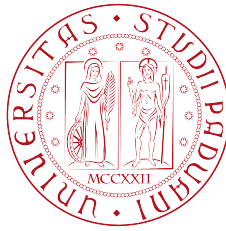


Università degli studi di Padova
Dipartimento di Scienze Statistiche
Corso di laurea magistrale in
Scienze Statistiche



Wavelet Based Models for Copy Number Variation Profiling

Relatore Prof. Chiara Romualdi

Dipartimento di Biologia

Correlatore Prof. Lieven Clement

Leuven Biostatistics and Statistical Bioinformatics Centre (Leuven, BE)

Department of Applied Mathematics, Computer Science and Statistics, Ghent
University (Gent, BE)

Laureando: Nicola Stufano

Matricola N 1018902

Anno Accademico 2012/2013

Contents

Introduction	5
1 Background	7
1.1 'Omics' in Biomedical Research	8
1.2 Copy Number Variation	9
1.3 The Array-CGH Platform	10
1.4 Data Analysis Methods	12
1.4.1 Preprocessing, Normalization and Calling	13
1.4.2 Calling and Segmentation Models	14
1.4.3 Multiple Testing	16
2 Case Study and Objectives	17
2.1 Background	17
2.2 Data Description	18
2.3 Preprocessing	20
2.3.1 Balanced Design	20
2.3.2 GC Content Normalization	21
2.3.3 Imputation of Missing Values	22
2.4 Data Overdispersion	22
2.5 Objectives and Outline	24
3 Data exploration using Wavelets	27
3.1 Introduction	27
3.2 Wavelet-based Smoothing	28
3.2.1 Computation of Wavelet Coefficients	29

3.2.2	Advantages and Disadvantages in the use of Wavelets	33
3.3	Wavelet Based Denoising	34
3.3.1	Wavelet Thresholding	34
3.3.2	Sparse Principal Component Analysis	35
3.3.3	Cluster Analysis	39
3.4	Results	39
4	Wavelet Based Functional Models	47
4.1	False Discovery Rate	47
4.1.1	Bayesian FDR	50
4.1.2	Local False Discovery Rate	51
4.2	Functional Models	53
4.2.1	LFDR-Based Thresholding	54
4.2.2	MAP Thresholding	55
4.3	Results	56
4.4	Multiple Testing	64
5	Wavelet Based Mixed Models	69
5.1	Mixed Model	69
5.2	Functional Mixed Model	70
5.2.1	Marginal Model	71
5.2.2	Prediction	71
5.3	Results	73
5.4	Testing	79
6	Conclusion and Discussion	83
A	R-code	87
A.1	Data Exploration	87
A.2	Wavelet Based Functional Model	94
A.3	Wavelet Based Mixed Model	104
	Bibliography	111

Introduction

Due to the growing avalanche of data in today's biological and biomedical research statistics has become key for enabling further research progress. The current high-throughput platforms can capture variation in genomic sequence, gene expression and genome-protein interactions at an increasing resolution. This enabled researchers to consider genomic profiles as a function along the genomic coordinate. From a statistical perspective, this opens perspectives for functional data analysis approaches where gene expression and copy number variation (CNV), for instance, can be considered as a non-parametric regression problem. Before implementing traditional functional data analysis approaches within the genomics realm, the methods have to be tuned towards the specific application. The choice of the method has to be driven by biological knowledge and the specific research questions.

In the context of copy number variation, for instance, biologists and biomedical researchers consider the underlying copy number profile for an individual patient to be a piecewise constant function along the chromosome. Hence, a piecewise constant representation seems favorable and also will allow a straightforward segmentation of the copy number profile in duplicated, normal and deleted regions. In this specific context, biologists are thus not interested in a smooth representation of CNV profiles, but, in methods that allow for abrupt changes at the boundaries of deleted (duplicated) regions and which can impose sparsity. Wavelets appear to be very useful for this purpose. A wavelet transform has a decorrelating properties and concentrates most of the structure of the signal in relatively few large wavelet coefficients while distributing white noise equally over all wavelet coefficients. Denoising can thus be done by thresholding the smallest wavelet coefficients or shrinking them towards zero.

The compact support of the wavelet basis functions allows for a discontinuity-preserving denoising and leads to a sparse representation of the profile. In this thesis, we develop new methods for modeling CNV using a case study on Gastrointestinal Stromal Tumors. We first adopt existing wavelet approach within the genomic context and extend them into a mixed model framework in order to model group profiles and subject specific CNV profiles, simultaneously.

The goal of the thesis is to assess the use of wavelet based functional data analysis approaches for modeling copy number variation and can be considered as a first step in a developing novel framework in this context.

Chapter 1

Background

The genetic information of most organism's is stored using the macromolecule deoxyribonucleic acid (**DNA**). The DNA is located in a cell's nucleus and encodes for all cellular processes and structures. The DNA first needs to be interpreted before it can be translated into actions. The central dogma of biology states that DNA is first transcribed in ribonucleic acid (**RNA**), which is biologically active. RNA can travel from the nucleus to the cytoplasm where it can be further translated into proteins. The DNA is a polynucleotide, a long chain of nucleotides (molecules with three functional groups) with the ability of catalyze reactions helpful to production of molecules of the same catalyzer. Moreover, polynucleotides can drive directly the formation of exact copies of their sequence.

Usually DNA is identified with a chain of nitrogenous basis, that represent one of the three functional groups (the other two are deoxyribose sugar and phosphate group). DNA could be seen as a language of 4 nitrogenous basis: Adenine (A), Guanine (G), Cytosine (C) and Timine (T). DNA has a double helix structure with single helix paired. It is known that Guanine pairs only with Cytosine while Adenine pairs with Timine, since Adenine and Guanine differs in the atomic form from Cytosine and Timine. A and G are called *Purines*, and their atom looks like a pair of rings fused together, while C and T are *Pyrimidines*, composed by only one ring.

DNA replicates itself through the separation of the double helix. On the two filaments the complementary helix is syntesized by DNA-polymerase enzyme.

DNA of eukaryote cells is compacted in series of **Chromosomes**, inside the cell nucleus.

As mentioned above, DNA could be seen as a language. Nitrogenous basis are read three at time and each triplet (*codon*) corresponds to a known *aminoacid*. A chain of aminoacids is a **Protein**, the fundamental constituent of every animal or plant cell. The information for the protein synthesis is contained in the **mRNA**, a polynucleotide created during the process of transcription. RNA has slight differences from DNA, e.g. the presence of the nitrogenous basis Uracil (U) in spite of Thymine.

1.1 'Omics' in Biomedical Research

The term 'Omics' refers to fields of study in biology such as genomics, transcriptomics, proteomics and metabolomics. It is used to address research on the genome (the entire collection of the genetic information stored in DNA), gene transcription (transcriptome or the entire RNA landscape), proteome (the collection of all proteins).

In the last decades, there is a growing consensus that many diseases are driven by aberrations in the genome, transcriptome and/or proteome. In the dissertation we focus on copy number variation, which can be considered to be a structural variation in the genome. Hence, we will focus on genomics.

Genomics refers to the studies of the genomes of organisms. **Genome** indicates the totality of the genetic material in an organism, and spans coding (*Exons*) and noncoding (*Introns*) regions. A **Gene** is a unit of the genomic code that contains all the informations necessary to synthesize a protein.

Genomics are studied through **Sequencing**. Genomic Sequencing is the starting point for a better understanding of the functioning and the evolution of organisms. Sequencing drives to the knowledge of the genomic structure, identifying genes and gene families. But also to genomic aberrations such as point mutations, translocations, copy number variations.

The study on the specific DNA code and aberrations is also referred to as *structural genomics* or genotyping. **Genotyping** is well established in biomedical re-

search for increasing the understanding of the genomic basis of disease, for finding biomarkers and it has an increasing use in diagnostic testing. *Functional genomics* is another branch, that focuses on gene function, gene interactions and also gene expression patterns under varying conditions. Both, structural and functional variation are known to be associated with diseases.

In the latter of the thesis, we focus on CNV genotyping.

1.2 Copy Number Variation

Genomewide studies have uncovered a considerable number of variants throughout the human genome. Some studies have confirmed that these alterations are often correlated with disease [1].

These alterations can be classified in four different types:

- **Deletion:** during the replication a genomic area is not copied;
- **Duplication:** during the replication a genomic area is duplicated;
- **Inversion:** during the replication a genomic area is replicated with the basis inverted in order;
- **Translocation:** during the replication a segment or a whole chromosome is interchanged with another one.

During the process of replication DNA could be affected by an alteration of the number of copies. Deletions and Duplications are located as copy number variations (CNV). The considerable number of genes that fall within these variable regions make CNV very likely to have functional consequences.

An alteration of the number of copies could be caused by a disease. I.e. Copy Number Variation (CNV) is often observed in tumor states: consequently, studying if the copy number variation is significant could be a considerable indicator of the genes activated when a tumor affects cells. Few cytogenetic techniques¹ have been

¹Cytogenetic is a branch of genetics that studies morphology of chromosomes and karyotype, i.e. the set of chromosomes in a cell. Cytogenetic Techniques were developed in the late 1960s. They are able to differentially stain chromosomes, in order to differentiate chromosomes of equal size and consequentially locate breakpoints and better understood deletions within the chromosome.

developed to give a measurement of the copy number variation, such as:

- **Fluorescent in Situ Hybridization (FISH)** : identifies the presence/absence of some specific DNA sequences. Fluorescent probes are bended to specific chromosome regions, via fluorescence microscopy.
- **Chromosome-Based Comparative Genomic Hybridization (CGH)** : analysis of copy number changes in the DNA content of a given subject's DNA.
- **Array-Based Comparative Genomic Hybridization (aCGH)**: detects changes at a higher resolution level than Chromosome-based technique.
- **Virtual Karyotyping**: Virtual Karyotyping is the digital information reflecting a karyotype, detecting CNV at a high resolution level. Virtual Karyotyping could be furnished using aCGH or SNP (Single Nucleotide Polymorphism) Arrays.²
- **CNV-seq**: It is a method for detecting CNV using high-throughput sequencing. The method is based on a robust statistical model that describes the complete analysis procedure and allows the computation of essential confidence values for detection of CNV. CNV-seq favors the next-generation sequencing methods that rapidly produce large amount of short reads [2].

1.3 The Array-CGH Platform

Microarrays allow for the quantitative measurement of thousands of biochemical reactions in parallel. They are commonly used for detecting genomic mutations and for analyzing RNA levels or gene transcription. The technology is based on complementary base-pairing of DNA or *hybridization*.

Microarrays consist of thousands of short pieces of DNA, probes that are immobilized on a support. Each of the probes is complementary to a specific DNA

²SNP arrays, due to their nature, are used to detect minor changes between whole genomes. That skill of this particular type of DNA Microarray finds its application in studies of genetic abnormalities in cancer, e.g. Loss of Heterozygosity, that occurs when one allele is damaged by the mutation and it cannot develop tumor suppressors [3].

sequence in the genome and can probe for the presence of specific DNA sequences in a sample. The samples are first processed and labeled using fluorescent markers. Next, the labeled DNA library is hybridized on the microarray. DNA in the library that is complementary to the probes will hybridize. After an incubation period, the array surface is washed to remove the remaining unhybridized and labeled molecules. Finally, a microarray readout is made by measuring the fluorescence intensity of the labeled molecules that are hybridized to the array. Higher concentrations in the sample, typically will lead to more hybridization and thus a higher intensity signal.

Array-CGH (**aCGH**) is an evolution of the CGH Platform. The probes are BACs (genomic DNA sequences) that are mapped on the genome. The signal has a spatial coherence that can be handled by specific statistical tools. Array-CGH profile can be viewed as a succession of segments that represent homogeneous regions in the genome whose BACs share the same relative copy number on average. The differences between subject and reference are analyzed by fluorescence. The fluorescence intensity of the subject sample and reference sample is then measured in order to calculate the ratio between them.

the CGH platform is a two-channel array. A test and a reference sample are hybridized to the same slide. Both samples are labeled with a different dye. The differences between subject and reference are analyzed by measuring the fluorescence. The fluorescence intensity between the dyes is measured and the ratio between them can be calculated.

CGH experiments suffer a loss of precision: it is possible to detect copy loss for regions' length of 5-10 Mb ³, while the detection of amplification is known to be sensitive down to less than 1 Mb. This imbalance can be overcome with the use of array CGH.

In aCGH, equal amounts of labeled genomic DNA from a test and a reference sample are co-hybridized to an array containing the DNA targets. The higher

³Mega Base pairs. A Base Pair is a couple of complementary nucleotides in an hybridized probe. Bp have a measurement system equivalent to the digital information, where a Base Pair corresponds to a byte.

resolution and throughput are the most significant advantages of aCGH over all cytogenetic methods except CNV-seq. In addition, there is no need for cell culture, making the turn around time shorter than cytogenetic methods. Most clinical aCGH platforms require only a few micrograms of genomic DNA, and whole-genome amplification procedures enable further reduction of the amount needed for analysis. Summarizing, aCGH has revealed in the last years clinically unsuspected genomic unbalances, leading researchers to focus more in whole-genome approach than locus-specific methods.

Even aCGH technology suffers some limitations. There are still the problems related with platforms that cover sequentially the entire genome with high resolution: these technologies are expensive and they are more likely to detect imbalances without a clear meaning [4]. At last aCGH profiles contain a wave bias. The hybridization potential of some probes is higher than others and some genomic regions will be preferentially amplified in the library preparation step. This bias can obscure the interpretation of the profile and induces additional challenges for the data analysis.

1.4 Data Analysis Methods

In an Array-CGH experiment, a CNV **profile** is measured at specific genomic locations that are spanned by **Clones**. For each profile, an intensity is measured by each clone and data can thus be structured in a data array. An example of CNV data can be found in Table 1.1. A considerable number of techniques to extract information from aCGH data have already been suggested. Few problems have been considered and discussed, starting with preprocessing on data and finishing with testing significant results.

Table 1.1: Example of CNV data

ID	Obser- vation	ID profile	Chromosome	Clone Name	CNV	Clone Po- sition
1		25	13	RP-11	0.845	10
2		36	13	RP-122	-0.45	23
3		31	13	RP-122	1.32	23
4		17	17	RP-982	0.88	23
5	

1.4.1 Preprocessing, Normalization and Calling

In aCGH studies, intensities are typically measured for test and a reference sample. They are commonly transformed into a \log_2 ratio of the test and reference intensity, which reflect the relative copy number in the test sample compared to the reference sample.

Usually a normalization is done. **Normalization** corrects for experimental artifacts in order to make different hybridizations comparable. For normalization mode subtraction is the favourite method, but other methods have been proposed, e.g. fitting Lowess curve or a Ridge regression to calibrate the two-channel intensity density plot. All the methods mentioned above are within-array normalization, but in some cases also between-array normalization is carried out, in order to give the potentially large proportion of aberrations in a DNA sample [3].

Another correction has been often imposed for GC Content.

GC Content is the percentage of GC nitrogenous basis on a DNA molecule on the total. It is known that there is a relation between the measure of the GC content and the bias of the response, so this measurement is used in preprocessing to correct bias on data.

Picard *et al.* [5] proposed a quadratic regression scheme as GC Correction:

$$Y_j(t) = \mu_{js} + b(t) + \alpha_1 GC_t + \alpha_2 GC_t^2 + E_j(t), \forall t \in]t_{s+1}^j, t_s^j] \quad (1.1)$$

Where Y is the processed signal, μ denotes the original signal, j the profile, s the segments (segmentation will be explained in 1.4.2), t the clone position, b is the probe effect, E a Gaussian white noise, GC is expressed as percentage.

We only have a relative measure at our disposal. Hence the log ratio between test and reference sample has to be transformed in a estimate of the CNV status, i.e. in Table 1.2.

Table 1.2: Conversion of the number of copies into an absolute measure and corresponding \log_2 values of expression

Loss	0-1 copies	$\log_2 y < 0$
Normal	2 copies	$\log_2 y = 0$
Gain	3-4 copies	$0 < \log_2 y \leq 2$
Amplification	more than 4 copies	$\log_2 y > 2$

The detection of an absolute measure is called **Calling** [6].

1.4.2 Calling and Segmentation Models

There are two main frameworks of approaches to deal with CNV in aCGH technology. *Calling* models the state of each probe using calling, i.e. giving an absolute measure to each probe. These methods can make use of the dependence between neighboring clones with Hidden Markov Models, where the true copy number values are the latent states in the HMM design. Van de Wiel *et al.* [6] performed an accurate algorithm, named CGHcall, where they counteract fluctuations of loss/normal/gain levels and combined a mixture model with six states (l indicates the states)

$$Y_{tsj} \sim N(\mu_{sj}, \sigma_j^2) \quad (1.2)$$

with

$$\mu_{sj} \sim \sum_{l=1}^6 p_l N(\gamma_l, \tau_l^2) \quad (1.3)$$

Hence, in this model the signal depends on a mixed proportion of the states (indicated by p_l for each l state).

Magi *et al.* [7] simulated the shifts in the mean with an algorithm, called Shifting Level Model, that simulates a noisy sequential process based on a Hidden Markov Model procedure. Parameters are then simulated using an EM algorithm.

The second approach exploits the use of segmentation models. *Segmentation* models are based on the need of identifying segments with common means, separated by breakpoints between neighboring clones, and estimate the mean of these regions. Picard *et al.* [8] suggested that CGH profile is supposed to be a gaussian signal and two types of changes can be considered: changes in mean and variance and changes only in mean. Hence, two different segmentation models based on these changes could be provided. Later, Picard *et al.* [5] proposed a model to extend segmentation to all profiles simultaneously (Joint Segmentation). Nobody said Calling and Segmentation can not be crossed. van Wieringen, Van de Wiel et Ylstra [13] showed calling consequentially to segmentation in preprocessing.

Some Models related to segmentation could solve one of the greatest problems of the aCGH data, the presence of “wave bias” [10] technical artifacts in the profiles. Some researchers began to consider aCGH data with different approaches. In 2005 Lai *et al.* [9] tried to perform eleven different algorithms on aCGH data, obtaining the best results with Quantile Regression, Lowess and Maximum Overlap Discrete Wavelet Transform (MODWT). In 2009 also van de Wiel *and al.* [10] began to study aCGH data in the continuous domain, purposing to shrink data applying a ridge regression on the signal smoothed with loess curve. Novak *et al.* [11] proposed the Fused Lasso Latent Feature Model (FLLat) to provide a statistical framework for modeling multi-sample aCGH data and identifying regions of CNV. The procedure involves modeling each sample of aCGH data as a weighted sum of a fixed number of features, then identifies regions through FLLat applied to each feature. Baladandayuthapani *et al.* [12] proposed a hierarchical Bayesian random segmentation approach for modeling aCGH data that uses information across profiles from common population to discover segments of shared copy number changes. These changes allow comparing different population aCGH profiles. The Posteriori simulation is done via MCMC.

Van de Wiel and van Wieringen [14] showed the need of reducing the dimensional-

ity for aCGH optimizing the information loss. Dimensionality is reduced creating regions of clones, using a threshold for the differences between neighboring clones. The distance function must express a value under the threshold within all the possible couples of clones in the region and over the threshold between clones neighboring in the breakpoints. The number of regions selected will depend on the threshold selected. With Ylstra [15] was proposed also a clustering algorithm for called aCGH, introducing weights for regions and clones by prior informations about chromosomes and the specific disease studied.

1.4.3 Multiple Testing

Since a segmentation model is developed, testing is necessary to verify if there are significant copy gains/losses between sample and reference. In the context of segmentation testing could be straightforward if the scientific question is only about copy gain or loss, because under the null hypothesis we have a straight line on the horizontal axis ($CNV = 0$). Hence, simple linear hypothesis tests could be provided with an estimation of mean and variance of the segments. The real problems are the number of simultaneous tests involved and the correlation between neighboring clones. It is often necessary resort to techniques of multiple testing, as FDR and Local FDR or correction of Holm-Sidak for p-values.

In testing differences between groups, one can prefer to perform tests on regions using aberration calls (i.e. a loss of a gain of one chromosome at least) rather than \log_2 ratios, as Van de Wiel et van Wieringen do [14] using a Wilcoxon two-sample test. This is to simplify the interpretation, because in the aberration case rejecting the null hypothesis one can conclude that the aberration levels differ. In the segmentation case, one can say there are significant differences between mean \log_2 -ratios, but this statement does not bring to a clear interpretation.

However, literature is very poor about testing, especially in the case of multiple profiles. For multiple profiles the Van de Wiel, Van Wieringen et Ylstra[15] used clustering and a model-based classification approach, as mentioned above, but there are not formal procedures for testing.

Chapter 2

Case Study and Objectives

2.1 Background

Gastrointestinal Stromal Tumors (GIST) are tumors that affect mesenchyme, a loose connective tissue that is derived from the mesoderm, one of the three primary germ layers in the embryo for the gastrointestinal tract. The majority (60 %) of these tumors is located in the stomach, 25% in the small intestine. the remaining 15% affects the large intestine or the esophagus. The presence of a GIST is usually associated with the mutation of the Mast/stem cell growth factor receptor, well known as KIT gene. GIST family has been identified only after the discovery of the link between KIT antigen and tumors. Prior they were recognized as muscular tissue tumors. In some cases where KIT mutation is not identified alpha-type platelet-derived growth factor receptor (a protein encoded by the PDGFRA gene) is mutated. KIT and PDGFRA mutations are mutually exclusive.

GIST tumors identification is not straightforward, since the lack of symptoms. Beside the difficult identification, they are also dangerous due to the emergence of metastasis, which happens in half of the cases. Moreover, these tumors cannot be treated with chemotherapy or radiotherapy. The opportunities to save a patient are surgery or the use of Imantinib, an inhibitor for KIT and PDGFRA[16].

2.2 Data Description

The researchers have used two-channel aCGH microarrays to assess the copy number variation for the chromosome 13 under three different GIST types:

1. PDGFRA Mutations ;
2. Gastric KIT Mutations;
3. Non-Gastric KIT Mutations.

60 profiles have been collected, using the dye-flip technique. Dye flip is performed for avoiding bias due to differences in fluorescence in the different channels. Some profiles consist of samples that are labeled with Cy3 and where Cy5¹ is used for the reference, for other profiles the dyes are reversed. Profiles for chromosome 13 have 97 contiguous clone positions. All profiles read the same clones in the same order, hence they are comparable.

For every observation in the dataset the following features are recorded:

- **Clone:** the name of the clone used for hybridization. The prefix "RP11" indicates that all clones come from the human male BAC² library. Except a few small ones, clones are extended in order to 130-225 Kb.
- **Response value :** To build the response value four measurements were collected: the intensities of the sample and the reference and the background intensities of the sample and the reference. The background intensities of the spot are used for a background correction. Response value is given subtracting the logarithm of the reference intensity to the sample intensity, both corrected with the background intensity. In other words, it is the \log_2 ratio between sample and reference intensities.

