \delta_0(\theta_g) & \text{otherwise,} \end{cases} \\ p(\xi) &\sim \mathsf{N}(0, \nu^2). \end{aligned}$$

Although previous studies (Hyman *et al.*, 2002; Pollack *et al.*, 2002) suggest that the association parameter θ_g, if non-zero, will be more likely positive than negative, our hierarchical setting does allow some of the θ_g to be less than zero. Also, this hierarchical prior for θ forces some dependency among all the θ's. This prior should guide the θ's to follow the overall trend ξ as the data suggest, but also accounts for heterogeneity of the genes. We specify non-informative priors on the remaining parameters, the intercept α_g and the error variance σ_d^2:

$$p(\alpha_g) \propto 1, \quad p(\sigma_d^2) \propto \sigma_d^{-2}.$$

It is straight forward to show that the posterior distribution is proper although the prior distributions are improper.

Let $\mathbf{C} = \{c_{ij}\}$ and $\mathbf{D} = \{d_{gj}\}$ denote the set of CGH and expression data. To present the likelihood we introduce the following notation. Given the segmentation parameter $(\mathbf{a}, \mathbf{b})$, we define

$$\mathbf{C}_j^k = \{c_{ij}^k, i = 1, ..n_k^c\}, \quad \mathbf{D}_j^k = \{d_{gj}^k, g = 1, ..n_k^d\}$$

as the collections of data points "covered" by segment k for patient j for CGH and expression data, respectively, with n_k^c and n_k^d denoting the corresponding numbers of data points in segment k. We also define

$$\mathbf{C}^0 = \mathbf{C} \setminus \cup_{jk} \mathbf{C}_j^k \quad \mathbf{D}^0 = \mathbf{D} \setminus \cup_{jk} \mathbf{D}_j^k$$

as the data points not "covered" by any segment, where $f(x_i) = 0$. Now the likelihood may be written as

$$\begin{aligned} & L(\boldsymbol{\beta}, \boldsymbol{\mu}, \mathbf{a}, \mathbf{b}, \sigma^2, \sigma_l^2, \sigma_d^2, \theta, \alpha \mid \mathbf{C}, \mathbf{D}) \\ \propto & \left\{ \prod_{c_{ij} \in \mathbf{C}^0} \mathsf{N}(c_{ij}; 0, \sigma^2) \prod_{d_{gj} \in \mathbf{D}^0} \mathsf{N}(d_{gj}; \alpha_g, \sigma_d^2) \right. \\ \times & \left. \prod_{jk} \prod_{c_{ij} \in \mathbf{C}_j^k} \mathsf{N}(c_{ij}; \beta_{jk}, \sigma^2) \prod_{d_{ij} \in \mathbf{D}_j^k} \mathsf{N}(d_{gj}; \alpha_g + \theta_g \beta_{jk}, \sigma_d^2) \right\}. \end{aligned}$$

4.3 Methods

Inference for parameters is done by sampling from the joint posterior distribution using a Reversible Jump Markov Chain Monte Carlo (RJ-MCMC) algorithm (Green, 1995), which consists of four types of steps: birth, death, split and merge. As the names suggest, in a birth/death step, we propose to add/delete a segment of copy number gain/loss; in the split step, we propose to split a segment in two; and in the merge step, we propose to merge two segments into one. The RJ-MCMC algorithm allows us to jump between models with different dimensions, in our case, models with different numbers of copy number gain/loss segments. The detail of the RJ-MCMC algorithm is in Appendix C.

The results presented in the following section are based on 5,000 steps of the RJ-MCMC algorithm after a burn-in period of 5,000. The trace plots of the parameters

suggest good mixing for the MCMC chain after 5,000 steps. After we collect the samples from the posterior distribution, summary statistics for parameters of interest can be obtained. For example, we can provide an estimate the posterior probability of $\theta_g \neq 0$ for each probe as a measurement of the degree of association between the CGH and expression data. The investigator can then focus on the probes with high posterior probability of $\theta_g \neq 0$.

Notice that our model does not require the CGH and expression data to be observed to be at the same location. Missing data can also be easily treated in our model by adding a step in the algorithm to sample the missing data from its posterior distribution.

4.4 Results

4.4.1 Simulation study

We designed a simple simulation study to validate our model. In this simulated data the CGH and expression data are aligned, i.e. $\{x_i\} = \{x_g\}$ and there are no missing data.

We have $\{x_i\} = \{x_g\}$ located at 128 equally spaced points in (0,1). There are 30 patients. For simulated CGH data, the true function f is a step function with 7 gain/loss segments. The patient-specific coefficients β's are drawn from the prior with $\sigma_l^2 = 0.7$. For the expression data, the association parameter θ_g is either 0 (no correlation) or a non-zero value drawn from $\mathsf{N}(\theta_1, 1)$, where θ_1 takes on three positive values: 1,3 and 5. In addition, when x_g is not covered by any gain/loss segment, we set θ_g to be zero. In our simulated data, the non-zero θ's make up about one-third of the probes (42 out of 128). The intercept α_g are randomly drawn from $\mathsf{N}(0, 1)$. Random noise is added to both sets of data to make the signal-to-noise ratio around 2. We generate three data sets with different values of θ_1 and run the analysis on the

same data set with 3 different hyperparameters τ^2 to test the sensitivity of the prior specification. Other hyperparameters are fixed: $\nu^2 = 1$, $q = 0.5$, $\delta = 1$ and $c = 30$, the sample size.

We reported our results on the posterior probability of $\theta_g \neq 0$ in the following tables. The results do not appear to be sensitive to the hyperparameter τ^2. From the tables we can see that as the signal gets stronger (higher θ_1), our algorithm does a better job separating those relevant probes. If we consider the median probability model (Barbieri and Berger, 2004), we can detect about 90% of the probes with positive association levels when $\theta_1 = 3$ or 5.

Table 4.1: Posterior probability of $\theta \neq 0$ from simulation study, $\theta_1 = 1$.

		$P(\theta \neq 0)$			
		0-0.1	0.1-0.25	0.25-0.5	> 0.5
$\tau^2 = 1$	$\theta_g = 0$	65	15	2	4
	$\theta_g \neq 0$	5	0	3	34
$\tau^2 = 3$	$\theta_g = 0$	65	5	8	8
	$\theta_g \neq 0$	3	2	5	32
$\tau^2 = 10$	$\theta_g = 0$	73	5	4	4
	$\theta_g \neq 0$	8	4	2	28

Table 4.2: Posterior probability of $\theta \neq 0$ from simulation study, $\theta_1 = 3$.

		$P(\theta \neq 0)$			
		0-0.1	0.1-0.25	0.25-0.5	> 0.5
$\tau^2 = 1$	$\theta_g = 0$	81	5	0	0
	$\theta_g \neq 0$	2	0	3	37
$\tau^2 = 3$	$\theta_g = 0$	81	3	2	0
	$\theta_g \neq 0$	4	1	1	36
$\tau^2 = 10$	$\theta_g = 0$	84	1	1	0
	$\theta_g \neq 0$	2	1	2	37

Table 4.3: Posterior probability of $\theta \neq 0$ from simulation study, $\theta_1 = 5$.

		$P(\theta \neq 0)$			
		0-0.1	0.1-0.25	0.25-0.5	> 0.5
$\tau^2 = 1$	$\theta_g = 0$	86	0	0	0
	$\theta_g \neq 0$	2	0	1	39
$\tau^2 = 3$	$\theta_g = 0$	86	0	0	0
	$\theta_g \neq 0$	1	0	3	38
$\tau^2 = 10$	$\theta_g = 0$	86	0	0	0
	$\theta_g \neq 0$	1	1	2	38

The upper panel of Figure 4.1 shows the sample mean and the fitted mean function of the simulated CGH data, along with the true mean function. The lower panel shows the posterior probability of copy number gain/loss (solid line), as well as the posterior probability that $\theta_g \neq 0$ for each probe. We correctly identify all the regions with copy number aberrations, including the short one on the right hand side of the plot. We can also separate the probes that are associated with copy number changes (black points) and those that do not (circles). The 95% pointwise predictive interval (not shown) covers about 90% of the observed data. Notice that in the model the association is only detectable when there is a copy number aberration present. Since there is another layer of uncertainty about the location of the copy number changes, the problem becomes more difficult. However, those probes coming out on top despite the noise will likely be the most relevant, since both the copy number change levels and its association with the expression level have to be strong enough for θ_g to stay consistently in the model.

4.4.2 Breast cancer cell lines

Hyman *et al.* (2002) analyzed the gene expression and copy number of 14 breast cancer

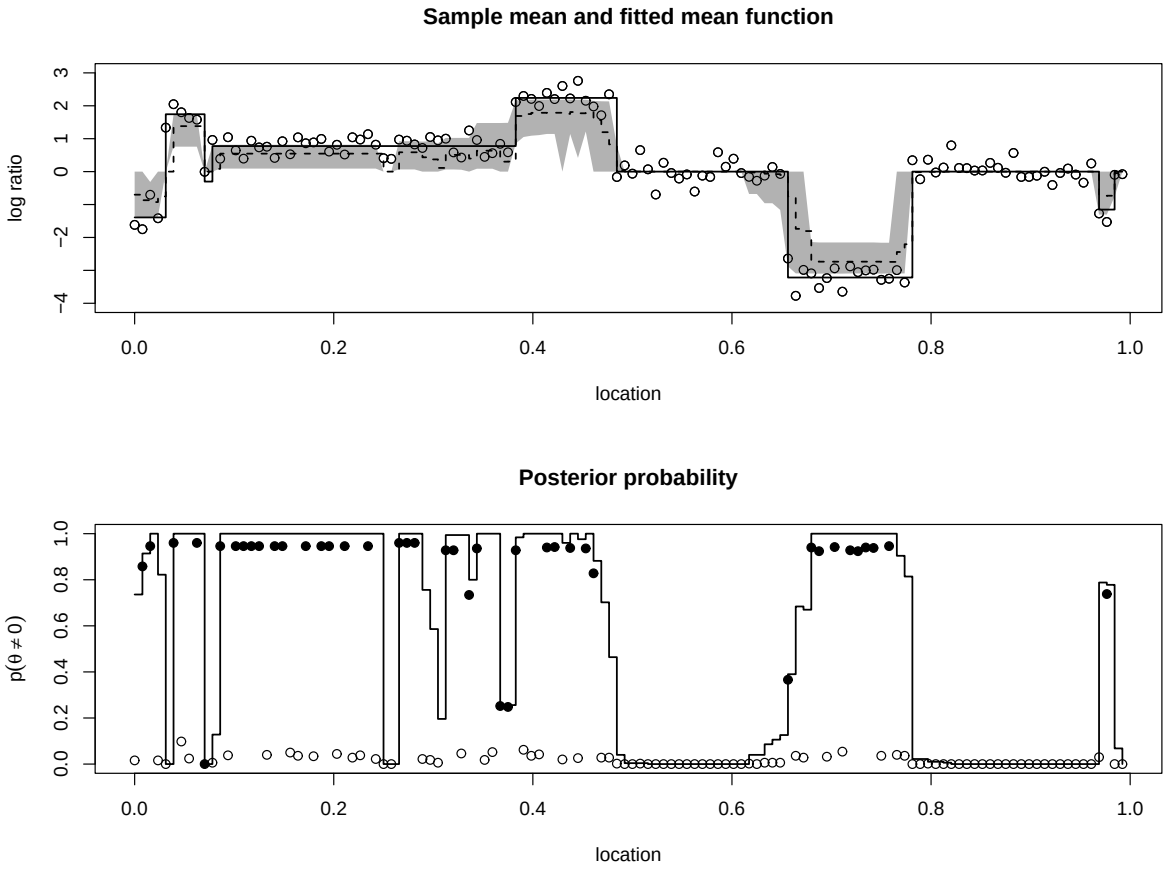

Figure 4.1: The upper panel shows the sample mean, the fitted mean function (dash line) and the true mean function (solid line) for a simulation run with $\theta_1 = 5$ and $\tau^2 = 10$. The lower panel shows the posterior probability of $\theta_g \neq 0$. The black dots are the genes of interest with true $\theta_g \neq 0$ and the circles are the genes with true $\theta_g = 0$. The solid line is the posterior probability of DNA copy amplification/deletion.

