Paola M.V. Rancoita

Stochastic methods in cancer research. Applications to genomics and angiogenesis

To my family, my husband and all friends that supported me

Contents

Acronyms

Lists of abbreviations:

aCGH	array comparative genomic hybridization
BPCR	Bayesian Piecewise Constant Regression
BRC	Bayesian Regression Curve
BRCAk	Bayesian Regression Curve Averaging over k
CLT	central limit theorem
CN	copy number
CNP	copy number polymorphism
FISH	fluorescent in situ hybridization
FPR	false positive rate
gBPCR	genomic Bayesian Piecewise Constant Regression
Het	heterozygous SNP
HMM	Hidden Markov Model
Hom	homozygous SNP
IBD	identical by descendent
kb	kilobases, i.e. one thousand base-pairs
LOH	loss of heterozygosity
MAP	maximum a posteriori
Mb	megabases, i.e. one million base-pairs
mBPCR	modified Bayesian Piecewise Constant Regression

mBRC	modified Bayesian Regression Curve
PCR	polymerase chain reaction
SNP	single nucleotide polymorphism
SNR	signal to noise ratio
TPR	true positive rate
UPD	uniparental disomy

Introduction

In this thesis, I study three stochastic methods that can be applied for the analysis of data in cancer research and, in particular, to cancer genomic data and to images of angiogenic processes. Cancer is a multistep process where the accumulation of genomic lesions alters cell biology. The latter is under control of several pathways and thus, cancer can arise via different mechanisms affecting different pathways. Due to the general complexity of this disease and the different types of tumors, the efforts of cancer research cover several research areas such as, for example, immunology, genetics, cell biology, angiogenesis. Moreover, in recent years, interactions between mathematicians and biomedical researchers have increased due to both the awareness of the complexity of the biological/medical issues and the development of new technologies, producing "large" data rich of information. Biomathematics is applied in many areas, such as epidemiology, clinical trial design, neuroscience, disease modeling, genomics, proteomics, and thus in several areas of cancer research.

The thesis is divided into two parts. In the former, I propose two Bayesian regression methods for the analysis of two types of cancer genomic data. In the latter, I study the properties of two estimators of the intensity of a stationary fibre process, which can be applied for the characterization of angiogenic and vascular processes.

PART I: “Bayesian integrative genomics”.

Tumors are largely related to chromosomal lesions and some of them can be detected by using specific types of microarrays. In Chapter 1, I introduce some basic knowledge of genetics, in order to describe the principal types of genomic alterations that can occur in genetics diseases, like cancer. DNA contains (in the genes) the information for coding proteins and the functions in a living cell depend on proteins, hence DNA aberrations can lead to a different production of proteins, changing the behavior of the cell.

One type of aberration is the change of DNA copy number in one or more regions of the genome. Apart from the sex chromosomes, in a healthy cell the copy number is two because we inherit a copy of each chromosome (called *homolog*) from each of our parents, but in a tumor cell the genome can present regions of deletions (copy number one or zero), gains (copy number three or four) or amplifications (copy number greater than four). Another type of mutation regards the status of the two homologs of the chromosomes, measured via the genotype of the *single nucleotide polymorphisms* (SNPs). A SNP is a base-pair location in the genome where the nucleotide can assume two or three different bases (called *alleles*) out of the four: thymine, adenine, cytosine and guanine. The status of the base-pair of nucleotides at a SNP position is called genotype and a SNP can be classified as either homozygous, if its two copies consist of equal alleles, or heterozygous, otherwise. Alterations of the homozygous status are often displayed by unusual long stretches of homozygous SNPs (called *loss of heterozygosity*, LOH) in regions with normal copy number and they can be explained with several genomic events such as uniparental disomy [40] or autozygosity [43]. I denote this type of aberrations with IBD/UPD. In literature, a relationship between some tumors and both types of mutations have been shown [4, 5, 10, 19, 20, 22, 54, 71, 87].

In Chapter 1, I also briefly explained the SNP microarrays, which are able to measure simultaneously both DNA copy number and genotype at hundred thousands of SNP positions. However, the raw copy number data (i.e. the copy number data obtained by the microarray) generally are very noisy, due to both technical and biological reasons. Then, an important issue is to define a method which can estimate well the number of regions

with different copy number, the endpoints of these regions (called *breakpoints*) and their copy number. In practice, the raw copy number is not an integer, since it is measured from the DNA extracted in a sample of cells, that does not contain only tumor cells. Moreover, among the tumor cells we can have cells that belong to different stages of the disease and thus can carry different mutations. Hence, the raw and the estimated copy number usually assume real values and the DNA copy number along the aberrated genome can be represented as a *piecewise constant* function.

An advantage of SNP microarray is the possibility to measure both DNA copy number and genotype at each SNP position considered. In this way, several types of "abnormalities" of the genome (regarding both DNA copy number and LOH status) can be observed and integrated for a better identification of the events occurred. For example, when a copy of a chromosomic segment is deleted, we detect a long stretch of homozygous SNPs (since the microarray is unable to distinguish between the presence of only one allele and the presence of two equal alleles), but, in general, the same genotype can also occur for other reasons, such as uniparental disomy. In this situation, the knowledge of both types of data can lead to the correct interpretation of the phenomenon, while it would not be possible with only the genotyping data or the copy number data. Several relationships between detected genotype and underlying copy number event (gain, amplification, loss of one or two copies) can be found.

Several methods were developed to solve the problem of copy number estimation from the raw data. We can roughly divide them into two classes: the ones that estimate the copy numbers as a piecewise constant function (such as CBS [59], CGHseg [63], GLAD [31] and HMM [21]) and the ones that are smoothing methods and estimate the copy numbers as a continuous curve (such as quantreg [18] and wavelet [28]). Other algorithms have been developed for the discovery of LOH regions, without distinguishing if they are caused by either the loss of one copy or other genomic events (IBD/UPD). Among them, two well-known (and most used) methods are dChip [8] and CNAT [2]. In literature, only one method [72] has been developed for the integration of these two types of data, with the purpose of estimating both copy number and LOH aberrations, and it uses HMM. In Chapter 2, I explain how generally the copy number data are

modeled and I describe some of the well-known methods for copy number or LOH estimation: CBS, HMM, CGHseg, quantreg and dChip.

In Chapter 3, I propose both a piecewise constant and two continuous regression methods, where the solution of the regression is found using Bayesian statistics, which is more suitable when we have regions with few data [65]. These methods build on and improve the methods presented by Hutter in [32, 33]: the Bayesian Piecewise Constant Regression (BPCR) and the Bayesian Regression Curve (BRC). In both, the data are assumed to be noisy observations of a piecewise constant function of which we want to estimate the number of segments, the boundaries of these segments and the value of the function in each interval (called level). The noise is assumed to be normally distributed and data points belonging to the same segment are conditionally independent of the level in that segment. The prior distributions of the parameters involved are defined in the following way. The number of segments is uniformly distributed in the interval $\{1, \ldots, k_{max}\}$. Given the number of segments, the boundaries are uniformly distributed in the set of all possible boundaries. The levels are independent and identically normally distributed. Finally, the hyper-parameters of the model (i.e. the variance of the noise and the mean and variance of the levels) are estimated from the data in an empirical Bayes way. The regression procedures require the computation of the posterior distributions of the parameters that we want to estimate, which needs the computation of the likehood over all possible number of segments, boundaries and levels. This can be done in an efficient way by using the dynamic programming presented in [32, 33].

In the original formulation, BPCR estimated the number of segments and each boundary with the maximum a posterior (MAP) estimator (which minimizes the posterior expected 0-1 error) and the segment levels with the posterior mean. However, the first two estimators failed to properly determine the corresponding parameters. In particular, the boundary estimator did not take into account the dependency among the boundaries and could estimate more than one breakpoint at the same position, losing segments. In the chapter, I define different segment number and boundary estimators to enhance the fitting procedure, by changing the error to minimize with respect to the posterior distribution. I also propose an alternative estima-

tor of the variance of the segment levels, which is useful in case of data with high noise. I compare my methods with other well-known and recent methods existing in literature on artificial data, showing that they generally outperform all the others. I also validate some results obtained by applying the piecewise constant regression (called mBPCR) on real data.

In cancer research, the accuracy in the DNA copy number estimation is crucial for the correct determination of the mutations that characterize the disease. In particular, the estimation of the breakpoints must be precise to detect correctly which genes are affected by these genomic aberrations. In this context, the use of mBPCR can highly improve the disease investigation, because it accurately determines breakpoints, is less sensitive to high noise and generally outperforms all the methods considered. Consequently, after its publication, mBPCR has been used as standard algorithm for the analysis of copy number data at the Laboratory of Experimental Oncology, Oncology Institute of Southern Switzerland (IOSI) [10, 68, 71].

In Chapter 4, I propose a Bayesian piecewise constant regression (called gBPCR), which infers the type of aberration occurred (high amplification, gain, loss of one copy, loss of two copies, IBD/UPD, normal state), taking into account all the possible influences in the microarray detection of the genotype, resulting from an altered copy number level [66, 67]. Namely, I model the distributions of the detected genotype given a specific genomic alteration and I estimate the parameters involved on public reference datasets. The prior distribution of the copy number alterations is derived from the copy number profile of the sample, while the probability of heterozygosity for each SNP is retrieved from the annotation file of the microarray used. The regression procedure is performed similarly to mBPCR, slightly changing the breakpoint estimator. I show the goodness of the method by applying it to both artificial and real data. I also compare it with two well-known methods for LOH estimation: dChip and CNAT. This comparison shows that they perform equally well on data with medium and low noise, while gBPCR outperforms the others on data with high noise. The model proposed is also more complete than the one in [72], since the latter cannot be applied to data, whose DNA sample come from a mixture of cell populations (which is usually the case for samples of patients affected by cancer). Moreover, since both types of data are inher-

ently noisy, a fully Bayesian model can include prior information that can lead to a better identification of the aberrations.

PART II: "Estimators of the intensity of stationary fibre processes applied to angiogenesis".

In solid tumors, cell proliferation is helped by the formation of a vascular network around the tumor. The vessels supply nutrient which allows the growth of the tumor. Hence, a challenge in cancer research is to find an antibody which is able to inhibit the formation of vessels. In order to quantify the effect of a specific antibody in the inhibition of this process from images of the vessels, we can estimate one or more parameters that characterize their geometry. We model them as a stationary planar fibre process and, for the purpose of antibody comparison, we estimate the *intensity* (i.e. mean length per unit area) of the corresponding processes.

In Chapter 5, I illustrate the basic concepts of the theory of fibre processes and some intensity estimators present in the literature. A fibre in $\mathbb{R}^2$ is a curve of class $\mathscr{C}^1$ defined on a bounded and closed interval. To model more complex objects, we can define a fibre system, i.e. a locally finite union of fibres, which can have only endpoints in common. Then, a fibre process is a random variable taking values in the space of fibre systems, endowed with a suitable σ-algebra. In particular, I consider *stationary* fibre processes, which are invariant under translations, and one of the main characteristics of this kind of processes is the intensity.

In literature, several estimators have been proposed (see for example [78]). The simplest estimator is the ratio between the fibre length in the window of observation and the area of the window. This estimator is unbiased and, if the fibre process is ergodic, it is also strongly consistent (the asymptotic properties of the intensity estimators are usually defined for a sequence of increasing windows of observation, which tends to $\mathbb{R}^2$). In practice, this estimator is not easy to compute. Usually the window of observation is a digitized image and the easiest way to measure the fibre length is to count the pixels belonging to the fibres. Since the pixel is a two-dimensional set, while a fibre is one-dimensional, in order to obtain a correct estimate we have to adjust the raw estimation with a factor which represents the mean length of the fibre in a pixel.

Another kind of estimation has been proposed by Ohser and Stoyan [58, 78], by intersecting the sample with a deterministic test fibre system with finite length (e.g. segments or circles). Using the properties of the point process derived from the intersection, it is possible to define an unbiased estimator of the intensity. This estimator consists in a counting measure defined on the point process of the intersections. But the resulting point process cannot have asymptotic properties, because of the finiteness of the length of test fibre system, and thus we cannot apply any ergodic theorem. In order to have an estimator easy to be computed and with good asymptotic properties, I intersected the sample with an independent, stationary and isotropic fibre process [53, 64]. Thus, I defined two estimators based on counting measures of the marked point process arisen from the intersection. These estimators are unbiased and, if the point process is ergodic, they are also strongly consistent. The main difficulties are in deriving their asymptotic normality.

As a consequence, in Chapter 6, I study conditions, regarding both the point process and the sequence of enlarging windows, under which the estimators are asymptotically normal. Penrose and Yukich [61] derived several central limit theorems for functionals of two types of point processes in $\mathbb{R}^n$, having independent increments: the Poisson and the binomial point processes. This property of independent increments (i.e. the points of the process, which fall in disjoint Borel sets, are independent) is crucial for their proofs. Since the fibres have not-null length, the points of intersection located at distance lower than the maximum length of the fibres (if it exists) are correlated. Therefore, the point process of intersection of fibre processes has in general not independent increments. Nevertheless, if the fibres are generated independently and have a.s. finite maximum length l, then at least the intersection points at a distance greater than l are independent.

Trying to mimicking the proofs in [61], in Chapter 6, I derive a central limit theorem which involves a general positive functional defined on a point process independent at distance l, for particular sequences of enlarging windows. Due to its general formulation, the theorem can be applied in a more general framework in the theory of point processes. Moreover,

in the chapter, I also deduce from this central limit theorem the asymptotic normality of some estimators of the intensity presented in Chapter 5.