¹Cy3 and Cy5 are the most popular Cyanine dyes used, providing respectively green and red fluorescence. Cyanine is the synthetic dye family used in biotechnology.

²BAC stands for Bacterial Artificial Chromosome. It is an artificial DNA vector able to transport regions up to 300 kb. The resulting protocol "BAC by BAC" has been used to sequence the human genome, alternatively to "shotgun" approach. BAC has been used for the Human Genome Project.

- **ID**: a number between 1 and 60 to indicate the source profile.
- **Group** a number between 1 and 3 to indicate the treatment group. The corresponding treatments are those mentioned above in this section
- **Starting Position** : indicates the start position of the clone in the chromosome.
- **Ending Position** : indicates the ending position of the clone in the chromosome.
- **GC Content**: GC content for the probe, it is used in section 2.3 for pre-processing.

We show below the CNV data of this dataset are presented in R.

	CLONE	CHROMOSOME	response1	Id_array	Group	START
1	RP11-76K19	13	0.41150354	1	1	20238604
2	RP11-187L3	13	-0.20778623	1	1	20841155
3	RP11-110K8	13	0.99019630	1	1	22085693
4	RP11-26A3	13	0.18916734	1	1	23210664
5	RP11-760M1	13	-0.19574522	1	1	23990527
6	RP11-556N21	13	-0.03875384	1	1	25025512
	END	GC_content				
1	20239297	0.4273188				
2	21021905	0.4635838				
3	22262824	0.3885880				
4	23211399	0.3742748				
5	24166746	0.4217345				
6	25204118	0.4207925				

2.3 Preprocessing

Giving a quick glance to data we can notice that the design is not balanced. There are 16 profiles for PDGFRA Group, 21 for Gastric-KIT and 23 for Non-Gastric KIT. Moreover, profiles are often affected by missingness and some probes and profiles are more affected than others. The reason why some data are missing stays in a technical artifact: it could happen that background intensity is bigger than spot intensity and negative values are obtained. Hence, it is impossible transform to log-intensity. Some preprocessing is needed to obtain clean data.

Furthermore, response has not been retained as it was, but It has been corrected by the median value of the autosomal³ chromosomes. The reason is because these data were used in the past for other studies with different chromosomes analyzed.

2.3.1 Balanced Design

We want to have a balanced design and exclude profiles that are too much affected by missingness. Balanced design is needed to calculate further a Mean Absolute Deviation (MAD) that is balanced between groups. We can mix these two operations excluding profiles that have at least more than 3 observations missing. Since missingness affects more the second and third, we have already balanced a little bit the groups. In order to obtain 14 profiles for each group, we have selected at random a few profiles in second and third group between profiles with exactly 3 missing values.

It is necessary for the wavelet transform to have entire profiles, without missing values. With this preprocessing we are going to impute 3 observation per profile at least, hence eventual bias in imputation will not affect too much estimates.

We had the raw data not at our disposal. Because the scope of the thesis is a first evaluation of the use of wavelet based functional mixed models to CNV applications, we consider a simple work-around. In real applications, the data could be extracted again and other strategies could be considered to circumvent the log intensity problem, e.g. the use of the generalized logarithm. For estimation purposes, we will also work with a balanced design. Because the methods

³Chromosomes that are not sex chromosomes.

are regression based, they can easily deal with non-balanced designs in practical applications.

2.3.2 GC Content Normalization

The response is first normalized by GC content. The response is also affected by the potential group effect. We therefore propose a model with a smoother for the group effect and for the GC content:

$$Y = \beta_0 + \sum_{j=1}^p f_j(u_j) \quad (2.1)$$

The response is corrected subtracting the prediction obtained with a Generalized Additive Model with p basis function $f_j(x)$ that represent the smooth effect for GC content and group effect.

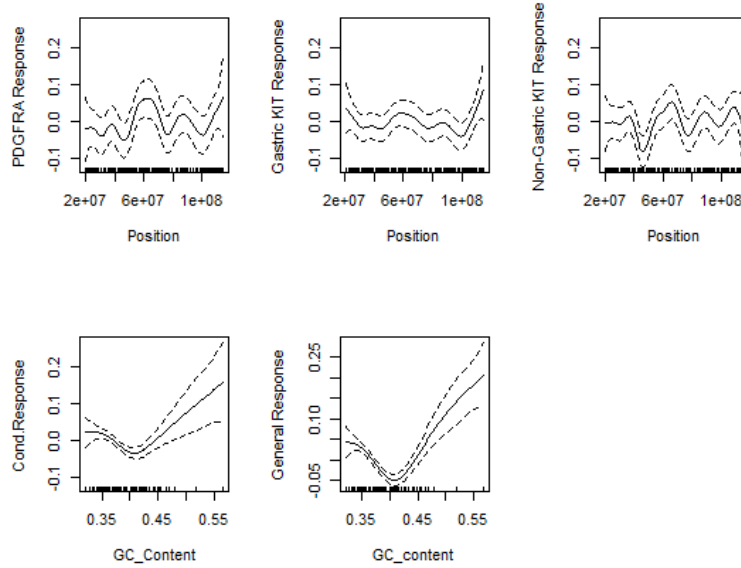


Figure 2.1: Decomposition of the signal in GC Content and Group Effects by GAM for Chromosome 13. The upper plots represent the groups' response while the lower are the response conditioned for groups and the unconditional response.

2.3.3 Imputation of Missing Values

To impute missing observations we use the algorithm MI, that stands for **Multiple Imputation**. The concept behind Multiple Imputation refers to impute each missing value more than once in order to complete data. Hence, suppose there is a matrix Y with y^o observed data and y^m missing values. Through Multiple Imputation we obtain M (the number of cycles of imputation) completed data set. A vector $\hat{\theta}$ of parameters is estimated for each dataset and the parameter vector final estimation is given by the arithmetic mean of the estimates.

Summarizing [17]:

- 1 Draw each Y_i^{m*} from $f(y_i^m|y_i^o)$;
- 2 Using completed data $Y^c = (Y^o, Y^{m*})$ estimate the parameter of interest $\hat{\theta} = \hat{\theta}(Y^c)$.
- 3 repeat steps 1 and 2 for M times and calculate $\hat{\theta}^* = \frac{1}{M} \sum_{j=1}^M \hat{\theta}^{(j)}$.

Hence, The imputation of missing data will be conducted hand in hand with estimates.

2.4 Data Overdispersion

It appears that some profiles between the 42 selected suffer some problems of overdispersion. The reasons of these cases of overdispersion are unknown, as we can see in figure 2.2 where the most serious cases of overdispersion are collected. Both profiles similar and non-similar to their corresponding group mean profile (group mean profiles are shown in Chapter 4) could suffer problems of overdispersion, hence it is not straightforward comprehend which is the reason. Overdispersion could be due to technical artifacts, or to data specific irregularities, but one explanation does not exclude the other. Overdispersed profile in Figure 2.2 are compared with profiles with regular dispersion in Figure 2.3.

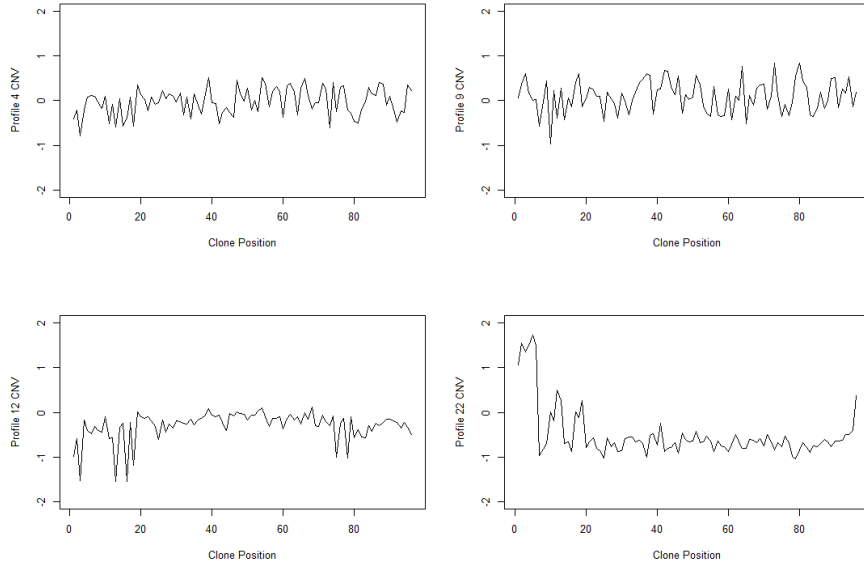


Figure 2.2: Overdispersed Raw Data Profiles 4,9,12 and 22.

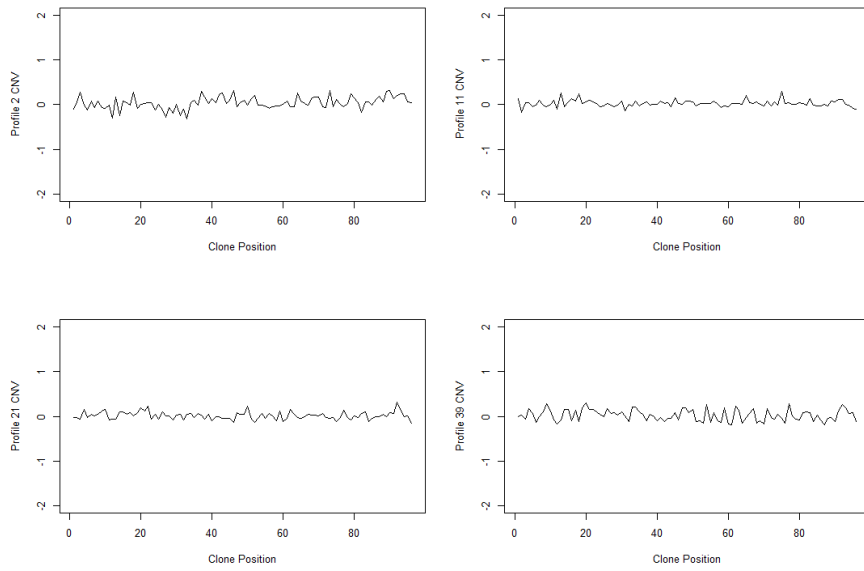


Figure 2.3: Raw Data Profiles 2,11 ,21 and 39.

2.5 Objectives and Outline

A first analysis of array-CGH in GIST was conducted by Wozniak *et al.* [18]. That work aimed to identify with aCGH technology copy number gains and losses in the chromosomes. Chromosome 13 was one of the chromosomes that showed more genomic losses, equal to 29 % of chromosomal array. Two minimal overlapping regions of deletion were found, at 13q14.11-q14.2 and 13q32.3-q33.1. The former includes a tumor suppressor gene, *RB-1*, while the latter has not been assigned to anything.

In this thesis we will provide a functional representation of these data in chapter 3. Data are very noisy and a sparse representation is needed for inferring on the underlying copy number status. But some problems arise in the choice of a credible technique of thresholding before backtransforming. Smoothing needs to be strong in order to obtain few big regions of clones that can be interpreted as linked to known genes in an enrichment analysis.

Further, in chapter 4, wavelet coefficients will be processed with a functional model and thresholded with two different empirical Bayes approaches : the first uses a Jeffreys' Noninformative Prior to provide MAP (Maximum a Posteriori) Thresholding. The second one selects coefficients of interest calculating their LFDR (Local False Discovery Rate) and selecting an appropriate threshold for p-values.

Looking to profiles, other problems arise when there are profiles completely different from the other profiles of the same group. Making inference only on group mean profiles could provide incomplete results and shows good differences between groups, but does not show differences within groups. Moreover, in the context of personalized medicine it is very useful for the clinicians to dispose of a method for interpreting both group mean profiles as well as individual profiles, simultaneously. A mixed model approach is considered in chapter 5 for modeling the group mean profile as well as individual profiles.

Wavelet based methods have been developed for genomic applications, e.g. Clement *et al.* [19]. Their method was developed for gene expression studies with tiling arrays. **Tiling arrays** are a microarray platform. Tiling Arrays are a subtype of microarray chips for the measurement of differential expression that differ from the usual microarrays in the nature of the clones: tiling arrays clone for sequences that are known to exist in a contiguous region, in order to characterize regions. Hence, Tiling Arrays allows to functional analysis. Clement adopted a wavelet-based functional model to this context with a fast empirical bayes method to provide adaptive regularization of the functional effects. The functions elaborated for that work were included in the package **WaveTiling** developed by De Beuf, Pipelers and Clement for Tiling Array transcriptome analysis. This work exploits the package **WaveTiling** using the functions `wavebacktransformK`, `MapMarImpEstJ` and `WaveMarEstVarJ`.

Chapter 3

Data exploration using Wavelets

3.1 Introduction

aCGH data are particularly noisy. The noise is assumed to be normally distributed. Regularization is not performed in the original domain but in the **wavelet** domain, which has advantageous properties for denoising.

However, smoothing is not straightforward. Many techniques have been tested in order to obtain correct smoothness. It is necessary to combine models with piecewise representation to proceed with a segmentation model that could answer to scientific questions in genomics context.

A **Sparse representation** meets some of these requirements. A sparse parameterization allows to represent the data with a relative few coefficients and reduces the dimensionality. Here, we exploit a sparse representation in the wavelet domain. One can use thresholding for this purpose, i.e. we first transform the data to the wavelet space and use a threshold to set some of the coefficients to zero. In Chapter 4 we will show advanced thresholding techniques based on empirical Bayes methods. Another way is to perform a decomposition which allows to reduce the dimensionality by defining recombining the wavelet coefficients in a more efficient basis, such by adopting a sparse principal component analysis (PCA) on the wavelet coefficients. The aCGH profiles will then be represented by a limited number of PCs. Upon regularization in the wavelet domain, we will evaluate the different approaches by backtransforming the denoised profiles to the original do-

main. Finally, we will show how clustering helps the identification of differences between profiles and its properties linked with wavelets.

3.2 Wavelet-based Smoothing

A **Wavelet** is a wave-like oscillation with a definite amplitude around zero. The Wavelet is a kind of representation that can be usually found in time frequency representation for continuous time. This kind of representation is related with harmonic analysis. But, it is possible to use a wavelet decomposition in any functional context. In practice we never dispose of continuous infinite observation of the function, but, we observe a discrete realization of it. Within a wavelet context, it is computationally efficient to use a discrete representation.

Mathematically, Wavelets use complete orthonormal basis to represent functions in an alternative way. The Wavelet representation allows to show a phenomenon with his time (or position, in a FDA¹ context) and frequency localization [20]. The wavelet decomposition is a linear projection of the data with as many coefficients as observations and it does not lead to an information loss. Therefore, the wavelet coefficients have to be further manipulated for obtaining a sparse representation. In summary, a **Wavelet- Based smoothing** procedure follows three steps:

- 1) compute the coefficients of the signal ($Y = DW^T$);
- 2) alter the coefficients D and obtain B^*
- 3) backtransform modified coefficients for obtaining $[Y^* = (W^T B^*)^{-1}]$.

Wavelet transformation could be done in a continuous domain (Continuous Wavelet Transform) or based on the grid that is spanned by the observations (Discrete Wavelet Transform). For many reasons, we adopt **DWT**:

- DWT is computationally simpler and easier to understand than CWT;

¹Functional Data Analysis

- DWT appears as a piecewise-constant representation, that is more familiar for our purpose, because it identifies chromosomal regions and breakpoints of a segmentation model;
- DWT coefficients tend to be less correlated than original data [22].

3.2.1 Computation of Wavelet Coefficients

The CNV on clone positions can be taught as realizations of a function f along genomic coordinates. If we indicate these coordinates with $t_i = i/T$,

$$Y_i = f(t_i) + \epsilon_i \quad i = 1, \dots, T. \quad (3.1)$$

With the error terms ϵ_i normally distributed random variables with zero mean and variance σ^2 . The Wavelet basis are generated by translations and dilations of one or more scaling function known as *Wavelet Father*. We adopt the Haar basis, which produces a piecewise-constant representation. The father wavelet is a constant and denoted by $\phi(t)$ and it is an identity function:

$$\phi(t) = I(t \in [0, 1]) \quad (3.2)$$

A wavelet father space is called *reference space* [20] and it is usually indicated with V_0 . The index l indicates the width of dilations. l is equal to 0 for the reference space because there is no dilation for the reference space. If l increases, the shape of the corresponding space is tighter.

Hence, we can impose l dilatations to the reference space that form orthonormal basis $V_l \supset V_{l-1} \supset \dots \supset V_0$ and each V_l is spanned by:

$$\phi_{l,m}(t) = 2^{l/2} \phi(2^l t - m) \quad (3.3)$$

and m is an integer that represents the translations of the dilations. There is clearly a dependence between l and m :

$$m \in [0, 2^l - 1] \quad (3.4)$$

The Haar Wavelet Mother function splits its support in two intervals:

$$\psi(t) = \begin{cases} 1 & 0 < t \leq 1/2 \\ -1 & 1/2 < t \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (3.5)$$

To reach a function for further dilations, we can consider W_l as the orthogonal complement of V_l to V_{l+1} . Then it is $V_{l+1} = V_l \oplus W_l$, calling W_l as *detail*. The orthonormal basis W_l are generated by the *Wavelet Mother*:

$$\psi_{l,m}(t) = 2^{l/2} \psi(2^l t - m) \quad (3.6)$$

It is straightforward notice that an orthonormal basis for W_0 is calculated as $\psi(t) = \psi(2t) - \psi(2t - 1)$.

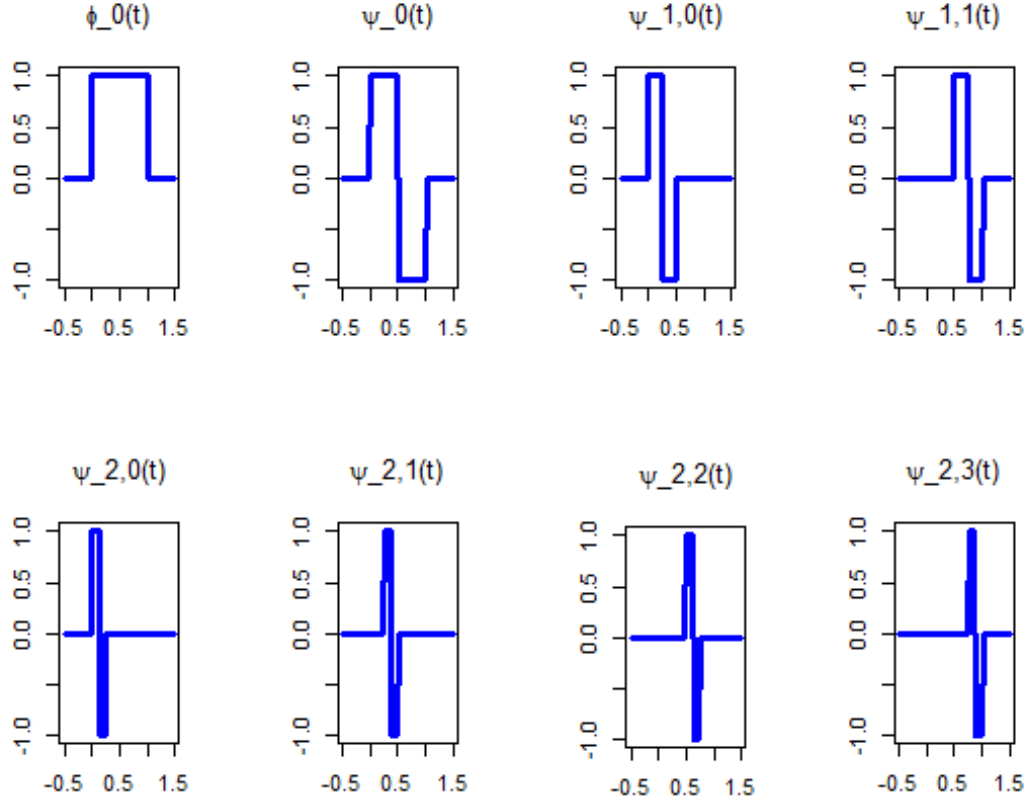


Figure 3.1: Wavelet basis functions obtained for a vector of dimension $N=8$. $\phi_0(t)$ represents the wavelet father, while $\psi_{l,m}$ are the wavelet mothers.

The wavelet transformation is a linear projection which can be denoted using the **Haar Transform Matrix** (W). The Haar Transform Matrix is presented as an $N \times N$ matrix, and N must be divisible by 2^L . The ratio $N/2^L = r$ indicates the number of wavelet father involved. If there is more than one wavelet father, the matrix W will be a block diagonal matrix with $2^L \times 2^L$ blocks, $T = 2^L$, one Wavelet father presented as:

$$\phi_0(t) = \frac{1}{\sqrt{2^L}} \quad 0 < t \leq 1 \quad (3.7)$$

And $2^L - 1$ wavelet mothers defined as:

$$\psi_{l,m}(t) = \frac{1}{\sqrt{2^L}} \begin{cases} 2^{l/2} & \frac{m}{2^l} < t \leq \frac{m+0.5}{2^l} \\ -2^{l/2} & \frac{m+0.5}{2^l} < t \leq \frac{m+1}{2^l} \\ 0 & \text{otherwise} \end{cases} \quad (3.8)$$

Hence a new matrix D is obtained that contains the empirical wavelet coefficients, i.e. the projections of the data on the space spanned by the wavelet basis functions.

$$Y = DW^T \quad (3.9)$$

The construction of the orthonormal basis and the estimation of the matrix D are done simultaneously in **R** following an algorithm called **Fast Wavelet Transform**. In the next section we are going through the algorithm and show which elements are involved to understand how a Wavelet Transform works.

Fast Wavelet Transform

FWT is a recursive algorithm designed by Mallat in 1989 to decompose a waveform into a sequence of coefficients based on orthonormal basis and reconstruct after processing.

First, an Haar Matrix 2×2 is created. We will call this matrix the *filter* W and the first row will be the *Low-Pass Filter*, as the second row will be the *High-Pass Filter*.

$$W = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \quad (3.10)$$

This Haar Matrix, used in combination with a vector y of N elements, could come up with the contrast (Low-Pass) and the grand mean (High-Pass) of each couple of adjacent values. The results of the first iteration are two $N/2$ -dimensional vectors of means and contrasts. The vector of means is allocated in the first $N/2$ positions of the resulting vector d , while the vector of contrasts is used for the second iteration. The vector of means resulting from the second iteration is allocated in queue to the resulting vector, while the vector of contrasts is iterated

another time, and so on.

The new vector d will be composed by $\sum_{l=1}^L N/2^l$ *Wavelet Mother* coefficients, while the remaining $N/2^L$ contrasts will be allocated in queue to the vector d as *Wavelet Father* coefficients.

After processing on Wavelet Coefficients, data will be reconstructed following the inverse procedure.

3.2.2 Advantages and Disvantages in the use of Wavelets

Wavelets differ from other functions that support smoothing in that they do not impose regularisation by smoothness but by signal sparsity for capturing discontinuities and isolated spikes[20].