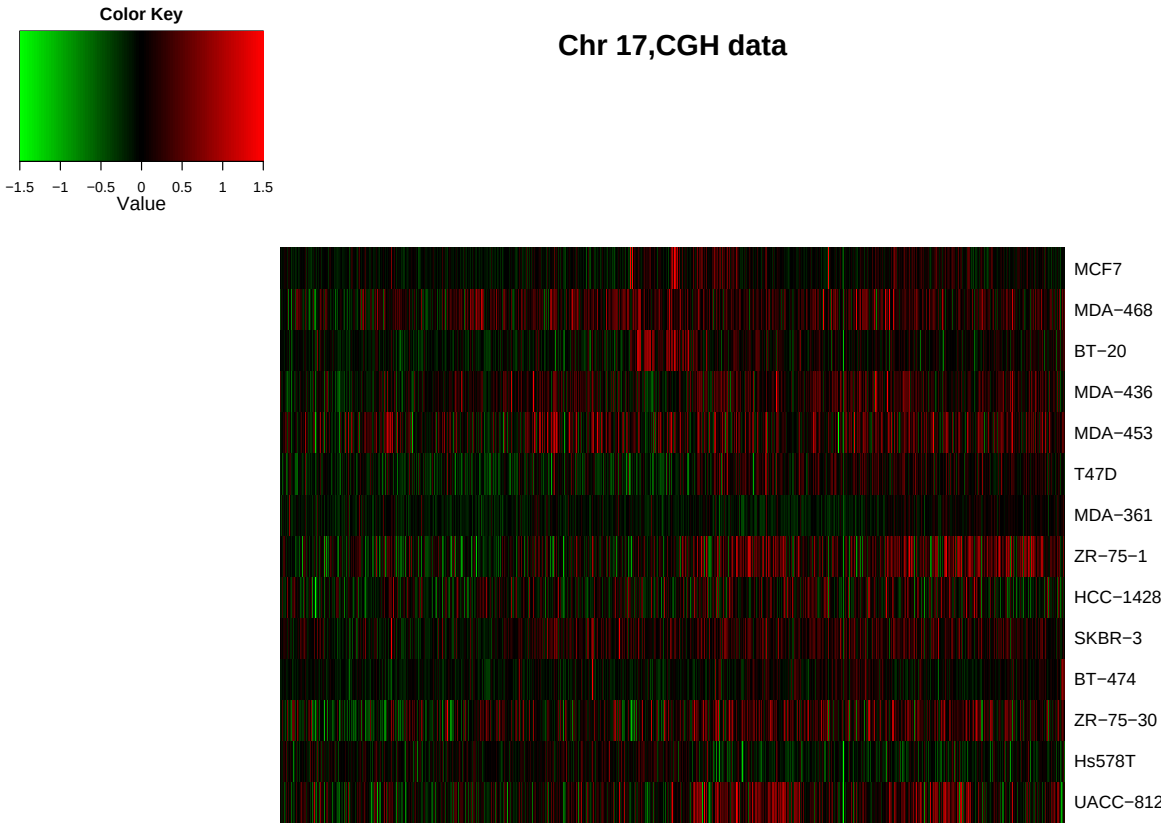

Figure 4.2: The heatmap illustrates the patterns of copy number ratios from chromosome 17 of 14 breast cancer cell lines in Hyman *et al.* (2002). Each gene occupies a grid and the the locations on the map do not correspond to the actual physical location of the genes.

cell lines with cDNA microarrays. Berger *et al.* (2006) used the same data set for their generalized singular value decomposition (GSVD) algorithm. After preprocessing, each cell line contains 11,994 CGH and cDNA observations. The cDNA and CGH copy number ratios have been $\log_2$ transformed and missing values were inferred prior to our analysis by Hyman *et al.* (2002). The cDNA and CGH data are observed at the same locations ($\{x_i\} = \{x_g\}$). Figure 4.2 and 4.3 show the pattern of copy number ratios and expression levels from the breast cancer cell line data.

For the breast cell line data set, the hyperparameter τ^2 and ν^2 are set to be

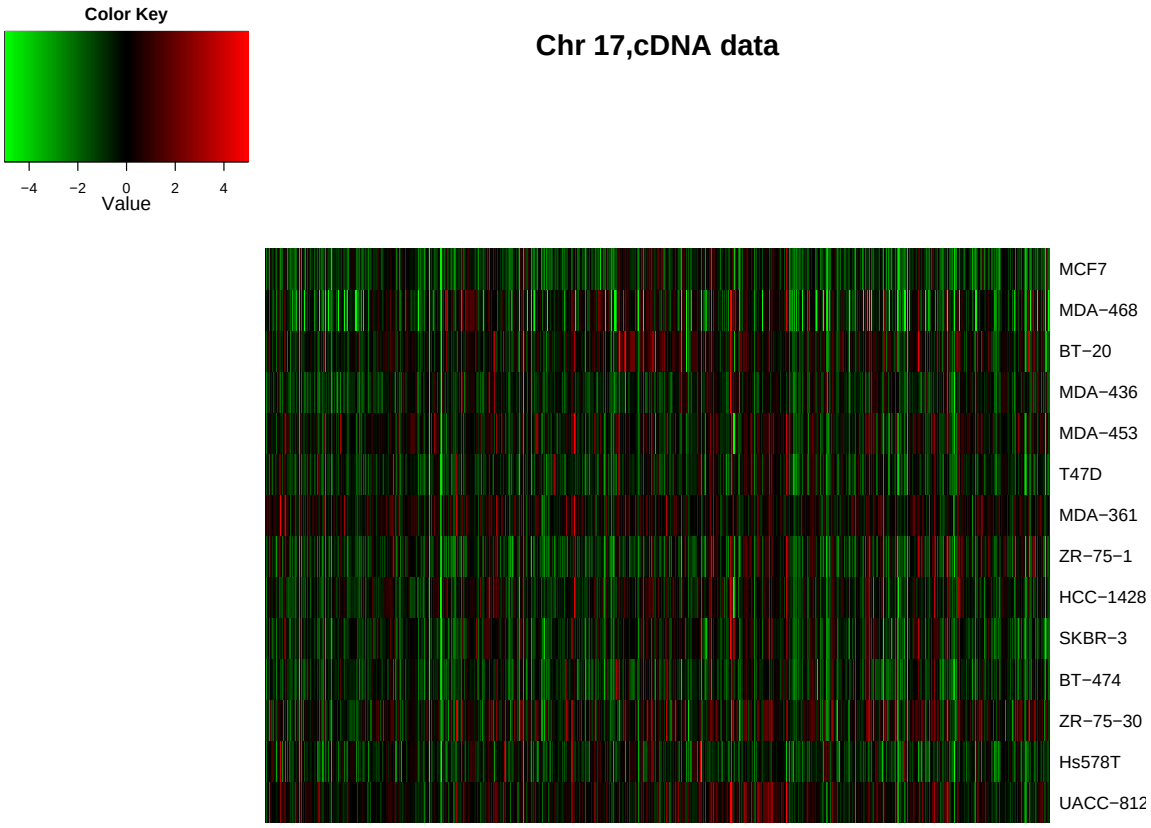

Figure 4.3: The heatmap illustrates the patterns of expression levels from chromosome 17 of 14 breast cancer cell lines in Hyman *et al.* (2002). Each gene occupies a grid and the the locations on the map do not correspond to the actual physical location of the genes.

$\tau^2 = 3, \nu^2 = 1$, so the marginal prior distribution for θ will cover a reasonable range implied from the exploratory analysis by Hyman *et al.* (2002) and Pollack *et al.* (2002). The prior probability of zero coefficients π is set at 0.56, which is the proportion of highly amplified genes that do not show overexpression in Hyman *et al.* (2002). We plot the posterior summary for chromosome 17 in Figure 4.4 and Figure 4.5. The upper panel of Figure 4.4 shows the posterior probability that $\theta_g \neq 0$ for each gene. We can clearly separate the genes whose expression levels are influenced by their copy numbers and the ones that are not. The ones that come out on top include ERBB2 gene (99.4%), which has been linked to the progression of some type of breast cancer. From Figure 4.2 we can see that three of the cell lines (SKBR3, BT474, UACC812) all have identifiable amplified regions around that gene, which makes it easier to detect the association between two data sets. Notice that in the upper panel of Figure 4.5 the sample mean for the CGH data shows an elevated region around 40M. It is driven by three cell lines out of 14 so our model does not consider it a common gain region across the patients. This indicates that the model is robust against outliers. We also found HOXB7 (96%), which is also clinically associated with cancer progression (Hyman *et al.*, 2002). There are other additional genes with high posterior probabilities of $\theta_g \neq 0$ that have not been previously studied. Table 4.4 lists posterior summaries of θ for some of the genes at 17q11-12 that are identified in Chin *et al.* (2006) as potential drug targets as they have high-level copy number amplifications that are associated with expression numbers. Most of them are confirmed in our study as having high posterior probabilities that $\theta \neq 0$, including the ones that have been functionally validated in literature (PPARPB, ERBB2, GRB7) and the ones considered druggable (PNMT, ERBB2, NR1D1). Though it is not clear if these genes play a role in breast cancer, we believe they are good candidates for future investigation.

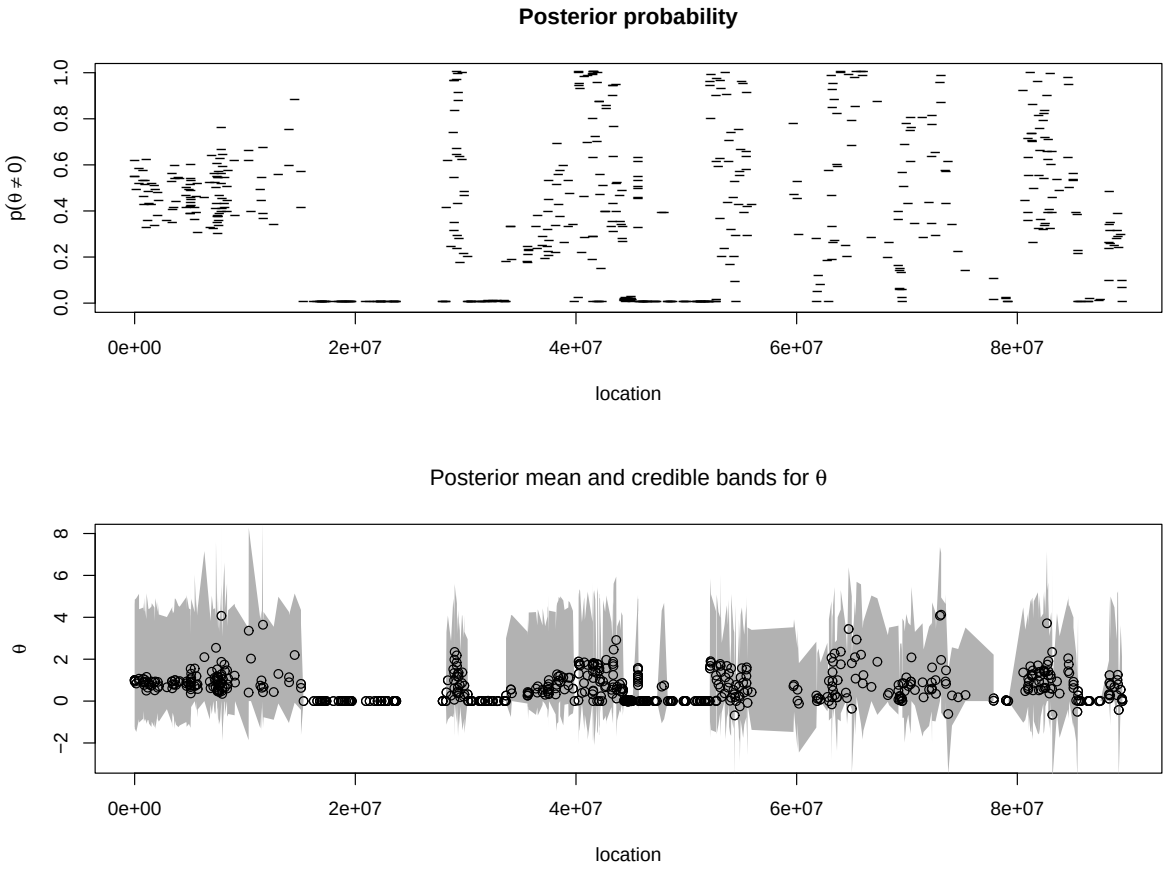

Figure 4.4: The upper panel shows the posterior probability of $\theta_g \neq 0$ for each gene in chromosome 17; the lower panel shows the posterior mean (points) and 95% Bayesian credible band for θ_g (gray area).

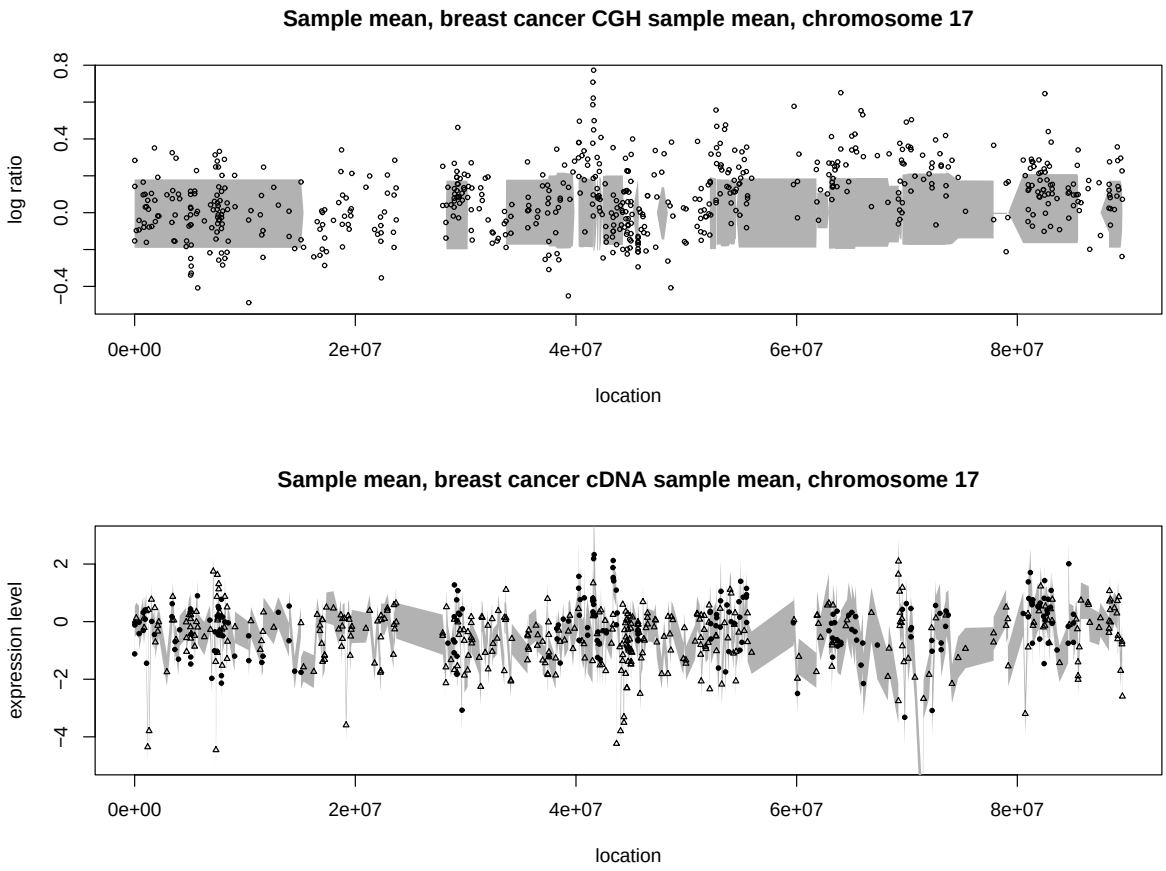

Figure 4.5: The sample mean from 14 breast cancer cell lines for CGH data (upper panel) and cDNA data (lower panel) from chromosome 17, plotted against the fitted mean function for all the MCMC draws. The black dots for the lower panel indicate the genes with over 50% posterior probability of $\theta_g \neq 0$. The expression levels of those genes are more likely to be associated with DNA copy number changes. The triangles represent the genes with less than 50% posterior probability of $\theta_g \neq 0$. The gray areas are the 95% Bayesian credible bands for the mean function of CGH data and expression data, respectively.