In Chapter 7, I verify empirically the asymptotic properties of the estimators on simulated data. I use different choices for the shape of the fibres of both the process under study Φ_1 and the test process Φ_2, and several values for the other parameters that characterize Φ_2. In this way, I can observe whether and how the speed of convergence of the estimator depends on the characteristics of Φ_1 and Φ_2. In fact, the variance of our estimators depends both on the dimension of the window of observation and the intensities of the two fibre processes. Therefore, by suitably choosing Φ_2, we can reduce the variance of the corresponding estimator and thus obtain an accurate estimate, especially in case of a small window of observation. I also derive a method to approximate the variance of the estimator via its upper bound, when only one window of observation is available and it has a small size.

Since in real applications the window of observation is usually a digital image, I verify that the asymptotic properties of the estimators hold also on simulated images of fibre processes (that is I use the 2D-box approximation given by the pixels to represent the fibres). Finally, I apply some intensity estimators to some images of angiogenesis in eyes of mice, in order to determine in a quantitative way which antibody was more able to inhibit the angiogenic activity of the protein VE-cadherin.

Part I
Bayesian integrative genomics

Chapter 1
Genetics and microarray

Abstract Cancer is a complex disease characterized by the accumulation of genomic lesions, that alter cell biology [25, 60, 82]. Among the DNA aberrations, we recall copy number changes, translocations, regions of loss of heterozygosity (LOH). The identification of causative mutations involved in tumorigenesis is very important for prognosis and the creation of drug therapies. Moreover, the genetic study of tumor diseases can also determine subtypes of the diseases, leading to the comprehension of the reasons for differences in drug-resistance and to the creation of more specific drug therapies.
Microarray technology allows to measure biological quantities at DNA or RNA level [17, 39, 75]. Single nucleotide polymorphism (SNP) microarrays are able to measures the copy number and the genotype at thousands/millions of SNP positions for the analysis of both copy number changes and LOH along the genome. Therefore, they are used in cancer research to identify genes or genomic regions bearing these types of aberrations that are correlated with the disease.
The chapter is divided into two parts: in Section 1.1 we describe some basic knowledge of genetics and in Section 1.2, we introduce SNP microarray technology.

1.1 Basic biology of genetics

We introduce some basic concepts of genetics that will be useful for understanding the procedure of SNP microarrays and, consequently, for the comprehension of the random variables involved in the statistical models described in Chapters 2, 3 and 4. Moreover, they also motivate the importance of the development of good statistical tools for the analysis of microarray data, especially in cancer research.

1.1.1 DNA and RNA: structure and role

The DNA is a double-strand molecule, which entwines to achieve the shape of a double helix (Figure 1.1). Each strand consists of a sequence of nucleotides. Each nucleotide contains both a segment of the backbone of the molecule (a phosphate group and a sugar molecule), which holds the chain together, and a nitrogen-containing base, which interacts with the other DNA strand. There are four types of bases (and thus four types of nucleotides): adenine (A), cytosine (C), guanine (G) and thymine (T). The two strands are joints by hydrogenous bonds between complementary bases (*base-pairs*): A with T and C with G. The complementarity of the bases is the principle of both cell reproduction and gene expression. Moreover, the strands have a direction, which is one the opposite of the other.

The genetic material of each cell/individual is organized in chromosomes, which consist of DNA and proteins. The number of chromosomes is a characteristic of any species and, for human beings, the genome consists of 23 pairs of chromosomes. Each pair is made of two copies of the same chromosome (called *homologues*), one inherited from the mother and one from the father. Each homologue is divided in two parts by the centromere and, since the centromere is not placed exactly in the middle of the chromosome, they are called *long* and *short arm*.

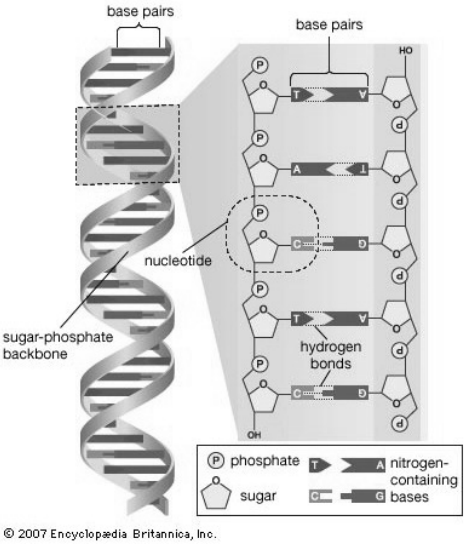

Fig. 1.1 DNA structure. [Adapted from Encyclopædia Britannica, Inc., copyright (2007)]

Every function in the living cell depends on proteins. A protein consists of a sequence of amino acids jointed by peptide bonds. Only 20 types of amino acids can be part of a protein and typically a protein contains between 100 and 1000 amino acids. The sequence of the proteins is determined on the basis of the information contained in the DNA. Each triplet of contiguous nucleotides (called *codon*) corresponds to a specific amino acid. Because the amino acids are 20 and the possible codons are $4^3 = 64$, the genetic code is redundant, i.e. different codons may codify the same amino acid. The sequences of DNA that encode proteins are inside the *genes*. Genes are sequences of nucleotides consisting of protein coding regions (called *exons*) interspaced by segments of noncoding regions (called *introns*). Any gene is preceded by a promoter region, which is a binding site for proteins (called *transcription factors*) that influence the transcription machinery. Therefore, this sequence controls the conditions under which the gene will be transcribed.

The process of protein synthesis involves RNA molecules. The RNA is similar to DNA with three differences: 1) it consists of only one strand of nucleotides, 2) the thymine is substituted by the uracil (U) and 3) it contains molecules of ribose sugar instead of deoxyribose sugar. The messenger RNA (mRNA) is synthesized using the DNA as a template (*transcription* process) and is then used for *translation* into protein (Figure 1.2). Each mRNA transcript represents a copy of only the exons of the gene, corresponding to the protein to be synthesized, and each molecule of protein requires and consumes one transcript. Therefore the rate of synthesis of a protein can be estimated by quantifying the abundance of corresponding mRNA transcripts, although the correspondence is not exact since some transcripts are degraded before protein translation.

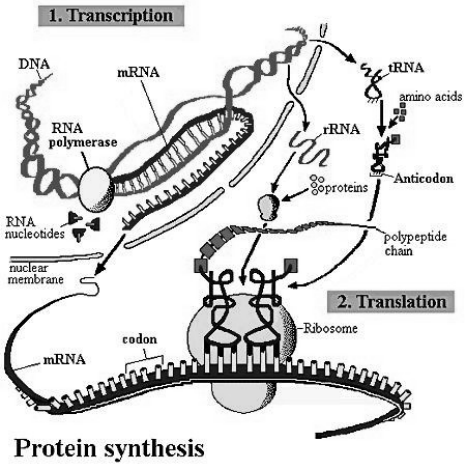

Fig. 1.2 Protein synthesis [Adapted from http://beckysroom.tripod.com/].

In general, all cells of the same organism contain the same DNA. Nevertheless, the cells differ from each other in function and their function can change over time. These functional differences are determined by differences in the abundance of the various types of proteins. Therefore, DNA

aberrations can lead to a different expression of specific genes, that can change the behavior of the cell. We refer to the *over-expression* or *under-expression* of a gene if the abundance of its transcript is, respectively, higher or lower with respect to the one in a reference cell (usually a normal cell of the same tissue).

1.1.2 DNA polymorphisms

Due to the high number of nucleotides in the DNA strands of the genome, we can have a high number of different genomes. Actually, most of human genome is equal for all individuals and the DNA markers that can show differences among individuals are called *DNA polymorphism*. Examples are *single nucleotide polymorphisms* (SNPs) and *copy number polymorphisms* (CNPs) or *copy number variations*. The alternative variants of a DNA polymorphism are called *alleles*. In order to be classified as a polymorphism, a genomic variant must have a minor allele frequency greater than 1% in a given population to avoid that it represents a rare mutation.

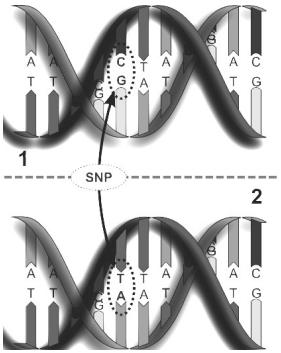

Fig. 1.3 Single nucleotide polymorphism. [Adapted from David Hall, http://en.wikipedia.org/wiki/File:Dna-SNP.svg, available under Creative Commons Attribution 2.5 Generic]

A SNP is a single base-pair position where the nucleotide can assume two possible bases (alleles) out of the four (Figure 1.3). In general, since

we have two copies of each chromosome, the *genotype* at any SNP can be: *AA*, *BB* or *AB*, where *A* and *B* represent the two possible alleles. In an individual, a SNP is said *homozygous* if the homologuos chromosomes carry the same allele (AA or BB), otherwise is said *heterozygous* (AB). It has been estimated that the number of SNPs in the human population is about 10 millions [47] and about 3 millions SNPs have been already identified [48] by the HapMap Consortium. In biomedical research we do not need to know the value of all SNPs. In fact, SNPs on a small chromosomal segment tend to be transmitted as a block, forming a haplotype. This correlation between alleles at nearby sites is known as linkage disequilibrium and enables the prediction of the genotypes at a large number of SNP loci from known genotypes at a smaller number of representative SNPs, called tag SNPs.

Instead, CNPs are defined as portions of the genome that can be deleted or present in extra copies. Their width ranges from 1 kb (one thousand base-pairs) to 1 Mb (one million base-pairs).

As a consequence of the existence of DNA polymorphisms, genes may presents alternative forms (alleles). Due to the redundancy of the genetic code, different alleles of a gene may or may not code for different amino acid sequences, sometime with drastic effects. Usually, harmful alleles are *recessive*, i.e. the organism need to carry those alleles on both homologues to express them. Therefore, one of the key goal in studying DNA polymorphisms is to identify gene variants associated with a disease.

1.1.3 Cell cycle, mitosis and meiosis

The formation of a complex organism from the zygote implies cell replication, growth and differentiation. The mechanism of replication is called *mitosis* (Figure 1.4). In the mitotic division, the chromosomes are separated in the two daughter cells so that each daughter cell contains the same genome of the mother cell. After their birth, the daughter cells grow, provide to the duplication of the chromosomes and perform their function inside the tissue in which they are (this phase is called *interphase*). The

period from the birth up to the next mitosis is called *cell cycle* (see Figure 1.5). Several key events in cell cycle are monitored and, when defects are identified, the progression through the cell cycle is halted at a *checkpoint*. For example, if a cell reports a DNA damage, it can be repaired or self-destroyed by a mechanism called *apoptosis*. Once the cell is fully differentiated, it becomes *quiescent* and stops dividing.

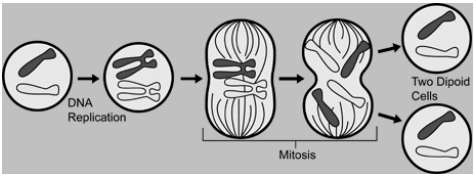

Fig. 1.4 Mitosis. [Adapted from National Library of Medicine (NLM) website, http://www.ncbi.nlm.nih.gov/]

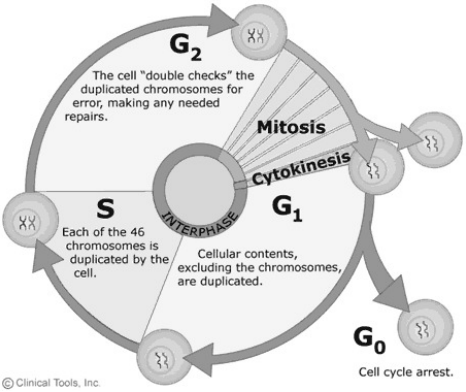

Fig. 1.5 Cell cycle. [Adapted from Clinical Tools, Inc., available under Creative Commons Attribution-Share Alike 3.0 United States License]

Another mechanism of cell division is the one that lead to the formation of the gametes. This process is called *meiosis* and can be done only by the germ cells. The meiosis is divided in two principal phases (see Figure 1.6). In the first, the chromosomes are duplicated, undergo the crossover (mechanisms of exchange of genetic material between ho-

mologue chromosomes) and are separated in two daughter cells. In the second phase, in each daughter cell, the homologues of the chromosomes are separated and the cell is divided in two cells.

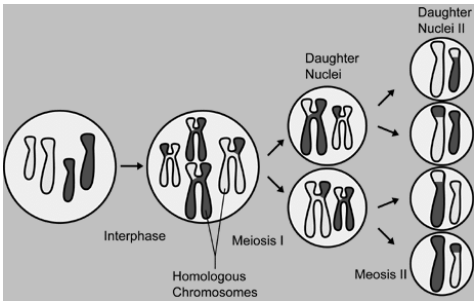

Fig. 1.6 Meiosis. [Adapted from National Library of Medicine (NLM) website, http://www.ncbi.nlm.nih.gov/]

During the cell cycle, most of the biochemical and metabolic activities of the cell are carried out by proteins. Each of these activities corresponds to one or more metabolic pathways. A metabolic pathway is a set of chemical reactions that take place in a definite order to convert a particular starting molecule in one or more specific products. Each of its steps is regulated by an enzyme (and thus by its corresponding gene).