In the context of DWT, Nyguen *et Al.* [21] objected that DWT creates artifacts around the discontinuities of the input signal. Moreover, Wavelet Fathers and Mothers are located depending on where the series start, hence starting only one clone further in the series could change the locations of the breakpoints. Nyguen *et al.* preferred in their work the use of Maximum Overlap DWT (MODWT). The Maximum Overlap DWT has NJ coefficients, with a redundant ratio of $(J+1):1$. The MODWT has good properties, but the wavelet coefficients become dependent from each other, that is disadvantageous for testing. DWT is also shift-variant, i.e. if wavelet father is shifted brings to different results, while MODWT is shift-invariant, because it uses a father wavelet for each position.

It should be recalled that wavelet transform needs complete data, without missingness, and a array of data with dimensionality proportional to 2^J to be performed. Note, however, that the detailed wavelet coefficients will also compensate an improper of the large wavelet functions and partially alleviate the problems reported by Nuyguen *et al.*

3.3 Wavelet Based Denoising

3.3.1 Wavelet Thresholding

Computed the DWT Coefficients, the next step is to process coefficients in order to shrink data. Many ways could be covered, mostly because the decorrelation property suggests processing the coefficients independently of each other [22]; Now we will focus on the classical choices for shrinkage, based on thresholding, proposed by Donoho and Johnstone [23] . Since typically the structure of the signal is concentrated in few large coefficients, and the remaining coefficients only capture noise, The *Hard Thresholding* simply sets all the coefficients under an arbitrary threshold λ to zero:

$$\delta_{\lambda}^{hard}(\omega) = \begin{cases} 0 & |\omega| \leq \lambda \\ \omega & |\omega| > \lambda \end{cases} \quad (3.11)$$

While the *Soft Thresholding* sets the coefficients below the threshold λ to zero and shrinks the remaining coefficients towards zero by subtracting the thresholds:

$$\delta_{\lambda}^{soft}(\omega) = \begin{cases} 0 & |\omega| \leq \lambda \\ \text{sgn}(\omega)(|\omega| - \lambda) & |\omega| > \lambda \end{cases} \quad (3.12)$$

In Figure 3.2 the difference between hard and soft thresholding is clearer.

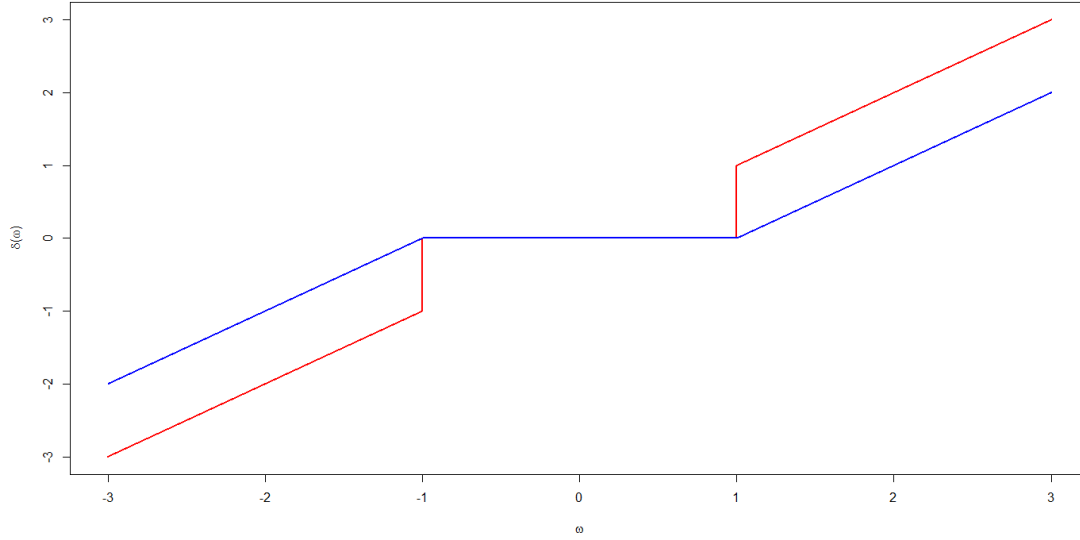


Figure 3.2: Representation of Hard and Soft Thresholding shrinking ω with $\lambda = 1$. Blue line is the soft thresholding representation, while red line is the hard thresholding.

An important aspect to keep in mind is that thresholding is applied only on wavelet mother coefficients. Threshold the wavelet father coefficients creates bias because they are the reference space for the wavelet mother coefficients.

3.3.2 Sparse Principal Component Analysis

Functional Principal Component Analysis

Ordinary Principal Component Analysis (PCA) is a well-known technique used in the field of multivariate statistics to reduce the dimensionality of data. The mathematical procedure behind PCA uses an orthogonal transformation to convert a set of multivariate observations into a set of values of linearly uncorrelated variables. The transformation is defined in such a way that the first component has the largest possible variance under the constraints mentioned above. The reduction of dimensionality is obtained selecting a limited number of components (the principals, indeed), less than the original number of variables.

This concept could be generalized to functions, operating on a set of continuous curves rather than discrete vectors.

Deeper through, we have a set of functional curves $f_t(x), t = 1, \dots, n$ and through an orthogonal transformation we want to obtain a set of $k \leq n$ principal components denoted by $\varphi_k(x)$.

The first FPCA $\varphi_1(x)$ is given as:

$$\beta_{t,1} = \int \varphi_1(x) f_t(x) dx \quad (3.13)$$

Where $\beta_{t,1}$ is the **Score** vector, the new coordinates for each t basis function, maximizing $\sum_t \beta_{t,k}^2$ subject to

$$\int \varphi_1(x)^2 dx = \|\varphi_1(x)\| = 1 \quad (3.14)$$

The second component has the Score vector $\beta_{t,2}$ with the additional constraint

$$\int \varphi_1(x) \varphi_2(x) dx = 0 \quad (3.15)$$

Hence, each component k is given as $\beta_{t,k} = \int \varphi_k(x) f_t(x) dx$ subject to $\int \varphi_k(x)^2 dx = 1$ and $\prod_{n=1}^k \int \varphi_n(x) dx = 0$.

Zou, Hastie and Tibshirani [24] discussed a regression approach to PCA. Hence, they demonstrate that Sparse PCA could be performed as an elastic-net regression problem and they built an algorithm to obtain a numerical solution for the **Loadings**, the measurement of the contribute of each basis to the principal components.

PCA Regression Approach

Suppose we are performing a functional PCA on p basis and we need to obtain $k \leq p$ principal components. The data matrix \mathbf{X} is decomposed using Singular Value Decomposition.

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad (3.16)$$

Where \mathbf{U} is the matrix of eigenvectors of the covariance matrix $\mathbf{X}\mathbf{X}^T$, $\mathbf{\Sigma}$ is a

rectangular diagonal $p \times k$ matrix and \mathbf{V} is the matrix of eigenvectors of the matrix $\mathbf{X}^T \mathbf{X}$. Then $\mathbf{Z} = \mathbf{U}\mathbf{\Sigma}$ are the principal components, while \mathbf{V} are the loadings.

Hence PCA could be seen as a regression problem, where scores are estimated by least squares criterion.

$$\hat{\beta} = \arg \min_{\beta} ||Z - \mathbf{X}\beta||^2 \quad (3.17)$$

Sparse Principal Components based on the SPCA criterion

Zou, Hastie and Tibshirani [24] use a *naive* elastic net to obtain a Sparse Representation for PCA. It requires a positive parameter λ for ridge penalty and a positive parameter λ_1 for LASSO penalty. Then we consider the following optimization problem:

$$\hat{\beta} = \arg \min_{\beta} ||Z_i - \mathbf{X}\beta||^2 + \lambda ||\beta||^2 + \lambda_1 ||\beta||_1 \quad (3.18)$$

Where $||\beta||_1 = \sum_{j=1}^p |\beta_j|$ is the 1-norm of β . Then $\hat{V}_i = \frac{\hat{\beta}}{||\hat{\beta}||}$ is the approximation of the i th loadings and $\mathbf{X}\hat{V}_i$ is the i th approximated principal component. Clearly, large enough penalties give a sparse representation.

But 3.18 depends on the results of PCA. Hence, Zou, Hastie and Tibshirani [24] presented a "self-contained" regression criterion to derive Principal Components. Let \mathbf{x}_i be the i th vector of the data matrix \mathbf{X} , then we can derive the whole sequence of principal components:

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \arg \min_{\mathbf{A}, \mathbf{B}} \sum_{i=1}^n ||\mathbf{x}_i - \mathbf{A}\mathbf{B}^T \mathbf{x}_i||^2 + \lambda \sum_{j=1}^k ||\beta_j||^2 + \sum_{j=1}^k \lambda_{1,j} ||\beta_j||_1 \quad (3.19)$$

$$\text{subject to } \mathbf{A}^T \mathbf{A} = I_{k \times k}$$

Then, $\hat{\beta}_j \propto V_j$ for $j = 1, 2, \dots, k$.

General SPCA Algorithm

The next step is to solve the optimization problem 3.19. Since there are two matrices, $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ to estimate, Zou, Hastie and Tibshirani [24] proposed an alternating algorithm. Before running the algorithm, we will show that

$$\sum_{i=1}^n \|\mathbf{x}_i - \mathbf{A}\mathbf{B}^T \mathbf{x}_i\|^2 = \|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^T\|^2 \quad (3.20)$$

Since \mathbf{A} is orthonormal, there exist an orthonormal matrix \mathbf{A}_\perp such that $[\mathbf{A}; \mathbf{A}_\perp]$ is $p \times p$ orthonormal. Then:

$$\begin{aligned} \|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^T\|^2 &= \|\mathbf{X}\mathbf{A}_\perp\|^2 + \|\mathbf{X}\mathbf{A} - \mathbf{X}\mathbf{B}\|^2 \\ &= \|\mathbf{X}\mathbf{A}_\perp\|^2 + \sum_{j=1}^k \|\mathbf{X}\alpha_j - \mathbf{X}\beta_j\|^2 \end{aligned} \quad (3.21)$$

1. To get \mathbf{B} given \mathbf{A} we should minimize:

$$\arg \min_{\mathbf{B}} \sum_{j=1}^k \{ \|\mathbf{X}\alpha_j - \mathbf{X}\beta_j\|^2 + \lambda \|\beta_j\|^2 + \lambda_{1,j} \|\beta_j\|_1 \} \quad (3.22)$$

\mathbf{A} is initialized with the loadings $\mathbf{V}[1 : k]$ calculated from the SVD of \mathbf{X} .

2. To get \mathbf{A} given \mathbf{B} the penalized part is not necessary and we should minimize only 3.21. The solution is obtained computing SVD for $(\mathbf{X}^T \mathbf{X})\mathbf{B}$ and set $\hat{\mathbf{A}} = \mathbf{U}\mathbf{V}^T$.
3. repeat steps 1 and 2 until convergence.
4. Normalize $\hat{V}_j = \frac{\beta_j}{\|\beta_j\|}, j = 1, \dots, k$.

Reconstruction of Profiles

SPCA algorithm estimates loadings, our goal is to provide a sparse representation of profiles. Now we will show how to get a representation centered by column means (means of the response values for each clone).

1. Calculate the scores on the sparse PCA: $\hat{B} = D\hat{V}$

2. Backtransform loadings for $i = 1, \dots, k$ Principal Components: $\hat{V}_i = W^{-1}V_i^*$
3. Calculate sparse profiles: $\sum_{i=1}^k V_i^* \hat{B}_i$

3.3.3 Cluster Analysis

The objective of a cluster analysis is to group the observations into a limited number of groups (clusters) so that the observations in the same cluster are similar, and observations in different clusters are dissimilar. From this description it may become clear that the definitions of similar and dissimilar are very important. In our thesis we will use a hierarchical clustering to group profiles, based on their wavelet coefficients. A distance matrix based on euclidean distances will be created, and we will present a dendrogram using Complete Linkage Clustering (the maximum of coefficients' distances) as distance function.

The Cluster Analysis in this context is useful to detect possible categories of profiles, and helps us to show that Wavelet Transform is invariant: it translates all the profiles in a different domain without changing relationships between profiles.

3.4 Results

In this section we will illustrate the performance of the wavelet denoising method introduced in section 3.3 with plots. We use three randomly chosen profiles from each group. The PDGFRA Group is represented in red, Gastric KIT in green and Non-Gastric KIT with a blue color. In Figure 3.3 we can see profiles 12, 22 and 39 as they appear after preprocessing and normalization. Note that the preprocessed data are very noisy, which makes it extremely difficult to interpret the profiles and confirms the strong need for denoising the profiles, first.

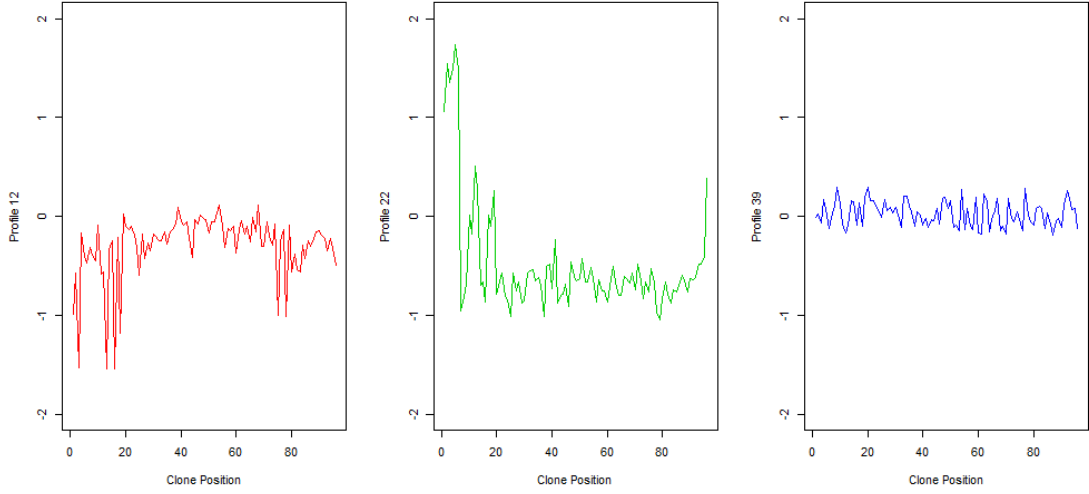


Figure 3.3: Profiles 12 (left), 22 (center) and 39 (right) represented after preprocessing and normalization

As introduced in Section 3.2.1 the discrete wavelet transform needs a dyadic data series. In our dataset we have $N = 97$ clones per profile, hence from now on we will suppress the last clone, *RP11-245B11* to obtain our Wavelet Transform. It is an affordable trade-off for the analysis, since that clone is affected by missingness in the 20 % of the profiles. Hereupon there are 96 coefficients and $L=5$.

First we will show Hard Thresholding In Figure 3.4 and Soft Thresholding in Figure 3.5.

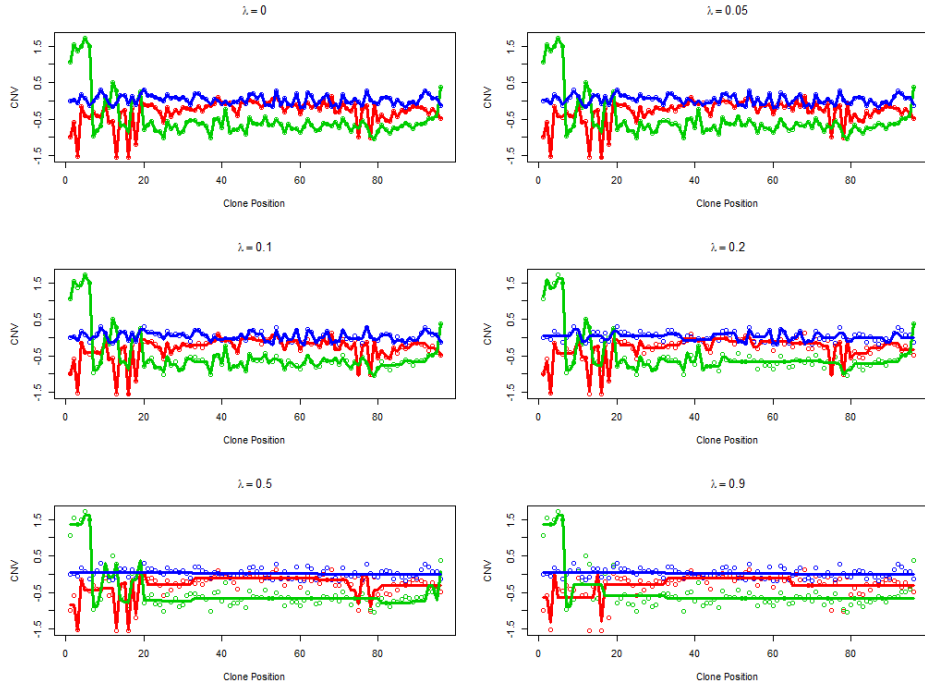


Figure 3.4: Profiles 12 (red line), 22 (green line), 39 (blue line) represented with Hard Thresholding. Lines are the obtained profiles, dots the raw data points. 6 different values of λ are selected and compared.

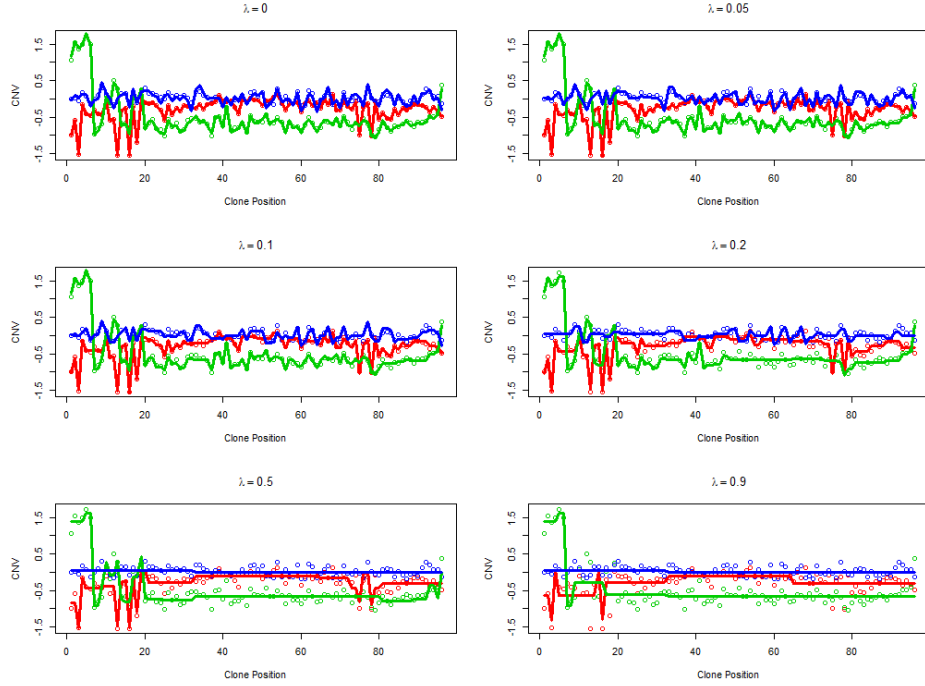


Figure 3.5: Profiles 12 (red line), 22 (green line), 39 (blue line) represented with Soft Thresholding. Lines are the obtained profiles, dots the raw data points. 6 different values of λ are selected and compared.

Obviously a larger λ gives rise to a more sparse representation. In this setting differences between hard and soft thresholding are minimal. The SPCA representation has another advantage in addition to the sparse representation of profiles. Biplots can be constructed which can be used to flag abnormal profiles. We produced a biplot using the first two sparse principal components. In Figure 3.6 it can be observed that the majority of abnormal profiles are belonging to the Non-Gastric Kit Group.

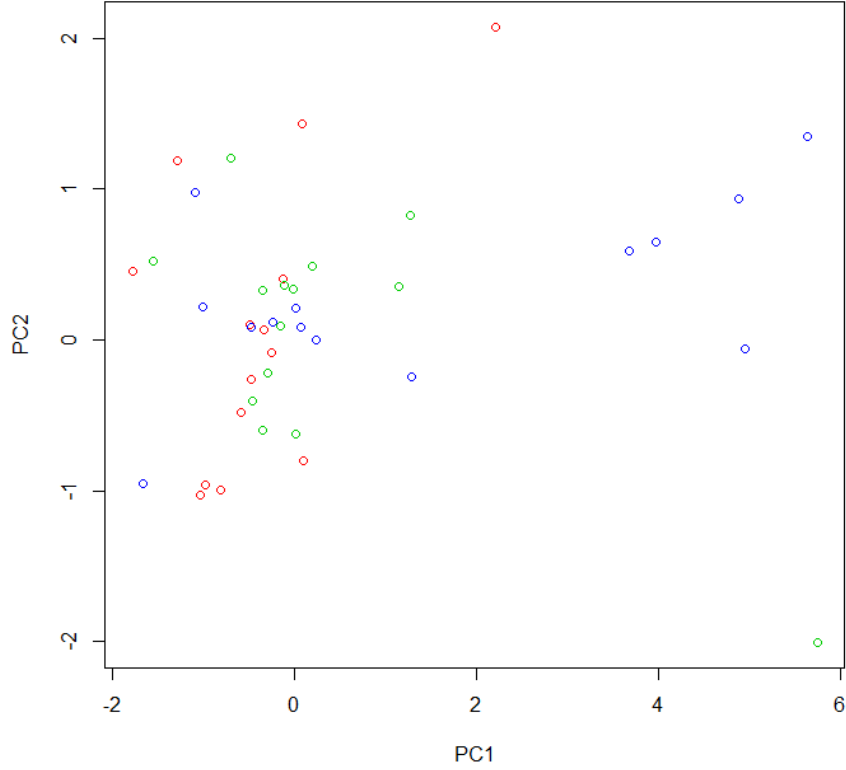


Figure 3.6: Scores of Principal Components for Wavelet Coefficients. Scores are red dots for PDGFRA profiles, green dots for Gastric KIT, blue dots for Non-Gastric KIT. PC1 expresses the 50.2 % of the variance, PC2 the 7.7%.

We will use an effective SPCA representation up to six principal components, with $\lambda = 10^{-6}$, while $\lambda_1 = 1$ for each PC. The SPCA function are a linear combination of wavelets and form a novel orthogonal basis. We will reconstruct the signal in the original space starting using the first principal component function and by expanding the basis with one additional PC function until we use all six PCs. Obviously, the precision of representation will increase with the number of PCs because they explained a larger part of the variance.