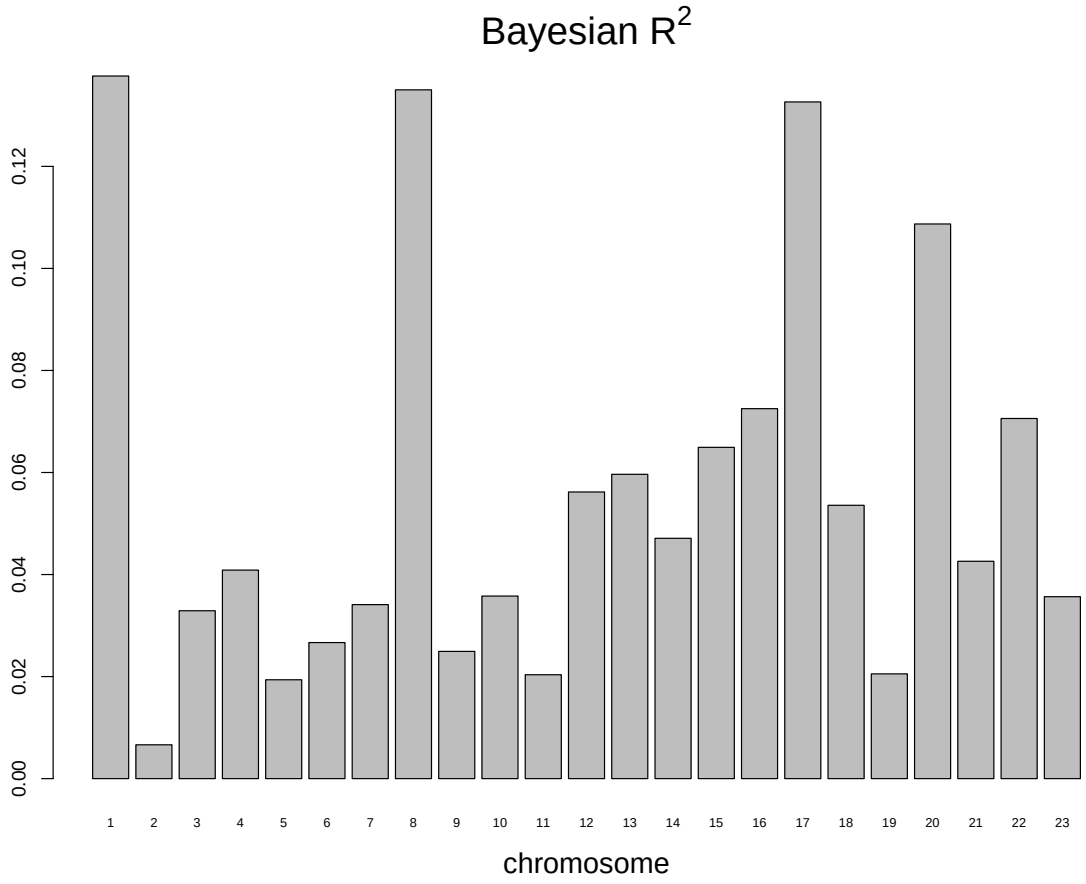

Figure 4.6: Bayesian R^2 for each chromosome based on posterior mean estimate. R^2 indicates the percentage of observed variation in expression level that can be directly explained by variation in copy numbers.

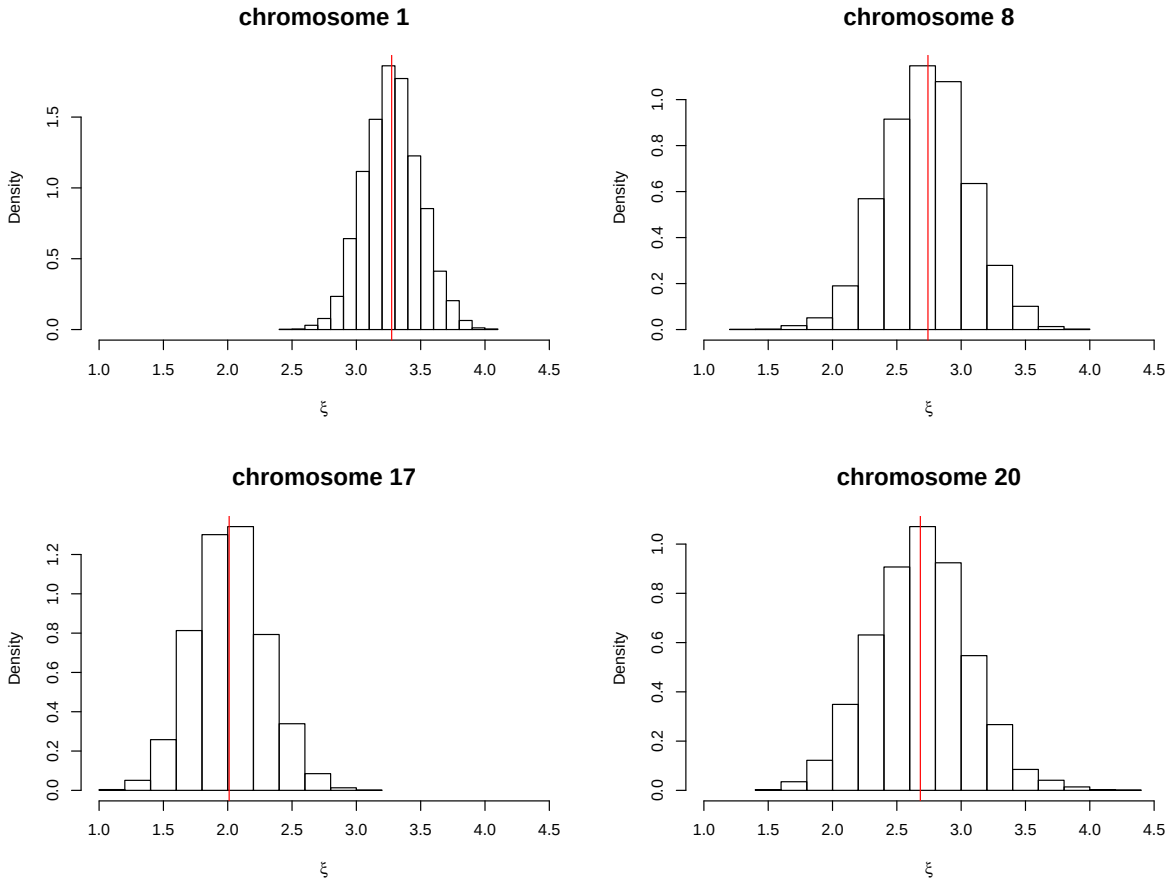

Figure 4.7: Histogram of samples from posterior distribution of ξ for selected chromosomes. The red lines indicate the empirical posterior median.

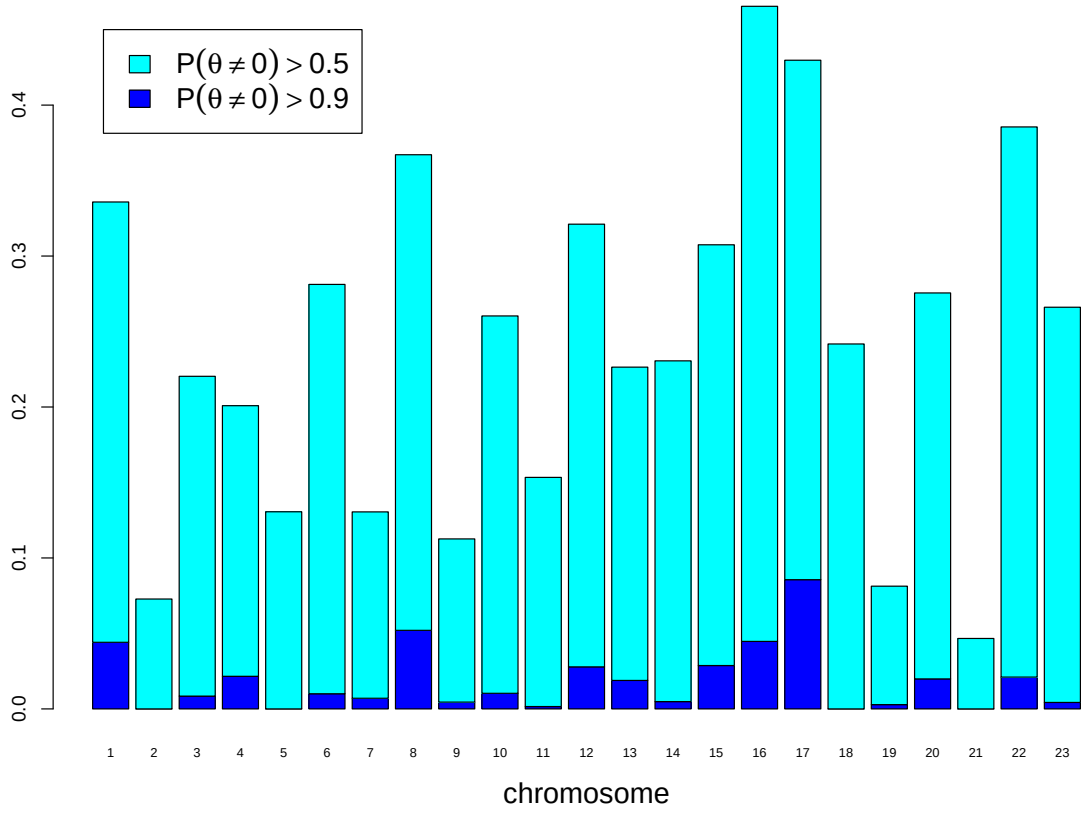

Figure 4.8: Percentages of genes showing high probability of association between expression levels and copy number.

The lower panel of Figure 4.4 shows the marginal posterior distribution of θ. Even though the prior is centered at zero the posterior is mostly supported on the positive side, as expected. Figure 4.5 shows the sample mean and fitted mean functions of the two data sets. Compared with Figure 4.4 we can clearly see the region with high probability of amplification/deletion and the region with no copy number change. The posterior probabilities for the no-change areas are all close to zero as we would not be able to estimate θ there. We calculate the Bayesian R^2 based on posterior mean estimates, which gives the percentage of variation in expression level directly explained by variation of copy numbers. The Bayesian R^2 has this following form

$$R^2 = 1 - \frac{\sum_{gj}(d_{gj} - \hat{\alpha}_g - \hat{\theta}_g \hat{f}_j(x_g))^2}{\sum_{gj}(d_{gj} - \hat{\alpha}_g)^2}, \tag{4.5}$$

where $\hat{\alpha}, \hat{\theta}_g$ and $\hat{f}$ are the posterior mean estimate for α, θ and f, respectively. From Figure 4.6 we can only see variation explained in the chromosomes showing amplicons in breast cancer cell lines as documented by Hyman *et al.* (2002), namely chromosome 1,8,17 and 20. For those chromosomes the R^2 are between 7 and 13 percents, which is consistent with the findings by Pollack *et al.* (2002). The association between two types of data at the population level ξ, appears to be positive (see Figure 4.7). At the individual level only 22% of the genes (2686 genes) have over 50% probability of $\theta \neq 0$ and only 2% (241 genes) have over 90% probability. Those genes are concentrated on the amplified regions, as expected (See Figure 4.8). The overall mean for θ is 0.725, also consistent with the results from Pollack *et al.* (2002), who found that 2-fold change in DNA copy number is associated with 1.4-1.5 fold change in expression level. As they pointed out, this is likely an underestimation of true impact of the DNA copy number on expression level, mostly due to the measurement error caused by the non-tumor cells contained in the tumor samples. There is also a lot of variability in θ, as the overall 95% credible interval for θ ranges from -0.39 to 4.79. If we only

include the genes with over 50% probability of $\theta \neq 0$, the overall mean for θ is about 1.5.

Table 4.4: Posterior summary of θ of selected genes with high-level copy number amplifications associated with gene expression levels as found in (Chin *et al.*, 2006). The last column indicates the genes for which drugs have been developed (Vogel *et al.*, 2002) (Trastuzumab; for ERBB2) or considered to be druggable (Russ and Lampel, 2005).