We can observe that, during the cell life, several mechanisms can be altered. Among the most important ones, we can recall the inhibition of apoptotis, the increase of cell growth over the normal or the defects in the DNA repair pathways. Each cell activity can be interrupted by alterations at different steps in the corresponding pathway. Therefore, cancer is highly related to DNA aberrations, which can appear during meiosis (*germ line* lesions) or mitosis (*somatic* lesions). For example, in meiosis the homologous recombination of repeated sequences (which belong to different genomic regions) can lead to the deletion and/or duplication of the genetic material between the repeats [44]. Moreover, since the same phenotype of the disease can be achieved by different genomic alterations targeting different genes involved in the same pathway, patients affected by the same disease can present heterogeneous DNA lesions.

1.1.4 Some genetic lesions in cancer

Cancer is a group of diseases that are characterized by the uncontrolled growth of the cells as a result of mutations that effect a limited number of genes. In the majority of the cases, the genetic changes are only in somatic cells (i.e. the disease is not familial). Cancer cells share several properties not found in normal cells:

- loss of contact inhibition (i.e. process that allows the inhibition of growth and division through cell-to-cell contact),
- loss of growth-factor dependence,
- insensitivity to anti-growth signals,
- evasion of apoptosis,
- immortality (no cell senescence),
- ability to metastasize and invade other tissues,
- sustained angiogenesis (i.e. formation of new vascular network which supplies more nutrient to tumor cells).

Tumor formation is a multi-step process in which normal cells evolve into cells with increasing neoplastic phenotype, through a sequence of randomly occurring alterations of DNA. Some DNA aberrations can occur at level of nucleotide sequence (e.g. single point mutations) or at level of chromosomal structure (e.g *translocations* and *inversions*) and number (e.g. deletions and amplifications); Figure 1.7.

A translocation is a chromosomal aberration resulting from the interchange of parts between non-homologous chromosomes. Translocations can be formed by interchange of parts between two broken chromosomes or by recombination between copies of repeated DNA sequences present in two non-homologous chromosomes. We call inversion a chromosomal region in which the linear order of a group of genes is the reverse of the normal order. An inversion can be formed, for example, by two break events in a chromosome in which the middle segment is reversed in orientation before breaks are healed. *Copy number (CN) changes* are defined as genomic regions where the number of DNA copies is different from the normal copy number (which is two, for human beings). We can divide these

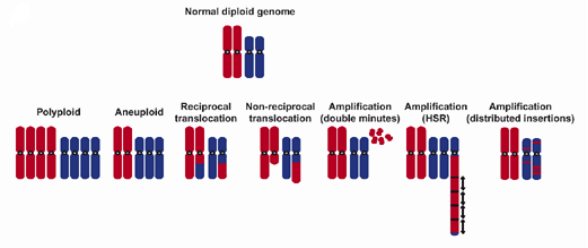

Fig. 1.7 Schematic illustration of some mechanisms by which chromosomal aberrations arise. [Adapted by permission from Macmillan Publishers Ltd: *Nature genetics* [3], copyright (2003)]

aberrations in four categories: *homozygous deletion* (CN=0), *loss* (CN=1), *gain* (CN=3 or 4) and *high amplification* (CN>4). For example, a deletion (CN=0 or 1) can occur by chromosome breakage and reunion.

Especially in solid tumors, two types of genes are the major targets of the aberrations in the multi-step cancer progression: *proto-oncogenes* and *tumor-suppressor genes*. The former are genes that, through a DNA aberration, are improperly enhanced to be expressed and their expression promotes cell proliferation or inhibits apoptosis. Examples of alterations that can activate a proto-oncogene are: amplification (e.g. of the genes FGFR and EGFR), mutation (e.g. targeting the gene Ras) and chromosomal translocations. In contrast, tumor-suppressor genes are genes that normally negatively control cell proliferation or activate the apoptotic pathway. Examples of lesions that inactivate this kind of genes are: homozygous deletion (e.g. of CDKN2A) and loss of the remaining normal allele, after the mutation of one copy of the gene (e.g. of TP53 and BRCA2). The last type of event can be detected as a *loss of heterozygosity* (LOH) of polymorphic markers in the region of that gene.

In general LOH can arise by several mechanisms (see Figures 1.8 and 1.9), such as deletion, somatic or germ-line recombinations resulting in uniparental disomy, autozygosity and chromosomal nondisjunction. When the LOH occurs without a change in copy number, it is referred as *copy-neutral LOH* and a possible cause of this alteration can be uni-

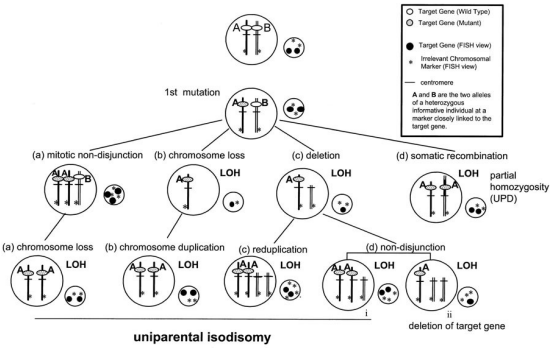

Fig. 1.8 Mechanisms leading to loss of heterozygosity due to uniparental disomy. Strategies for possible recombination events leading to loss of heterozygosity and their detectable sequelae. Informative microsatellite alleles are A and B and are in the region of the tumor suppressor gene. White and grey markers distinguish specific alleles of the tumor suppressor gene. Solid black signals indicate the FISH interphase appearance with a specific tumor suppressor gene probe that cannot distinguish alleles. The asterisk (*) is a distal FISH chromosomal marker (or a centromeric probe). Note that several mechanisms can give rise to a loss of heterozygosity, but only 2b, 2c, and 3di and 3dii are associated with a copy number deletion. Cells with any of these events may be selected during clonal outgrowth. [Reprinted by permission from Macmillan Publishers Ltd: *Modern Pathology* [54], copyright (2002)]

parental disomy or autozygosity. Uniparental disomy (UPD) describes the inheritance of a pair of homologous chromosomes (or only a portion of them) from a single parent [40, 80]. Either the presence of both homologues (heterodisomy), or two copies of one homologue (isodisomy), or a mixture of both are possible, due to meiotic recombinations. Similar events can also happen during the mitosis. In cancer cells, uniparental isodisomy can also occur when an homologue of a part of a chromosome is lost and the remaining homologue is duplicated. In cancer biology, the relevance of copy-neutral LOH aberrations, derived from meiotic or mitotic isodisomy, could for example lie in the inactivation of one allele of a tumor suppressor gene, which is then duplicated (while the remaining normal

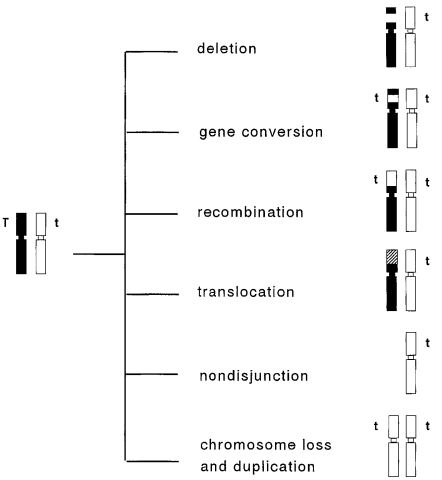

Fig. 1.9 A model for mechanisms which may contribute to LOH. In cells carrying one wildtype allele (T) and one mutant allele (t) of a tumor suppressor gene, LOH events result in the expression of the recessive mutation. It has been postulated that mechanisms like deletion, gene conversion, recombination, non-disjunction, and non-disjunction followed by duplication of the remaining chromosome may result in LOH. In addition translocation may also result in LOH [Reprinted by permission from John Wiley & Sons, Inc.: *Genes, chromosomes and cancer* [46], copyright (1998)]].

allele is lost). Instead, autozygosity describes a situation where the homologues are identical by descendent (IBD), because they are inherited from a common ancestor. Among human beings, inbreeding is usually uncommon because of social conventions and laws, although in small isolated populations (like religious communities, isolated villages) it does occur, mainly between relatives more distant than second cousins (remote relatives). The most common type of close inbreeding is between first cousins and the effect is an increase in the frequency of homozygous genotypes for rare, harmful recessive allele.

In human acute leukemia (cancer that arises in white blood cells), the initial genetic events usually are not alterations in cell-cycle regulation or checkpoints. Up to 65 percentage of cases arise as a consequence of chromosomal translocation involving genes that play a role in blood cell development. The breakpoints can occur in introns of two genes on different

chromosomes, producing a gene that encodes a chimeric protein (*fusion gene*) which may interfere with the normal cell development. Since one of the two genes is normally expressed in the cell, also the fusion gene will be expressed. Nevertheless, both the uniqueness of these chimeric proteins and the fact that they are present only in cancer cells, make them a target for drugs. If one could successfully attack the cells expressing those proteins, then one could selectively kill all cancer cells.

1.2 Microarrays for DNA profiling

In the last decade, several types of microarray have been developed, for example, to measure simultaneously the abundance of transcripts of thousands of genes or the copy number at thousands/millions of DNA positions. The main difference with other conventional biological techniques is that, in short time, they are able to measure simultaneously a specific quantity at several units (e.g. genes or SNP positions). For example, the *fluorescent in situ hybridization* (FISH) technique can be used to measure the copy number of specific DNA sequences. Using the bonds caused by the complementarity of the bases, the fluorescent probes bind the specific target sequences of DNA (*hybridization*) and a fluorescent microscopy is used to find out where the probes bound. Therefore, this technique cannot be massively applied to find the copy number of thousands of genes.

In the following, we will consider only microarrays that measure DNA copy number and/or genotyping. The first type of array are called *array comparative genomic hybridization* (aCGH) microarray. In general, an aCGH microarray consists of a glass slide (called also *chip* or *array*) with spotted DNA probes (i.e. single-strand DNA segments). The probes are attached on the slide to form a matrix in order that they can be uniquely identified and, for each target sequence, more than one probe is used. The technique is based on the complementarity property of the basis. Several types of aCGH arrays exist and they differ for the type of probe used, the resolution (i.e. how many positions of DNA are measured), the design of the matrix and the biochemical process used.

In the next applications (in Chapters 3 and 4), we will consider only oligonucleotide-based microarray of Affymetrix (Santa Clara, CA, USA), whose target sequences are oligonucleotides (sequence of 25 nucleotides) containing a SNP position. Briefly, the biochemical process is the following (Figure 1.10):

- the DNA is extracted from a sample of cells of a patient,
- using a restriction enzyme, the genomic DNA is cut in specific fragments to reduce the genomic complexity,
- using the *polymerase chain reaction* (PCR), each fragment is amplified, augmenting its abundance,
- the fragments are cut again (in pieces which are related to the probes), labeled with a fluorescent tag and hybridized on the chip,
- the microarray is "washed" to eliminate the fragments that did not hybridize,
- the array is stimulated with the laser and the intensities of the probes are measured through a scanner.

At the end, all probes on the microarray, corresponding to a target sequence, should be bound to a quantity of labeled DNA that is proportional to the copy number of the target sequence for that patient. Therefore, by measuring the intensity of label bound to the probes, we can obtain an estimate of the copy number of the target sequences.

In the Affymetrix GeneChip Mapping 10K Array, the probes are designed in the following way. For each SNP, we have probes for both alleles, setting as position of the SNP the center of the oligonucleotide or a position -4, -1, 1 and 4 around the center. For each of these sequences, we have also probes for both directions of the strand (Figure 1.11). In this way, in the 10K Array (11,464 SNP positions), there is a total of 40 probes interrogating the same SNP, while, in the 250K Nsp Array (262,217 SNP positions), only a subset of 24 probes is used in order to measure more SNP positions. Moreover, each probe (called *perfect match*, PM) is paired with a *mismatch* one (MM), having a wrong nucleotide in the center. The MM probes are assumed to bind to non specific sequences at the same rate as the PM probes and thus they can be used to correct the intensity measured by the PM probe for background nonspecific hybridization. However, the

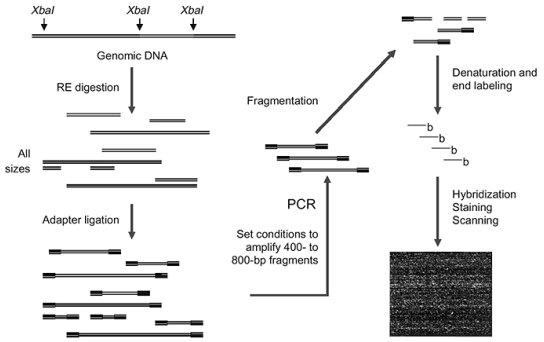

Fig. 1.10 Scheme of Affymetrix GeneMapping 10K Array [Adapted by permission from Macmillan Publishers Ltd: *Nature biotechnology* [37], copyright (2003)].

MM probes can also bind to differently-specific labeled subsequences in the sample and thus some preprocessing methods do not take into account the background adjustment given by MM probes (see, for example, [12]). The probes interrogating the same SNP position are usually spotted sparse in the matrix to reduce the effects in the measurement of possible artifacts.

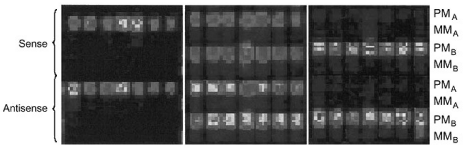

Fig. 1.11 SNP miniblock showing the hybridization in three individuals, demonstrating the three possible genotypes: AA (left), AB (middle), BB (right) [Adapted by permission from Macmillan Publishers Ltd: *Nature biotechnology* [37], copyright (2003)].