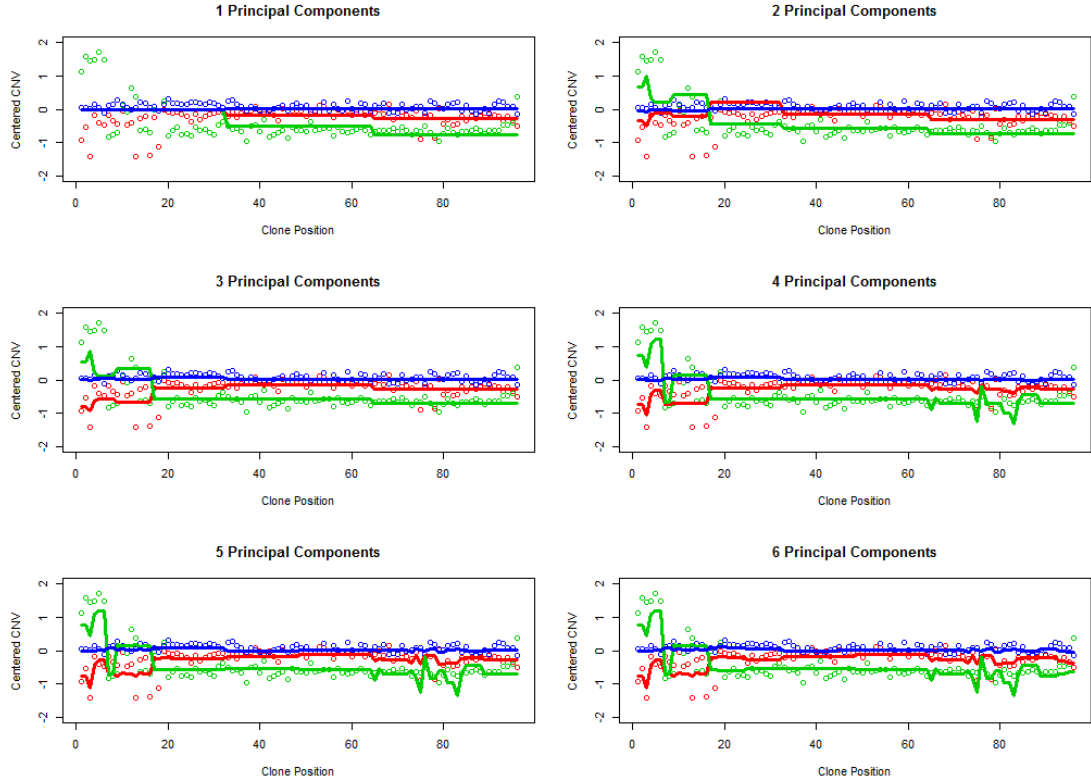


Figure 3.7: SPCA Representation for centered CNV expression for Profiles 12 (red line), 22 (green line), 39 (blue line) with a variable number of principal components, from 1 up to 6.

We conclude the Section on wavelet denoising with a Cluster Analysis using a distance matrix based on Complete Linkage between Wavelet Coefficients. The cluster analysis also shows evidence on linked abnormalities in Non-Gastric KIT Group, and gives us the opportunity to show another important property of wavelet coefficients: if we carry out the analysis on Raw Data, we will obtain exactly the same results. This could be expected because the wavelet transform is a variance preserving rotation of the data.

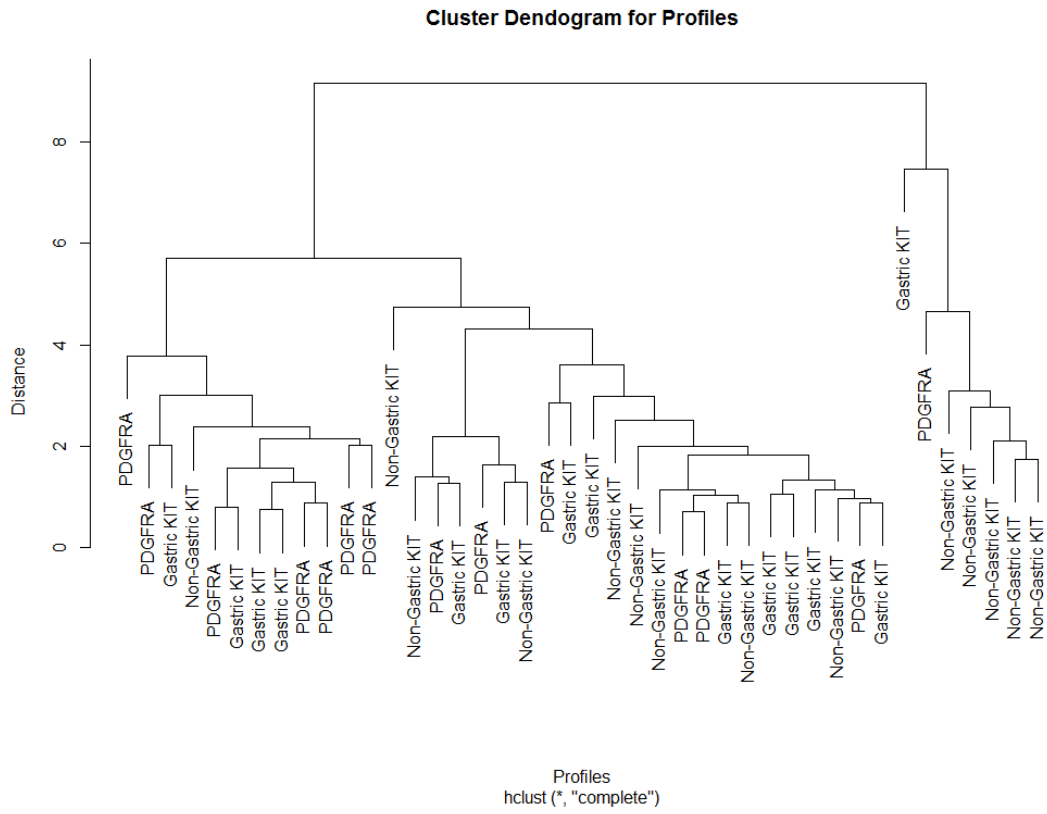


Figure 3.8: Cluster Analysis on Empirical Wavelet Coefficients of profiles.

Chapter 4

Wavelet Based Functional Models

4.1 False Discovery Rate

When the significance level is set to $\alpha = 0.05$ for N independent tests, the consequence is a probability of 0.05 to declare a test significant by chance under the null hypothesis.

Hence, the probability to declare at least one test significant for N independent tests is $1 - (1 - \alpha)^N$. This leads us to deal with a problem: if N is large, there will be for sure a huge number of false positive tests.

This problem usually arises in genomics, when thousands of simultaneous tests are carried out. Since the development of bioinformatics, many techniques have been provided to solve this problem. First the focus has been on methods of correction of the significance levels, controlling a quantity called **Family Wise Error Rate (FWER)** : FWER is defined as the probability to have at least one false positive test between the N tests. But, the FWER method has the disadvantage to be too conservative in the context of genomics. Benjamini and Hochberg introduced the false discovery approach in 1995. [26]. This approach controls the expected ratio of false discoveries on the total number of discoveries. In case of a considerable fraction of positive feature, the Benjamini-Hochberg FDR procedure provides a conservative estimate of the FDR. We first consider the different outcomes in the situation of multiple testing. Next we introduce the BH-FDR and the improved procedure of Storey and Tibshirani [25].

Definition

Table 4.1 gives the confusion table for a multiple testing problem with m simultaneous tests. Let V be the number of false positives, S the number of true positives and $R = V + S$ the total number of significant features. Then m_0 corresponds to the true-null or negative features and m_1 to the positive features.

	#declared non-significant	#declared significant	Total
#true null	U	V	m_0
#non-true null	T	S	m_1
Total	$m - R$	R	m

Table 4.1: Table of confusion for m tests

The p-values guarantee that the expected number of false positives $E[V] \leq 0.05m$, that is an expectation which is large in a typical genomics experiment with a vast amount of features m . In a large scale testing problem controlling the FWER will lead to conservative procedure. For biologists, a list of positive features that contains a small number of false positives is very useful to setup validation experiments. This rational has lead to controlling a measure which is related to the false positive rate:

$$\frac{\#false\ positive\ features}{\#significant\ features} = \frac{V}{V + S} = \frac{V}{R} \quad (4.1)$$

The **False Discovery Rate (FDR)** is basically the expected value of this quantity.

The FDR analogue for the p-value is the q-value, which is the the minimum false discovery rate at which a particular test can be called significant.

Benjamini-Hochberg procedure

To get more insight in the Benjamini-Hochberg procedure we first order the p-values from small to large, with (i) indicating the rank of the p-value. Hence, FDR is that proportion of significant features whose p-value is less or equal than some threshold t , with $0 < t \leq 1$. Denote the ordered p-values by p_1, \dots, p_m , hence:

$$V(t) = \#\{\text{null } p_{(i)} \leq t; i = 1 \dots m\} \quad R(t) = \#\{p_{(i)} \leq t; i = 1 \dots m\}. \quad (4.2)$$

and we want to estimate:

$$FDR(t) = E \left[\frac{V(t)}{R(t)} \right] \quad (4.3)$$

If the observed $R(t)$ (the number of observed p-values $\leq t$) is an easy estimate for $R(t)$, estimating $V(t)$ requires also the estimate of the number of truly null features m_0 . But it is more interpretable the estimate of the proportion of the truly null features, that we indicate with $\hat{\pi}_0 = \frac{m_0}{m}$.

In their seminal method Benjamini and Hochberg proposed to control the FDR at the α -level by setting $\pi_0 = 1$. Their approach proceeds as follows:

1. order p-values: $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$
2. find the value \hat{t} that $\hat{t} = \max\{t : p_{(t)} \leq \frac{t\alpha}{m}\}$
3. If \hat{t} exists, reject all null hypotheses corresponding to $p_{(1)}, \dots, p_{(\hat{t})}$.
4. If no such \hat{t} exists, accept all null hypotheses.
5. $\tilde{p}_{(i)} = \min_{j=i, \dots, m} \left(\frac{m}{j} p_{(j)}, 1 \right)$

Their method is a “step-up procedure” because it moves from small (less significant) to larger test statistics. Note, that BH-FDR assumes mutually independent tests. If the assumptions hold, it guarantees

$$FDR \leq \frac{m_0}{m} \alpha \leq \alpha \quad (4.4)$$

FDR estimation

If m_0 is known, a less conservative methods can be used by applying BH and controlling it on the level $\alpha m_0/m$. Note, that $\pi_0 = m_0/m$. Storey and Tibshirani, REF, developed an improved procedure by estimating the proportion of true null features π_0 .

Under the hypothesis that p-values are uniformly distributed, we can form a reasonable estimate, that involves the use of a tuning parameter λ .

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1, \dots, m\}}{m(1 - \lambda)} \quad (4.5)$$

Setting the parameter λ we select how many features are involved in the estimate of $\hat{\pi}_0$. The rational behind this estimating procedure is that the p-values of the null features are uniformly distributed in $[0, 1]$. Taking λ too large results in a conservative estimate for $\hat{\pi}_0$, and for small λ one runs the risk to include positive features in the estimate for π_0 . Storey and Tibshirani [25] proposed to calculate $\hat{\pi}(\lambda)$ at many different values for λ and to smooth $\hat{\pi}(\lambda)$ in function of λ . Then they proposed to use the smoothed estimate $\hat{\pi}(0.5)$.

Going back to 4.3, we replace the estimates in the formula:

$$F\hat{D}R(t) = \frac{\hat{\pi}_0 m \cdot t}{\#\{p_i \leq t\}} \quad (4.6)$$

Hence, the q-value is the minimum FDR that can be obtained when calling the feature i significant [25].

$$\hat{q}(p_i) = \min_{t \geq p_i} (F\hat{D}R)(t) \quad (4.7)$$

4.1.1 Bayesian FDR

The FDR can also be obtained from a Bayesian perspective. Efron [27] considered a mixture model (**two-groups model**) for the m hypothesis test: $F(z) = p_0 F_0(z) + (1 - p_0) F_1(z)$, with p_0 the marginal probability that a feature is negative and $1 - p_0$ the marginal probability that a feature is positive and the cumulative distributions of the corresponding z-value of the tests are $F_0(z)$ and $F_1(z)$ under

the null and the alternative hypothesis, respectively. The Bayes' rule yields the posteriori probability of a feature being in the null group given its z-value smaller than a threshold as:

$$Fdr(z) \equiv Pr \{null|Z \leq z\} = \frac{p_0 F_0(z)}{F(z)} \quad (4.8)$$

The researcher select a proper control level and finds the maximum z-value that satisfies the rule. First, The joint cumulative density $F(z)$ can be estimated as $\bar{F}(z) = \#\{z_i \leq z\}/N$, while the theoretical null distribution $N(0, 1)$ can be used for $F_0(z)$. Hence, $\bar{Fdr}(z_0)$ is the maximum value of z satisfying:

$$\bar{Fdr}(z_0) \leq q \quad (4.9)$$

Where q is the level to control the posterior probability. Note, that by setting p_0 to 1, a conservative estimate is obtained for the Bayesian FDR and that the Bayesian FDR becomes equivalent to the Benjamini-Hochberg approach.

4.1.2 Local False Discovery Rate

It is also possible to perform a similar derivation using densities. Hence the Bayes rule gives:

$$fdr(z) \equiv Pr \{null|Z = z\} = \frac{p_0 f_0(z)}{f(z)} \quad (4.10)$$

Which is also referred to as the bayesian **Local False Discovery Rate (LFDR)**. There is a straightforward connection between FDR and Local FDR. FDR is the mixture average of $fdr(Z)$ for $Z \leq z$:

$$Fdr(z) = E_f\{fdr(Z)|Z < z\} \quad (4.11)$$

The estimate of LFDR comes from an empirical Bayes approach. $f_0(z)$ could be assumed as the theoretical null, or better for simultaneous testing situations *empirical null* $N(\delta_0, \sigma_0^2)$. These parameters are estimated in combination with p_0 , using the zero assumption, stating that most of the z-values near 0 come from null tests for f_0 , the theoretical null distribution can be used. However, the massive

parallel structure of the large scale testing problem, also allows to estimate f_0 by the empirical null $N(\hat{m}u, \hat{\sigma})$. These parameters are estimated in combination with p_0 , using the assumption that most of the z-values near 0 come from null distribution. . Two different estimation methods are proposed by Efron [27], analytical and geometrical, but we will focus only on the analytical one, even if the `locfdr` algorithm developed in R shows both.

The analytic method assumes that the nonnull density $f_1(z)$ is supported outside a given interval $[a, b]$ containing zero. Define the Poisson P_0 :

$$P_0(\delta_0, \sigma_0) = \Phi\left(\frac{b - \delta_0}{\sigma_0}\right) - \Phi\left(\frac{a - \delta_0}{\sigma_0}\right) \quad (4.12)$$

and

$$\theta = p_0 P_0 \quad (4.13)$$

Then we get the desired estimates through the likelihood function of \mathbf{z}_0 , the vector of N_0 z-values in $[a, b]$. ($f_0(z) = \varphi(z)$).

$$f_{\delta_0, \sigma_0, p_0}(\mathbf{z}_0) = [\theta^{N_0} (1 - \theta)^{N - N_0}] \cdot \left[\prod_{z_i \in \mathbf{z}_0} \frac{\varphi_{\delta_0, \sigma_0}(z_i)}{P_0(\delta_0, \sigma_0)} \right] \quad (4.14)$$

The convolution of the two distributions belonging to the exponential family provides the following maximum likelihood estimates $(\hat{\delta}_0, \hat{\theta}_0, \hat{p}_0)$. Thus, To estimate fdr we only need to estimate $f(z)$ in 4.10

Now to estimate LFDR we need only $f(z)$ in 4.10. In the `locfdr` algorithm $f(z)$ is estimated by means of a standard Poisson GLM. The z-values are binned, giving counts $y_k = \#\{z_i \text{ in bin } k\}$ with $k = 1, \dots, K$. Hence the y_k are defined independent Poisson:

$$y_k \sim P_0(v_k), k = 1, \dots, K. \quad (4.15)$$

with v_k proportional to density $f(z)$ at midpoint x_k of the k th bin.

Hence, $f(z)$ could be modeled by an exponential family of j parameters:

$$f(z) = \exp \left\{ \sum_{j=0}^n \beta_j z^j \right\} \quad (4.16)$$

While the y_k counts are used for power diagnostic:

$$y_k^{(1)} = [1 - f\hat{d}r(x_k)]y_k \quad (4.17)$$

That indicates the nonnull counts, hence

$$E(f\hat{d}r^{(1)}) = \frac{\sum_{k=1}^K y_k^{(1)} f\hat{d}r_k}{\sum_{k=1}^K \hat{y}_k^{(1)}} \quad (4.18)$$

is the expected nonnull LFDR.

4.2 Functional Models

We will construct a wavelet functional model for inferring on the group mean profiles. The models allow for assessing and calling individual group mean profiles as well as on contrasts between the group profiles. There are two questions of scientific interest:

- 1) What are the significant aberrations for the groups?
- 2) Do the groups have significant different mean profiles?

Similar to Clement *et al.* [19], we estimate a linear model in the wavelet space:

$$D = XB^* + E^*, \quad (4.19)$$

with D the matrix of empirical wavelet coefficients, X a matrix with dummy variables for the group means, B^* the group mean profiles in the wavelet space and E^* the errors in the wavelet space, which are assumed to be i.i.d. normally distributed within each wavelet scale. The coefficients are regularised in the wavelet space and upon estimation, the estimated profiles are backtransformed to the original space. To infer on contrasts, two different methods are evaluated. On one hand, the contrasts are calculated in the wavelet space, and the wavelet coefficients of the contrasts are denoised and backtransformed. On the other hand, a similar regularization is imposed on the group mean profiles and contrasts are calculated after

backtransformation. Each procedure will lead to similar estimates, but they will differ in smoothness.

As mentioned in section 3.1, different thresholding techniques could be used. We will describe LFDR-based in 4.2.1, and MAP-based in 4.2.2.

4.2.1 LFDR-Based Thresholding

The linear model on the $j = 42$ profiles return estimates of the coefficients B^* and their standard error $SE(B^*)$. These estimates could be used to provide a simple t-test $t_{i,k} = \frac{B_{i,k}^*}{SE(B_{i,k}^*)}$ $i = 1, \dots, 96$, $k = 1, 2, 3$. and $j - k$ degrees of freedom. Hereupon the degrees of freedom are $42-3=39$. These $t_{i,k}$ tests are converted to z-values, hence LFDR is calculated as described in 4.1.2. We will use the LFDR values to retain few coefficients, and wavelet father coefficients will not be considered but automatically retained for the reasons mentioned in 3.3.1.

There is not a general rule for the selection of the cutoff point for retaining coefficients or shrink them to zero, and it depends on the level of smoothing the researcher wants to obtain. For CNV data, a reasonable cutoff point is obtained at $q = 0.5$. Hence:

$$\delta_{0.5}(B^*) = \begin{cases} B_{i,k}^* & fdr(z_{i,k}) \leq 0.5 \\ 0 & fdr(z_{i,k}) > 0.5 \end{cases} \quad (4.20)$$

Once selected the estimated coefficients of interest, it is possible to backtransform in the original domain with the inverse wavelet function and obtain the group mean estimates.

$$\hat{Y}^* = (W^T B^*)^{-1} \quad (4.21)$$

where \hat{Y}^* is an $i \times k$ matrix. k group mean profiles will be obtained and the retained group coefficients in the wavelet space will drive to a piecewise constant representation where joined "pieces" are mentioned as **Segments**.

4.2.2 MAP Thresholding

One of the drawbacks of the thresholding methods is the presence of tuning parameters. Figueiredo [22] proposes an alternative approach, using empirical Bayes estimation imposing a Jeffreys' noninformative prior on the variance parameter. Returning to the 3 steps described in 3.2, in the second step it will be obtained a Bayes Estimate of \hat{B}^* .

\hat{B}^* is interpreted as the minimizer of the a posteriori expected loss, then:

$$\hat{Y}^* = W^{-1} \arg \min_{\tilde{B}^*} \int L(B^*, \tilde{B}^*) p(B^*|D) dB^* \quad (4.22)$$

Where $L(.)$ is a loss function that minimize the discrepancy between the parameter and any possible estimate. It is proved [22] that the estimate of \hat{Y}^* correspond to a Bayesian criterion in the signal domain and the loss adopted corresponds to $L(Y^*, \tilde{Y}^*)$ under orthogonal transformations as the Wavelet.

The 0/1 loss leads to the MAP (*Maximum a Posteriori*, the mode of the posterior distribution) criterion:

$$L_{0/1}(B^*, \tilde{B}^*) = L_{0/1}(WY^*, W\tilde{Y}^*) = L_{0/1}(Y^*, \tilde{Y}^*) \quad (4.23)$$

because this loss function is invariant under orthogonal transformations.

Thus, the inverse DWT estimate of the coefficients is the MAP estimate of the signal domain.

For the decorrelation property, we model the coefficients as mutually independent. We approximate decorrelation to independence [22] to state that also *a posteriori* coefficients are mutually independent. Hence, under the MAP criterion, coefficient can be computed separately:

$$\hat{B}_{MAP}^* = \arg \max_D p(B^*|D) = \left[\arg \max_{D_1} p(B_1^*|D_1), \dots, \arg \min_{D_N} p(B_N^*|D_N) \right] \quad (4.24)$$

where $N = IK$.

The estimate comes from a hierarchical bayesian model, where coefficients are zero-mean gaussian ($B^*|\phi^2 \sim N(0, \phi^2)$) and the improper Jeffreys' prior $p(\phi^2) \propto$

$1/\phi^2$ express amplitude scale invariance, that means that the inference procedure is invariant under changes of amplitude scale.

The hierarchical bayesian model allows the use of an empirical Bayes technique for the estimation of B^* and ϕ^2 . Consider $D|B^* \sim N(B^*, \sigma^2)$:

$$\hat{\phi}^2 = \left(\frac{D^2}{3} - \sigma^2 \right)_+ \quad (4.25)$$

With estimates to zero when $(.) \leq 0$, Thus:

$$\hat{B}^* = \frac{\hat{\phi}^2}{\hat{\phi}^2 + \sigma^2} D \quad (4.26)$$

Notice there is any tuning parameter, as mentioned above.

4.3 Results

In this section we will first explain the practical solution to the missingness problem. We will calculate wavelet coefficients of the matrix D using multiple imputation. Missingness does not affects heavily estimations, for the following regions:

- 1) Missingness affects only 12 Clone positions of 4032;
- 2) Wavelet Coefficients are not directly used for final inference but preprocessed with a strong smoothing and backtransforming in all models. Thus, final estimates of Y^* will not be influenced by the technique of imputation used.
- 3) Moreover, missingness is mainly introduced by an improper preprocessing and can be avoided in practice.

However, the Y_{ik}^o clone vectors are assumed as normally distributed, with mean τ_{ik} and variance ϖ_{ik} . Hence we will simulate Y_{ik}^m from a random normal for $M = 10$ times and proceed with the algorithm, obtaining the matrix D .

Then, we will estimate the group mean model following LFDR-Thresholding procedure. Following the multiple imputation procedure in 2.3.3, we obtain a matrix D of Empirical Wavelet Coefficients. We can use this matrix for the group

mean model, using the R function `lmFit`. With the 96×3 estimates of coefficients and standard error we can obtain the z-values necessary for LFDR thresholding.