Gene	P($\theta \neq 0$)	Median of θ	95% Credible interval	Cancer function	Druggable
PSMB3	0.994	1.842	(-1.133,5.136)		
PIP5K2B	0.936	1.216	(-2.017,5.147)		
FLJ20291	0.946	1.683	(-1.381,5.339)		
PPARBP	0.98	1.776	(-1.841,5.227)	Zhu *et al.* (2000)	
TCAP	0.282	0	(-0.54,3.799)		
PNMT	0.624	0.292	(-1.577,4.74)		Yes
ERBB2	0.996	1.796	(-1.449,4.899)	Slamon *et al.* (1989)	Yes
GRB7	0.996	1.76	(-2.096,5.296)	Tanaka *et al.* (2000)	
PSMD3	0.718	1.068	(-1.379,4.836)		
NR1D1	0.696	0.767	(-1.403,5.453)		Yes

4.5 Discussion

Joint analysis of different type of microarray data can give us more accurate predictions and a more coherent view of the information from different sources of data. In this chapter, we take a model-based approach to integrate CGH copy number data with gene expression data in order to understand the relation of these two different genomic measurements. We believe that our model can adequately capture the complicated interaction between the two types of data. We propose a procedure based on a Bayesian hierarchical model for identifying regions with gains or losses of genomic DNA copy numbers and genes whose expression levels are associated with such copy

number changes.

From the simulation study, our method can correctly identify the regions with copy number gains or losses without having to choose an arbitrary threshold. Our method also has very strong performance picking out the genes whose expression levels are associated with copy numbers. Those genes may provide some insight on the mechanism behind such genomic abnormality and are good candidates for further research. Our analysis of the breast cancer cell lines successfully identifies most of the genes that have been clinically associated with cancer in the literature (Hyman *et al.*, 2002). At the same time it also gives an estimate of the underlying copy number of each gene and the level of uncertainty at each location from the Bayesian credible bands.

In conclusion, combining information from more than one kind of genomic measurement, such as CGH and gene expression levels, can lead to more accurate estimation and prediction. Our method provides an efficient way to jointly analyze the CGH and expression data to identify the patterns of the data. Hopefully it will lead to better understanding of the mechanism behind the cancer development and effective therapeutic measures.

Chapter 5

Discussion

5.1 Summary

In this thesis we have developed a Bayesian approach for function estimation in an overcomplete wavelet dictionary setting.

In Chapter 2 we described the overcomplete wavelet dictionary setting in a function estimation context. The model allows for arbitrary location and scale parameters instead of fixing them on a grid; it also allows the data to select the number of elements. The advantages of our methods include: 1. It can be applied to non-equally spaced data without interpolation or other kind of data manipulation. 2. It gives sparser representation and better performance in prediction. However, the MCMC algorithm requires evaluation of wavelet functions at arbitrary locations. Without the benefit of the cascade algorithm the wavelet function evaluation is more computationally intensive than it would be if we only evaluated it on an equally-spaced grid.

There is one exception. The Haar wavelet can be evaluated easily at any point. Therefore for data where a step function model is appropriate, we can model it with an overcomplete Haar wavelet dictionary without the heavy computation of the Daubechies-Lagarias algorithm. Therefore the CGH data for Chapter 3, which are typically modeled as a step function in the current available model-based methods, is a prefect candidate for our overcomplete Haar wavelet dictionary model. The dictionary has been slightly modified to improve computational efficiency. For example, the dictionary elements that represents segments of copy number alterations are mutually exclusive, instead of laying on top of each other as in Chapter 2. The elements

will still span the same space of step function but the mixing improves significantly. The hierarchical model for patient grouping is built on the CWD which helps us tackle important issues beyond copy number estimation, such as patient classification, prediction, and identifying the genes relevant to the group difference. We use simulation studies and an ovarian cancer data set to demonstrate the performance of our method.

In Chapter 4, we extended the model to incorporate gene expression data. Our goal is to detect the genes of which the expression level is related with its copy number change, as those genes are likely associated with cancer. Building on the model in Chapter 3, we conduct a joint analysis with both types of data and separate the potentially relevant genes from the rest. The results from the breast cancer data confirm the association of two genomic measurement in genes that have been linked to cancer development in the literature. Even though the examples in the chapter do not involve some of the model features discussed in Chapter 3, such as patient classification and prediction, our joint model does not preclude that type of analysis and can be easily extended to accommodate it if more patient information becomes available.

5.2 Future Work

Here we discuss possible improvements and extensions to my research.

5.2.1 Improving Computational Efficiency

First of all, we plan to make our code available for the public in the future as an R package. For that purpose, the speed of the RJ-MCMC algorithm needs to be improved. For example, we can port part of the sampling function for Chapter 3 and 4 from R to C, utilizing the C code for a single function estimation in Chapter

2. Even though we have bypassed the most time-consuming part of the program (Daubechies-Lagarias algorithm) for the CGH data analysis by modeling it as step function, the processing time of the Markov chain needs to be reduced. This has become especially important as higher-resolution CGH array becomes available.

In addition, the models for Chapter 3 and 4 can take advantage of parallel computing. Since the only parameter common to all chromosomes is σ^2, we can estimate the parameters for each chromosome in parallel on 23 nodes in a computer cluster, then collect all the fitted values to sample σ^2. This will increase the speed of our algorithm significantly.

5.2.2 Model Specification for CGH Data

Our model for CGH data in Chapter 3 assumes the same segmentation for all patients in the same group. Even though the assumption is reasonable and our model specification still accounts for the heterogeneity within the group, there are still ways to extend the model to be more flexible. For example, putting a mixture prior with some point mass at zero for β will shrink some coefficients to zero therefore allowing different segmentation and number of segments within a group. Though the focus of our method is to analyze the group as a whole, the features that only appear in some particular patients can be of interest as well. For that purpose, further extensions of model that introduce even more flexibility for segmentation is also possible, if computationally feasible.

5.2.3 Modeling Interval Data

One important assumption in our model in Chapter 4 is that the locations of the probes x are measured without error. In practice it is not always the case. Some CGH microarrays are from lower-resolution bacterial artificial chromosomes (BACs)

clones which are about 250KB long and spaced roughly at 1MB intervals, so that a number of genes may "fall under" a probe. There will also be genes near a probe, but not under it. Some of the probes cannot be accurately located. In addition, each of these probes span a certain interval, with some of the intervals overlapping. Potentially we could have multiple gene expression probes corresponding to one CGH probe, or vice versa. Our model can be extended to accommodate these misaligned arrays by redefining the endpoints of gain/loss segments **a** and **b**.

For example, let $z_i = [x_i^L, x_i^R]$ represent the CGH data and $z_g = [x_g^L, x_g^R]$ represent the expression data observed on intervals. We still have the same model as in (4.1) and (4.4)

$$c_{ij} = f_j(z_i) + \epsilon_{ij}$$

$$d_{gj} = \alpha_g + \theta_g f_j(z_g) + \epsilon_{gj}.$$

Now c_{ij} represents the CGH data observed at interval z_i and d_{gj} represents the expression data observed at interval z_g for patient j.

Define $A_c = \{x_i^L; \forall i\}$, the collection of left endpoints of all the CGH probes. Also define $A_d = \{x_g^L; x_g^L \notin z_i, \forall g, i\}$, the collection of left endpoints of all the expression probes that do not fall under any CGH probe. Similarly we define $B_c = \{x_i^R; \forall i\}$ and $B_d = \{x_g^R; x_g^R \notin z_i, \forall g, i\}$. In the RJ-MCMC algorithm, when we propose a new segment, the candidate for left endpoint a^* are drawn from $A = A_c \cup A_d$ and the candidate for left endpoint b^* are drawn from $B = B_c \cup B_d$ such that (1) $b^* > a^*$ and (2) $[a^*, b^*) \cup \{z_i\} \neq \emptyset \quad \forall i$. The second condition is added to make sure the segment $[a^*, b^*)$ covers at least one CGH probe. If the segment only covers expression probes it might not change the likelihood at all since the expression levels do not always depend on the copy number changes. We only need to check the condition if both endpoints are from expression probes, $a^* \in A_d$ and $b^* \in B_d$.

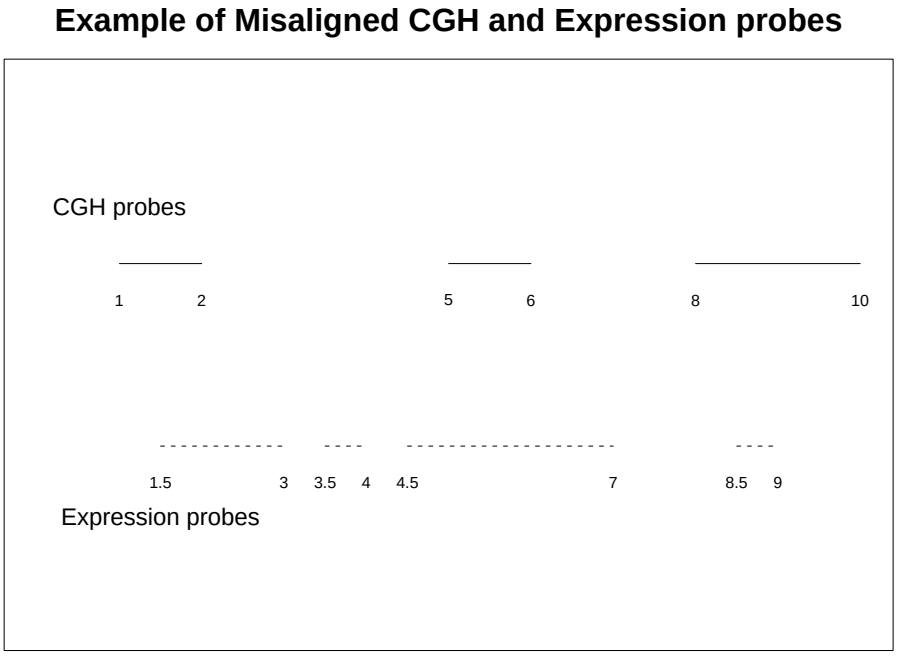

Figure 5.1: Illustration of misaligned intervals for CGH and expression probes. The solid lines are for CGH probes and the dashed lines are for expression probes.

For example, in Figure 5.1, we have $A_c = \{1, 5, 8\}$, $A_d = \{3.5, 4.5\}$, $B_c = \{2, 6, 10\}$, $B_d = \{3, 4, 7\}$. Any interval (a, b) with $a \in A$ and $b \in B$ will be a valid candidate for a new segment. For example, $[1, 3), [3.5, 6)$ and $[5, 10)$. However, intervals such as $[1.5, 3), [3.5, 4)$ and $[8.5, 9)$ will not be proposed because they do not contain a whole CGH probes or they break a CGH probe in half. Even though theoretically a breakpoint can occur anywhere, in practice if a breakpoint occurs in the middle of a CGH probe it will be very difficult to get a reliable estimate for the segments that cover that probe, as the expression data only contain weak information about the underlying copy number. Therefore, it is reasonable to assume that all points on one CGH probe correspond to the same underlying copy number. The assumption reduces the number of possible number of configurations and simplifies the computation

significantly.

The function $f_j(x)$ is modeled as the following

$$f_j(x) = \begin{cases} \beta_{jk} & \text{if } x \in [a_k, b_k) \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

In practice, the paired CGH and expression data are often misaligned and the data have usually been preprocessed and missing values inferred to make a full matrix. While the CGH probes and expression probes do not always correspond to the same set of genes, this kind of measurement error has not been considered. In the future we plan to extend our model to interval data and hopefully more accurate inference and prediction will result from it.

Appendix A

Appendix to Chapter 2

A.1 Reversible Jump MCMC

We follow the general framework by Green (1995) and Denison *et al.* (1998) and include three types of movement in the MCMC algorithm:

1. Birth step.(add a wavelet)
2. Death Step. (delete a wavelet)
3. Update Step. (move a wavelet)

As in Green (1995), the birth and death probabilities are chosen to be

$$p_b(K) = c\min\{1, p(K+1)/p(K)\},$$

$$p_d(K) = c\min\{1, p(K)/p(K+1)\},$$

where $c < 0.5$ is some constant. For the birth step, we propose to add a wavelet coefficient β_{K+1} with scale a_{K+1} and b_{K+1} from some joint proposal $q(\beta, a, b)$. Let $\hat{f}(x)$ be the mean estimate for the current model and $\tilde{f}(x)$ be the mean estimate for the proposed model, then the likelihood ratio:

$$LR = \frac{\mathsf{N}(\mathbf{Y}; \tilde{f}(x), \sigma^2 I)}{\mathsf{N}(\mathbf{Y}; \hat{f}(x), \sigma^2 I)}. \tag{A.1}$$

The acceptance ratio for the birth step is

$$LR \times \text{prior ratio} \times \text{proposal ratio} \times \text{Jacobian},$$

where the prior ratio is

$$\frac{p(K+1)p(\beta_{1:K+1}, a_{1:K+1}, b_{1:K+1})}{p(K)p(\beta_{1:K}, a_{1:K}, b_{1:K})}$$

$$= \frac{p(K+1)(K+1)!\prod_{k=1}^{K+1} p(\beta_k, a_k, b_k)}{p(K)K!\prod_{k=1}^{K} p(\beta_k, a_k, b_k)}$$

$$= \frac{p(K+1)(K+1)p(\beta_{K+1}, a_{K+1}, b_{K+1})}{p(K)},$$

and the proposal ratio is

$$\frac{p_d(K+1)q'(\beta_{K+1}, a_{K+1}, b_{K+1} \mid K+1)}{p_b(K)q(\beta_{K+1}, a_{K+1}, b_{K+1} \mid K)},$$

where $q'(\beta, a, b)$ is the proposal for the death step. In particular, $q'(\beta_{K+1}, a_{K+1}, b_{K+1} \mid K+1)$ is the probability of proposing to delete $\{\beta_{K+1}, a_{K+1}, b_{K+1}\}$ given that the current model has $K+1$ wavelets.