Due to their design, SNP microarray are able to measure simultaneously the copy number and the genotype at several SNP positions. In fact, by comparing the intensity of the probes targeting the different alleles of

the SNP, it is possible to determine its genotype. Moreover, we can also identify its copy number by taking into account all the information given by the probes of both alleles. The use of these arrays for the identification of DNA alterations in cancer has three major advantages: 1) their genotyping ability allows for analysis of LOH, 2) they determine the copy number of each interrogated SNP and 3) the density of SNP loci being interrogated allows for very high-resolution analysis (for example, in comparison to probes targeting only genes). Moreover, as recently shown [36], SNP microarrays can also potentially detect the breakpoints involved in unbalanced translocations, allowing the identification of fusion genes (described in Subsection 1.1.4).

1.2.1 Image analysis and preprocessing of Affymetrix SNP microarray data

The acquisition of the so called raw data is not automatic from the scanning of the array. After the microarray has been scanned, an image file is created, storing all pixel-level intensities. Obviously, there are many more pixels than probes, thus an image analysis is needed to process and convert the pixel level data into measures of the probe intensities. The main step of image analysis are:

- gridding, i.e. overlay a rectangular grid onto the pixels in order to isolate the spots corresponding to different probes,
- segmentation, i.e. inside each cell, identify the pixels belonging to the probe (*foreground region*),
- intensity extraction, i.e. for each probe, the intensity is estimated as the 75^{th} percentile of the pixel intensities in the foreground region.

Due to the complexity of the microarray procedure, these estimated intensities are not biologically reliable. A preprocessing method is needed to correct the intensity value for technical noise, such as fragment length, nucleotide content in the fragment, lab effect. The preprocessing procedure is essential for allowing the comparison among the data obtained in

different microarray (i.e. different patients). Moreover, for the genotyping, a clustering algorithm (usually called *genotyping calling algorithm*) is needed for the classification of the SNPs as AA, BB or AB. The two main methods that solve this issue are: BRLMM [1] and CRLMM [12].

Chapter 2
The problem of copy number and LOH estimation

Abstract Lesions at DNA level represent the cause of cancer and of many congenital or hereditary disorders. The change of the number of copies of DNA in a genomic region is one of the most common aberrations. In normal cells each genomic segment is present in two copies, but, for example, in tumor cells the genome can present regions of deletions (copy number one or zero), gains (copy number three or four) or amplifications (copy number greater than four). Thus, in general, the DNA copy number along the genome can be represented as a piecewise constant function.
With microarray technology it is possible to simultaneously measure the copy number along the genome at hundred thousands of positions (see for example [30]). However, raw copy number data are generally very noisy. Hence, it is important to define a method which allows to estimate the number of regions with different copy number, the endpoints of these regions (called *breakpoints*) and their copy number.
In Section 2.1, we explain how the copy number data are commonly modeled and in Section 2.2 we present some well-known methods for copy number estimation.
Another kind of lesion of DNA is the loss of heterozygosity (LOH), i.e. the transformation of the SNPs from heterozygous to homozygous due to genomic events, such as the loss of one copy or uniparental disomy (see

Subsection 1.1.4). SNP microarrays are also able to detect the homozygosity status at hundred thousands of SNP loci. In Section 2.3, we describe two methods which are currently the most used for the estimation of the LOH profile.

2.1 The general hypotheses of the copy number inference

Since the genome can be divided in regions of constant copy number (or $\log_2$ratio of the copy number), the DNA copy number profile can be modeled as a piecewise constant function. Therefore, many methods model the problem in the following way.

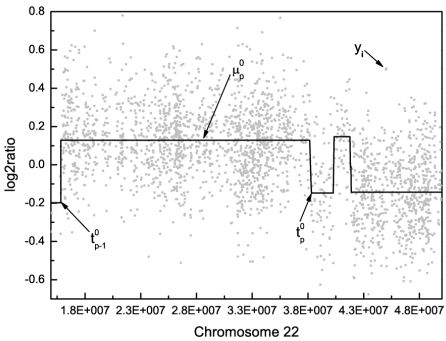

Fig. 2.1 Example of piecewise constant profile. We used the $\log_2$ratio data of chromosome 22 of cell line JJN-3 (unpublished).

In general, all cells of the same organism contain the same DNA. Nevertheless, the cells differ from each other in function and their function can change over time. These functional differences are determined by differences in the abundance of the various types of proteins. Therefore, DNA aberrations can lead to a different expression of specific genes, that can change the behavior of the cell. We refer to the *over-expression* or *under-expression* of a gene if the abundance of its transcript is, respectively,

higher or lower with respect to the one in a reference cell (usually a normal cell of the same tissue).

Let $Y \in \mathbb{R}^n$ be a random vector, such that each component (called *data point* or, since Y represents a quantity measured on part of the genome, *probe*) is conditionally normally distributed and conditionally independent:

$$Y_i|_{\widetilde{\mu}_i^0,\sigma^2} \sim \mathcal{N}(\tilde{\mu}_i^0, \sigma^2), \qquad i = 1,\ldots,n. \tag{2.1}$$

Let us assume also that Y represents a noisy observation of a piecewise constant function, which consists of k_0 horizontal segments. Then, the segment level at a generic position i ($\widetilde{\mu}_i^0$) does not assume different values for each i, but the data are divided into k_0 intervals (with boundaries $0 = t_0^0 < t_1^0 < \cdots < t_{k_0-1}^0 < t_{k_0}^0 = n$) where $\widetilde{\mu}_{t_{q-1}+1}^0 = \ldots = \widetilde{\mu}_{t_q}^0 =: \mu_q^0$ for each $q = 1$, …, k_0 (see Figure 2.1). Hence, μ_q^0 represents the level of the q^{th} segment. Given this setting, the joint distribution of Y (i.e., the likelihood) is

$$\begin{aligned} p(y\,|\,\mu^0,t^0,k_0,\sigma^2) &= \prod_{p=1}^{k_0} p(y_{t_{p-1}^0,t_p^0}\,|\,\mu_p^0,t^0,k_0,\sigma^2) \\ &= \prod_{p=1}^{k_0} \prod_{i=t_{p-1}^0+1}^{t_p^0} p(y_i\,|\,\mu_p^0,\sigma^2) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{\frac{1}{2\sigma^2}\sum_{p=1}^{k_0}\sum_{i=t_{p-1}^0+1}^{t_p^0}(y_i-\mu_p^0)^2\right\}, \end{aligned}$$

where $y_{i,j} = (y_{i+1}, \ldots, y_j)$.

The goal is to estimate the levels $\mu^0 = (\mu_1^0, \ldots, \mu_{k_0}^0)$ of all the segments. In order to do that, we need to estimate also the number of the segments k_0 and the partition of the data t^0.

Usually, the vector $Y \in \mathbb{R}^n$ represents the vector of the observed $\log_2$ratio of the copy number at n positions along the genome, and the triplet (k_0,t^0,μ^0) identifies the true piecewise constant $\log_2$ratio profile along the genome. The $\log_2$ratio scale of the data (the ratio is computed with re-

spect to a normal reference sample) is usually assumed to be normally distributed (see Figure 2.2). We refer to the whole model just described as the *piecewice constant formulation* and the methods that assume this model are sometimes called segmentation methods. There are other methods that only suppose a normal distribution of the data points $Y \in \mathbb{R}^n$, without assuming (taking into account) that the vector of the means $\widetilde{\mu}^0$ (i.e., the true $\log_2$ratio) is piecewise constant.

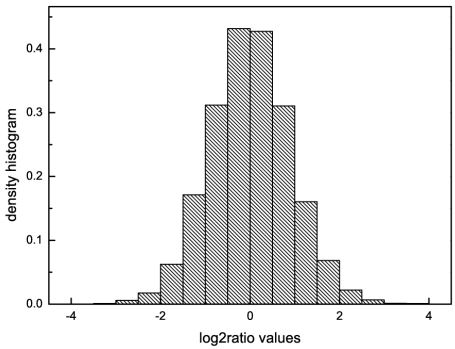

Fig. 2.2 Density histogram of the raw $\log_2$ratio values of cell line JEKO-1, obtained by using the Affymetrix GeneChip Mapping 10K Array [69].

2.2 Estimation of copy number profile in literature

Several methods have been developed to infer the copy number profile, using the formulations described in Section 2.1 (see Table 2.1). We can roughly subdivide all of these methods into two classes: the ones that estimate the copy numbers as a piecewise constant function and the others that estimate the copy numbers as a continuous curve. The methods belonging to the latter group are called *smoothing methods*.

Among the methods belonging to the first class, we can find the following. The Circular Binary Segmentation (CBS) approach is a recursive

Table 2.1 The table summarizes some of the well-known and recent methods for copy number estimation.

type	name	reference
piecewise constant method	CBS	[59]
piecewise constant method	CGHseg	[63]
piecewise constant method	GLAD	[31]
piecewise constant method	HMM	[21]
piecewise constant method	BioHMM	[45]
piecewise constant method	Rendersome	[57]
smoothing method	wavelet	[28]
smoothing method	quantreg	[18]
smoothing method	smoothseg	[29]

method in which the breakpoints are determined on the basis of a test of hypothesis, with null hypothesis that in the interval considered there is no change in copy number [59]. Picard et al. [63] used a piecewise constant regression model, where the parameters are estimated by maximizing a penalized likelihood (i.e. the likelihood with the addition of a penalty function). This method is usually denoted with the abbreviation CGHseg. The GLAD method is another piecewise constant regression method, but in this case the parameters are estimated by maximizing a weighted likelihood [31]. Fridlyand et al. [21] applied Hidden Markov Models (HMM), while Marioni et al. [45] defined an HMM method which takes into account the distance among the data points (BioHMM). Recently, Nilsson et al. [57] derived a segmentation method based on total variation minimization, called Rendersome. It is optimized for gene expression data, but the authors affirm that it can be used also on copy number data.

Among the smoothing methods, Hsu et al. [28] used a wavelet regression method with Haar wavelet. Eilers and de Menez [18] applied a quantile smoothing regression (quantreg), with the solution found by minimizing a loss function based on the L_1 norm, to obtain a flatter curve. Huang et al. [29] proposed smoothseg, i.e. a smooth segmentation method based on a doubly heavy-tailed random-effect model.

In this section, we describe four well-known algorithms among the previously cited: CBS, HMM, CGHseg and quantreg.

2.2.1 The circular binary segmentation (CBS) method

The *Circular binary segmentation* procedure (CBS) is based on the likelihood ratio test for testing the null hypothesis that all data points have the same mean against that there is one change at the unknown position t (see [59, 74]). Let us assume to test:

$$H_0: \ \tilde{\mu}_1 = \tilde{\mu}_2 = \cdots = \tilde{\mu}_n$$
$$H_1: \ \tilde{\mu}_1 = \cdots = \tilde{\mu}_t \neq \tilde{\mu}_{t+1} = \cdots = \tilde{\mu}_n, \text{ for some } t \in \{1, \ldots, n-1\}.$$

Since we assume that the data points are normally distributed with variance σ^2 and unknown mean, the likelihood-ratio test statistic is given by

$$\frac{p(Y \mid H_0)}{\sup_{1 \leq t \leq n-1} p(Y \mid H_t)} = \frac{(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \overline{Y})^2\right)}{\sup_{1 \leq t \leq n-1} \left\{(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{t} (Y_i - \overline{Y}_{0,t})^2 + \sum_{i=t+1}^{n} (Y_i - \overline{Y}_{t,n})^2\right]\right)\right\}}, \quad (2.2)$$

where $H_t = \{\tilde{\mu}_1 = \cdots = \tilde{\mu}_t \neq \tilde{\mu}_{t+1} = \cdots = \tilde{\mu}_n\}$ and

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$
$$\overline{Y}_{i,j} = \frac{1}{j-i} \sum_{h=i+1}^{j} Y_h.$$

After simplifications, Equation (2.2) becomes

$$\inf_{1 \leq t \leq n-1} \exp\left(-\frac{1}{2\sigma^2 (t^{-1} + (n-t)^{-1})} \left(\overline{Y}_{0,t} - \overline{Y}_{t,n}\right)^2\right)$$

$$= \exp\left(\sup_{1\le t\le n-1} \left| \frac{\overline{Y}_{0,t} - \overline{Y}_{t,n}}{(t^{-1}+(n-t)^{-1})^{1/2}} \right| \right).$$

and we can use $Z := \sup_{1\le t\le n-1} \left| \frac{\overline{Y}_{0,t}-\overline{Y}_{t,n}}{(t^{-1}+(n-t)^{-1})^{1/2}} \right|$ as a statistic for the test. If the statistic exceeds the upper α^{th} quantile of the null distribution of Z, we reject the null hypothesis and estimate the location of the change-point as

$$\arg\max_{1\le t\le n-1} \left| \frac{\overline{Y}_{0,t} - \overline{Y}_{t,n}}{(t^{-1}+(n-t)^{-1})^{1/2}} \right|.$$

The *binary segmentation procedure* applies the test recursively in each detected segment until no more change-points are found. The major problem of this procedure is the estimation of a single change-point at a time, which is not suitable for the detection of aberrations of small width (i.e. when two consecutive change-points are close). In the *circular binary segmentation* defined in [59], the two endpoints of the interval are joint together (to form a circle) and we test

$$\begin{aligned} H_0 &: \ \tilde{\mu}_1 = \tilde{\mu}_2 = \cdots = \tilde{\mu}_n \\ H_1 &: \ \tilde{\mu}_{t_1+1} = \cdots = \tilde{\mu}_{t_2} \neq \tilde{\mu}_{t_2+1} = \cdots = \tilde{\mu}_n = \tilde{\mu}_1 = \cdots = \tilde{\mu}_{t_1}, \\ &\quad \text{for some } t_1 < t_2, t_1, t_2 \in \{1,\ldots,n-1\}. \end{aligned}$$

Therefore, the statistic is given by

$$Z = \sup_{1\le t_1<t_2\le n} \left| \frac{\overline{Y}_{t_1,t_2} - \frac{(n-t_2)\overline{Y}_{t_2,n}+t_1\overline{Y}_{0,t_1}}{n-t_2+t_1}}{[(t_2-t_1)^{-1}+(n-t_2+t_1)^{-1}]^{1/2}} \right|$$

and, if we reject the null hypothesis, we estimate the two change-points as

$$\arg\max_{1\le t_1<t_2\le n} \left| \frac{\overline{Y}_{t_1,t_2} - \frac{(n-t_2)\overline{Y}_{t_2,n}+t_1\overline{Y}_{0,t_1}}{n-t_2+t_1}}{[(t_2-t_1)^{-1}+(n-t_2+t_1)^{-1}]^{1/2}} \right|.$$

The advantage of the second procedure is that it looks at two change-points at a time and we can obtain also single changes if t_1 or t_2 is equal to an endpoint of the segment. Nevertheless, the application of this procedure can lead to an edge effect in the estimation. In fact, if t_1 is estimated "close" to 1 and t_2 "close" to n, there might be only one change-point. To avoid this issue, in [59], the authors added to the procedure two tests: the first one to verify if t_1 is a change-point in the segment $[1, t_2]$, the second one to verify if t_1 is a change-point in the segment $[t_1, n]$. Due to the difficulties in the definition of the "closeness" of the estimated boundaries to the endpoints of the interval, the two tests are performed for all paired estimated change-points.