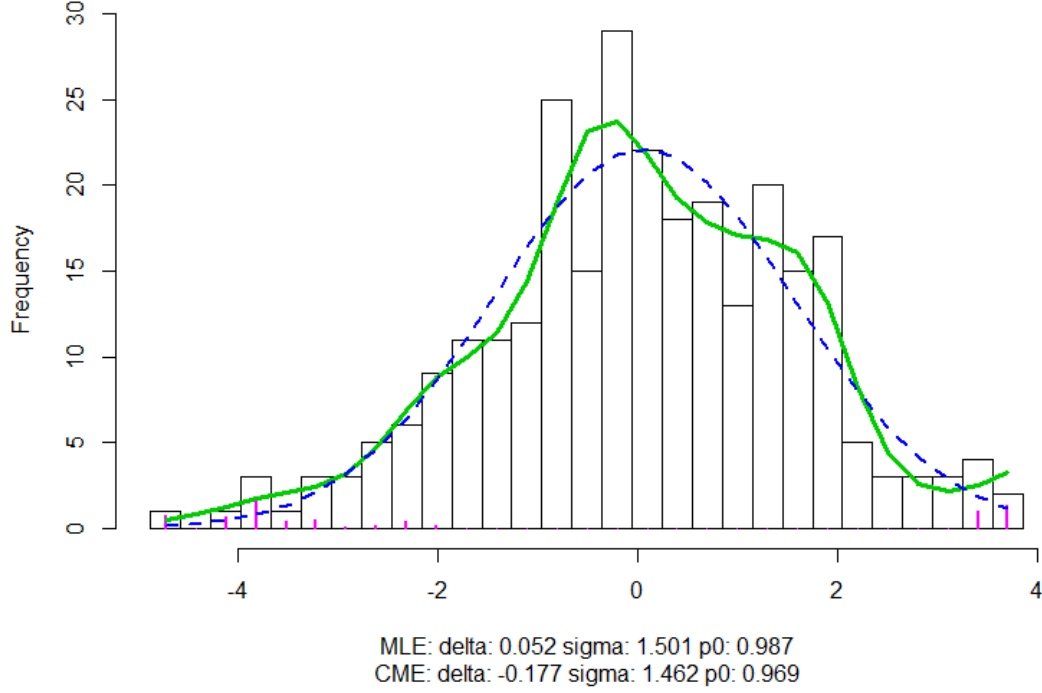


Figure 4.1: Histogram of LFDR for the group mean model. Green line is the estimated $d(z)$, while blue line is the empirical null. Pink bars are the estimated nonnull counts. The horizontal axis is labeled by z-values.

The Figure 4.1 shows the estimate of $f(z)$ cited in equation 4.16 with the green line, while the blue dotted line is The empirical null. The estimate of $f(z)$ is done with 10 parameters for the exponential family and space is binned in $k=30$ bins. The `locfdr` algorithm has estimated $\hat{\delta}_0=0.052$ and $\hat{\sigma}_0=1.501$, giving an empirical null f_0 similar to the theoretical in mean but larger in variance. $\hat{p}_0 = 0.987$ indicates a strong influence on $fdr(z)$ of the empirical null. Pink bars are the estimated quantities of nonnull counts ($\hat{y}_k^{(1)}$). The measure of $E(\hat{fdr}^{(1)})$ could be used as comparison between different thresholding. The bigger is this expected value, the more profiles will result smoothed. In this case $E(\hat{fdr}^{(1)}) = 0.513$.

Using the threshold discussed in 4.2.1 we can backtransform group mean estimates and their standard errors and obtain the plot of group mean profiles. As we see in plot 4.2 and 4.3 in this model does not seem group mean estimates will result significantly different and it could be that there are not significant CNV aberrations in all the group profiles. In Figure 4.3 same sparsity has been imposed as additional constraint, shrinking $B_{i,k}^*$ to 0 if $\min_i[fdr(z_{i,k})] > 0.5$. The same sparsity imposed with the minimum LFDR for the i -th position increases the total number of segments considered. Indeed, in Figure 4.2 there are 24 segments among the groups, while 48 are the segments identified in Figure 4.3.

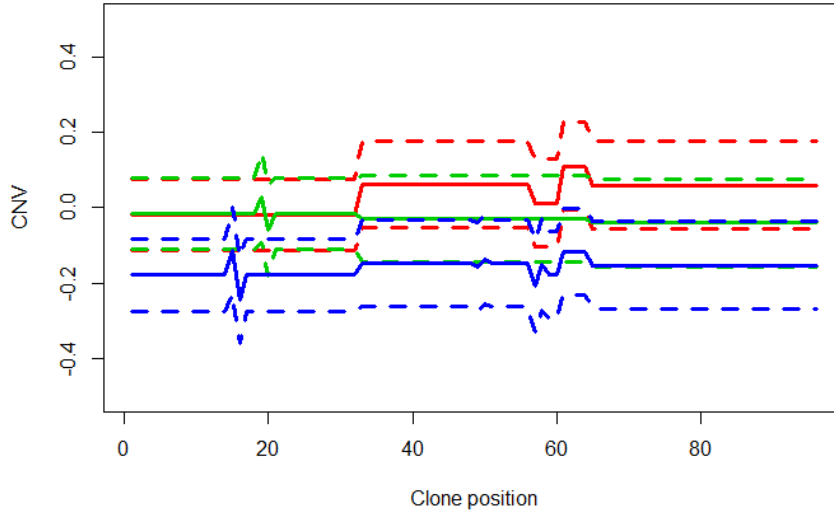


Figure 4.2: Group Profiles estimated with LFDR thresholding. Red is PDGFRA estimate, green is the gastric KIT, blue is the non-gastric KIT. Dotted lines are the confidence intervals.

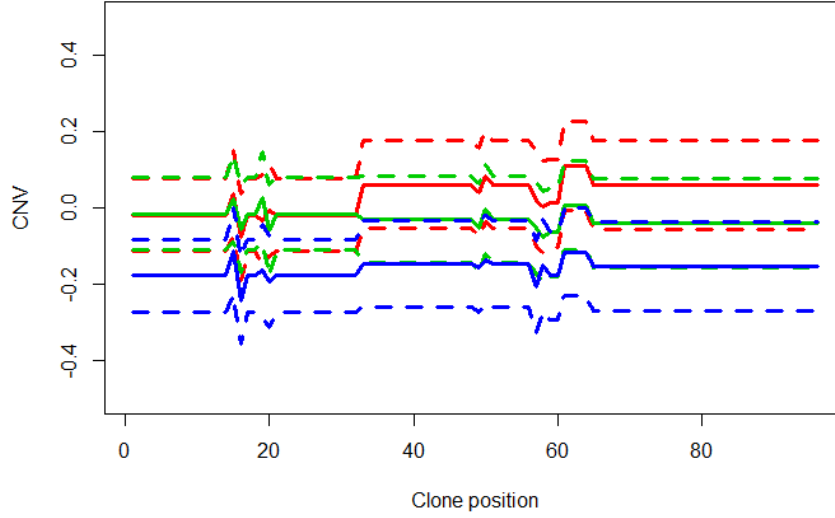


Figure 4.3: Group Profiles estimated with LFDR thresholding imposing the same sparsity.

The Model based on MAR algorithm needs an estimate of σ^2 . In these study we propose **MAD (Mean Absolute Deviation)** as contribute. Mean absolute deviation is a robust measurement for the variability. Indeed, compared to standard deviation, it is less sensible to outliers. Given a sample X with J values, MAD is:

$$MAD = \text{median}[X_j - \text{median}(X)] \quad (4.27)$$

MAD for a matrix of coefficients in the wavelet domain is calculated among the coefficients of the wavelet space representing the means of adjacent observations. MAP Thresholding seems not sufficiently able to remove wave bias, as we see in Figure 4.4.

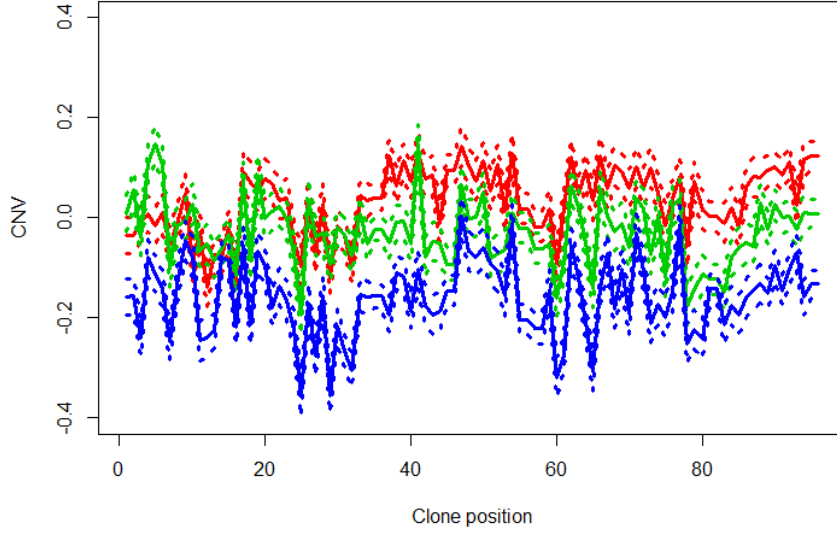


Figure 4.4: Group Profiles estimated with MAP thresholding. Red is PDGFRA estimate, green is the gastric KIT, blue is the non-gastric KIT. Dotted lines are the confidence intervals.

Two approaches based on contrasts between profiles will be discussed. A contrast matrix could be applied on backtransformed matrix Y^* with the same sparsity imposed or the same matrix could be imposed to the group mean estimates in the wavelet space and then the obtained contrasts between estimates in the wavelet space could be thresholded using LFDR technique and then backtransformed. It is possible to obtain three contrasts' combinations, to which a "contrast profile" will correspond. From now on dots and lines in cyan will be for contrasts between PDGFRA and Gastric KIT, in pink for contrasts between Gastric and Non-Gastric KIT, and in yellow for contrasts between PDGFRA and Non-Gastric Kit. In Figure 4.5 we see the profiles deriving from the first approach used with LFDR Thresholding and in Figure 4.6 with MAP thresholding, while Figure 4.7 shows the results of the second approach. Contrasts profiles were smoothed using LFDR Thresholding in Figure 4.8. Looking to plots we can notice that PDGFRA and Gastric KIT balance themselves on the whole profile, while they show contrasts with Non-Gastric KIT. Moreover, after the 30th clone position the contrast

between KIT and Non-Gastric KIT turns similar to PDGRA and Gastric KIT, while the contrast of PDGFRA and Non-Gastric KIT appears larger. MAP still catches noisy profiles.

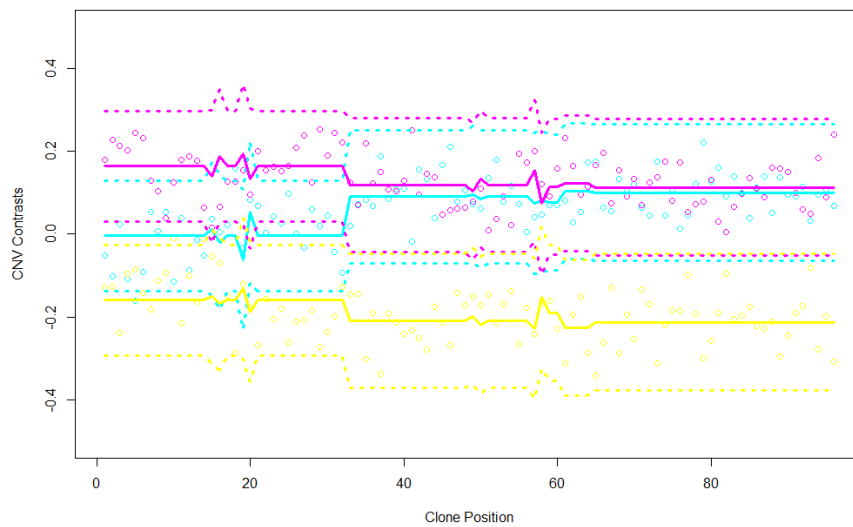


Figure 4.5: Contrast Profiles with their confidence intervals calculated in data domain after LFDR thresholding. Cyan indicates PDGFRA vs. gastric KIT, pink is gastric KIT vs. non-gastric KIT, yellow is PDGFRA vs. non-gastric KIT. Dots are the raw data contrasts.

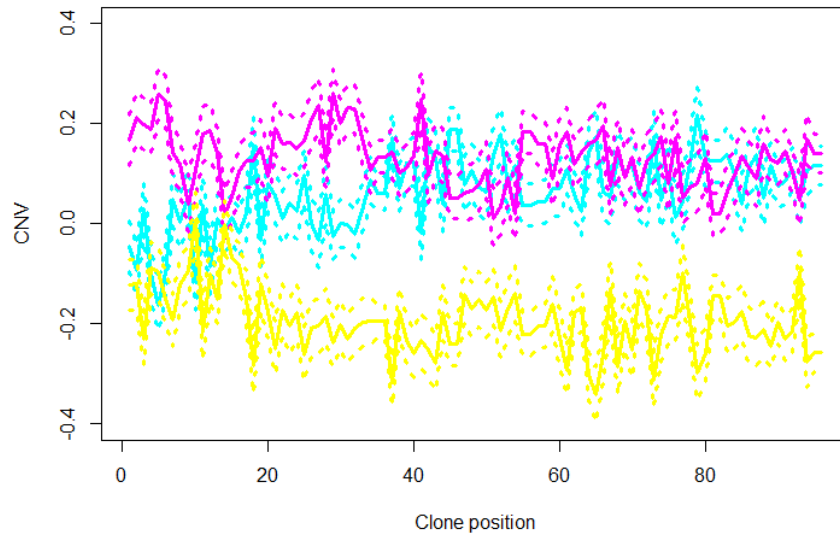


Figure 4.6: Contrast Profiles with their confidence intervals calculated in data domain after MAP thresholding.

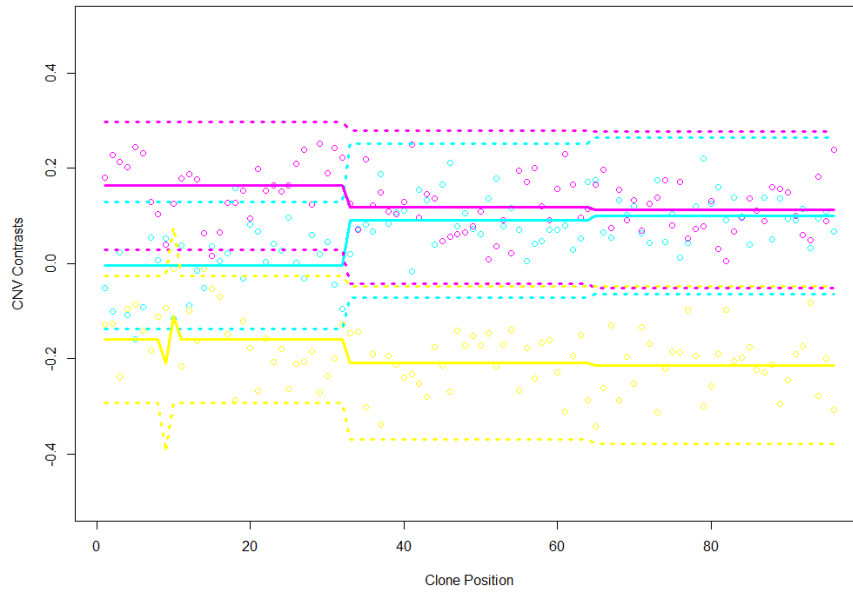


Figure 4.7: Contrast Profiles with contrasts calculated in wavelet domain and backtransformed with LFDR thresholding.

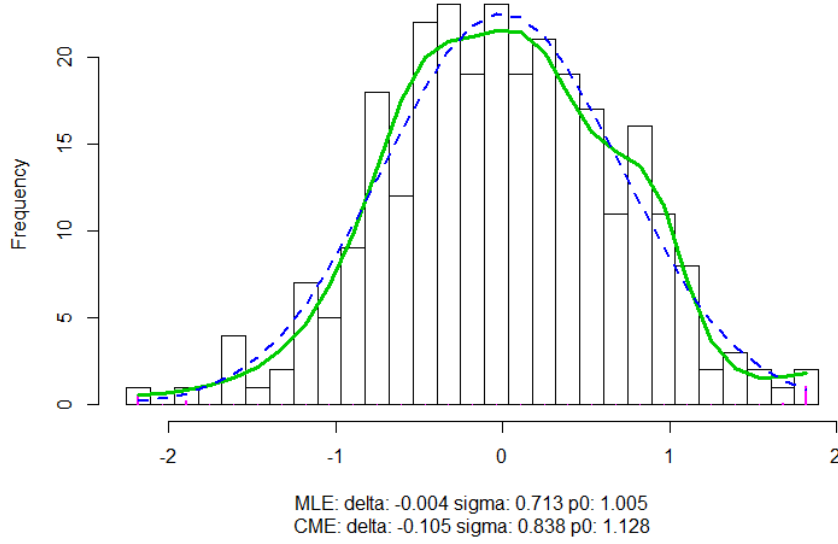


Figure 4.8: Histogram of LFDR for wavelet contrast coefficients z -values. Green line is the estimated $f(z)$, blue line is the empirical null, pink bars are the estimated nonnull counts.

In Figure 4.7 we see stronger smoothing than in Figure 4.5 (same number of parameters for $f(z)$ and bins have been imposed). The reason is clear in Figure 4.8: a limited number of nonnull counts is identified, hence a limited number of coefficients is retained imposing strong smoothness also $E(\hat{f}dr^{(1)}) = 0.552$, greater than the previous case.

4.4 Multiple Testing

We are in the field of the multiple testing, but selecting which technique is the more appropriate for testing is not straightforward, since the limited and variable number of tests done. In the functional model, group case or contrasts case, theoretically we could have $96 \times 3 = 288$ different tests. These mentioned above are segmentation models and since we have segmented profile, it makes sense testing segments rather than clones. Hence, we have a variable and relatively small number of tests for models with LFDR thresholding: in the functional model we have 24

segments, that turn to 48 if we impose the same sparsity, while for the contrast models turn to 57 in the first case and down to 17 in the second case. Even if the number of tests depend of the kind of test done (in group mean functional model testing on differences between profiles involves automatically more tests than the case with same sparsity) FDR seems a suitable approach for all the functional models used.

We will not provide tests for MAP. The level of sparsity reached was not enough to obtain segments, thus we do not have segments in practice and testing will be not useful for biological applications. Referring to questions mentioned at the beginning of section 4.2, we will answer to them model per model, when it is possible. Complete results will be shown in the appendix.

CNV Aberrations

CNV= 0 indicates that sample intensity balances the reference, hence it is straightforward to test the presence of significant copy number gains/losses as indicated in 4.28.

$$\begin{cases} H_0 : Y_{ik}^* = 0 \\ H_1 : Y_{ik}^* \neq 0 \end{cases} \quad (4.28)$$

This hypothesis test is considered for the following cases.

- *Group Profiles with LFDR thresholding imposing different sparsity* : a simple z-test converted to FDR is applied to segments. It results almost the whole Non-Gastric KIT group suffering Copy Number Loss. Only FDR values in $Y_{15,3}^*$ and $Y_{61-64,3}^*$ result non-significant, while the other two profiles does not present significant aberrations.
- *Group Profiles with LFDR thresholding imposing same sparsity*: The number of tests is doubled but the results are not that different. The only difference is the presence of an additional non-significant test in $Y_{50,3}^*$.

Differences between groups

The differences between groups are tested with two similar approaches based on a contrast matrix. One can test if the segment estimate of a clone position in a specific group a is significantly different to the corresponding clone positions on the other groups, hence the hypothesis tested is:

$$\begin{cases} H_0 = Y_{ia}^* - \frac{\sum_{k \neq a} Y_{ik}^*}{k-1} = 0 \\ H_1 = Y_{ia}^* - \frac{\sum_{k \neq a} Y_{ik}^*}{k-1} \neq 0 \end{cases} \quad (4.29)$$

Similarly, one could be interested in differences between two groups a and b in the i -th clone position.

$$\begin{cases} H_0 = Y_{ia}^* - Y_{ib}^* = 0 \\ H_1 = Y_{ia}^* - Y_{ib}^* \neq 0 \end{cases} \quad (4.30)$$

In both cases a row of the appropriate contrast matrix $L_{k \times k}$ designed to reduce test to linear hypothesis is considered. Σ indicates the matrix of the errors of the estimates Y^* .

$$t_{i,k} = \frac{L_k Y_i^*}{L_k \Sigma_i L_k^T} \sim \chi_1^2 \quad (4.31)$$

Then, we consider the first hypothesis for the estimates of the group mean profiles and the second hypothesis for the estimates of the contrasts profiles.

- *Group mean Profiles with LFDR Thresholding imposing different sparsity:* Any test resulted significant already in the χ^2 domain. Hence FDR is not necessary.
- *Group mean Profiles with LFDR thresholding imposing the same sparsity:* Same situation of the previous case.
- *Contrast Profiles in original domain with LFDR thresholding :* Even in this case, FDR will not be necessary because all test are already not significant in the χ^2 domain.

- *Wavelet Contrasts Profiles*: Here the contrast were applied to coefficients estimated in the wavelet domain, hence thresholding is provided. But tests results still non-significant and that is what we expected, since the only difference between these profiles and the profiles calculated applying a contrast matrix to final estimates is the level of smoothing reached and wavelet contrast profiles are more smoothed.

Chapter 5

Wavelet Based Mixed Models

5.1 Mixed Model

Among the extensions of the linear models, mixed models allow to incorporate random effects. Mixed models are useful in presence of profiles, e.g. for LDA¹ and FDA: with mixed models it is possible to decompose the signal of one specific profile in fixed effect for all the profiles (or for one subset) and subject-specific random effect.

Indicating with j the j -th profile, the general linear mixed model is described as:

$$Y_j = X_j\beta + Z_jb_j + \epsilon_j$$

Basically, it is the combination of a two-stage model:

$$\begin{cases} Y_j = Z_j\beta_j + \epsilon_j \\ \beta_j = K_j\beta + b_j \end{cases}$$

Where the first stage describes a linear regression model for each profile, while the second stage explains variability of the regression coefficients in the first stage.

These are the notations used:

¹Longitudinal Data Analysis

- $Y_j = (Y_{j1} \dots Y_{jn_i})$: Response for the j th subject with n_i covariates
- Z_j : matrix ($n_i \times q$) of known covariates
- K_j : matrix ($q \times p$) of known covariates
- X_j : product of the matrices Z and K
- β_j : q -dimensional vector of the subject regression parameters
- β : p -dimensional vector of the unknown regression parameters
- b_j : vector of the random effects, normally distributed with mean 0 and covariance matrix G
- ϵ_j : error term

Mixed models have the interesting property that they allow for conditional inference and marginal inference. The conditional model has a hierarchical interpretation.

$$Y_j | b_j \sim N(X_j \beta_j + Z_j b_j, \Sigma_j)$$

Marginally, Y_j is distributed as:

$$Y_j \sim N(X_j \beta_j, ZGZ_j^T + \Sigma_j)$$

Hence, it is possible to distinguish two components of the variance. The first component describes the variance between profiles; the second one is the variance within. A large variance between profiles compared to the variance within profiles indicates that there is a considerable variation between the individual profiles that belong to a certain treatment group, indicating the need for the incorporation of the random effects.