The Jacobian here is 1 since once we propose the added wavelet we will have one-to-one mapping to the new model space. And if we take Green's birth and death probabilities, those will cancel with the prior ratio and the acceptance rate will become:

$$AR = LR \times \frac{p(\beta_{K+1}, a_{K+1}, b_{K+1})(K+1)q'(\beta_{K+1}, a_{K+1}, b_{K+1} \mid K+1)}{q(\beta_{K+1}, a_{K+1}, b_{K+1} \mid K)}. \tag{A.2}$$

Notice that with normal prior for β as in (2.8), the full posterior for β_{K+1} is also normal

$$p(\beta_{K+1} \mid \beta_{1:K}, a_{1:(K+1)}, b_{1:(K+1)}, \mathbf{Y}) \sim \mathsf{N}(\hat{\beta}, \hat{\sigma^2}_\beta), \tag{A.3}$$

where

$$\hat{\sigma^2}_\beta = (\frac{1}{ca^{-\delta}} + \frac{\psi_{a_{K+1},b_{K+1}}{}'\psi_{a_{K+1},b_{K+1}}}{\sigma^2})^{-1},$$

and

$$\hat{\beta} = \frac{\hat{\sigma^2}_{\beta}}{\sigma^2}\psi_{a_{K+1},b_{K+1}}{}'(\mathbf{Y} - \hat{f}),$$

which we can take as a Gibbs-step like proposal, though it is not really a Gibbs step as it can still be rejected along with a and b, it can improve the acceptance rate. This normal proposal can also apply to heavy-tailed priors, with which the the posterior for β does not have a close-form.

Similarly, the acceptance rate for a death step is:

$$AR = LR \times \frac{q(\beta_k, a_k, b_k \mid K-1)}{p(\beta_k, a_k, b_k)Kq'(\beta_k, a_k, b_k \mid K)}, \tag{A.4}$$

where $q'(\beta_k, a_k, b_k \mid K)$ is given in (2.11).

In an update step, we randomly pick an index k from Unif$(1:K)$, and propose a scale a_k and location b_k from a random-walk proposal and propose the wavelet coefficient β_k from (A.3) so that

$$q(\tilde{\beta}_k, \tilde{a}_k, \tilde{b}_k) = p(\beta_k \mid \beta_{-k}, a_{1:K}, b_{1:K}, \mathbf{Y})\mathsf{N}([\tilde{a}_k, \tilde{b}_k]; [a_k, b_k], [\sigma_a^2, \sigma_b^2]^T I_2). \tag{A.5}$$

The second part cancels the reverse proposal so that the acceptance rate

$$AR = LR \times \frac{p(\tilde{\beta}_k, \tilde{a}_k, \tilde{b}_k)}{p(\beta_k, a_k, b_k)} \times \frac{p(\beta_k \mid \beta_{-k}, a_{1:K}, b_{1:K}, \mathbf{Y})}{p(\tilde{\beta}_k \mid \beta_{-k}, a_{1:K}, b_{1:K}, \mathbf{Y})}. \tag{A.6}$$

The reversible jump algorithm goes as follows:

1. Initially, select K_0 wavelet coefficients and scale and location parameters $\{\beta, a, b\}_0$.
2. Find the mean estimates $f(x|\{\beta, a, b\}_0)$.
3. Generate a uniform (0,1) random number u,
   (i) If $u < p_b(K)$, perform the birth step.

(ii) If $p_b(K) < u < p_b(K) + p_d(K)$, perform the death step.

(iii) If $u > p_b(K) + p_d(K)$, perform the update step.

4. Update σ^2 by Gibb Sampling:

$$\sigma^2_{\text{new}} \sim IG(n/2, 2/SSE)$$

where $SSE = \sum_{i=1}^{n}(Y_i - \hat{f}(x_i))^2$.

5. Repeat steps 2-4.

A.2 Proof of Detailed Balance

Define $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \mathbf{a}, \mathbf{b}, K\}$ as the set of parameters of interest in CWD and $\mathcal{C}$ as the parameter space, which is the union of parameter subspaces $\mathcal{C}_k$ with dimension $K = k$. When the current state is $\boldsymbol{\theta}$, we propose a type of move $m \in \{b, d, u\}$ which moves the state to $d\boldsymbol{\theta}'$ with a proposal probability $q_m(\boldsymbol{\theta}, d\boldsymbol{\theta}')$. Let $\alpha_m(\boldsymbol{\theta}, \boldsymbol{\theta}')$ be the acceptance rate for such move and $\pi(\boldsymbol{\theta} \mid \mathbf{Y})$ be the posterior distribution for $\boldsymbol{\theta}$ which is our target distribution. Then we can define α

$$\begin{aligned}
\alpha_m(\boldsymbol{\theta}, \boldsymbol{\theta}') &= \min\left\{1, \frac{\pi(d\boldsymbol{\theta}' \mid \mathbf{Y}) q_m(\boldsymbol{\theta}', d\boldsymbol{\theta})}{\pi(d\boldsymbol{\theta} \mid \mathbf{Y}) q_m(\boldsymbol{\theta}, d\boldsymbol{\theta}')}\right\} \\
&= \min\left\{1, \frac{p(d\boldsymbol{\theta}') L(d\boldsymbol{\theta}' \mid \mathbf{Y}) q_m(\boldsymbol{\theta}', d\boldsymbol{\theta})}{p(d\boldsymbol{\theta}) L(d\boldsymbol{\theta} \mid \mathbf{Y}) q_m(\boldsymbol{\theta}, d\boldsymbol{\theta}')}\right\}, \qquad \text{(A.7)}
\end{aligned}$$

where p and L are the prior distribution and likelihood for $\boldsymbol{\theta}$, respectively. Then we can write the transition kernel as

$$T(\boldsymbol{\theta}, \mathbf{B}) = \sum_m \int_{\mathbf{B}} q_m(\boldsymbol{\theta}, d\boldsymbol{\theta}') \alpha_m(\boldsymbol{\theta}, \boldsymbol{\theta}') + \sum_m \int_{\mathcal{C}} q_m(\boldsymbol{\theta}, d\boldsymbol{\theta}')(1 - \alpha_m(\boldsymbol{\theta}, \boldsymbol{\theta}')) I(\boldsymbol{\theta} \in \mathbf{B}) \qquad \text{(A.8)}$$

for any Borel set $\mathbf{B}$ in $\mathcal{C}$. The move from $\boldsymbol{\theta}$ into $\mathbf{B}$ results from either an acceptance of proposal to $\mathbf{B}$ or from a rejection from any proposal if $\boldsymbol{\theta}$ is already in $\mathbf{B}$.

Theorem A.1. *The MCMC algorithm described in Chapter 2 and Appendix A.1 satisfies the detailed balance condition and induces a reversible Markov chain with π as its invariant distribution.*

Proof: The detailed balance condition requires that for any measurable set $\mathbf{A}$ and $\mathbf{B}$ in $\mathcal{C}$:

$$\begin{aligned}
\int_{\mathbf{A}} T(\boldsymbol{\theta}, \mathbf{B})\pi(d\boldsymbol{\theta}) &= \int_{\mathbf{A}}\int_{\mathbf{B}} T(\boldsymbol{\theta}, d\boldsymbol{\theta}')\pi(d\boldsymbol{\theta}) \\
&= \int_{\mathbf{A}}\int_{\mathbf{B}} T(\boldsymbol{\theta}', d\boldsymbol{\theta})\pi(d\boldsymbol{\theta}') \\
&= \int_{\mathbf{B}}\int_{\mathbf{A}} T(\boldsymbol{\theta}', d\boldsymbol{\theta})\pi(d\boldsymbol{\theta}') \\
&= \int_{\mathbf{B}} T(\boldsymbol{\theta}, \mathbf{A})\pi(d\boldsymbol{\theta}). \qquad \text{(A.9)}
\end{aligned}$$

Plugging in (A.8), we get

$$\begin{aligned}
&\sum_m \int_{\mathbf{A}} \pi(d\boldsymbol{\theta}) \int_{\mathbf{B}} q_m(\boldsymbol{\theta}, d\boldsymbol{\theta}')\alpha_m(\boldsymbol{\theta}, \boldsymbol{\theta}') + \sum_m \int_{\mathbf{A}\cap\mathbf{B}} \pi(d\boldsymbol{\theta}) q_m(\boldsymbol{\theta}, d\boldsymbol{\theta}')(1 - \alpha_m(\boldsymbol{\theta}, \boldsymbol{\theta}')) \\
= &\sum_m \int_{\mathbf{B}} \pi(d\boldsymbol{\theta}') \int_{\mathbf{A}} q_m(\boldsymbol{\theta}', d\boldsymbol{\theta})\alpha_m(\boldsymbol{\theta}', \boldsymbol{\theta}) + \sum_m \int_{\mathbf{A}\cap\mathbf{B}} \pi(d\boldsymbol{\theta}') q_m(\boldsymbol{\theta}', d\boldsymbol{\theta})(1 - \alpha_m(\boldsymbol{\theta}', \boldsymbol{\theta})).
\end{aligned}$$

For this condition to hold it is sufficient that

$$\int_{\mathbf{A}} \pi(d\boldsymbol{\theta}) \int_{\mathbf{B}} q_m(\boldsymbol{\theta}, d\boldsymbol{\theta}')\alpha_m(\boldsymbol{\theta}, \boldsymbol{\theta}') = \int_{\mathbf{B}} \pi(d\boldsymbol{\theta}') \int_{\mathbf{A}} q_m(\boldsymbol{\theta}', d\boldsymbol{\theta})\alpha_m(\boldsymbol{\theta}', \boldsymbol{\theta}) \qquad \text{(A.10)}$$

for all $m, \mathbf{A}, \mathbf{B}$.

Without loss of generality, we consider only birth and death move here. The update move does not involve a change in dimensionality and it is straight-forward to show the detailed balance condition for this standard Metropolis-Hasting algorithm.

Consider the birth move from $\boldsymbol{\theta} \in \mathcal{C}_k$ to $\boldsymbol{\theta}' \in \mathcal{C}_{k+1}$. For any Borel set $\mathbf{A} \in \mathcal{C}_k$ and $\mathbf{B} \in \mathcal{C}_{k+1}$ we can define a probability measure ξ_b on $\mathcal{C} \times \mathcal{C}$

$$\begin{aligned}
\xi_b(\mathbf{A} \times \mathbf{B}) &= \xi_b(\mathbf{B} \times \mathbf{A}) \\
&= \xi_b((\mathbf{A} \cap \mathcal{C}_k) \times (\mathbf{B} \cap \mathcal{C}_{k+1})) + \xi_b((\mathbf{A} \cap \mathcal{C}_{k+1}) \times (\mathbf{B} \cap \mathcal{C}_k)) \\
&= P(\boldsymbol{\theta} : \boldsymbol{\theta} \in \mathbf{B}), \qquad \text{(A.11)}
\end{aligned}$$

where P is a $k+1$ dimensional probability measure on $\mathcal{C}_{k+1}$. It is easy to see that ξ_b is symmetric and a symmetric measure ξ_d for the death step can be defined similarly. Therefore, to show that

$$\begin{aligned}
\int_{\mathbf{A}} \pi(d\boldsymbol{\theta} \mid \mathbf{Y}) \int_{\mathbf{B}} q_m(\boldsymbol{\theta}, d\boldsymbol{\theta}') \alpha_m(\boldsymbol{\theta}, \boldsymbol{\theta}') &= \int_{\mathbf{A}} \int_{\mathbf{B}} \xi_m(d\boldsymbol{\theta}, d\boldsymbol{\theta}') \pi(\boldsymbol{\theta} \mid \mathbf{Y}) q_m(\boldsymbol{\theta}, \boldsymbol{\theta}') \alpha_m(\boldsymbol{\theta}, \boldsymbol{\theta}') \\
&= \int_{\mathbf{B}} \int_{\mathbf{A}} \xi_m(d\boldsymbol{\theta}', d\boldsymbol{\theta}) \pi(\boldsymbol{\theta}' \mid \mathbf{Y}) q_m(\boldsymbol{\theta}', \boldsymbol{\theta}) \alpha_m(\boldsymbol{\theta}', \boldsymbol{\theta}) \\
&= \int_{\mathbf{B}} \pi(d\boldsymbol{\theta}' \mid \mathbf{Y}) \int_{\mathbf{A}} q_m(\boldsymbol{\theta}', d\boldsymbol{\theta}) \alpha_m(\boldsymbol{\theta}', \boldsymbol{\theta}),
\end{aligned}$$

will only require

$$\pi(\boldsymbol{\theta} \mid \mathbf{Y}) q_m(\boldsymbol{\theta}, \boldsymbol{\theta}') \alpha_m(\boldsymbol{\theta}, \boldsymbol{\theta}') = \pi(\boldsymbol{\theta}' \mid \mathbf{Y}) q_m(\boldsymbol{\theta}', \boldsymbol{\theta}) \alpha_m(\boldsymbol{\theta}', \boldsymbol{\theta}), \qquad \text{(A.12)}$$

due to the symmetry of ξ_m. Substituting (A.7) it is clear that the condition in (A.12) is satisfied.

Since the parameter space $\mathcal{C}$ is connected, following Theorem 6.2.2 in Robert and Casella (1999, pg. 235) we can see that π is the invariant density of the reversible Markov chain.