Moreover, to deal with the real copy number data, which are very noisy, other two modifications were added to the procedure:

1. a smoothing of the outliers before the segmentation,
2. a test to delete the estimated change-points not biologically meaningful (for example, they can be related to technical bias).

In the smoothing, for any fixed position i, they consider a region from $i-R$ to $i+R$ ($R \in \{1, 2, 3, 4, 5\}$) and they defined

$$m_i = \overline{Y}_{i-R-1,i+R}$$
$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y}_{0,n})^2.$$

If Y_i is the maximum or the minimum value in the region, they take

$$j = \underset{i-R \le h \le i+R}{\arg\min} |Y_h - Y_i|$$

and, if $|Y_i - Y_j| > 2\hat{\sigma}$, then Y_i is substituted by $m_i + \text{sign}(Y_i - Y_j)3\hat{\sigma}$.

In order to eliminate meaningless breakpoints, they perform a test. Assuming that $k-1$ breakpoints are found, they define $SS(k-1)$ as the sum of squares of the data points with respect to their segment average. Then, they compute $SS(1), \ldots, SS(k-2)$, using the best set of breakpoints of size, respectively, $1, \ldots, k-2$, choosing the change-points among the es-

timated ones. Finally, they determine k' such that

$$k' = \min_{1 \le p \le k-1} \frac{SS(p)}{SS(k-1)} - 1 < \gamma, \qquad \gamma = 0.05 \text{ or } 0.10,$$

and the final breakpoints are the ones use to calculate $SS(k')$.

We can observe that, due to the smoothness procedure, CBS is not able to detect small segments in the profile. Nevertheless, copy number changes with small width can be possible, especially using microarrays with a low resolution or high density microarray with probe at CNP positions.

2.2.2 The Hidden Markov model (HMM) method for copy number estimation

Another commonly used algorithm models the copy number data with a *Hidden Markov Model* (HMM), see [21, 84]. In this setting, the observations are assumed to depend (in a probabilistic way) on the value of the "state" at that position. In other words, we have an underlying sequence of random variables (called *states*), which is not observable (it is hidden) and it is observable only trough another sequence of random variables, which produces the observations. The states are assumed to belong to a discrete set $\{s_1, \ldots, s_{N_S}\}$. In our case, the states are type/degree of copy number changes (S) and the observations are the raw data (Y).

The model is characterized by the definition of the following quantities:

1. N_S, the number of the possible states,
2. the *initial state distribution* $\pi = \{\pi_l\}_{l=1}^{N_S}$, where $\pi_l = \mathrm{P}(S_1 = s_l)$, $l = 1, \ldots, N_S$,
3. the *transition probability distribution* $A = \{a_{pq}\}_{p,q=1}^{N_S}$, where $a_{pq} = \mathrm{P}(S_i = s_q \,|\, S_{i-1} = s_p)$, $p, q = 1, \ldots, N_S$, i.e. at each time the probability to pass from one value of the state to another,

4. the *emission distribution* $B = \{b_l(y)\}_{l=1}^{N_S}$, where $b_l(y) = p(y|m_l, \Sigma_l)$, i.e. the distribution of any observation knowing the value of the corresponding state.

The authors assume that, given the l states, y has a multivariate normal distribution with mean vector m_l and covariance matrix Σ_l. Notice that, in order to satisfy the standard probability constrains, the elements of the matrix A have to fulfil the following properties:

$$\begin{aligned} a_{pq} &\geq 0, \qquad p,q = 1,\ldots,N_S, \\ \sum_{q=1}^{N_S} a_{pq} &= 1, \qquad p = 1,\ldots,N_S. \end{aligned}$$

Moreover, assuming that from each state value it is possible to reach any other state value, then $a_{pq} > 0$, for all $p,q = 1, \ldots, N_S$.

In [21], the parameters are initialized as following. The initial state distribution is defined to have a high probability at the value corresponding to the "normal" copy number and equal probabilities for the remaining values. The matrix A is assigned in order to have high probability to remain in the same state value. In this way, the HMM has all the states connected. Finally, fixed a number of state values N_S, they estimate the emission probabilities by partitioning the observations in N_S groups, using the *partitioning among medoids* (PAM) algorithm [35], and the mean of each state is estimated as the median of the observations allocated in that state. The common initial variance is estimated similarly.

To fit the HMM with N_S state values, the authors use first the Forward-Backward procedure to calculate the likelihood. To identify the optimal sequence of state associated to the vector of observations, for each observation y_i they choose the state s_l which is individually most likely. Finally, they re-estimate the parameters (π, A, B) to maximize the likehood, by using the EM algorithm.

It remains to define or estimate the number of state values N_S. For this purpose, the HMM algorithm in [21] first fits the model for $N_S = 1$, …, $N_{S,\max}$(=5) and then chooses the optimal number of state, using the *Akaike's information criterion* (AIC), i.e.

$$N_{S,opt} = \underset{1 \le N_S \le N_{S,\max}}{\arg\min} \; -2\log(p(Y \mid A,B,\pi)) + \frac{2q_{N_S}}{n},$$

where q_{N_S} is the number of the parameters corresponding to the model with N_S state values. Moreover, to improve the estimation, the authors added the following recursive procedure: if $N_S \neq 1$, the algorithm identifies the two states values which have the closest medians (of their corresponding observations) and if this distance is lower than a fixed threshold, then the two state values are merged.

2.2.3 The CGHsegmentation method

The CGHsegmentation algorithm (or CGHseg, see [63]) consists in a maximum likelihood estimation of a piecewise constant function (e.g. the true profile of the $\log_2$ratio values). Similar to the HMM algorithm in [21], the model is estimated for several number of segments k and then the optimal k is chosen as the one that minimize a penalty criterion.

Fixed k, the log-likelihood of the data points is given by,

$$\mathscr{L}_k = -\frac{1}{2}\sum_{p=1}^{k}\sum_{i=t_{k-1}+1}^{t_k}\left[\log(2\pi\sigma^2) + \left(\frac{y_i - \mu_p}{\sigma}\right)^2\right]$$

and μ and σ^2 are estimated with the corresponding maximum likelihood estimators. Also the boundaries t are estimated with the maximum likelihood estimator, but, for its efficient computation, the use of a dynamic programming is necessary. $\widehat{\mathscr{L}_k}$ is the log-likelihood evaluated at the estimated parameters.

After estimating the model for $k = 1, \ldots, k_{\max}$, the optimal k is chosen as the value for which a penalized version of the log-likelihood is maximized,

$$\widetilde{\mathscr{L}_k} = \widehat{\mathscr{L}_k} - \beta 2k$$

$$\Rightarrow \hat{k} = \underset{1 \le k \le k_{\max}}{\arg\max}\, \widetilde{\mathscr{L}_k}.$$

The penalty function is similar to the one given by the AIC criterion, but the penalty constant β is defined in an adaptive way. Actually, the aim of the penalty criterion is to find k such that $\widehat{\mathscr{L}_k}$ ceases to increase significantly (because $\widehat{\mathscr{L}_k}$ increases with k, due to the overfitting). If we define,

$$\beta_k = \frac{\widehat{\mathscr{L}_{k+1}} - \widehat{\mathscr{L}_k}}{2(k+1) - 2(k)}, \qquad k = 1, \ldots, k_{\max},$$

they represent the slopes between the points $\{(2k, \widehat{\mathscr{L}_k})\}_{k=1}^{k_{\max}}$. Thus, if we want to see when $\widehat{\mathscr{L}_k}$ ceases to increase significantly, we look for breaks in the slope of the curve, i.e. difference between the βs,

$$\begin{aligned} l_k &= \beta_k - \beta_{k-1} \\ &= \frac{\widehat{\mathscr{L}_{k+1}} - \widehat{\mathscr{L}_k}}{2(k+1) - 2(k)} - \frac{\widehat{\mathscr{L}_k} - \widehat{\mathscr{L}_{k-1}}}{2(k) - 2(k-1)} \\ &= \frac{\widehat{\mathscr{L}_{k+1}} - 2\widehat{\mathscr{L}_k} + \widehat{\mathscr{L}_{k-1}}}{2} \\ &=: \frac{D_k}{2}. \end{aligned}$$

We can observe that l_k is the second derivative of the log-likelihood, computed with the finite difference. Finally, the optimal k is chosen as the maximal number of segments such that the second derivative is lower than a fixed threshold,

$$\hat{k} = \max(k \in \{1, \ldots, k_{\max}\} \,|\, D_k < -0.5n).$$

2.2.4 The quantreg method

Given a regression problem

$$Z = B\alpha + \varepsilon,$$

the 0.5-quantile (or median) regression [38] estimates the parameters α, by minimizing the objective function

$$S(0.5) = \frac{1}{2}\sum_{i=1}^{m}\left| z_i - \sum_{j=1}^{p} b_{ij}\alpha_j \right|. \tag{2.3}$$

where m and p are the dimensions of the vectors Z and α, respectively, and $\{b_{ij}\}_{i,j}$ are the elements of B.

Eilers and de Menezes [18] thought to estimate the copy number profile with a smoothing method, which minimizes

$$Q_1 = \sum_{i=1}^{n} |y_i - \mu_i| + \lambda \sum_{i=2}^{n-1} |\mu_i - \mu_{i-1}|. \tag{2.4}$$

The objective function Q_1 is an L_1 norm version of the one presented in [83], which uses L_2 norm. The first term in the equation measures the goodness of the fitting, the second term is a penalty that discourages changes in μ. The authors found that the L_2 norm has the effect of smoothing the data with a roundish function. This is not suitable for the estimation of copy number data, since the copy number changes are represented as flat plateaus, which are instead enhanced by the use of L_1 norm analogously to *Lasso* [26].

The problem of finding μ, which minimizes Q_1, is equivalent to solving a 0.5-quantile regression with $m = n$, $p = 2n - 1$, $\alpha = \mu$,

$$Z = \begin{pmatrix} y \\ 0 \end{pmatrix}, B = \begin{pmatrix} I \\ \lambda D \end{pmatrix}, D = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix}$$

and I is the $n \times n$ identity matrix. In fact, by substituting the previous values of the parameters in Equation (2.3), we obtain $Q_1/2$ (2.4). This convex optimization problem is solved by linear programming and in [18] the authors use the implementation given in the R package `quantreg`.

2.3 Estimation of LOH in literature

Since the LOH consists in the alteration of the homozygosity status of the SNPs, the inference of the LOH profile usually requires the knowledge of either the matched normal sample (from the same patient) or a large normal reference dataset. The former type of analysis is called *paired*, the latter *unpaired*. In cancer studies, the normal sample is usually not available for all patients, therefore the analysis of the data is performed with algorithms which use a normal reference dataset (usually the HapMap dataset, see e.g. [47]), to retrieve the information regarding the homozygosity of the SNPs in normal conditions. Nowadays, the two unpaired methods, which are the mostly used, are: dChip [8] and CNAT 4.01 [2] and both employ HMM. The HMM used in CNAT 4.01 is similar to the one of dChip. They have the same observed variables and unobserved states. Moreover, the philosophy behind the definition of the initial, transition and emission probabilities is similar to the one of dChip, but the explicit formulas are not provided in [2]. Therefore, we will describe only dChip.

2.3.1 The dChip algorithm

Beroukhim et al. [8] derived an HMM algorithm for inferring the LOH of unpaired samples which is usually referred to as dChip. As we saw in Subsection 2.2.2, HMM needs the specification of the unobserved states and observed variables, the emission probabilities, the transition probabilities and the initial probabilities.

The observed variables are the SNP calls classified as homozygous (Hom), heterozygous (Het) and "NoCall". The unobserved states are the LOH status, which are defined as loss (LOSS) if there is an LOH, and retention (RET), otherwise. The aim is to estimate the unobserved states from the observations (see Figure 2.3).

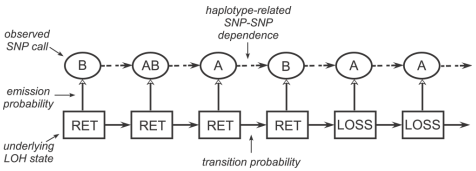

Fig. 2.3 Scheme of the HMM used in dChip [Adapted from *PLOS Computational Biology* [8], copyright (2006), available under Creative Commons Attribution License].