5.2 Functional Mixed Model

It is clear that three different kind of tumors could have a different trend in copy number variations, but there can also be important deviations from the group

mean profiles in individual profiles. Hence, a study of single profiles might be very interesting in the light of personalized medicine. We build a mixed model with a fixed group effect and a random subject specific effect for each profile.

5.2.1 Marginal Model

The mixed model has a great advantage in his structure. Indeed, the signal coming from each profile is decomposed in group effect and subject-specific effect using appropriate design matrices. Data will be modeled in the wavelet domain and sparsity will be imposed using LFDR thresholding as criterion for the coefficients :

$$D_{j(i,k)} = X_j B_{(i,k)}^* + Z_j b_j + \epsilon_{j(i,k)} \quad (5.1)$$

Where X is the design matrix for groups and Z is a $N \times N$ identity matrix. Marginally, $D_{j(i,k)}$ is distributed as

$$D_{(j,k)} \sim N(X_j B^*, Z_j G Z_j^T + \Sigma_{j(i,k)}) \quad (5.2)$$

Hence, the functional mixed model shares the same mean with the linear model. In this case it is more interesting to observe the differences within the groups: the variance of each subject is decomposed in error term (Σ) and subject-specific effect ($Z G Z^T$, or just G since Z is an identity matrix). The matrix G is a diagonal matrix, since it is made the strong assumption that there is no correlation between the copy number variation of the coefficients.

5.2.2 Prediction

BLUP Estimator

Best Linear Unbiased Prediction (BLUP) estimation is provided for this model. The BLUP minimizes the expected prediction error, with a least squares criterion for B^* that provides the same estimates of the functional model in 4.2. But this model furnish also en estimate of the random effect. It is possible to estimate all

the parameters together [28]:

$$\hat{\theta} = \begin{bmatrix} \hat{B}^* \\ \hat{b} \end{bmatrix} = (C^T C + \sigma_\epsilon^2 T)^{-1} C^T D \quad (5.3)$$

With

$$C = [XZ] \quad (5.4)$$

and

$$T = \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} \end{bmatrix} \quad (5.5)$$

Estimation of Covariance Matrix

Covariance matrix estimates are:

$$\hat{\Sigma}_{\hat{\theta}} = (C^T C \frac{1}{\sigma_\epsilon^2} + \hat{T})^{-1} \quad (5.6)$$

Variance Components estimates

The variance estimation is not straightforward. As there are no replicate profiles available we only dispense of one wavelet coefficient per sample. We can however borrow strength among the observations within the same profile to check a relevant presence of the within profile variance, e.g. MAD estimator in the functional mixed model can be used for that scope, as constraint to select clone position estimates with the presence of within variance. One estimator of the within variance is the difference between the total variance and the error term.

MAD is estimated at the most detailed wavelet space of D (hence, the first N/2 columns of the matrix) as the error term σ_ϵ , while σ_i^2 is the variance estimator obtained in the functional model of the chapter 4.

Hence, to every i is imposed the following constraint

$$|\hat{\sigma}_i^2 - \hat{\sigma}_\epsilon^2|_+ \quad \hat{\sigma}_\epsilon = MAD(D_{[1:N/2]}) \quad (5.7)$$

If σ_i^2 is greater than σ_ϵ^2 , random effects exist and will be estimated as in 5.3 and 5.2.2. Otherwise, there is no random effect for the i -th clone and the common least squares estimates will be calculated.

$$\hat{B}^* = (X^T X)^{-1} X^T D \quad (5.8)$$

$$\hat{\sigma}^2 = (X^T X)^{-1} \sigma_\epsilon^2 \quad (5.9)$$

LFDR Thresholding

LFDR Thresholding imposed to Wavelet Mixed Model to alter coefficients will be similar to the thresholding discussed in 4.2.1. Some slight differences need to be discussed.

We will provide separated thresholds for estimates of the group coefficients and estimates of the random effects. Estimates of the group coefficients are theoretically the same as in 4.3, but minor changes are produced by the constraint commented above, hence the resulting group estimates could appear slightly different (or not). One can think to sum the estimates of group mean and random effects in the wavelet domain and then threshold coefficients and represent profiles, but this operation could erase the advantages of the mixed model, since it will not be possible to make inference on random effects anymore. Hence, we preferred to threshold separately group mean estimates and mixed effects estimates, obtaining respectively Y^* and y^* , then reconstruct profiles summarizing.

5.3 Results

We will show how the group mean profiles estimates change in the mixed model, hence how we can represent profile in a more intriguing way. We can see that LFDR histogram in Figure 5.1 for group mean estimates is different from figure 4.1 and the estimate of $\hat{\sigma}_0 = 1.339$. $E(\hat{f}dr^{(1)}) = 0.528$ indicates a slightly stronger thresholding than the functional model.

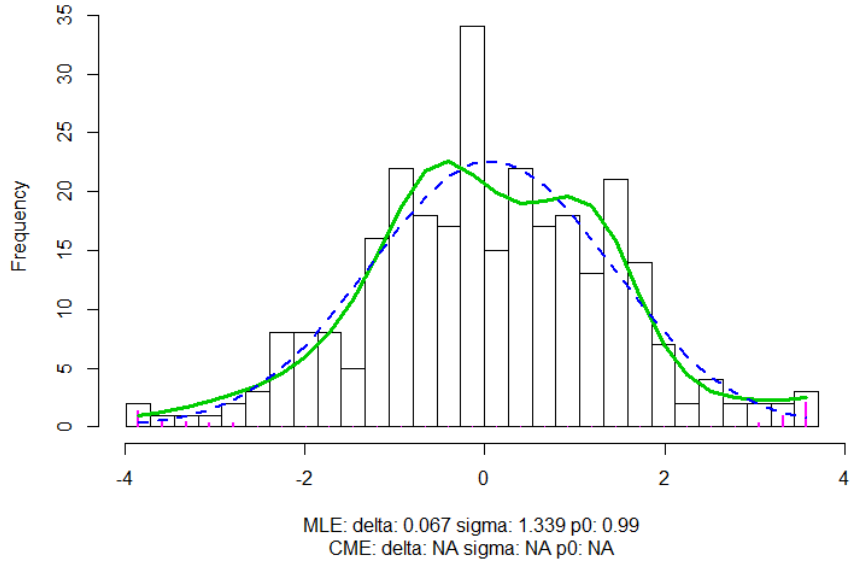


Figure 5.1: LFDR for group effect estimates calculated with a mixed model on wavelet coefficients. z-values dispersion is represented by the histogram, green line is the estimated $f(z)$, blue dotted line is the empirical null, pink bars are the estimated nonnull counts.

Figure 5.2 shows that Non-gastric KIT profile has some slight differences compared to Figure 4.2, more than the other profiles. This is an indication of what will be clearer with the representation of the random effects, i.e. the non-gastric KIT group benefits more than the other groups of the mixed model and the representation of his profiles will be more influenced by the presence of random effects.

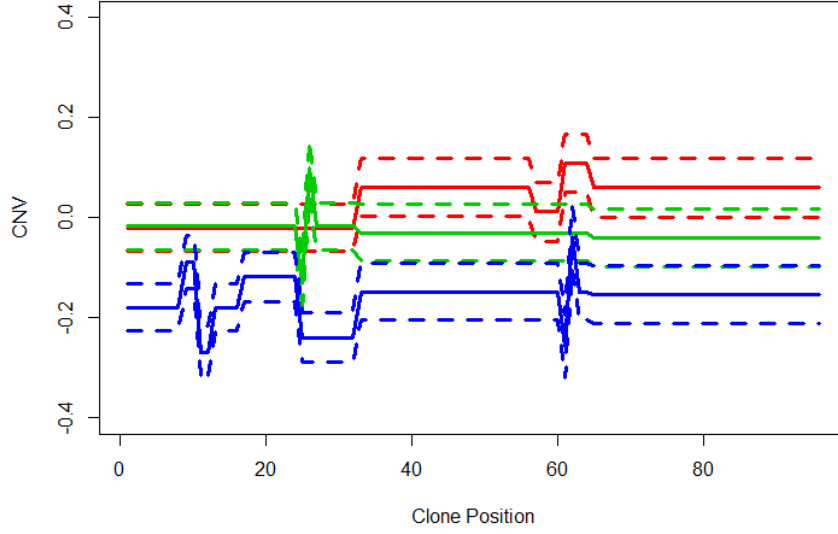


Figure 5.2: Group Profiles of PFDGRA (green), gastric-KIT (red), non-gastric KIT (blue). The dotted lines are the confidence intervals.

The second step consists of calculating the LFDR for the random effects in the wavelet space. Subsequently the random effects can be denoised using the LFDR, . We can notice in Figure 5.3 that LFDR increases the number of simultaneous tests results more similar to FDR: The histogram of z-values tends to be distributed as a normal and $\hat{p}_0 = 0.851$. We will not use LFDR for inference, since results of LFDR are valid when the number of tests is in the order of thousands.

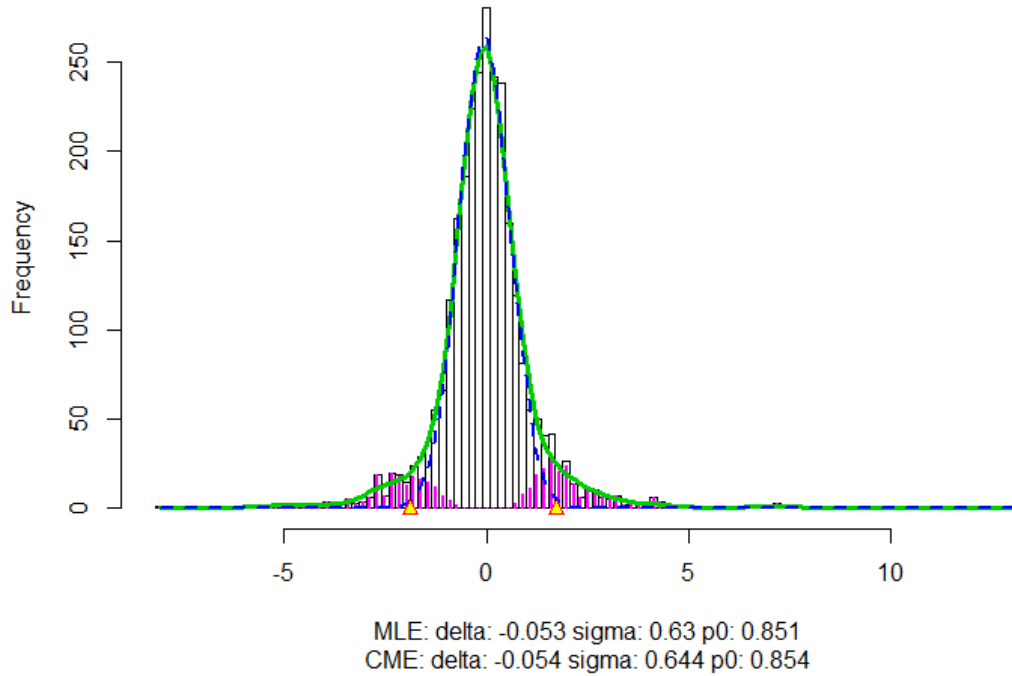


Figure 5.3: HLFDR for random effect estimates calculated with a mixed model on wavelet coefficients.

Now we can show how to deal with profiles using a mixed model. We will show the estimated profiles first, and the decomposition of profiles in group effect and mixed effect in a second time. The group effect will be more clear now. For our purpose we will take profiles already used in chapter 3.4, to compare results obtained in that section with these results.

In contrast with Figure 3.3 of raw data, the modeled Profiles 12, 22 and 39 in Figure 5.4 now provide an estimate of the subject-specific effect without capturing too much noisy details. In Figure 5.5 we depict the group effects and the captured deviations from the group mean profile.

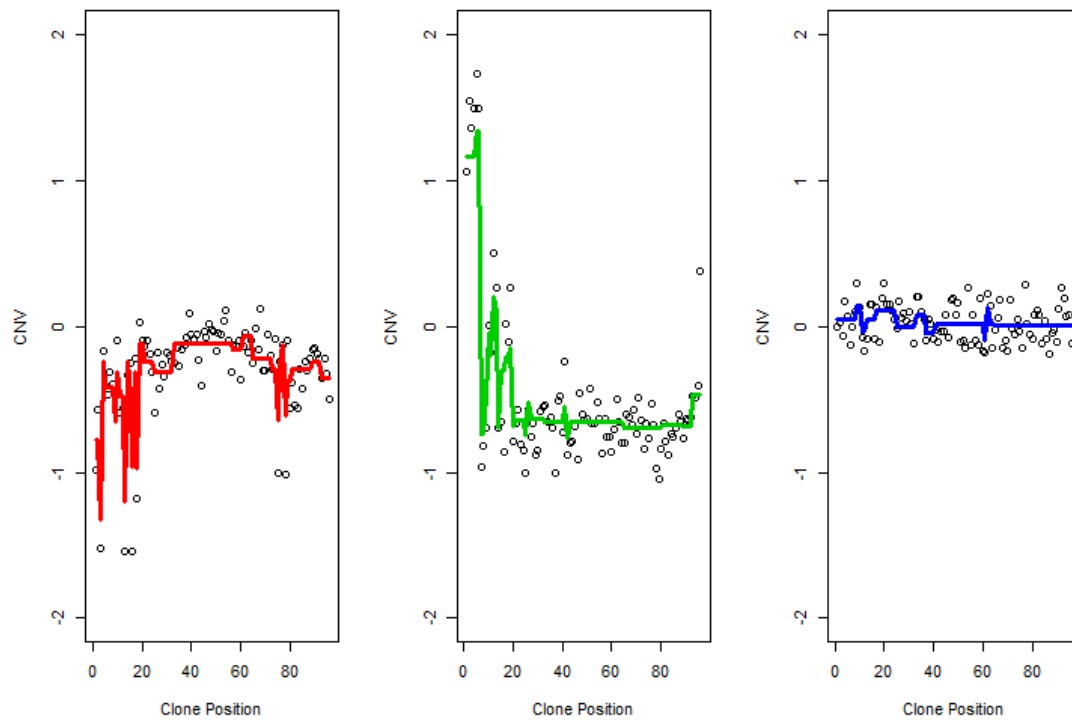


Figure 5.4: Profiles 12 (left), 22 (center), 39 (right) after processing of wavelet coefficients with a mixed model. Dots are the Raw Data Points

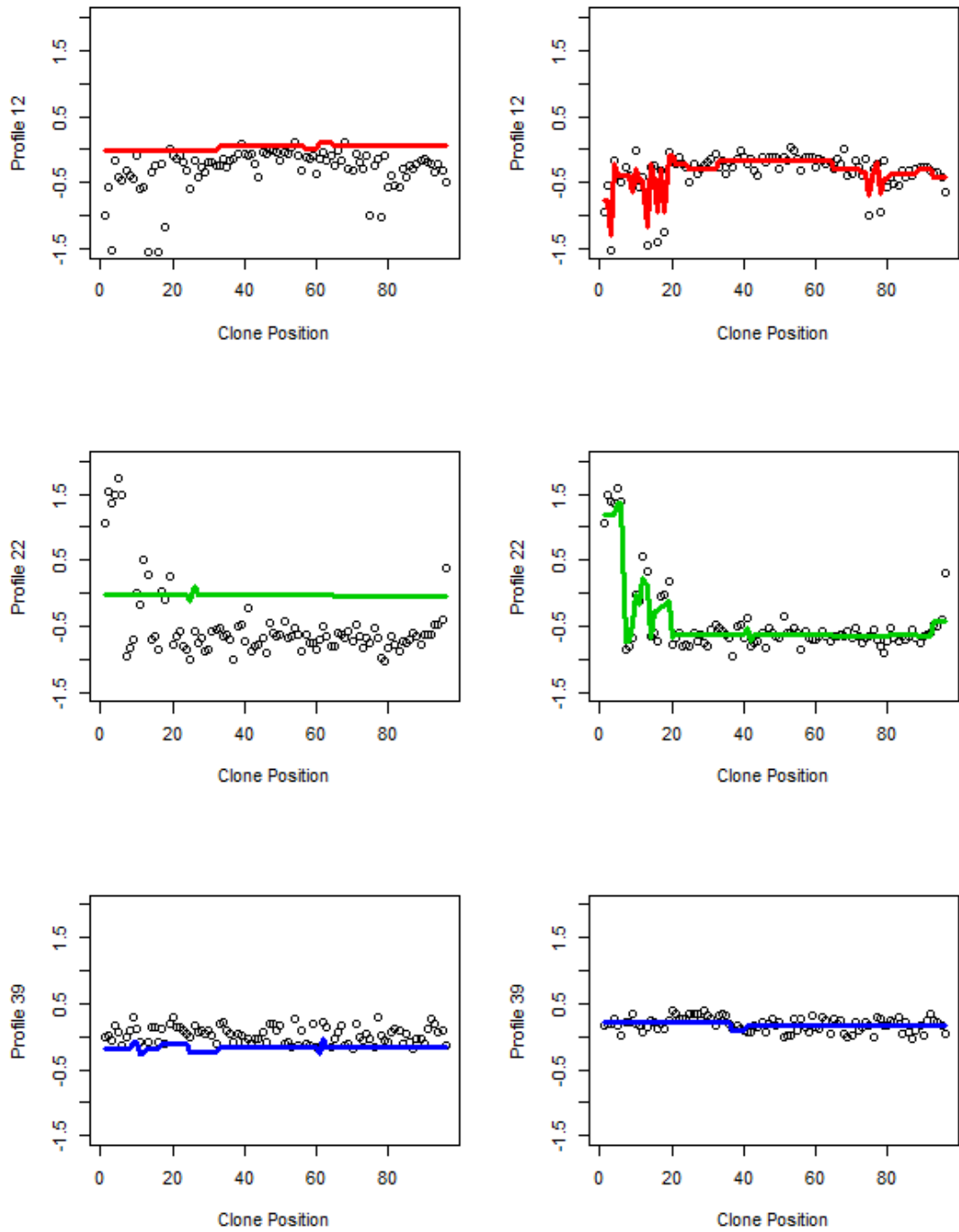


Figure 5.5: Profiles 12 (top), 22 (center), 39 (bottom) after processing of wavelet coefficients with a mixed model. At left, we see the group mean estimates in the domain of raw data, at right the mixed effect in the domain of the deviation of profiles from the group mean.

5.4 Testing

In this chapter we still conduct tests on CNV aberrations for group mean profiles: results are the same of the section 4.4. More interesting is to provide tests for random effects and see where they significantly deviate their profile from the group mean.

Hence, z-tests for random effects were calculated and significant clone-positions were identified under the following hypothesis test:

$$\begin{cases} H_0 : \hat{y}_{ij}^* = 0 \\ H_1 : \hat{y}_{ij}^* \neq 0 \end{cases} \quad (5.10)$$

with the FDR correction for multiple testing. The results show us that the first two groups have more or less regular profiles, almost ever similar to the corresponding group profile. On the other side, the non-gastric KIT has more than one irregular path. In the plots 5.6, 5.7, 5.8 we will show profiles with their significant mixed effects indicated with red dots.

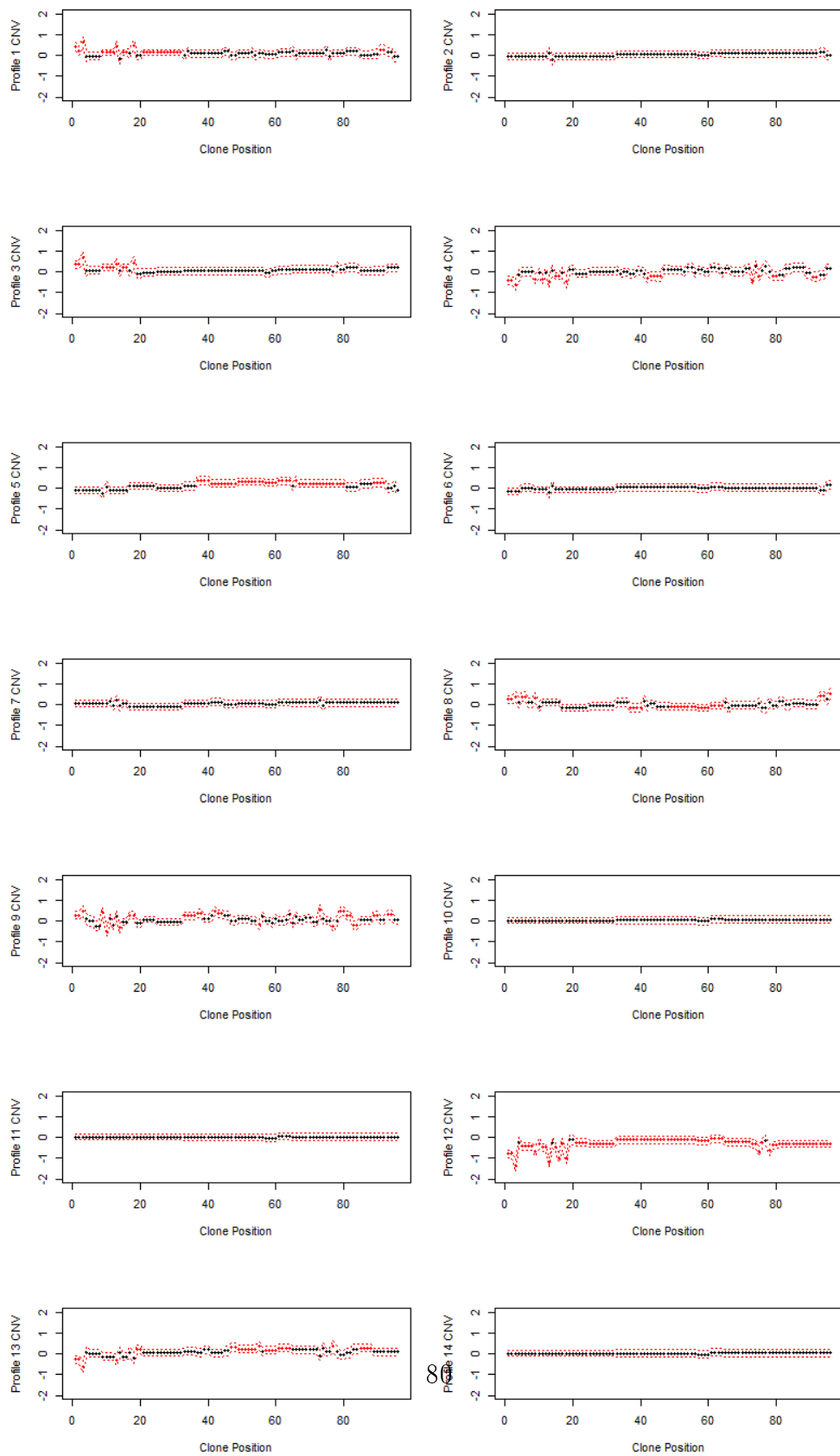


Figure 5.6: PDGFRA profiles. The red dots are the clone positions significantly different from the group mean.