Appendix B

Appendix to Chapter 3

This is an outline of the RJ-MCMC algorithm in Chapter 3.

- For $l = 1, \ldots, L$, sample $(\mathbf{a}, \mathbf{b}, \boldsymbol{\beta}, \boldsymbol{\mu} \mid \sigma^2, \sigma_l^2, \mathbf{Y})$ jointly.

    1. Sample $(\mathbf{a}, \mathbf{b} \mid \sigma^2, \sigma_l^2, \mathbf{Y})$ using Metropolis-Hasting step.

        The RJ-MCMC includes four kinds of step:

        1. birth step.
        2. death step.
        3. split step.
        4. merge step.

        For birth step, we propose two new location from $q_b(a_{sl}, b_{sl})$.

        For death step, we propose the element to kill from $q_d(h_{kl})$.

        For split step, we propose a new location from $q_s(c_s)$.

        For merge step, we propose two adjacent elements to kill from $q_m((h_{kl}, h_{(k+1)l}))$.

        p_b, p_d, p_s, p_m are the proposal probabilities for each kind of step.

        For the birth step where we propose (a_{kl}, b_{kl}), the acceptance rate for the birth step is

        $$\text{likelihood ratio} \times \text{prior ratio} \times \text{proposal ratio}.$$

        Here the likelihood ratio is based on the marginal likelihood $f(\mathbf{a}, \mathbf{b} | \mathbf{Y}, \sigma^2, \sigma_l^2)$, which has a closed form since $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ both have normal priors and we can

integrate them out. Notice that we only need to calculate the likelihood ratio on the probes that involve a change from the birth proposal.

The proposal ratio is

$$\frac{p'_d q'_d(h_{kl})}{p_b q_b(a_{kl}, b_{kl})},$$

and the prior ratio is

$$\frac{p(a_{kl}, b_{kl}, \mathbf{a}_l, \mathbf{b}_l)p(k = K_l + 1)}{p(\mathbf{a}_l, \mathbf{b}_l)p(k = K_l)}.$$

Similarly for the split step where we propose a new jump point c_{kl} between a_{kl} and b_{kl} we can calculate the marginal likelihood ratio over the probes affected by the split proposal.

The proposal ratio is

$$\frac{p'_m q'_m(h_{(k^-)l}, h_{(k^+l)})}{p_s q_s(c_{kl})},$$

and the prior ratio is

$$\frac{p(c_{kl}, \mathbf{a}_l, \mathbf{b}_l)p(k = K_l + 1)}{p(\mathbf{a}_l, \mathbf{b}_l)p(k = K_l)},$$

where

$$\frac{p(c_{kl}, \mathbf{a}_l, \mathbf{b}_l)}{p(\mathbf{a}_l, \mathbf{b}_l)} = p(c_{kl}|\mathbf{a}_l, \mathbf{b}_l) = \frac{\sum_{s=1}^{K} I(a_{sl} < c_{kl} < b_{sl})I(c_{kl} \in \{x_1, ...x_n\})}{\sum_{s=1}^{K_l}\sum_{i=1}^{n} I(a_{sl} < x_i < b_{sl})}.$$

The death and merge steps are the reverse of the birth and split steps, respectively.

2. Sample $(\boldsymbol{\mu} \mid \mathbf{a}, \mathbf{b}, \mathbf{Y}, \sigma^2, \sigma_l^2)$ from the posterior

$$f(\mu_{kl}|\mathbf{a}, \mathbf{b}, \mathbf{Y}, \sigma^2, \sigma_l^2) \sim \mathsf{N}(\hat{\mu}_{kl}, \hat{\sigma}_{kl}^2), \tag{B.1}$$

where

$$\hat{\mu}_{kl} = \frac{J_l \bar{\mathbf{Y}}^k_{\cdot l} c(b_{kl} - a_{kl})^\delta}{\frac{\sigma^2}{n_{kl}} + \sigma_l^2 + J_l c(b_{kl} - a_{kl})^\delta},$$

and

$$\hat{\sigma}^2_{kl} = \frac{(\frac{\sigma^2}{n_{kl}} + \sigma_l^2) J_l c(b_{kl} - a_{kl})^\delta}{\frac{\sigma^2}{n_{kl}} + \sigma_l^2 + J_l c(b_{kl} - a_{kl})^\delta}.$$

Let $\bar{\mathbf{Y}}^k_{jl}$ be the average of y_{ijl} between a_{kl} and b_{kl}

$$\bar{\mathbf{Y}}^k_{jl} = \frac{\sum_{i=1}^n y_{ijl} I(a_{kl} \le x_i < b_{kl})}{\sum_{i=1}^n I(a_{kl} \le x_i < b_{kl})} = \frac{\sum_{i=1}^n y_{ijl} I(a_{kl} \le x_i < b_{kl})}{n_{kl}}.$$

$\bar{\mathbf{Y}}^k_{\cdot l}$ be the average of $\bar{\mathbf{Y}}^k_{jl}$ for $j = 1, \ldots, J_l$ and $n_{kl} = \sum_{i=1}^n I(a_{kl} \le x_i < b_{kl})$, be the number of probes between a_{kl} and b_{kl}.

3. Sample $(\boldsymbol{\beta} \mid \boldsymbol{\mu}, \mathbf{a}, \mathbf{b}, \mathbf{Y}, \sigma^2, \sigma_l^2)$ from its posterior

$$\beta_{jkl} | \mathbf{Y}, \sigma^2, \sigma_l^2, \mathbf{a}_l, \mathbf{b}_l, \mu_{kl} \sim \mathsf{N}\left(\frac{\bar{\mathbf{Y}}^k_{jl} \sigma_l^2 + \mu_{kl} \sigma^2 / n_{kl}}{\sigma_l^2 + \sigma^2 / n_{kl}}, \frac{\sigma^2 \sigma_l^2}{\sigma_l^2 n_{kl} + \sigma^2} \right). \quad \text{(B.2)}$$

- For $l = 1, \ldots, L$, sample $(\sigma_l^2 \mid \boldsymbol{\beta}, \boldsymbol{\mu})$.

$$\sigma_l^2 \sim IG\left(\alpha + \frac{J_l * K_l}{2}, \gamma + \frac{\sum_{j=1}^{J_l} \sum_{k=1}^{K_l} (\beta_{jkl} - \mu_{kl})}{2} \right).$$

- Sample $(\sigma^2 \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\beta}, Y)$

$$\sigma^2 \sim IG(n/2, SSE/2),$$

where $SSE = \sum_l \sum_{j=1}^{J_l} \sum_{i=1}^n (Y_{ijl} - \hat{f}_{jl}(x_i))^2$.

The proposal distribution in RJ-MCMC algorithm is very important since it often suffers from slow convergence if the proposal rarely leads to parameter values with high likelihood. Here our proposal is based on current configurations and residuals, which can improve convergence in practice.

In Green (1995), the birth and death probabilities are chosen to be

$$p_b(K) = c \min\{1, p(K+1)/p(K)\},$$

$$p_d(K) = c \min\{1, p(K)/p(K+1)\},$$

where $c \leq 0.5$ is some constant.

Therefore we can let

$$p_b + p_s = c \min\{1, p(K+1)/p(K)\},$$

$$p_d + p_m = c \min\{1, p(K)/p(K+1)\},$$

for some constant $c = 1/(p_b + p_d + p_s + p_m)$.

For the birth and split step, we increase the dimension by 1, so we can consider these two kind of steps together. Same for the death and merge step. First we pick a "seed" point z with probability proportional to the magnitude of the current residual. The reason is that proposing a birth at the region with large residuals can potentially reduce the residual the most, therefore increase the likelihood the most. If the seed point has been occupied ($\sum_{k=1}^{K_l} I(a_{kl} \leq z < b_{kl}) = 1$), we propose a split step. If it's not been occupied ($\sum_{k=1}^{K_l} I(a_{kl} \leq z < b_{kl}) = 0$), we propose a birth step.

Therefore we have

$$p_b = \min\{1, p(K_l+1)/p(K_l)\} \left(1 - \frac{\sum_{i=1}^{n} \sum_{k=1}^{K_l} v_{il} I(a_{kl} \leq x_i < b_{kl})}{\sum_{i=1}^{n} v_{il}}\right),$$

and

$$p_s = \min\{1, p(K_l+1)/p(K_l)\} \left(\frac{\sum_{i=1}^{n} \sum_{k=1}^{K_l} v_{il} I(a_{kl} \le x_i < b_{kl})}{\sum_{i=1}^{n} v_{il}} \right),$$

where v_{il} is the magnitude of the current residuals. Since we do not propose μ and β at step 1, the proposal for (a_{kl}, b_{kl}) cannot depend on μ and β. So we define the current "residual" as follows:

$$v_{il} = \begin{cases} \sum_{j=1}^{J_l} |Y_{ijl} - \bar{\mathbf{Y}}_{jl}^k| & \text{if} \quad Y_{ijl} \in \mathbf{Y}_l^k \\ \sum_{j=1}^{J_l} |Y_{ijl}| & \text{if} \quad Y_{ijl} \in \mathbf{Y}_0. \end{cases}$$

If it's a birth step, define the proposal probability of the "seed" location z:

$$q_z^b(x_i) = \frac{v_{il}}{\sum_{s=1}^{n} v_{sl}(1 - I(a_{kl} \le x_s < b_{kl}))},$$

and pick another point z' uniformly from $[b_{(k-1)l}, a_{kl})$. Then our birth proposal will be (z, z') or (z', z), depending on which one is larger.

So the birth proposal

$$q_b(a_{kl}, b_{kl}) = \frac{q_z^b(a_{kl}) + q_z^b(b_{kl})}{\sum_{i=1}^{n} I(b_{(k-1)l} \le x_i < a_{(k+1)l}) - 1}.$$

Similarly, for the split step,

$$q_s(c_{kl} = x_i) = \frac{v_{il}}{\sum_{s=1}^{n} v_{sl}(I(a_{sl} \le x_i < b_{sl}))}.$$

The proposal of the death or merge step is inversely proportional to the difference of the coefficients between two adjacent regions. If either coefficient is zero, we propose a death step and kill the other one. If both coefficients are non-zero, we propose a merge step of these two elements. For the death step, we want to kill

the one with small coefficient, so we do not remove the significant ones. Similarly, it makes more sense to merge two adjacent segments with coefficients close to each other, which do not change the likelihood dramatically therefore more likely to get accepted.

First we label the jump points $\mathbf{c}_l = \{c_{0l}, ...c_{Ml}\}$. We define $c_{0l} = 0$, $c_{Ml} = 1$, $c_{ml} < c_{(m+1)l}$ for all m and $\mathbf{c}_l = \{c_{0l}, c_{Ml}, \mathbf{a}_l, \mathbf{b}_l\}$. And let $\theta_{1l}, ...\theta_{Ml}$ be the coefficient for the regions separated by the jump points. Then define $\hat{\theta}$ as the mean of the data at each segments, which is the MLE of the coefficients given a set of $\mathbf{a}, \mathbf{b}$, since the proposal cannot be dependent on the actual coefficients:

$$\hat{\theta}_{ml} = \begin{cases} \bar{\mathbf{Y}}_{\cdot l}^{m} & \text{if} \quad c_{(m-1)l} = a_{kl}, c_{ml} = b_{kl} \quad \text{for some k} \\ 0 & \text{otherwise.} \end{cases}$$

Therefore

$$p_d = \min\{1, p(K_l)/p(K_l+1)\} \frac{\sum_{m=1}^{M-1} 1/|\hat{\theta}_{(m+1)l} - \hat{\theta}_{ml}| I(\hat{\theta}_{ml}\hat{\theta}_{(m+1)l} = 0)}{\sum_{m=1}^{M-1} 1/|\hat{\theta}_{(m+1)l} - \hat{\theta}_{ml}|},$$

$$p_m = \min\{1, p(K_l)/p(K_l+1)\} \left(\frac{\sum_{m=1}^{M-1} (1/|\hat{\theta}_{(m+1)l} - \hat{\theta}_{ml}|) I(\hat{\theta}_{ml}\hat{\theta}_{(m+1)l} \neq 0)}{\sum_{m=1}^{M-1} (1/|\hat{\theta}_{(m+1)l} - \hat{\theta}_{ml}|)} \right),$$

and

$$q_d(h_{kl}) = \frac{1/|\bar{\mathbf{Y}}_{\cdot l}^{k}| (I(a_{kl} \neq b_{(k-1)l}) + I(b_{kl} \neq a_{(k+1)l}))}{\sum_{m=1}^{M-1} (1/|\hat{\theta}_{(m+1)l} - \hat{\theta}_{ml}|) I(\hat{\theta}_{ml}\hat{\theta}_{(m+1)l} = 0)},$$

$$q_m(h_{kl}, h_{(k+1)l}) = \frac{(1/|\bar{\mathbf{Y}}_{\cdot l}^{k} - \bar{\mathbf{Y}}_{\cdot l}^{k+1}|) I(b_{kl} = a_{(k+1)l})}{\sum_{m=1}^{M-1} (1/|\hat{\theta}_{(m+)l1} - \hat{\theta}_{ml}|) I(\hat{\theta}_{ml}\hat{\theta}_{(m+1)l} \neq 0)}.$$

Appendix C

Appendix to Chapter 4

This is an outline for the Reversible Jump Markov chain Monte Carlo algorithm used in Chapter 4 for sampling from the posterior distribution of the parameters. The following algorithm will apply to the simulation study and the breast cancer data in the paper, where the CGH and expression data are observed at the same locations and there is no missing data. The algorithm can be easily extended to accommodate missing and/or misaligned data.