The emission probabilities are the probabilities of the observed calls, given an unobserved state. To define them, the authors considered a SNP having observed calls Hom and Het as a random variable with a different distribution with respect to the NoCall SNPs. In fact, they assumed that actually a NoCall SNP can be Hom or Het, independently of the corresponding LOH status. Thus, $P(NoCall|LOSS) = 1$ and $P(NoCall|RET) = 1$. For the others SNPs, it is sufficient to set $P(Het|RET)$ and $P(Het|LOSS)$ and then, $P(Hom|LOSS) = 1 - P(Het|LOSS)$ and $P(Hom|RET) = 1 - P(Het|RET)$. For any SNP_i (SNP_i denotes the i^{th} SNP interrogated), the probability of being heterozygous under the RET state is estimated with the average heterozygosity rate in a normal population ($P(Het|RET) = p_{het}$). Instead, the probability of being heterozygous under the LOSS state is related to the genotyping error (which is 0.01 [37]). Thus, $P(Het|LOSS) = 0.01$.

The initial probabilities are the prior distribution of the LOH status at any SNP_i: $\mathrm{P}_0(Het)$ and $\mathrm{P}_0(Hom)$ = 1 - $\mathrm{P}_0(Het)$. Using basic probabilities rules, we can see that

$$\begin{aligned}\mathrm{P}(Het) &= \mathrm{P}(Het|LOSS)\mathrm{P}_0(LOSS) + \mathrm{P}(Het|RET)\mathrm{P}_0(RET)\\ &= 0.01\mathrm{P}_0(LOSS) + p_{het}\mathrm{P}_0(RET)\\ &\approx p_{het}\mathrm{P}_0(RET)\\ \Rightarrow \mathrm{P}_0(RET) &\approx \frac{\mathrm{P}(Het)}{p_{het}},\end{aligned}$$

by considering the SNP genotyping error negligible. As a consequence, $\mathrm{P}_0(RET)$ is estimated as the ratio between the proportion of heterozygous SNPs in the sample and the heterozygosity rate in the normal population.

The transition probabilities are the conditional distribution of the LOH state of two consecutive SNPs. Since nearby SNPs tend to have the same LOH status, while distant markers not, first the authors defined the probability θ that SNP_{i-1} does not influence SNP_i. They used a function which increases with the distance d (in Megabases, Mb) between the markers: $\theta = (1 - e^{-2d})$. Therefore, the probability of a LOSS at SNP_i, given the LOSS at SNP_{i-1}, is decomposed in the probability that SNP_i is not influenced by SNP_{i-1} and SNP_i is LOSS, and the probability that SNP_i is influenced by SNP_{i-1},

$$\mathrm{P}(LOSS \text{ at } SNP_i|LOSS \text{ at } SNP_{i-1}) = \theta\mathrm{P}_0(LOSS) + (1-\theta).$$

Similarly, the probability of a LOSS at SNP_i, given the RET status at SNP_{i-1}, is equal to the probability that SNP_i is not influenced by SNP_{i-1} and SNP_i is LOSS,

$$\mathrm{P}(LOSS \text{ at } SNP_i|RET \text{ at } SNP_{i-1}) = \theta\mathrm{P}_0(LOSS).$$

Obviously,

$$\begin{aligned}\mathrm{P}(RET \text{ at } SNP_i|LOSS \text{ at } SNP_{i-1}) &= 1 - \mathrm{P}(LOSS \text{ at } SNP_i|LOSS \text{ at } SNP_{i-1})\\ &= \theta\mathrm{P}_0(RET)\end{aligned}$$

$$\begin{aligned} \mathrm{P}(RET \text{ at } SNP_i | RET \text{ at } SNP_{i-1}) &= 1 - \mathrm{P}(LOSS \text{ at } SNP_i | LOSS \text{ at } SNP_{i-1}) \\ &= \theta \mathrm{P}_0(RET) + (1-\theta). \end{aligned}$$

Chapter 3
New statistical methods for copy number estimation

Abstract As we saw in Chapter 2, the copy number profile can be estimated with either a piecewise constant function or a continuous curve. In [32, 33], Hutter proposed two Bayesian regression methods that can be applied for the inference of the copy number profile: the Bayesian Piecewise Constant Regression (BPCR) and the Bayesian Regression Curve (BRC).
BPCR is a Bayesian regression method for data that are noisy observations of a piecewise constant function. The method estimates the unknown segment number, the endpoints of the segments and the value of the segment levels of the underlying piecewise constant function. BRC estimates the same data with a smoothing curve. However, in the original formulation, some estimators failed to properly determine the corresponding parameters. For example, the boundary estimator did not take into account the dependency among the boundaries and estimated more than one breakpoint at the same position, losing segments.
Therefore, in Section 3.1, we present an improved version of the BPCR (called mBPCR), changing the segment number estimator and the boundary estimator to enhance the fitting procedure. We also propose an alternative estimator of the variance of the segment levels, which is useful in case of data with high noise. In Section 3.2 we deduce two improved versions

of BRC: mBRC and BRCAk.
In literature, some methods estimate the copy numbers as a piecewise constant function, while other algorithms estimate them as a continuous curve (Chapter 2). Hence, we compare the original and the modified version of BPCR to the former group of methods (Subsection 3.1.7), while the the original and the modified version of BRC to the latter (Subsection 3.2.3). On artificial data, we show that mBPCR and the improved versions of BRC generally outperformed all the others. We observe that similar results were obtained also on real data. The choice of using Bayesian statistics, although it has higher computational complexity, appears appropriate especially for the estimation of regions containing only few data points.
In Section 3.3, we describe a dynamic programming for the computation of the quantities involved in the estimation, since it is not possible to find them analytically. In Section 3.4, we show a further change of mBPCR, in order to reduce the false discovery rate of the breakpoint estimator in presence of only one segment.
Our method (already published in [65]) was implemented in R and the corresponding R package (called mBPCR) can be downloaded from the Bioconductor website (http://www.bioconductor.org/).
Regarding notations, we will not indicate explicitly the random variable to which a distribution is referred, if it is clear from the context. For example, $p_K(k) \equiv p(k)$ or $p_{Y,\mathrm{M}}(y,\mu) \equiv p(y,\mu)$.

3.1 Piecewise constant estimation: the mBPCR method

The *Bayesian Piecewise Constant Regression* (BPCR) estimates a piecewise constant function using a Bayesian regression [32, 33]. Since copy number data can be modeled as a piecewise constant function, this algorithm can be used for their profile estimation. Nevertheless, we found that some estimators defined in the procedure gave practical and/or theoretical problems. Therefore, we propose improved Bayesian estimators. The new version of BPCR is called *modified Bayesian Piecewise Constant Regression* (mBPCR) and has been presented in [65].

In Subsection 3.1.1, we define the complete model, which is common to both BPCR and mBPCR. In Subsection 3.1.2, we briefly describe the estimation of the parameters in the original BPCR, in order to show how we changed some of these estimators in Subsections 3.1.3, 3.1.4 and 3.1.5. In Subsection 3.1.6, we select the best performing estimators on the basis of the results obtained on artificial datasets. A further selection is performed in Subsection 3.1.7, on the basis of the comparison with other methods on artificial data. Finally, in Subsection 3.1.8, we define mBPCR and, in Subsection 3.1.9, we compare it with other methods on real data.

3.1.1 Priors and posteriors

Given the "piecewise constant setting" defined in Section 2.1, we want to estimate the number of the segments k_0, the partition of the data t^0 and the levels of all the segments μ^0. From a Bayesian point of view, μ^0, t^0 and k_0 are treated as random variables, hence we denote them with the corresponding upper case letters (M, T and K). Moreover, because of their randomness, we need to define a prior distribution for each of them to complete the model.

For the number of segments and the boundaries, we assume noninformative prior distributions:

$$p(k) = \frac{1}{k_{\max}}, \qquad k \in \mathbb{K} \tag{3.1}$$

$$p(t \,|\, k) = \frac{1}{\binom{n-1}{k-1}}, \qquad t \in \mathbb{T}_{k,n}, \tag{3.2}$$

where $\mathbb{K} = \{1, \ldots, k_{\max}\}$ and $\mathbb{T}_{k,n}$ is the subspace of $\mathbb{N}_0^{k+1}$ such that $t_0 = 0$, $t_k = n$ and $t_q \in \{1, \ldots, n-1\}$ for all $q = 1, \ldots, k-1$, in an ordered way and without repetitions.

About M, we assume that all its components are mutually independent and identically normally distributed,

$$\mathrm{M}|_{\nu,\,\rho^2,\,K=k} \sim \mathcal{N}(\nu, \rho^2 \mathbb{I}), \tag{3.3}$$

where $\nu \in \mathbb{R}^k$, such that $\nu_q = \nu$ for each $q = 1, \ldots, k$, and $\mathbb{I} \in \mathbb{R}^{k \times k}$, such that $\mathbb{I}_{p,q} = \delta_{p,q}$ for each $p, q = 1, \ldots, k$.

Instead of these assumptions, we could assume a Cauchy distribution for each Y_i or M_q in order to model an observation whose noise has heavier tails, as previously done by Hutter [32, 33].

In general, the Bayesian estimation procedure requires the computation of the posterior distributions of the unknown parameters. In particular, in BPCR and/or in mBPCR we need to calculate the posterior distribution of the number of segments, the posterior joint or marginal distribution of the boundaries and the posterior marginal distribution of the levels M. We obtain the desired distributions by using Bayes's rule and conditioning with respect to the other parameters,

$$\begin{aligned}
&p(k \,|\, y, \sigma^2, \nu, \rho^2) \\
&\quad = \frac{p(y \,|\, k, \sigma^2, \nu, \rho^2) p(k)}{p(y \,|\, \sigma^2, \nu, \rho^2)} \\
&\quad = \frac{\sum_{t \in \mathbb{T}_{k,n}} p(y \,|\, t, k, \sigma^2, \nu, \rho^2) p(t \,|\, k) p(k)}{p(y \,|\, \sigma^2, \nu, \rho^2)} \\
&\quad = \frac{p(k) \sum_{t \in \mathbb{T}_{k,n}} p(t \,|\, k) \left(\int_{\mathbb{R}^k} p(y \,|\, \mu, t, k, \sigma^2) p(\mu \,|\, k, \nu, \rho^2) d\mu \right)}{p(y \,|\, \sigma^2, \nu, \rho^2)} \\
&\quad = \frac{p(k) \sum_{t \in \mathbb{T}_{k,n}} p(t \,|\, k) \left(\int_{\mathbb{R}^k} p(y \,|\, \mu, t, k, \sigma^2) p(\mu \,|\, k, \nu, \rho^2) d\mu \right)}{\sum_{k=1}^{k_{\max}} p(k) \sum_{t \in \mathbb{T}_{k,n}} p(t \,|\, k) \left(\int_{\mathbb{R}^k} p(y \,|\, \mu, t, k, \sigma^2) p(\mu \,|\, k, \nu, \rho^2) d\mu \right)}
\end{aligned}$$

$$\begin{aligned}
p(t \,|\, y, k, \sigma^2, \nu, \rho^2) &= \frac{p(y \,|\, t, k, \sigma^2, \nu, \rho^2) p(t \,|\, k)}{p(y \,|\, k, \sigma^2, \nu, \rho^2)} \\
&= \frac{p(t \,|\, k) \int_{\mathbb{R}^k} p(y \,|\, \mu, t, k, \sigma^2) p(\mu \,|\, k, \nu, \rho^2) d\mu}{\sum_{t \in \mathbb{T}_{k,n}} p(t \,|\, k) \int_{\mathbb{R}^k} p(y \,|\, \mu, t, k, \sigma^2) p(\mu \,|\, k, \nu, \rho^2) d\mu}
\end{aligned}$$

$$
\begin{aligned}
p(t_p \,|\, y,k,\sigma^2,\nu,\rho^2) &= \sum_{\substack{\{t' \in \mathbb{T}_{k,n}:\\ t'_p = t_p\}}} p(t' \,|\, y,k,\sigma^2,\nu,\rho^2) \\
p(\mu_p \,|\, y,t,k,\sigma^2,\nu,\rho^2) &= p(\mu_p \,|\, y_{t_{p-1},t_p},t,k,\sigma^2,\nu,\rho^2) \\
&= \frac{p(y_{t_{p-1},t_p} \,|\, \mu_p,t,k,\sigma^2) p(\mu_p \,|\, \nu,\rho^2)}{p(y_{t_{p-1},t_p} \,|\, t,k,\sigma^2,\nu,\rho^2)} \\
&= \frac{p(y_{t_{p-1},t_p} \,|\, \mu_p,t,k,\sigma^2) p(\mu_p \,|\, \nu,\rho^2)}{\int_{\mathbb{R}^k} p(y_{t_{p-1},t_p} \,|\, \mu_p,t,k,\sigma^2) p(\mu_p \,|\, \nu,\rho^2) d\mu_p},
\end{aligned}
\tag{3.4}
$$

where the likelihood is defined in Equation (2.2) and the priors of the parameters are defined in Equations (3.1), (3.2) and (3.3). Notice that it is not possible to compute analytically the posteriors (since they require to sum the likelihood over all $t \in \mathbb{T}_{k,n}$ and $k \in \mathbb{K}$), therefore the estimation procedure uses a dynamic programming which is explained in Section 3.3.

In the following, we will not explicitly indicate the hyper-parameters σ^2, ν and ρ^2 in the posteriors and in the likelihood, to simplify the notations.

3.1.2 Original estimation: the BPCR method

The statistical procedure consists in a sequence of estimations due to the relationship among the parameters.