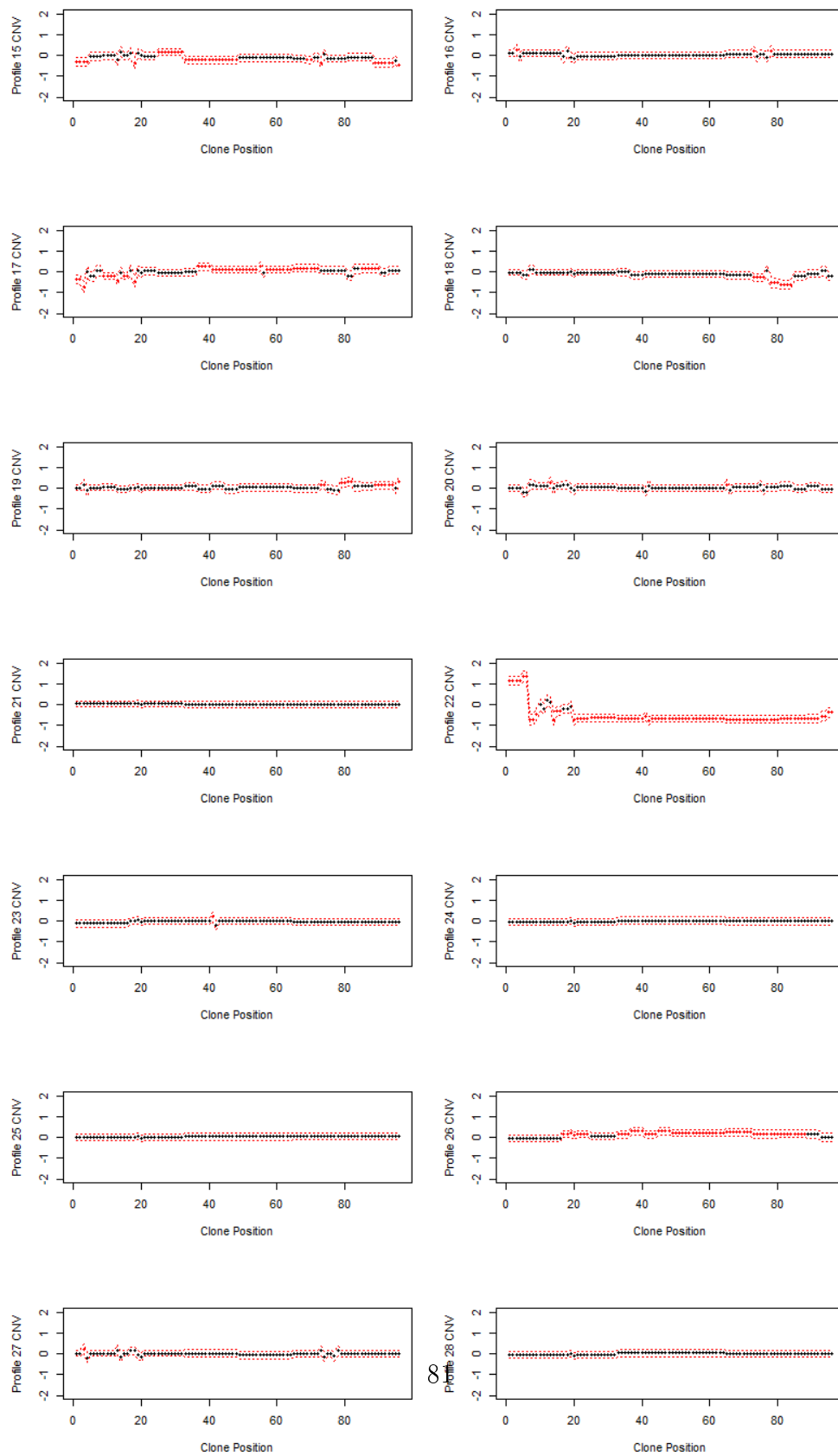


Figure 5.7: Gastric KIT profiles. The red dots are the clone positions significantly different from the group mean.

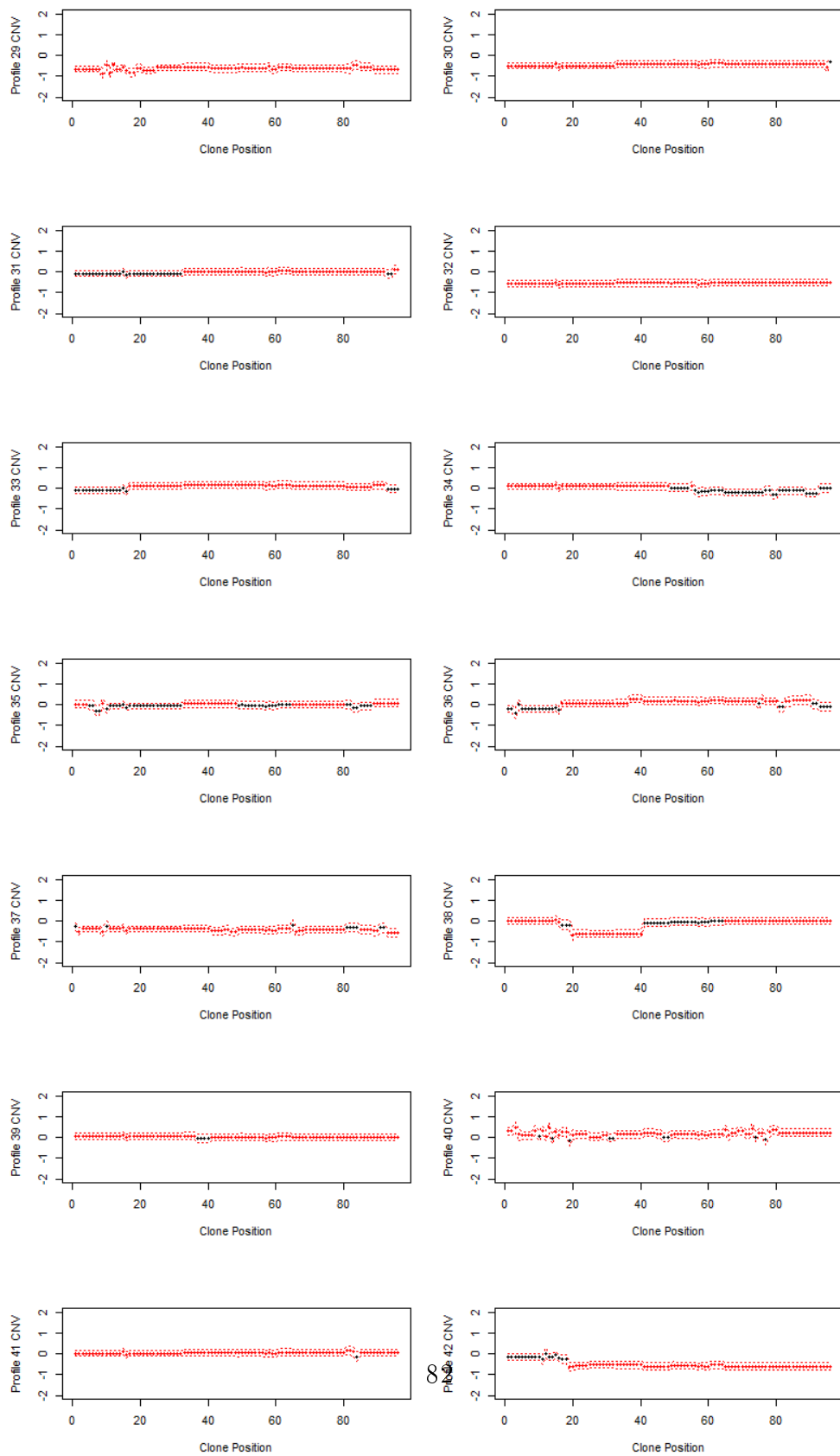


Figure 5.8: Non-Gastric KIT profiles. The red dots are the clone positions significantly different from the group mean.

Chapter 6

Conclusion and Discussion

We have seen in the previous chapters that denoising in the wavelet domain can serve different purposes. If a researcher aims at representing the smoothed individual profiles graphically, each profile can be denoised in the wavelet domain, individually. When inference is needed on the level of the individual group mean profiles, the mean model is estimated and denoised in the wavelet domain, and the inference is performed on the backtransformed denoised group mean profiles. If the interest is in getting subject-specific effects and group mean effects for each clone position, simultaneously, a mixed model can be used for assessing the research question.

The thresholding methods have been shown to be useful for developing piecewise constant models, but they suffer from the presence of a tuning parameter for thresholding. Even in LFDR thresholding, the cutoff $\lambda = 0.5$ was selected by empirical considerations. But, as Efron indicated [27], the new generation of high-throughput devices forces us to rethink basic topics in statistical theory. E.g. we have used the LFDR technique that is a technique for multiple testing that follows empirical Bayes consideration, but not for testing a significant hypothesis in the original domain. The profiles were too smooth and we did not have enough segments for estimating the empirical mixture and null distribution. The LFDR tests, with an appropriate cutoff, however, provide an intriguing way to smooth profiles.

MAP thresholding did not involve tuning parameters, but it revealed to break

down in this context. The sparsity imposed was too limited leading to noise prediction, and an underestimation of the variance. This was probably due to the MAD estimator that was used for calculating the threshold. The MAD only captures variability within profiles when smoothing the wavelet coefficients. Hence, the variability between the profiles was ignored. Leading to a considerable underestimation of the total variance that is needed for a proper denoising of the group mean profiles. Note, however, that a denoising based on the total variance is provided within the `waveTiling` package and can be further explored in the future. The LFDR approach, however, used the total variance estimate for regularization, which leads to a more efficient denoising.

The representation under SPCA analysis also has given interesting results, but it seems that a large number of PC are needed to provide an unbiased fit. In this specific case study the first PC explained a large part of variance (around 50%) and other components only explained a small fraction of the variability. Moreover, It has been observed that the `S pca` algorithm does not calculate subsequent PCs that explain a decreasing percentage of the variance. Hence, the algorithm suffers some convergence problems when a considerable number (say, $K = 10$) of components is required. Finally, the method also needs careful tuning of the regularization parameters.

The mixed model has offered advantages by providing inference of random effects as well as by providing good estimates of the group mean coefficients. The estimate of the variance for the wavelet estimates with MAD changed the LFDR thresholding of the group mean profiles. It imposes additional smoothness because the maximum of the MAD and the total variance is used for denoising of the group profiles. The mixed model further allowed us to draw conclusions about the nature of the mutation. It has been shown that profiles in the Non-Gastric KIT have a larger variability in the CNV profiles, indicating that the identification of these tumors is more difficult than for the other cases. The group mean profile for this group is also less representative. The functional model has showed that Non-Gastric KIT has a tendency to copy number losses, as already prospected in the Wozniak study discussed in 2.5. Since it is a segmentation model, these losses have not a straightforward interpretation in terms of exact copy number gains/losses as in a Calling Model, as mentioned in 1.4.3.

In chapter 4 we showed some different approaches for assessing contrasts between profiles, calculating the contrasts based on denoised profiles in the data domain first and calculating the contrasts in the wavelet domain based on the raw coefficients, denoising the contrasts before transforming them back to the original space. As we expected the estimates of the contrasts are similar, but the level of smoothing changes. In this specific case, since contrasts are never significant, the level of smoothing reached after applying contrasts in the wavelet domain is stronger. There is not a particular rule to decide which level of smoothing is better, and some techniques usually used in other contexts get into trouble in genetics. E.g., the complexity of a functional model is driven by the covariates considered: with low complexity, bias are larger but variance is limited; the situation reverses with the increase of complexity. Hence, the choice of the number of covariates is a trade-off between bias and variance [29]. Similarly, if two or more procedures return a different level of thresholding the mean squared error could measure and compare the goodness of fit. But in the context of genetics the mechanism of generation of data is often complicated and we cannot create artificial benchmarks simulating data, neither trust in a control group if we come across circumstances as the non-gastric KIT group. An alternative way could be considering CNV benchmarks of copy number gains/losses and conduct a classification analysis on them.

Appendix A

R-code

A.1 Data Exploration

Preprocessing

```
library(waveslim)
library(mgcv)
library(locfdr)
library(limma)
library(multtest)
library(elasticnet)
library(waveTiling)

#read data

alldatachr13updatebis <- read.delim("D:/statistica/Thesis/Practise/data/
alldatachr13updatebis.txt")
cnv13<- alldatachr13updatebis

#removing profiles too much affected by missingness.
#select profiles with at least 94 observation is a good optimum.
#Groups are not balanced. We can select al least 14 profiles par group.
```

```

full <- NULL
for(i in 1:60)
{
  full[i] <- sum(cnv13$Id_array==i) > 93
}

for(i in which(full==FALSE))
{
  cnv13 <- cnv13[-which(cnv13$Id_array==i),]
}

#we can retain at least 14 profiles par group

# remove profiles in surplus to get balanced data (preferring where
#there are more missing values)
set.seed(456)

sample(17:36,2)
#18,20

cnv13 <- cnv13[-which(cnv13$Id_array==17),]
cnv13 <- cnv13[-which(cnv13$Id_array==18),]
cnv13 <- cnv13[-which(cnv13$Id_array==20),]
cnv13 <- cnv13[-which(cnv13$Id_array==41),]
cnv13 <- cnv13[-which(cnv13$Id_array==48),]

# remove last observation to obtain a multiple of 2^n
cnv13 <- cnv13[-which(cnv13$STARTNEW==114753356)]

```



```

#GC content without considering groups
gamtestN<-gam(response1~s(GC_content),data=cnv13)
plot(gamtestN,ylab="General Response")

#GC content normalization considering groups
group=as.factor(cnv13$Group)
gamtest<-gam(response1~s(STARTNEW,by=group,k=96)+s(GC_content),data=cnv13)

hlp <-predict(gamtest,type="terms")
#estimated GC
#hlp[,4]

#correction
cnv13$response2 <- cnv13$response1 - hlp[,4]

#building a matrix of data (rawdata13)

cnv13$start2 <- as.factor(cnv13$STARTNEW)
lev <- levels(cnv13$start2)
rawdata13 <- matrix(NA,nrow=42,ncol=96)
conste <-cbind(as.numeric(levels(as.factor(cnv13$Id_array))),1:42)
for(j in 1:96)
{
  for(i in conste[,2])
  {
    if(any(cnv13$start2==lev[j] & cnv13$Id_array==conste[i,1]))
    {
      rawdata13[i,j] = cnv13$response2[which( (cnv13$start2)==lev[j] &
        cnv13$Id_array==conste[i,1])]
    }
    else{rawdata13[i,j]=NA

```

```

    }

}

```

Multiple Imputation

```

## Multiple Imputation : estimating the distribution of observed data
meanimputation <- matrix(NA,nrow=3,ncol=96)
sdimputation <- matrix(NA,nrow=3,ncol=96)
meanimputation[1,] <- colMeans(rawdata13[1:14,],na.rm=T)
sdimputation[1,] <- apply(rawdata13[1:14,],2,sd,na.rm=T)
meanimputation[2,] <- colMeans(rawdata13[15:28,],na.rm=T)
sdimputation[2,] <- apply(rawdata13[15:28,],2,sd,na.rm=T)
meanimputation[3,] <- colMeans(rawdata13[29:42,],na.rm=T)
sdimputation[3,] <- apply(rawdata13[29:42,],2,sd,na.rm=T)

#multiple imputation of Y and multiple estimate of D

D <- list()
Y <- list()
for(k in 1:10)
{
  Y[[k]] <- rawdata13
  for(i in 1:42)
  {
    for(j in 1:96)

```

```

    {
      if(is.na(Y[[k]][i,j])) {Y[[k]][i,j] <- rnorm(1,
        meanimputation[ceiling(i/14),j],sdimputation[ceiling(i/14),j])}
    }
  }

D[[k]] <- matrix(NA,42,96)

Y[[k]] <- cbind(Y[[k]])
J=5
D[[k]] <- t(apply(Y[[k]],1,wave.transform,n.levels=J))
}

#final estimation of D

D<-(D[[1]]+D[[2]]+D[[3]]+D[[4]]+D[[5]]+D[[6]]
+D[[7]]+D[[8]]+D[[9]]+D[[10]])
D <-D/10

#since the effect of missingness on Y is irrelevant, we will use
# the imputed Y[[1]] from now on as dataset.

```

Data Exploration Using Wavelets

```

#HARD THRESHOLDING

selected <- Y[[1]][c(12,22,39),]
J=5
Ybeta <- list(matrix(NA,3,96),matrix(NA,3,96),matrix(NA,3,96),

```

```

matrix(NA,3,96),matrix(NA,3,96),matrix(NA,3,96))

Draw <- t(apply(selected,1,wave.transform,n.levels=J))
vectlambda <- c(0,0.05,0.1,0.2,0.5,0.9)
Wv <- matrix(1,3,96)
Wv[,94:96] <- 10 #trick to retain wavelet father coefficients
for(i in 1:3)
{
  for(j in 1:length(vectlambda))
  {
    Ybeta[[j]][i,]<- wave.backtransformK(Draw[i,]*
      (abs(Draw[i,]*Wv[i,])>vectlambda[j]), J,order=1)
  }
}

#SOFT THRESHOLDING

Ybeta2 <- list(matrix(NA,3,96),matrix(NA,3,96),matrix(NA,3,96),
matrix(NA,3,96),matrix(NA,3,96),matrix(NA,3,96))

for(i in 1:3)
{
  for(j in 1:length(vectlambda))
  {
    Ybeta2[[j]][i,]<- wave.backtransformK(sign(Draw[i,])*abs(Draw[i,]
      -vectlambda[i])*(abs(Draw[i,]*Wv[i,])>vectlambda[j]), J,order=1)
  }
}

```

```

#FUNCTIONAL PCA

#calculating CA
ca <- spca(D,K=6,c(rep(1,6)))
#calculating and representing scores
scores <- D%*%ca$loadings

proc0 <- rep(0,96)
proc1 <- wave.backtransformK(ca$loadings[,1],J,order=1)
proc2 <- wave.backtransformK(ca$loadings[,2],J,order=1)
proc3 <- wave.backtransformK(ca$loadings[,3],J,order=1)
proc4 <- wave.backtransformK(ca$loadings[,4],J,order=1)
proc5 <- wave.backtransformK(ca$loadings[,5],J,order=1)
proc6 <- wave.backtransformK(ca$loadings[,6],J,order=1)
proc <- data.frame(proc0,proc1,proc2,proc3,proc4,proc5,proc6)
scoresel <- scores[c(12,22,39),]
coef <- matrix(0,96,7)

par(mfrow=c(3,2))
# reconstruction of profiles with a growing number of PC components
for(j in 1:6)

{
  matplot(t(selected)-colMeans(Y[[1]]),ylim=c(-2,2),col=c(2,3,4),
  pch=1,main= paste(j,"Principal Components",sep=" "),
  xlab="Clone Position",ylab="Centered CNV Expression")

  for(i in 1:(dim(selected)[1]))
  {
    for(k in 2:7)
    {
      coef[,k]<- proc[,k]*scoresel[i,k-1]
    }
  }
}

```

```

    }
    lines(apply(coef[,1:(j+1)],1,sum),col=i+1,lwd=3)
  }
}

#CLUSTER ANALYSIS

names <- as.character(c("PDGFRA","Gastric KIT","Non-Gastric KIT"))

rownames(D) <- c(rep(names[1],14),rep(names[2],14),rep(names[3],14))
plot(hclust(dist(D)),main="Cluster Dendogram for Profiles",xlab=
"Profiles",ylab="Distance")
plot(hclust(dist(Y[[1]])),main="Cluster Dendogram for Profiles",xlab=
"Profiles",ylab="Distance")
#same plot

```

A.2 Wavelet Based Functional Model

Models and Thresholding

```

#FUNCTIONAL MODEL

Xgroup <- matrix(0,nrow=42,ncol=3)
Xgroup[1:14,1]<- 1
Xgroup[15:28,2]<- 1
Xgroup[29:42,3] <-1
lmRaw <- lmFit(t(D),design=Xgroup)
RawCoef <- lmRaw$coefficients
seRaw<-matrix(lmRaw$sigma,ncol=1)%*%sqrt(diag(lmRaw$cov.coefficients))

```

```

#LFDR thresholding

#calculating t-tests
Rawpt <- RawCoef/seRaw
Rawdf <- lmRaw$df.residual
tdistRaw <- pt(-abs(Rawpt),Rawdf) #-abs solves some problems of boundary
# that often occur. The sign is corrected in the next operation
#in the z-domain
zdistRaw <- qnorm(tdistRaw)*sign(Rawpt)
#LFDR
Rawfdr <- locfdr(zdistRaw[1:93,],df=10,bre=30)
#LFDR matrix
RawfdrMat <- matrix(Rawfdr$fdr,nrow=96,ncol=3,byrow=FALSE)
RawfdrMat[94:96,] <- 1e-10 #thresholding is only on mother wavelet
#coefficients. We ever retain the wavelet father

betagroup <- matrix(NA,96,3)
varbetagroup <- matrix(NA,96,3)
for(i in 1:3)
{betagroup[,i]<- wave.backtransformK(RawCoef[,i]*(RawfdrMat[,i]<0.5),
  J,order=1)
  varbetagroup[,i]<- wave.backtransformK(seRaw[,i]^2*(RawfdrMat[,i]<0.5),
  J,order=2)
}

#LFDR imposing the same sparsity

unionfdr <- apply(RawfdrMat,1,min)
betagroup1 <-matrix(NA,96,3)
varbetagroup1 <- matrix(NA,96,3)

```

```

for(i in 1:3)
{betagroup1[,i]<- wave.backtransformK(RawCoef[,i]*(unionfdr<0.5),
  J,order=1)
  varbetagroup1[,i]<- wave.backtransformK(seRaw[,i]^2*(unionfdr<0.5),
  J,order=2)
}

```

#MAP model and thresholding

```

fit <-WaveMarEstVarJ(Y=Y[[1]],X=Xgroup,n.levels=5,wave.filt="haar",
D=D,var.eps="mad",prior="improper",tol=1e-6,saveall=TRUE,trace=T)

```

```

betaMAP <- matrix(NA,3,96)
varbetaMAP <- matrix(NA,3,96)
for(i in 1:3)