1. Sample $(\mathbf{a}, \mathbf{b}, \boldsymbol{\beta}, \boldsymbol{\mu} \mid \sigma^2, \sigma_l^2, \sigma_d^2, \boldsymbol{\theta}, \boldsymbol{\alpha}, \mathbf{C}, \mathbf{D})$ using the Metropolis-Hasting step.

   The RJ-MCMC includes four kinds of step: birth, death, split and merge. The probability of proposing each type of steps are p_b, p_d, p_s and p_m, respectively. At each step, we generate a uniform $(0, p_b + p_s + p_d + p_m)$ random number u,

   (i) If $u < p_b$, perform the birth step.

   (ii) If $p_b < u < p_b + p_s$, perform the split step.

   (iii) If $p_b + p_s < u < p_b + p_s + p_d$, perform the death step.

   (iv) If $p_b + p_s + p_d < u$, perform the merge step.

   For birth step, we propose two new location from $q_b(a_k, b_k)$ and new coefficients $\mu_k^*, \beta_k^* = \{\beta_{jk}^*; j = 1, ...J\}$.
   For death step, we propose the element to kill from $q_d(h_k)$.
   For split step, we propose a new location from $q_s(c_k)$ and new coefficients $\mu_k^1, \mu_k^2, \beta_k^1 = \{\beta_{jk}^1; j = 1, ...J\}, \beta_k^2 = \{\beta_{jk}^2; j = 1, ...J\}$ for the newly split segments.

For merge step, we propose two adjacent elements to kill from $q_m((h_{kl}, h_{(k+1)l}))$. The proposal q_b, q_d, q_s and q_m and the proposal probabilities p_b, p_d, p_s and p_m are the same from Appendix B. The proposal for μ and β for the birth and split steps are from those following proposals:

$$q(\beta_{jk}^*) \sim \mathsf{N}(\hat{\beta}_{jk}, \hat{\sigma}_{jk}^2), \tag{C.1}$$

where

$$\hat{\sigma}_{jk}^2 = \frac{\sigma^2\sigma_d^2\sigma_l^2}{n_k^c\sigma_l^2\sigma_d^2 + \sum_{d_{gj}\in\mathbf{D}_j^k,\theta_g\neq 0}\theta_g^2\sigma_l^2\sigma^2 + \sigma^2\sigma_d^2},$$

$$\hat{\beta}_{jk} = \frac{\sum_{c_{ij}\in\mathbf{C}_j^k} c_{ij}\sigma_l^2\sigma_d^2 + \sum_{d_{gj}\in\mathbf{D}_j^k,\theta_g\neq 0}(d_{gj}-\alpha_g)\theta_g\sigma_l^2\sigma^2 + \bar{\mathbf{C}}_j^k\sigma^2\sigma_d^2}{n_k^c\sigma_l^2\sigma_d^2 + \sum_{d_{gj}\in\mathbf{D}_j^k,\theta_g\neq 0}\theta_g^2\sigma_l^2\sigma^2 + \sigma^2\sigma_d^2},$$

$$q(\mu_k^* \mid \beta_{jk}^*) \sim \mathsf{N}\left(\frac{\sum_{j=1}^J \beta_{jk}^* c(b_k-a_k)^\delta}{Jc(b_k-a_k)^\delta + \sigma_l^2}, \frac{c(b_k-a_k)^\delta\sigma_l^2}{Jc(b_k-a_k)^\delta + \sigma_l^2}\right), \tag{C.2}$$

which is based on their posterior distribution, except we use the sample mean $\bar{\mathbf{C}}_j^k$ instead of μ_k when it's not available.

The acceptance rate for the birth step is

$$\text{likelihood ratio} \times \text{prior ratio} \times \text{proposal ratio}.$$

The likelihood ratio for the birth step where we propose a new segment k and coefficients β_{jk} for all j:

$$LR = \frac{\prod_{c_{ij}\in\mathbf{C}_j^k}\mathsf{N}(c_{ij};\beta_{jk}^*,\sigma^2)\prod_{d_{ij}\in\mathbf{D}_j^k}\mathsf{N}(d_{gj};\alpha_g+\theta_g\beta_{jk}^*,\sigma_d^2)}{\prod_{c_{ij}\in\mathbf{C}_j^k}\mathsf{N}(c_{ij};0,\sigma^2)\prod_{d_{ij}\in\mathbf{D}_j^k}\mathsf{N}(d_{gj};\alpha_g,\sigma_d^2)}$$

and the proposal ratio is

$$\frac{p'_d q'_d(h_k)}{p_b q_b(a_k, b_k) q(\mu_k^*, \beta_k^*)},$$

where $p'_d q'_d(h_k)$ is the probability of reverse proposal, i.e., the death proposal in which we kill the segment h_k that spans from a_k to b_k.

The prior ratio is

$$\frac{p(a_k, b_k, \mathbf{a}, \mathbf{b}) p(k = K+1) p(\mu_k^*, \beta_k^*)}{p(\mathbf{a}, \mathbf{b}) p(k = K)}.$$

For the split step where we propose a new jump point c_k between a_k and b_k the likelihood ratio is

$$LR = \frac{\prod_{c_{ij} \in \mathbf{C}_j^k} \mathsf{N}(c_{ij}; \beta_{jk}^{new}, \sigma^2) \prod_{d_{ij} \in \mathbf{D}_j^k} \mathsf{N}(d_{gj}; \alpha_g + \theta_g \beta_{jk}^{new}, \sigma_d^2)}{\prod_{c_{ij} \in \mathbf{C}_j^k} \mathsf{N}(c_{ij}; \beta_{jk}, \sigma^2) \prod_{d_{ij} \in \mathbf{D}_j^k} \mathsf{N}(d_{gj}; \alpha_g + \theta_g \beta_{jk}, \sigma_d^2)},$$

with

$$\beta_{jk}^{new} = \begin{cases} \beta_{jk}^1 & \text{if} \quad a_k \le x_i < c_k \\ \beta_{jk}^2 & \text{if} \quad c_k \le x_i < b_k, \end{cases}$$

where $\beta_{jk}^1, \beta_{jk}^2$ are the proposed coefficients for the split segments.

The proposal ratio is

$$\frac{p'_m q'_m(h_{(k^-)}, h_{(k^+)})}{p_s q_s(c_{kl}) q(\mu_k^1, \mu_k^2, \beta_k^1, \beta_k^2)}.$$

where $p'_m q'_m(h_{(k^-)}, h_{(k^+)})$ is the probability of the reverse proposal. In the reverse proposal we propose a merge of two adjacent segments $h_{(k^-)} = (a_k, c_k)$ and $h_{(k^+)} = (c_k, b_k)$.

The prior ratio is

$$\frac{p(c_k, \mathbf{a}, \mathbf{b})p(k = K + 1)p(\mu_k^1, \mu_k^2, \beta_k^1, \beta_k^2)}{p(\mathbf{a}, \mathbf{b})p(k = K)},$$

where

$$\frac{p(c_k, \mathbf{a}, \mathbf{b})}{p(\mathbf{a}, \mathbf{b})} = p(c_k|\mathbf{a}, \mathbf{b}) = \frac{\sum_{s=1}^{K} I(a_{sl} < c_k < b_s)I(c_k \in \{x_1, ...x_n\})}{\sum_{s=1}^{K}\sum_{i=1}^{n} I(a_s < x_i < b_s)}.$$

The death and merge steps are the reverse of the birth and split steps, respectively. And the likelihood ratio and the acceptance rate for those steps can be derived similarly.

2. Sample $(\boldsymbol{\mu} \mid \boldsymbol{\beta}, \mathbf{a}, \mathbf{b}, \mathbf{C}, \mathbf{D}, \sigma^2, \sigma_l^2, \sigma_d^2, \boldsymbol{\alpha}, \boldsymbol{\theta})$ from the posterior

   The posterior has the same form as equation (C.2). Though we already update μ and β in the above Metropolis-Hasting step, we only change them one at a time and that will result in slow mixing. Therefore, we might want to update the coefficients of the unchanged parts once in awhile. For example, we can update all the β and μ's every ten steps.

3. Sample $(\boldsymbol{\beta} \mid \boldsymbol{\mu}, \mathbf{a}, \mathbf{b}, \mathbf{C}, \mathbf{D}, \sigma^2, \sigma_l^2, \sigma_d^2, \boldsymbol{\alpha}, \boldsymbol{\theta})$ from the posterior

   The posterior for $\boldsymbol{\beta}$ has the same form as in equation (C.1) except for $\hat{\beta}_{jk}$ which is changed to:

$$\hat{\beta}_{jk} = \frac{\sum_{c_{ij} \in \mathbf{C}_j^k} c_{ij}\sigma_l^2\sigma_d^2 + \sum_{d_{gj} \in \mathbf{D}_j^k, \theta_g \neq 0}(d_{gj} - \alpha_g)\theta_g\sigma_l^2\sigma^2 + \mu_k\sigma^2\sigma_d^2}{n_k^c\sigma_l^2\sigma_d^2 + \sum_{d_{gj} \in \mathbf{D}_j^k, \theta_g \neq 0}\theta_g^2\sigma_l^2\sigma^2 + \sigma^2\sigma_d^2}$$

   Notice that if $\theta_g = 0$, then that point g does not enter the likelihood at all.

4. Sample $\boldsymbol{\alpha}$

Define an index z_g

$$z_g = \begin{cases} 1 & a_k \leq x_g < b_k \quad \text{for some k} \\ 0 & \text{otherwise.} \end{cases}$$

Assuming a flat prior on $\boldsymbol{\alpha}$, then

$$f(\alpha_g \mid \mathbf{D}, \mathbf{a}, \mathbf{b}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_d^2)$$

$$\begin{cases} \propto \mathsf{N}(\sum_j \frac{d_{gj} - \theta_g \beta_{jk}}{J}, \frac{\sigma_d^2}{J}) & \text{if} \quad z_g = 1 \\ \propto \mathsf{N}(\sum_j \frac{d_{gj}}{J}, \frac{\sigma_d^2}{J}) & \text{if} \quad z_g = 0. \end{cases}$$

5. Sample $\boldsymbol{\theta}$

Introduce a new variable γ where

$$\gamma_g = \begin{cases} 0 & \text{if} \quad \theta_g = 0 \\ 1 & \text{if} \quad \theta_g \neq 0. \end{cases}$$

Notice that if x_g is not covered by a segment, then θ_g does not enter the likelihood at all and we can not make any inference about it. Therefore we fix θ_g at zero in this situation. If $z_g = 1$, then θ_g has a mixture prior of a point mass at 0 and some distribution $\pi(\theta_g)$ centered at ξ, and ξ also has a normal prior distribution center at zero. Let q be the prior probability of $\gamma_g = 0$ and $\pi(\theta_g \mid \gamma_g = 1) \sim \mathsf{N}(\xi, \tau^2)$, $\pi(\xi) \sim \mathsf{N}(0, \nu^2)$ then

$$\begin{aligned} & p(\gamma_g = 0 \mid \mathbf{D}, \mathbf{a}, \mathbf{b}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_d^2, \xi) \\ = & \int p(\gamma_g = 0 \mid \mathbf{D}, \mathbf{a}, \mathbf{b}, \theta_g, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_d^2) \pi(\theta_g) d\theta_g \\ = & \begin{cases} \frac{q \prod_{jl, z_g = 1} \mathsf{N}(d_{gj}; \alpha_g, \sigma_d^2)}{q \prod_{j, z_g = 1} \mathsf{N}(d_{gjl}; \alpha_g, \sigma_d^2) + (1-q) \mathsf{N}(\mathbf{d_g}; \alpha_\mathbf{g} + \xi \boldsymbol{\beta}_\mathbf{g}, \sigma_\mathbf{d}^\mathbf{2} \mathbf{I} + \tau^\mathbf{2} \boldsymbol{\beta}_\mathbf{g} \boldsymbol{\beta}_\mathbf{g}')} & \text{if} \quad z_g = 1 \\ 1 & \text{if} \quad z_g = 0, \end{cases} \end{aligned}$$

where $d_g, \boldsymbol{\beta}_g$ and α_g are vectors with length J: $d_g = \{d_{gj}; \forall j\}$, $\boldsymbol{\beta}_g = \{\beta_{jk}; \forall j\}$, and α_g is α_g repeating J times.

$$\begin{aligned} & f(\theta_g \mid \gamma_g = 1, \mathbf{D}, \mathbf{a}, \mathbf{b}, \boldsymbol{\alpha}, \tau^2, \sigma_d^2, \xi) \\ = \; & \mathsf{N}\left(\frac{\tau^2 \sum_j \beta_{jk}(d_{gj} - \alpha_g) + \sigma_d^2 \xi}{\tau^2 \sum_j \beta_{jk}^2 + \sigma_d^2}, \frac{\sigma_d^2 \tau^2}{\sigma_d^2 + \tau^2 \sum_j \beta_{jk}^2} \right), \end{aligned}$$

then update ξ from its posterior:

$$\xi \sim \mathsf{N}\left(\frac{\nu^2 \sum \theta_g}{\nu^2 \sum I(\theta_g \neq 0) + \tau^2}, \frac{\tau^2 \nu^2}{\nu^2 \sum I(\theta_g \neq 0) + \tau^2} \right).$$

The hyperparameters τ^2 and ν^2 are chosen so the prior for θ covers a plausible range.