BPCR estimates the number of segments with the MAP (Maximum A Posteriori) estimate given the sample point y,

$$
\widehat{k} := \arg\max_{k \in \mathbb{K}} p(k \,|\, y), \tag{3.5}
$$

and, given $\hat{k}$, also each boundary is estimated separately with its corresponding MAP estimate,

$$\widehat{t}_p := \underset{h\in\{p,\,\ldots,\,n-(\widehat{k}-p)\}}{\arg\max} \mathrm{P}(T_p = h \,|\, y, \hat{k}) \tag{3.6}$$

for all $p = 1, \ldots, \hat{k} - 1$. Finally, the r^{th} moment of the level of the p^{th} segment is estimated with its posterior mean. Since its computation needs the knowledge of the number of segments and the partition of the data, we replace them with the estimated ones,

$$\widehat{\mu}_p^r := \mathrm{E}[\mathrm{M}_p^r \,|\, y, \widehat{t}, \widehat{k}], \tag{3.7}$$

for all $p = 1, \ldots, \hat{k}$. Equation (3.4) implies that the posterior probability of M_p is proportional to $p(y_{t_{p-1},t_p} \,|\, \mu_p, t, k, \sigma^2) p(\mu_p \,|\, \nu, \rho^2)$ and, assuming that Y and M are normally distributed (see (2.1) and (3.3)),

$$\begin{aligned} p(\mu_p \,|\, y, \widehat{t}, \widehat{k}) &\propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=\widehat{t}_{p-1}+1}^{\widehat{t}_p} (y_i - \mu_p)^2 - \frac{1}{2\rho^2}(\mu_p - \nu)^2 \right\} \\ &\propto \exp\left\{ -\frac{1}{2}\left(\frac{\widehat{t}_p - \widehat{t}_{p-1}}{\sigma^2} + \frac{1}{\rho^2} \right) \left[\mu_p - \frac{\frac{\sum_{i=\widehat{t}_{p-1}+1}^{\widehat{t}_p} y_i}{\sigma^2} + \frac{\nu}{\rho^2}}{\frac{\widehat{t}_p - \widehat{t}_{p-1}}{\sigma^2} + \frac{1}{\rho^2}} \right]^2 \right\}. \end{aligned}$$

Consequently, M_p has the following posterior normal distribution,

$$\mathrm{M}_p|_{y,\widehat{t},\widehat{k}} \sim \mathcal{N}\left(\frac{\rho^2 \sum_{i=\widehat{t}_{p-1}+1}^{\widehat{t}_p} y_i + \sigma^2 \nu}{(\widehat{t}_p - \widehat{t}_{p-1})\rho^2 + \sigma^2}, \frac{\sigma^2 \rho^2}{(\widehat{t}_p - \widehat{t}_{p-1})\rho^2 + \sigma^2} \right)$$

and Equation (3.7) imply that the estimate of the p^{th} level is

$$\widehat{\mu}_p = \frac{\rho^2 \sum_{i=t_{p-1}+1}^{t_p} y_i + \sigma^2 \nu}{(t_p - t_{p-1})\rho^2 + \sigma^2}, \tag{3.8}$$

for all $p = 1, \ldots, \hat{k}$. When the sample contains only one segment, the Bayesian estimation of the posterior distribution of the levels should theoretically lead to a normal distribution, similar to a Dirac delta function centered at $\hat{\nu}$, since the levels can assume only one value by knowing only the data. In fact, in this case, if we estimate ρ^2 only using the data (without using any prior or other information), then this value will be close to zero (the variance of a constant random variable, since M can assume only one value by knowing only the data) and thus the level will be estimated with $\hat{\nu}$, the mean of the data (see Equation (3.9)).

The probability distributions defined previously depend on the hyperparameters ν, ρ^2 and σ^2 (the mean and the variance of the segment levels and the variance of the noise, respectively). Hutter [32, 33] suggested the following estimators:

$$\widehat{\nu} := \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y} \tag{3.9}$$

$$\hat{\rho}^2 := \frac{1}{n}\sum_{i=1}^{n} (Y_i - \overline{Y})^2 \tag{3.10}$$

$$\widehat{\sigma}^2 := \frac{1}{2(n-1)}\sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2. \tag{3.11}$$

3.1.3 Improved estimators of the number of segments

To understand the real meaning of the MAP estimator $\widehat{K}$, we need to introduce the theory of the construction of a generic Bayesian estimator.

In general, a Bayesian estimator is defined in the following way. Let us suppose that Z is a random variable whose distribution depends on an unknown parameter θ, which we want to estimate. Since we cannot know exactly the true value of the parameter, we consider it as a random variable Θ with a given prior probability distribution. In order to measure the goodness of the estimation, we define an error (or *loss function*) and we

choose the estimator that minimizes the expected error given the sample Z,

$$\widehat{\Theta} := \arg\min_{\theta'} \mathrm{E}[err(\Theta, \theta') \,|\, Z]. \tag{3.12}$$

The 0-1 error (defined as $1 - \delta_{\theta,\theta'}$) is commonly used for a parameter which can assume only a discrete number of values. The estimator corresponding to this error is the MAP estimator,

$$\begin{aligned} \arg\min_{\theta'} \mathrm{E}[1 - \delta_{\Theta,\theta'} \,|\, Z] &= \arg\max_{\theta'} \sum_{\theta} \delta_{\theta,\theta'}\, p(\theta \,|\, Z) \\ &= \arg\max_{\theta'} p(\theta' \,|\, Z). \end{aligned} \tag{3.13}$$

Obviously, if we use different types of errors, we can obtain different estimators. In the following, we will use $\widehat{K}$ to denote any estimator of K, while $\widehat{K}_{01}$ to denote the estimator $\widehat{K}$ based on the 0-1 error.

Using the 0-1 error, we give the same penalty to each value different from the true value, whether it is close to or far away from the true one. To take into account the distance of the estimated value from the true one, we need to use other types of errors, which are based on different definitions of distance, such as,

$$\text{absolute error} := |\theta - \theta'| \tag{3.14}$$

$$\text{squared error} := (\theta - \theta')^2. \tag{3.15}$$

If the parameter $\theta \in \mathbb{R}$, then the estimators corresponding to these errors are the median and the mean of its posterior distribution, respectively. In our case, we denote these estimators of k_0 with $\widehat{K}_1$ and $\widehat{K}_2$.

3.1.4 Improved estimators of the boundaries

Similarly to the previous subsection, we derive alternative boundary estimators by considering different types of errors. We denote the MAP boundary estimator defined in Equation (3.6) with $\widehat{T}_{01}$.

The estimator $\widehat{T}_{01}$ is defined in such a way that each component minimizes the 0-1 error of the corresponding boundary, separately. Explicitly, given the sample point y and the segment number k_0, its estimate is

$$\widehat{T}_{01} = \left(0, \arg\max_{t_1 \in \mathbb{T}} p(t_1 \,|\, y, k_0), \ \ldots, \ \arg\max_{t_{k_0-1} \in \mathbb{T}} p(t_{k_0-1} \,|\, y, k_0), n\right),$$

where $\mathbb{T} = \{1, \ \ldots, n-1\}$. $\widehat{T}_{01}$ may be regarded as an approximation of the Bayesian estimator that minimizes the error which counts the number of wrongly estimated boundaries:

$$\text{sum 0-1 error} = \sum_{p=1}^{k_0-1} \left(1 - \delta_{t_p^0, t_p}\right) = k_0 - 1 - \sum_{p=1}^{k_0-1} \delta_{t_p^0, t_p}, \tag{3.16}$$

that is

$$\widehat{T}_{\text{sum}} = \arg\max_{t \in \mathbb{T}_{k_0,n}} \sum_{p=1}^{k_0-1} p(t_p \,|\, Y, k_0). \tag{3.17}$$

A problem of the estimator in Equation (3.17) is its computational complexity, because it needs the computation of all the ordered combinations of the boundaries. On the other hand, there are two reasons for which $\widehat{T}_{01}$ is not a suitable estimator of the boundaries. First, it does not take into account that the boundaries are dependent, because they have to be ordered, and second, in principle, it can have more than one component with the same value. As a consequence, a theoretically more correct way to estimate the boundaries is minimizing the 0-1 error with respect to the joint boundary probability distribution (this error is called *total 0-1 error*). Then, given k_0 and Y, the boundary estimator becomes

$$\widehat{T}_{\text{joint}} = \underset{t \in \mathbb{T}_{k_0,n}}{\arg\max}\, p(t \mid Y, k_0). \tag{3.18}$$

We must notice that the estimators considered until now have the same length of the true vector of the boundaries. In practice, the number of segments k_0 is unknown, so that we should use $\widehat{k}$. As a consequence, if $\widehat{k}$ is different from k_0, then, strictly speaking, we cannot minimize the previous types of error because the vectors have different length.

A way to solve this issue is to map each boundary vector into a vector $\tau \in \mathbb{R}_0^{n+1}$ in the following way:

$$t \mapsto \tau \text{ such that } \tau_i = \begin{cases} 1 \text{ if } \exists p \text{ such that } t_p = i \\ 0 \text{ otherwise.} \end{cases} \tag{3.19}$$

We denote with $\boldsymbol{\tau}_{k,n}$ the set of all the possible τ with $\tau_0 = 1$, $\tau_n = 1$ and $k-1$ of the other components equal to 1.

Now, for the new two vectors τ_0 (which has two-norm $\sqrt{k_0+1}$) and τ (which has two-norm $\sqrt{\widehat{k}+1}$), we need to define a suitable error. For example, we can consider the sum 0-1 error, the number of elements in τ_0 which differ from those in τ (i.e. the number of missed breakpoints) or the Euclidian distance, which are, respectively,

$$\sum_{i=1}^{n-1} (1 - \delta_{\tau_i^0, \tau_i}) = k_0 + \hat{k} - 2\sum_{i=1}^{n-1} \tau_i^0 \tau_i, \tag{3.20}$$

$$k_0 - 1 - \langle \tau^0, \tau \rangle = k_0 - 1 - \sum_{i=1}^{n-1} \tau_i^0 \tau_i, \tag{3.21}$$

$$\|\tau^0 - \tau\|_2 = \left(\|\tau_0\|_2 + \|\tau\|_2 - 2\sum_{i=1}^{n-1} \tau_i^0 \tau_i \right)^{1/2}. \tag{3.22}$$

The errors (3.20) and (3.21) are minimized by the same argument. Moreover, minimizing (3.21) is also the same of minimizing (3.22), because $\|\tau^0\|_2$ and $\|\tau\|_2$ are fixed. Furthermore, the error (3.21) is consistent with the Russell-Rao dissimilarity measure defined on the space of the binary

vectors, so we will call (3.21) the *binary error*.

$$\begin{aligned}\widehat{\boldsymbol{\tau}}_{\text{BinErr}} &:= \underset{\tau' \in \boldsymbol{\pi}_{k_0,n}}{\arg\min}\ \mathrm{E}\left[k_0 - 1 - \sum_{i=1}^{n-1} \boldsymbol{\tau}_i \tau_i' \,\middle|\, Y, k_0\right] \\ &= \underset{\tau' \in \boldsymbol{\pi}_{k_0,n}}{\arg\max}\ \mathrm{E}\left[\sum_{i=1}^{n-1} \boldsymbol{\tau}_i \tau_i' \,\middle|\, Y, k_0\right]. \end{aligned} \tag{3.23}$$

We can make the last expected value explicit, given the sample point y,

$$\begin{aligned}\mathrm{E}\left[\sum_{i=1}^{n-1} \boldsymbol{\tau}_i \tau_i' \,\middle|\, y, k_0\right] &= \sum_{\tau \in \boldsymbol{\pi}_{k_0,n}} \sum_{i=1}^{n-1} \tau_i \tau_i' p(\tau \,|\, y, k_0) \\ &= \sum_{i=1}^{n-1} \sum_{\tau \in \boldsymbol{\pi}_{k_0,n} : \tau_i = 1} \tau_i \tau_i' p(\tau \,|\, y, k_0) \\ &= \sum_{i=1}^{n-1} \sum_{\tau \in \boldsymbol{\pi}_{k_0,n} : \tau_i = 1} \tau_i' p(\tau \,|\, y, k_0) \\ &= \sum_{i=1}^{n-1} \delta_{\tau_i', 1} \sum_{\tau \in \boldsymbol{\pi}_{k_0,n} : \tau_i = 1} p(\tau \,|\, y, k_0) \\ &= \sum_{i=1,\ldots,n-1 : \tau_i' = 1} \mathrm{P}(\boldsymbol{\tau}_i = 1 \,|\, y, k_0)\end{aligned}$$

Moreover,

$$\begin{aligned}\mathrm{P}(\boldsymbol{\tau}_i = 1 \,|\, y, K = k_0) &= \mathrm{P}(\{t \in \mathbb{T}_{k_0,n} : \exists p \text{ such that } t_p = i\} \,|\, y, k_0) \\ &= \sum_{p=1}^{\min(i, k_0 - 1)} \mathrm{P}(T_p = i \,|\, y, k_0). \end{aligned} \tag{3.24}$$

Therefore, to find the value of $\widehat{\boldsymbol{\tau}}_{\text{BinErr}}$, we need to find the $\widehat{k}-1$ highest $\mathrm{P}(\boldsymbol{\tau}_i = 1 \,|\, y, k_0)$ (corresponding to the indices $i_1, \ldots, i_{\widehat{k}-1}$) and then take $\widehat{\tau}_{\text{BinErr}}$ such that $\widehat{\tau}_{i_p} = 1$ for $p = 1, \ldots, \widehat{k}-1$. Using the inverse of the function in (3.19), we obtain the corresponding value of the estimator $\widehat{T}_{\text{BinErr}}$.