```

```

{betaMAP[i,]<- wave.backtransformK(fit$beta_MAP[i,],fit$n.levels ,
  filt=fit$wave.filt,order=1)
  varbetaMAP[i,] <- wave.backtransformK(fit$varbeta_MAP[i,],\\
  n.levels=fit$n.levels, filt=fit$wave.filt,order=2)
}

```

#CONTRAST PROFILES

```

###1. Contrasts calculated on backtransformed data (after LFDR
# thresholding and same sparsity imposed)

```

```

xcont <- matrix(c(1,-1,0,0,1,-1,-1,0,1),3,3)
xvarcont <- matrix(c(1,1,0,0,1,1,1,0,1),3,3)

```



```

contral1 <- betagroup1%*%xcont
sdcontral1 <- sqrt(varbetagroup1%*%xvarcont)

```

###2. Contrasts calculated on backtransformed data (after MAP thresholding)

```

beta2MAP <- t(betaMAP)%*%xcont
varbeta2MAP <- t(varbetaMAP)%*%xvarcont

```

#3. Contrast calculated in the wavelet domain,
#hence LFDR thresholding of contrasts

```

contcoef <- RawCoef%*%xcont
contsigma <- (seRaw^2)%*%xvarcont

```

```

contbeta <- matrix(nrow=96,ncol=3)
sigmacontbeta <- matrix(nrow=96,ncol=3)

```

```

Conttpt <- contcoef/sqrt(contsigma)
Contdist <- pt(-abs(Conttpt),Rawdf)
zContdist <- qnorm(Contdist)*sign(Conttpt)
Contfdr <- locfdr(zContdist[1:93,],df=10,bre=30)

```

```

ContfdrMat<-matrix(Contfdr$fdr,nrow=96,ncol=3,byrow=FALSE)

```

```

ContfdrMat[94:96,]<-.01

```

```

for(i in 1:3)
{contbeta[,i]<- wave.backtransformK(contcoef[,i]*(ContfdrMat[,i]<0.5),J)
  sigmacontbeta[,i]<- sqrt(wave.backtransformK(contsigma[,i]*
(ContfdrMat[,i]<0.5), J,order=2))

```

```
}
```

Tests

```
####1. CNV ABERRATIONS
```

```
##a.model with LFDR thresholding
```

```
testgroup <- betagroup/sqrt(varbetagroup)
```

```
#some tests are repeated due to thresholding, we consider once  
#at time making some operations on the test matrix
```

```
for(j in 1:dim(testgroup)[2])  
{  
  for(i in 2 : dim(testgroup)[1]-1)  
  {  
  
    if(testgroup[i,j]==testgroup[i+1,j]) {testgroup[i,j] =NA}  
  }  
}
```

```
pdistgroup <- 2*(1-pnorm(abs(testgroup)))  
pgroup <- pdistgroup[is.na(pdistgroup)==FALSE &  
  (pdistgroup==1)==FALSE]
```

```
correctgroup <-p.adjust(pgroup,method="fdr")
```

```
#rebuliding the whole test matrix  
groupselect <- which(is.na(testgroup)==FALSE)  
groupfdrMat <- c(rep(NA,288))  
for (i in groupselect)
```

```

{
  groupfdrMat[i] <- correctgroup[which(groupselect==i)]
}

for (i in length(groupfdrMat):1)
{
  if(is.na(groupfdrMat[i]))
  {
    groupfdrMat[i] <-groupfdrMat[i+1]
  }
}
groupfdrMat<-matrix(groupfdrMat,nrow=96,ncol=3,byrow=FALSE)

##b. model with LFDR thresholding (same sparsity imposed)

testgroup1 <- betagroup1/sqrt(varbetagroup1)

for(j in 1:dim(testgroup1)[2])
{
  for(i in 2 : dim(testgroup1)[1]-1)
  {

    if(testgroup1[i,j]==testgroup1[i+1,j]) {testgroup1[i,j] =NA}
  }
}

pdistgroup1 <- 2*(1-pnorm(abs(testgroup1)))
pgroup1 <- pdistgroup1[is.na(pdistgroup1)==FALSE &

```

```

(pdistgroup1==1)==FALSE]

correctgroup1 <- p.adjust(pgroup1,method="fdr")

groupselect1 <- which(is.na(testgroup1)==FALSE)
groupfdrMat1 <- c(rep(NA,288))
for (i in groupselect1)

{
  groupfdrMat1[i] <- correctgroup1[which(groupselect1==i)]
}

for (i in length(groupfdrMat1):1)
{
  if(is.na(groupfdrMat1[i]))
  {
    groupfdrMat1[i] <- groupfdrMat1[i+1]
  }
}
groupfdrMat1<-matrix(groupfdrMat1,nrow=96,ncol=3,byrow=FALSE)

```

####2. TESTING DIFFERENCES BETWEEN GROUPS

##a. model with LFDR thresholding

```

betacallL <- matrix(NA,nrow=96,ncol=3)
sigmabetaL <- matrix(NA,nrow=96,ncol=3)
for(i in 1:3)
{
  L <- c(rep(-1/3,3))

```

```

L <-t(L)
L[,i] <- 2/3

Lp <- matrix(L,nrow=96,ncol=3,byrow=TRUE)

betacallL[,i] <- (L%%t(betagroup))

sigmabetaL[,i] <- sqrt((Lp*varbetagroup)%%t(L))

}

testgroupL <- betacallL/sigmabetaL

for(j in 1:dim(testgroupL)[2])
{
  for(i in 2 : dim(testgroupL)[1]-1)
  {

    if(testgroupL[i,j]==testgroupL[i+1,j]) {testgroupL[i,j] =NA}
  }
}

pdistgroupL <- 1-pchisq(abs(testgroupL),1)
pgroupL <- pdistgroupL[is.na(pdistgroupL)==FALSE &
(pdistgroupL==1)==FALSE]
#any test is significant already in the chisq domain.

##b. model with LFDR thresholding and same sparsity imposed

betacall1bis <- matrix(NA,nrow=96,ncol=3)
sigmabeta1bis <- matrix(NA,nrow=96,ncol=3)

```

```

for(i in 1:3)
{
  L <- c(rep(-1/3,3))
  L <-t(L)
  L[,i] <- 2/3

  Lp <- matrix(L,nrow=96,ncol=3,byrow=TRUE)

  betacall1bis[,i] <- (L%*%t(betagroup1))

  sigmabeta1bis[,i] <- sqrt((Lp*varbetagroup1)%*%t(L))
}

testgroup1bis <- betacall1bis/sigmabeta1bis


for(j in 1:dim(testgroup1bis)[2])
{
  for(i in 2 : dim(testgroup1bis)[1]-1)
  {

    if(testgroup1bis[i,j]==testgroup1bis[i+1,j])
    {testgroup1bis[i,j] =NA}
  }
}

pdistgroup1bis <- 1-pchisq(abs(testgroup1bis),1)
pgroup1bis <- pdistgroup1bis[is.na(pdistgroup1bis)==FALSE &
(pdistgroup1bis==1)==FALSE]

```

```

##TESTS ON CONTRASTS PROFILES
##a. contrasts in the data domain, after LFDR thresholding
testcontra <- contra1/sdcontra1
for(j in 1:dim(testcontra)[2])
{
  for(i in 2 : dim(testcontra)[1]-1)
  {

    if(testcontra[i,j]==testcontra[i+1,j]) {testcontra[i,j] =NA}
  }
}

pdisttestcontra <- 1-pchisq(abs(testcontra),1)
ptestcontra <- pdisttestcontra[is.na(pdisttestcontra)==FALSE &
(pdisttestcontra==1)==FALSE]

##b. Contrast calculated in the wavelet domain,
# hence LFDR thresholding of contrasts

testcont <- contbeta/sigmacontbeta
for(j in 1:dim(testcont)[2])
{
  for(i in 2 : dim(testcont)[1]-1)
  {

    if(testcont[i,j]==testcont[i+1,j]) {testcont[i,j] =NA}
  }
}

```

```

pdisttestcont <- 1-pchisq(abs(testcont),1)
ptestcont <- pdisttestcont[is.na(pdisttestcont)==FALSE
  & (pdisttestcont==1)==FALSE]

```

A.3 Wavelet Based Mixed Model

Model

```

N <- 42
NGROUP <- 3

madWav <- mad(D[,1:48])
X <- diag(42)
C <- cbind(Xgroup,X)
B <- matrix(0,N+NGROUP,N+NGROUP)
B[-(1:NGROUP),-(1:NGROUP)] <- diag(N)
sigma <- lmRaw$sigma

#blup

blupFun<-function(x)
{
  if (sigma[x]>madWav)
    solve(t(C)%*%C + madWav^2/(sigma[x]^2-madWav^2)*B) %*%t(C)
    %*%matrix(D[,x],ncol=1)
  else c(solve(t(Xgroup)%*%Xgroup)%*%t(Xgroup)%*%
    matrix(D[,x],ncol=1),rep(0,N))
}

estimates <- t(sapply(1:96,blupFun))

```



```

#standard error

varblupFun <- function(x)
{
  if(sigma[x]>madWav)
    diag(solve(t(C)%*%C/madWav^2 + B/(sigma[x]^2-madWav^2)))
  else c(diag(solve(t(Xgroup)%*%Xgroup)*madWav^2),rep(0,N))
}

standarderr <- sqrt(t(sapply(1:96,varblupFun)))

ttest2 <- estimates/standarderr

#LFDR for group mean estimates

tdistRaw2 <- pt(-abs(ttest2[,1:3]),N-NGROUP)
zdistRaw2 <-qnorm(tdistRaw2)*sign(ttest2[,1:3])

Rawfdr1 <- locfdr(zdistRaw2[1:93,],bre=30)
RawfdrMat1<-matrix(Rawfdr$fdr,nrow=96,ncol=3,byrow=FALSE)
RawfdrMat1[94:96,]<-.01

#representation of group mean profiles

Gmat <- matrix(NA,nrow=3,ncol=96)
Gmatvar <- matrix(NA,nrow=3,ncol=96)
Bmat <- matrix(NA,nrow=42,ncol=96)
Bmatvar <- matrix(NA,nrow=42,ncol=96)
for(i in 1:3)
{
  Gmat[i,] <- wave.backtransformK(estimates[,i]*(RawfdrMat1[,i]<0.5),
    J,order=1)

```

```

    Gmatvar[i,] <- wave.backtransformK((standarderr[,i]^2)*
    (RawfdrMat1[,i]<0.5),J,order=2)
  }
  Gmatvar[Gmatvar==0] = 1e-12

#LDFR for random effects

selectrand <- ttest2[,4:45]
constraint <- which(sigma>madWav)
randomdist <- pnorm(-abs(selectrand))
zrandom <- qnorm(randomdist)*sign(selectrand)
randomfdr <- locfdr(zrandom[is.na(zrandom)==FALSE],df=20)
helpmat <- matrix(randomfdr$fdr,nrow=length(constraint),
ncol=42,byrow=FALSE)
randomfdrmat<- matrix(1,nrow=96,ncol=42)
randomfdrmat[constraint,] <- helpmat
randomfdrmat[94:96,]<-.01
#vectorize mixed effects

for(i in 1:42)
{
  Bmat[i,] <-wave.backtransformK(estimated[,i+3]*
  (randomfdrmat[,i]<0.5),J,order=1)
  Bmatvar[i,] <- wave.backtransformK((standarderr[,i+3]^2)*
  (randomfdrmat[,i]<0.5),J,order=2)
}
Bmatvar[Bmatvar==0] =1e-12 #to avoid problems in calculating tests when
#random effect do not exist

##representation of completed profiles

profiles <- matrix(NA,nrow=42,ncol=96)

```

```

varprofiles <- matrix(NA,nrow=42,ncol=96)
for(i in 1:42)
{
  profiles[i,] <- Gmat[ceiling(i/14),]+Bmat[i,]
  varprofiles[i,] <- Gmatvar[ceiling(i/14),] + Bmatvar[i,]
}

```

Tests

#1. Test on CNV aberrations for group mean

```

testgroupR <- t(Gmat/sqrt(Gmatvar))

for(j in 1:dim(testgroupR)[2])
{
  for(i in 2: dim(testgroupR)[1]-1)
  {

    if(testgroupR[i,j]==testgroupR[i+1,j]) {testgroupR[i,j] =NA}
  }
}

pdistgroupR <- 2*(1-pnorm(abs(testgroupR)))
pgroupR <- pdistgroupR[is.na(pdistgroupR)==FALSE &
  (pdistgroupR==1)==FALSE]

correctgroupR <-p.adjust(pgroupR,method="fdr")

selectR <- which(is.na(testgroupR)==FALSE)

```

```

groupfdrmatR<- c(rep(NA,288))
for (i in selectR)

{
  groupfdrmatR[i] <- correctgroupR[which(selectR==i)]
}

for (i in length(groupfdrmatR):1)
{
  if(is.na(groupfdrmatR[i]))
  {
    groupfdrmatR[i] <-groupfdrmatR[i+1]
  }
}

groupfdrmatR <- matrix(groupfdrmatR,nrow=96,ncol=3,byrow=FALSE)
groupfdrmatR

##2. test on random effects

randomtest <-Bmat/sqrt(Bmatvar)

for(j in 1:dim(randomtest)[1])
{
  for(i in 2: dim(randomtest)[2]-1)
  {

    if(randomtest[j,i]==randomtest[j,i+1]) {randomtest[j,i] =NA}
  }
}

prandomtest <- 2*(1-pnorm(abs(randomtest)))

```

```

prandomtest <- prandomtest[is.na(randomtest)==FALSE]
correctrandom <- p.adjust(prandomtest,method="fdr")

selectrandom <- which(is.na(randomtest)==FALSE)
groupfdrandom <- c(rep(NA,4032))
for (i in selectrandom)

{
  groupfdrandom[i] <- correctrandom[which(selectrandom==i)]
}
groupfdrandom <- matrix(groupfdrandom,nrow=42,ncol=96,byrow=FALSE)

for (i in 1:nrow(groupfdrandom))
{
  for(j in ncol(groupfdrandom):1)
  {
    if(is.na(groupfdrandom[i,j]))
    {
      groupfdrandom[i,j] =groupfdrandom[i,j+1]
    }
  }
}
}

```


Bibliography

- [1] Gonzalez, J.R. , Subirana, I. , Escaramis, I. , Peraza, S. , Caceres, A. , Estivill, X. , Armengol, L. , (2009) "Accounting for uncertainty when assessing association between copy number and disease: a latent class model" . *BMC Bioinformatics 2009* ,10:172.
- [2] Xie, C. , Tammi, M.T. , (2009) "CNV-seq, a new method to detect copy number variation using high-throughput sequencing", *BMC Informatics 2009*, 10(1):80.
- [3] van de Wiel, M.A. , Picard, F. , van Wieringen, W.N. , Ylstra, B. , (2010) "Preprocessing and downstream analysis of microarray DNA copy number profiles" *Briefing in Bioinformatics Advance Access*, 12(1):10-21.
- [4] Shinawi, M. , Cheung, S.W. , (2008) "The array CGH and its clinical applications", *Drug Discovery Today*, Vol. 13, Issues 17-18, Pages 760-770.
- [5] Picard, F. , Lebarbier, E. , Hoebeke, M. , Rigai, G. , Thiam, B. , Robin, S. , (2011) "Joint segmentation, calling, and normalization of multiple CGH profiles", *Biostatistics 2011*,12:3, pages 413-428.
- [6] van de Wiel, M.A. , Kim, K.I. , Vosse, S.J. , van Wieringen, W.N. , Wilting, S.M. , Ylstra, B. (2007) "CGHcall: calling aberrations for array CGH tumor profiles", *Bioinformatics applications note*, Vol.23, no.7, pages 892-894.
- [7] Magi, A. , Benelli, M. , Marseglia, G. , Nannetti, G. , Scordo, M.R. , Torricelli, F. (2010) "A shifting level model algorithm that identifies aberrations in array-CGH data" *Biostatistics 2010* 11:2. pp.265-280.

- [8] Picard, F. , Robin, S. , Lavielle, M. , Vaisse, C. , Daudin, J.-J. , (2005) "A statistical approach for array CGH data analysis" , *BMC Informatics* 2005, 6:27
- [9] Lai, W.R. , Johnson, M.D. , Kucherlapati, R. , Park, P.J. (2005) "Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data", *Bioinformatics* 21, 3763-3770.
- [10] van de Wiel, M.A. , Brosens, R. , Eilers, P.H.C. , Kumps, C. , Meijer, G.A. , Menten, B. , Sistermans, E. , Speleman, F. , Timmerman, M.E. , Ylstra, B. (2009) "Smoothing Waves in array CGH tumor profiles", *Bioinformatics* Vol.25, no.9, pages 1099-1104.
- [11] Nowak, G. , Hastie, T. , Pollack, J.R. , Tibshirani, R. , (2011) "A fused lasso latent feature model for analyzing multi-sample aCGH data", *Biostatistics* 0,0, pages 1-16.
- [12] Baladandayuthapani, V. , Ji, Y. , Talluri, R. , Nieto Barajas, L.E. , Morris, J.S. (2010) "Bayesian Random Segmentation Models to identify Shared Copy Number Aberrations for Array CGH Data" *Journal of the american statistical association*, 105(492):1358-1375.
- [13] van Wieringen, W.N. , van de Wiel, M.A. , Ylstra, B. (2007), "Normalized, Segmentized or Called aCGH data?", *Cancer Informatics* 2007:3.
- [14] van de Wiel, M.A. , van Wieringen, W.N. , (2007) "CGHregions: Dimension reduction for Array CGH Data with Minimal Information Loss", *Cancer Informatics* 2007:3. .
- [15] van de Wiel, M.A. , van Wieringen, W.N. , Ylstra, B. (2008) "Weighted Clustering of called array CGH data", *Biostatistics* 2008 9:3, pp. 484-500.
- [16] Miettinen, M. , Lasota, J. , (2001) "Gastrointestinal stromal tumors: definition, clinical, histological, immunohistochemical, and molecular genetic features and differential diagnosis", *Virchows Arch* 438:1-12.
- [17] Verbeke, G. , Molkenbergs, G. , (2000) *Linear Mixed Models for Longitudinal Data*, Springer, New York.

- [18] Wozniak, A. , Sciot, R. , Guillou, L. , Pauwels, P. , Wasag, B. , Stul, M. , Vermeesch, J.R. , Vandenberghe, P. , Limon, J. , Debiec-Rychter, M. (2007) "Array CGH Analysis in Primary Gastrointestinal Stromal Tumors: Cytogenetic Profile Correlates with Anatomic Site and Tumor Aggressiveness, Irrespective of Mutational Status", *Genes, Chromosomes & Cancer*, 46:261-276.
- [19] Clement, L. , De Beuf, K. , Thas, O. , Vuylsteke, M. , Irizarry, R.A. , Crainiceanu, M.C. , (2012) "Fast Wavelet Based Functional Models for Transcriptome Analysis with Tiling Arrays" , *Statistical Applications in Genetics and Molecular Biology*, Volume 11, Issue 1, Article 4.
- [20] Hastie, T. , Tibshirani, R. , Friedman, J. , (2001) *The Elements of Statistical Learning*, Springer, New York.
- [21] Nguyen, N. , Huang, H. , Soontorn, O. , Vo, A. (2010), "Stationary Wavelet Packet Transform and Dependent Laplacian Bivariate Shrinkage Estimator for Array-CGH Data Smoothing", *Journal of Computational Biology*, Volume 17, Number 2, pp. 139-152.
- [22] Figueiredo, M.A.T. and Nowak, R. (2001) "Wavelet Based Image Estimation: An Empirical Bayes Approach Using Jeffreys' Noninformative Prior", *IEEE Transactions on Image Processing* Vol. 10, NO. 9.
- [23] Donoho, D. , Johnstone, I. , (1994) "Ideal adaptation via wavelet shrinkage", *Biometrika*, vol. 81, 425-455.
- [24] Zou, H. , Hastie, T. , Tibshirani, R. "Sparse Principal Component Analysis", *Journal of Computational and Graphical Statistics*, Volume 15, Number 2, Pages 265-286.
- [25] Storey, J.D. , Tibshirani, R. (2003) "Statistical Significance for Genomewide Studies" *PNAS*, vol.100, no.16.
- [26] Benjamini, Y. , Hochberg, Y. (1995) "Controlling false discovery rate: a practical and powerful approach for multiple testing", *Journal of the Royal Statistical Society, Series B* , 57(1), 289-300.

- [27] Efron, B. (2008). "Microarrays, Empirical Bayes and the Two-Groups Model", *Statistical Science*, Vol.23, No.1, 1-22
- [28] Ruppert, D. , Wand, M.P. , Carroll, R.J. ,(2003) *Semiparametric Regression*, Cambridge Univerisity Press, Cambridge.
- [29] Azzalini, A. , Scarpa, B. (2012) *Data Analysis and Data Mining*, Oxford University Press, Oxford.

Acknowledgements

To professor Clement, who followed me carefully and with extreme devotion in every step of this thesis.

Alla Professoressa Romualdi, che mi ha preziosamente aiutato nell'ultimo periodo, nonostante i tempi stretti.

A "Stats", l'ambiente più genuino nel quale mi sia trovato a studiare in vita mia.

In ordine di apparizione:

A Tiberio, nonostante appaia un cretino (e in parte lo è) ma mi ha fatto entrare "nel giro giusto" e si è dimostrato una brava persona.

A Zio Fabio, compagno di mille battaglie universitarie, col quale ho condiviso prime batoste e successi finali.

a Zia Frin, amica schietta e sincera, nonostante qualche piccolo problema di comunicazione reciproca.

ad Alice, e a quel suo modo unico di vivacizzare la banalità del quotidiano.

a mamma Maddalena, che col suo affetto mi ha sostenuto...come una mamma, appunto.

a Zia Eli, l'unica persona che c'è stata sempre, dall'inizio alla fine, il deus ex-machina della banda.

ad Anna, ed all'infinita disponibilità ed ospitalità che mi ha dimostrato.

ad Alessia, làlà, làlà, làlà.

a Zia Cawy, tutor nella Statistica e nella Vita.

a Miki, Manu, Ric e tutti gli altri "padovani simpatici".

a Dani e Sbenso, i migliori coinquilini che potessi desiderare, i quali hanno ravvivato in me l'amore rispettivamente per la statistica e per l'umanità.

ad Ema, uno degli ultimi dandy al mondo.

a tutti gli altri studenti coi quali ho stretto amicizia, dai Vecchi (Avanzi) ai Nuovi (Baresi).

al Maschio Alpha, a Scintilla e a tutte le altre entità oniriche che hanno vegliato sul mio studio.

Ai miei amici di sempre, tipo Tonia, Claudio, Mauro, Rosa (e tanti altri), distanti mille chilometri, mai allontanatisi di un centimetro.

Alla mia famiglia naturale, in particolare a mia sorella Vera, la persona più buona al mondo (al pari con Sbenson), che ha sempre creduto in me come nessun altro.

A Padova, la città dove ho Studiato.

*“...un ultimo sguardo commosso
all’arredamento e chi si è visto,
s’è visto.” (cit.)*