6. Sample $(\sigma_l^2 \mid \boldsymbol{\beta}, \boldsymbol{\mu})$

$$\sigma_l^2 \sim IG\left(\frac{J * K}{2} - \frac{1}{2}, \frac{\sum_{j=1}^{J} \sum_{k=1}^{K} (\beta_{jk} - \mu_k)^2}{2} \right).$$

7. Sample $(\sigma^2 \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\beta}, \mathbf{C})$

$$\sigma^2 \sim IG(nJ/2, SSE^c/2),$$

where $SSE^c = \sum_{j=1}^{J} \sum_{i=1}^{n} (c_{ij} - \hat{f}_j^c(x_i))^2$.

8. Sample $(\sigma_d^2 \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\beta}, \mathbf{D}, \boldsymbol{\alpha}, \boldsymbol{\theta})$

$$\sigma_d^2 \sim IG(nJ/2, SSE^d/2),$$

where $SSE^d = \sum_{j=1}^{J} \sum_{g=1}^{n} (d_{ij} - \alpha_g - \theta_g f_j(x_g))^2$.

Bibliography

Abramovich, F., Sapatinas, T., and Silverma, B. W. (2000). Stochastic expansions in an overcomplete wavelet dictionary. *Probability Theory and Related Fields*, **117**, 133–144.

Aguirre, A. J., Brennan, C., Bailey, G., Sinha, R., Feng, B., Leo, C., Zhang, Y., Zhang, J., Gans, J. D., Bardeesy, N., Cauwels, C., Cordon-Cardo, C., Redston, M. S., DePinho, R. A., and Chin, L. (2004). High-resolution characterization of the pancreatic adenocarcinoma genome. *Proc. Natl. Acad. Sci.*, **101**(24), 9067–9072.

Arteaga, C. L. (2001). The epidermal growth factor receptor: from mutant oncogene in nonhuman cancers to therapeutic target in human neoplasia. *J. Clin. Oncol.*, **19**, 32–40.

Baraud, Y. (2004). Confidence balls in Gaussian regression. *Ann. Stat.*, **32**, 528–551.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *Ann. Stat.*, **32**(3), 870–897.

Berg, A. P. and Mikhael, W. B. (1999). A survey of mixed transform techniques for speech and image coding. *Proceedings of the 1999 IEEE International Symposium on Circuits and Systems*, **4**, 106–109.

Berger, J. A., Hautaniemi, S., Mitra, S. K., and Astola, J. (2006). Jointly analyzing gene expression and copy number data in breast cancer using data reduction models. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **3**(1).

Brinkman, N. (1981). Ethanol – a single-cylinder engine study of efficiency and exhaust emissions. *SAE Transcations*, **90**, 1410–1424.

Broët, P. and Richardson, S. (2006). Detection of gene copy number changes in CGH microarrays using a spatially correlation mixture model. *Bioinformatics*, **22**(8), 911–918.

Brooks, S. and Giudici, P. (2000). Markov chain monte carlo convergence assessment via two-way analysis of variance. *Journal of Computational and Graphical Statistics*, **9**(2), 266–285.

Candés, E. J. and Donoho, D. L. (2004). New tight frames of curvelets and the optimal representation of objects with piecewise c^2 singularities. *Commun. Pure. Appl. Math*, **57**, 219–266.

Chen, S. S., Donoho, D. L., and Saunders, M. (2001). Atomic decomposition by basis pursuit. *SIAM Review*, **43**(1), 129–159.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W.-L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T., Chen, F., Feiler, H., Tokuyasu, T., Kingsley, C., Dairkee, S., Meng, Z., Chew, K., Pinkel, D., Jain, A., Ljung, B. M., Esserman, L., Albertson, D. G., Waldman, F. M., and Gray, J. W. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, **10**, 529–541.

Clyde, M., Parmigiani, G., and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika*, **85**, 391–401.

Coifman, R. R. and Meyer, Y. (1990). Orthonormal wave packet bases. Preprint.

Coifman, R. R., Meyer, Y., and Wickerhauser, M. (1992). Adapted waveform analysis, wavelet packets and applications. *ICIAM 1991, Proceedings of the Second International Conference on Industrial and Applied Mathematics*, pages 41–50.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Number 61 in CBMS-NSF Series in Applied Mathematics. SIAM, Philadelphia.

DeBrunner, V. E., Chen, L. X., and Li, H. J. (1997). Lapped multiple basis algorithm for still image compression without blocking effect. *IEEE Trans. Image. Proc.*, **6**, 1316–1322.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998). Automatic Bayesian curve fitting. *J. R. Statist. Soc. B*, **60**, 333–350.

Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. A., and Chinnaiyan, A. M. (2001). Delineation of prognostic biomarkers in prostate cancer. *Nature*, **412**, 822–826.

Donoho, D. and Johnstone, I. (1998). Minimax estimation via wavelet shrinkage. *Annals of Statistics*, **26**, 879–921.

Donoho, D., Elad, M., and Temlyakov, M. (2006). Stable recovery of sparse overcomplete representation in the presence of noise. *Information Theory, IEEE Transactions on*, **52**, 6–18.

Donoho, D. L. and Elad, M. (2003). Maximal sparsity representation via l_1 minimization. *Proc. Natl. Acad. Sci.*, **100**, 2197–2202.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Dutilleux, P. (1989). An implementation of the "algorithme á trous" to compute the wavelet transform. In J.-M. Combes, A. Grossman, and P. Tchamitchain, editors, *Wavelets: Time frequency methods and phase space*, Inverse problems and theoretical imaging, pages 298–304. Springer-Verlag, Berlin.

Fernández, C., Ley, E., and Steel, M. F. (2001). Benchmark priors for Bayesian model averaging. *J. Econometrics*, **100**, 381–427.

Gelman, A. and Rubin, D. P. (1992). Inference from iterative simualtion using mutliple sequences. *Statistical Science*, **7**, 457–511.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Gröchenig, K. (2001). *Foundations of Time-Frequency Analysis*. Birkhäuser, Boston.

Hill, B. M. (1965). Inference about variance components in the one-way model. *J. Am. Statist. Ass.*, **60**, 806–825.

Hobert, J. P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *J. Am. Statist. Ass.*, **91**(436), 1461–1473.

Hsu, L., Self, S. G., Grove, D., Randolph, T., Wang, K., Delrow, J. J., Loo, L., and Porter, P. (2005). Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, **6**(2), 211–226.

Hupe, P., Stransky, N., Thiery, J.-P., Radvanyi, F., and Barillot, E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**(18), 3413–3422.

Hyman, E., kauramniemi, P., Hautaniemi, S., Wolf, M., Mousses, S., Rozenblum, E., Ringnér, M., Sauter, G., Monni, O., Elkahloun, A., Kallioniemi, O.-P., and Kallioniemi, A. (2002). Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Research*, **62**, 6240–6245.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Stat.*, **32**, 1594–1649.

Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Stat.*, **33**, 1700–1752.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Statist. Ass.*, **90**(431), 928–934.

Lai, W. R., Johnson, M., Kucherlapati, R., and Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**(18), 3763–3770.

Lewicki, M. and Sejnowski, T. (1998). Learning overcomplete representations. *Neuron Computation*.

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2005). Mixtures of g-priors for Bayesian variable selection. ISDS Discussion Paper 2005. Submitted for publication.

Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Norton, H., and Brown, E. (1996). Expreesion monitoring by hybridizatin to high density oligonucleotide arrays. *Natural Biotechnology*, **14**, 1675–1680.

Lockhart, D. J. and Winzeler, E. A. (2000). Genomics, gene expression and DNA arrays. *Nature*, **405**, 827–836.

Lockwood, W. W., Chari, R., Chi, B., and Lam, W. L. (2006). Recent advances in array comparative genomic hybridization technologies and their appplications in human genetics. *European Journal of Human Genetics*, **14**, 139–148.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *Internat. Statist. Rev.*, **63**, 215–232.

Mallat, S. (1989a). Multisolution approximations and wavelet orthonormal bases of $\mathbb{L}^2(\mathbb{R})$. *Transs. Amer. Math. Soc.*, **315**, 69–87.

Mallat, S. (1989b). A theory for multiresolution signal decomposition. *IEEE trans. on Pratt. Anal. Mach. Intell.*, **11**, 674–693.

Mallat, S. (1998). *A wavelet tour of signal processing*. Academic Press, second edition.

Myers, C. L., Dunham, M. J., Kung, S. Y., and Troyanskaya, O. G. (2004). Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics*, **20**(18), 3533–3543.

Nason, G. (2002). Choice of wavelet smoothness, primary solution and thresholdin wavelet shrinkage. *Statistics and Computing*, **12**, 219–227.

Nason, G. P. and Silverman, B. W. (1995). The stationary wavelet transform and some statistical applications. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*, volume 103 of *Lecture Notes in Statistics*, pages 281–300. Springer-Verlag, New York.

Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1. *Vision Research*, **37**, 311–325.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2005). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Pensky, M. and Vidakovic, B. (1998). On non-equally spaced wavelet regression. Discussion Paper 98–06, ISDS, Duke University.

Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C. F., Lashkari, D., Shalon, D., Brown, P. O., and Botstein, D. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci.*, **96**, 9212–9217.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, Ø., Pergamenschikov, A., Williams, C., Zhu, S. X., Børresen-Dale, P. E. L. A.-L., Brown, P. O., and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.

Picard, F., Robin, S., Lavielle, M., Vaisse, C., and Daudin, J.-J. (2005). A statistical approach for array CGH analysis. *BMC Bioinformatics*, **6**(27).

Pinkel, D. and Albertson, D. G. (2005). Array comparative genomic hybridization and its application in cancer. *Nature Genetics*, **37**, 511–517.

Platzer, P., Upender, M. B., Wilson, K., Willis, J., Lutterbaugh, J., andJames K. V. Willson, A. N., Mack, D., Reid, T., and Markowitz, S. (2002). Silence of chromosomal amplification in colon cancer. *Cancer Research*, **62**, 1134–1138.

Pollack, J. R., Sørlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E., Tibshirani, R., Botstein, D., Børresen-Dale, A.-L., and Brown, P. O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl. Acad. Sci.*, **99**(20), 12963–12968.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag, New York, NY.

Ross, J. S. and Fletcher, J. A. (1999). The her-2/neu oncogene: prognostic factor, predictive factor and target for therapy. *Semin. Cancer Biol.*, **9**, 125–138.

Rueda, O. M. and Daz-Uriarte, R. (2006). A flexible, accurate and extensible statistical method for detecing genomic copy-number changes. *Manuscript*.

Russ, A. P. and Lampel, S. (2005). The druggable genome: An update. *Drug Discov. Today*, **10**, 1607–1610.

Sardy, S., Perceval, D. B., G., B. A., Gao, H., and Athestzle, W. (1999). Wavelet shrinklage for unequally spaced data. *Statistics and Computing*, **9**, 65–75.

Schena, M., shalon, D., Davis, R. W., and Brown, P. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Slamon, D., Godolphin, W., Jones, L., Holt, J., Wong, S., Keith, D., Levin, W., Stuart, S., Udove, J., Ullrich, A., and Press, M. (1989). Studies of the her-2/neu proto-oncogene in human breast and ovarian cancer. *Science*, **244**, 707–712.

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J. R. Statist. Soc. B*, **55**, 3–23.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisenh, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P. E., and Børresen-Dale, A.-L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.*, **98**, 10869–10874.

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lønning, P. E., Brown, P. O., Børresen-Dale, A.-L., and Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci.*, **100**, 8418–8423.

Sun, D., Tsutakawa, R. K., and He, Z. (2001). Propriety of posterior with improper priors in hierarchical linear mix models. *Statistical Sinica*, **11**, 77–95.

Tanaka, S., Sugimachi, K., Kawaguchi, H., Saeki, H., Ohno, S., and Wands, J. (2000). GRB7 signal transduction protein mediates metastatic progression of esophageal carcinoma. *J. Cell. Physiol.*, **183**, 411–415.

Velulescu, V., Zhang, L., Volgelstein, B., and Kinzler, K. (1995). Serial analysis of gene expression. *Science*, **270**, 484–487.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. Computational & Graphical Statistics. John Wiley & Sons, New York, NY.

Vogel, C. L., Cobleigh, M. A., Tripathy, D., Gutheil, J. C., Harris, L. N., Fehrenbacher, L., Slamon, D. J., Murphy, M., Novotny, W. F., Burchmore, M., Shak, S., Stewart, S. J., and Press, M. (2002). Efficacy and safety of trastuzumab as a single agent in first-line treatment of her2-overexpressing metastatic breast cancer. *J. Clin. Oncol.*, **20**, 719–726.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B., and Tibshirani, R. (2005). A method for calling gains and losses in array CGH data. *Biostatistics*, **6**(1), 45–58.

Wickerhauser, M. (1994). *Adapted wavelet analysis from theory to software*. Wellesley.

Willenbrock, H. and Fridlyand, J. (2005). A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, **21**(22), 4084–4091.

Wolfe, P. J., Godsill, S. J., and Ng, W.-J. (2004). Bayesian variable selection and regularisation for time-frequency surface estimation. *J. R. Statist. Soc. B*, **66**, 575–589.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243. North-Holland/Elsevier.

Zhang, J. and Feng, B. (2006). How to use cghMCR. *Manuscript*.

Zhu, Y., Kan, L., Qi, C., Kanwar, Y., Yeldandi, A., Rao, M., and Reddy, J. (2000). Isolation and characterization of peroxisome proliferator-activated receptor (ppar) interacting protein (prip) as a coactivator for ppar. *J. Biol. Chem*, **275**, 13510–13516.

Biography

Jen-hwa Chu was born on March 11, 1975 in Taipei, Taiwan (Republic of China). He received his bachelor degree in international business in June 1997 from National Taiwan University in Taipei, Taiwan. He received an M.S. in statistics from the Institute of Statistics and Decision Sciences, Duke University in May 2004.