Since we do not know the real value of k_0, we should replace it with $\widehat{k}$ to compute Equation (3.23). Doing this, we could amplify the error of the boundary estimation because of the addition of the error of the segment number estimation. A way to attenuate this issue is to integrate out the number of segments in the conditional expected value. Then the estimator becomes

$$\widehat{\boldsymbol{\tau}}_{\text{BinErrAk}} := \underset{\tau' \in \boldsymbol{\pi}_{\widehat{k},n}}{\arg\max}\, \mathrm{E}\left[\sum_{i=1}^{n-1} \boldsymbol{\tau}_i \tau_i' \,\middle|\, Y \right]. \tag{3.25}$$

and the explicit formula of the expected value, given the sample point y, is

$$\begin{aligned}
\mathrm{E}\left[\sum_{i=1}^{n-1} \boldsymbol{\tau}_i \tau_i' \,\middle|\, y \right] &= \sum_{k=2}^{k_{\max}} \sum_{\tau \in \boldsymbol{\pi}_{k,n}} \sum_{i=1}^{n-1} \tau_i \tau_i' p(\tau \,|\, y, k) p(k \,|\, y) \\
&= \sum_{k=2}^{k_{\max}} \sum_{i=1}^{n-1} \delta_{\tau_i', 1} \mathrm{P}(\boldsymbol{\tau}_i = 1 \,|\, y, k) p(k \,|\, y) \\
&= \sum_{i=1}^{n-1} \delta_{\tau_i', 1} \sum_{k=2}^{k_{\max}} \mathrm{P}(\boldsymbol{\tau}_i = 1 \,|\, y, k) p(k \,|\, y) \\
&= \sum_{i=1,\ldots,n-1 : \tau_i' = 1} \mathrm{P}(\boldsymbol{\tau}_i = 1 \,|\, y).
\end{aligned}$$

Analogously to the computation of $\widehat{\boldsymbol{\tau}}_{\text{BinErr}}$, the components of $\widehat{\tau}_{\text{BinErrAk}}$ are equal to the ones corresponding to the $\widehat{k}-1$ highest probabilities in $\{\mathrm{P}(\boldsymbol{\tau}_i = 1 \,|\, y)\}_{i=1}^{n-1}$. Again, using the inverse of the function in (3.19), we obtain the corresponding value of the estimator $\widehat{T}_{\text{BinErrAk}}$.

3.1.5 Properties of the hyper-parameter estimators and definition of new estimators

In order to study the properties of the hyper-parameter estimators defined in Equations (3.9), (3.11) and (3.10), first we need to compute joint distribution of any two data points Y_i and Y_j belonging to the same segment, conditioned only on the hyper-parameters ν, ρ^2, σ^2. In the following, n_q will denote the number of data points in the q^{th} segment.

At first, let us consider only the data which belong to the q^{th} segment. From the hypothesis of the model, we know that

$$
\begin{aligned}
&Y_j|_{\mathrm{M}_q,\,\sigma^2} \sim \mathcal{N}(\mathrm{M}_q,\,\sigma^2), \quad j = t_{q-1}+1,\,\ldots,\,t_q,\\
&\mathrm{M}_q|_{\nu,\,\rho^2} \sim \mathcal{N}(\nu,\,\rho^2),
\end{aligned}
$$

thus, the joint density of any two data points Y_i and Y_j belonging to the q^{th} segment is

$$
\begin{aligned}
&f(y_i,y_j|\nu,\rho^2,\sigma^2)\\
&= \int_{\mathbb{R}} f(y_i,y_j,\mu_q|\nu,\rho^2,\sigma^2)\,d\mu_q\\
&= \int_{\mathbb{R}} f(y_i|\mu_q,\sigma^2)\,f(y_j|\mu_q,\sigma^2)\,f(\mu_q|\nu,\rho^2)\,d\mu_q\\
&= \int_{\mathbb{R}} \left(\frac{1}{2\pi}\right)^{\frac{3}{2}} \frac{1}{\sigma^2\rho} \exp\left\{-\frac{1}{2}\left[\sum_{h=\{i,j\}}\left(\frac{y_h-\mu_q}{\sigma}\right)^2+\left(\frac{\mu_q-\nu}{\rho}\right)^2\right]\right\} d\mu_q\\
&= \frac{1}{2\pi\sigma^2\rho}\exp\left\{-\frac{1}{2}\left[-\frac{\sigma^2\rho^2}{\sigma^2+2\rho^2}\left(\frac{y_i+y_j}{\sigma^2}+\frac{\nu}{\rho^2}\right)^2+\frac{y_i^2+y_j^2}{\sigma^2}+\frac{\nu^2}{\rho^2}\right]\right\}\\
&\quad\cdot \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\left(\frac{2}{\sigma^2}+\frac{1}{\rho^2}\right)\left(\mu_q-\frac{\sigma^2\rho^2}{\sigma^2+2\rho^2}\left(\frac{y_i+y_j}{\sigma^2}+\frac{\nu}{\rho^2}\right)\right)^2\right]\right\} d\mu_q\\
&= \frac{1}{2\pi\sigma\sqrt{2\rho^2+\sigma^2}}\exp\left\{-\frac{1}{2}\left[\frac{(\rho^2+\sigma^2)(y_i^2+y_j^2)-2\rho^2y_iy_j}{\sigma^2(2\rho^2+\sigma^2)}+2\frac{\nu^2-\nu(y_i+y_j)}{2\rho^2+\sigma^2}\right]\right\}.
\end{aligned}
$$

The quantity in the square brackets can be written as a quadratic form in the following way,

$$\begin{bmatrix} y_i - \nu & y_j - \nu \end{bmatrix} \begin{bmatrix} \frac{\sigma^2+\rho^2}{\sigma^2(2\rho^2+\sigma^2)} & \frac{-\rho^2}{\sigma^2(2\rho^2+\sigma^2)} \\ \frac{-\rho^2}{\sigma^2(2\rho^2+\sigma^2)} & \frac{\sigma^2+\rho^2}{\sigma^2(2\rho^2+\sigma^2)} \end{bmatrix} \begin{bmatrix} y_i - \nu \\ y_j - \nu \end{bmatrix},$$

thus $(Y_i, Y_j)|_{\nu,\,\rho^2,\,\sigma^2} \sim \mathcal{N}_2(\nu, \Sigma)$, where

$$\Sigma = \begin{bmatrix} \sigma^2+\rho^2 & \rho^2 \\ \rho^2 & \sigma^2+\rho^2 \end{bmatrix}.$$

It follows that the covariance between two data points, which belong to the same segment, is

$$\mathrm{Cov}(Y_i, Y_j|\nu, \rho^2, \sigma^2) = \rho^2 \qquad i \neq j, \tag{3.26}$$

and

$$\mathrm{E}[Y_j|\nu, \rho^2, \sigma^2] = \nu \tag{3.27}$$

$$\mathrm{Var}(Y_j|\nu, \rho^2, \sigma^2) = \sigma^2 + \rho^2, \tag{3.28}$$

for each $j = 1, \ldots, n$, independently of the segment to which it belongs.

Furthermore, from the hypotheses of the model, knowing the segmentation t^0, data points belonging to different segments are independent.

Expected value and variance of the estimator $\widehat{\nu}$

The estimator of ν is defined as $\widehat{\nu} = \overline{Y}$ (see Equation (3.9)) and Equation (3.27) implies that this estimator is unbiased. To compute its variance, first, we need to calculate the second moment of the arithmetic mean.

From equations (3.27), (3.28) and (3.26), we can deduce that

$$\mathrm{E}[Y_i Y_j] = \begin{cases} \nu^2 & \text{if they are independent} \\ \nu^2 + \rho^2 & \text{if they are dependent and } j \neq i \\ \nu^2 + \rho^2 + \sigma^2 & \text{if } j = i \end{cases} \tag{3.29}$$

Using this fact and remembering that Y_i and Y_j are independent only if they belong to different segments, we can compute the expected values of the square arithmetic mean,

$$\begin{aligned}
\mathrm{E}[\overline{Y}^2] &= \frac{1}{n^2}\sum_{i,\,j=1}^{n}\mathrm{E}[Y_iY_j] \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathrm{E}[Y_i^2]+\sum_{p=1}^{k_0}\sum_{\substack{i,j=1\\ i\neq j}}^{n_p}\mathrm{E}[Y_i^pY_j^p]+\sum_{\substack{p,q=1\\ p\neq q}}^{k_0}\sum_{i=1}^{n_p}\sum_{j=1}^{n_q}\mathrm{E}[Y_i^pY_j^q]\right) \\
&= \frac{1}{n^2}\left[n\,(\nu^2+\rho^2+\sigma^2)+\sum_{p=1}^{k_0}n_p(n_p-1)(\nu^2+\rho^2)\right. \\
&+ \left.\sum_{p=1}^{k_0}n_p(n-n_p)\nu^2\right] = \nu^2+\frac{\sigma^2}{n}+\rho^2\frac{\sum_{p=1}^{k_0}n_p^2}{n^2}, \qquad (3.30)
\end{aligned}$$

where Y_i^p denotes the i^{th} data point of the p^{th} segment. For the last equation we used the fact that $\sum_{p=1}^{k_0}n_p = n$. Therefore,

$$\begin{aligned}
\mathrm{Var}[\overline{Y}] &= \mathrm{E}[\overline{Y}^2]-\mathrm{E}[\overline{Y}]^2 \\
&= \frac{\sigma^2}{n}+\rho^2\frac{\sum_{p=1}^{k_0}n_p^2}{n^2} \qquad (3.31) \\
&\geq \frac{\sigma^2}{n}+\rho^2\frac{\left\lceil\frac{n}{k_0}\right\rceil^2 k_0}{n^2} = \frac{\sigma^2}{n}+O\left(\frac{\rho^2}{k_0}\right).
\end{aligned}$$

Hence, the variance is always greater than $O\left(\frac{\rho^2}{k_0}\right)$, even if we use a denser sampling, i.e. we augment the number of data points in the interval in which we are estimating the piecewise constant function.

New definition of the estimator $\widehat{\sigma^2}$ and its expected value

A circular version of the σ^2 estimator defined in Equation (3.11) is

$$\widehat{\sigma^2} := \frac{1}{2n}\sum_{i=1}^{n}(Y_{i+1}-Y_i)^2, \tag{3.32}$$

where $Y_{n+1} := Y_1$. In order to compute the expected value of the estimator, first we compute the expected value for each term of the summation. We will consider two cases:

1. Y_i and Y_{i+1} belong to different segments,
2. Y_i and Y_{i+1} belong to the same segment.

In the first case, by using Equation (3.29), we obtain that

$$\begin{aligned}\mathrm{E}[(Y_{i+1}-Y_i)^2] &= \mathrm{E}[Y_{i+1}^2]+\mathrm{E}[Y_i^2]-2\mathrm{E}[Y_iY_{i+1}]\\ &= 2(\nu^2+\rho^2+\sigma^2)-2\nu^2 = 2(\rho^2+\sigma^2),\end{aligned}$$

and in the second case,

$$\mathrm{E}[(Y_{i+1}-Y_i)^2] = 2(\nu^2+\rho^2+\sigma^2)-2(\nu^2+\rho^2) = 2\sigma^2.$$

Then, the expected value of $\widehat{\sigma^2}$ is

$$\begin{aligned}&\mathrm{E}[\widehat{\sigma^2}]\\ &= \frac{1}{2n}\left\{\sum_{q=1}^{k_0}\sum_{i=1}^{n_q-1}\mathrm{E}[(Y_{i+1}^q-Y_i^q)^2]+\sum_{q=1}^{k_0-1}\mathrm{E}[(Y_1^{q+1}-Y_{n_q}^q)^2]+\mathrm{E}[(Y_n-Y_1)^2]\right\}\\ &= \frac{1}{2n}\left[(n-k_0)2\sigma^2+2k_0(\rho^2+\sigma^2)\mathbb{I}_{\{k_0\geq 2\}}+2\sigma^2\mathbb{I}_{\{k_0=1\}}\right]\\ &= \sigma^2+\rho^2\frac{k_0}{n}\mathbb{I}_{\{k_0\geq 2\}}.\end{aligned} \tag{3.33}$$

In the computation, we considered two situations: (a) when $k_0 = 1$, Y_1 and Y_n belong to the same segment (thus, they are dependent), (b) when $k_0 \geq 2$,

Y_1 and Y_n are independent, because we supposed that the first and the last segments have different levels. If the first and the last segments had the same level, then the two segments would be joined together and Y_1 and Y_n would be dependent. In this case, the expected value would be the same but with $k_0 - 1$ instead of k_0, since the number of segments would be $k_0 - 1$.

In any case, for $k_0 = 1$, the estimator $\widehat{\sigma^2}$ is unbiased, while for $k_0 \ll n$ but $k_0 \neq 1$, $\widehat{\sigma^2}$ is almost unbiased.

Expected value of the estimator $\widehat{\rho^2}$

To derive the expected value of the estimator $\widehat{\rho^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$, we first compute the expected value for each term of the summation,

$$\begin{aligned}
\mathrm{E}[(Y_i - \overline{Y})^2] &= \mathrm{E}[Y_i^2] - \frac{2}{n}\sum_{j=1}^{n}\mathrm{E}[Y_i Y_j] + \mathrm{E}[\overline{Y}^2] \\
&= \nu^2 + \rho^2 + \sigma^2 - \frac{2}{n}\left(\nu^2 + \rho^2 + \sigma^2 + (n_q - 1)(\nu^2 + \rho^2)\right. \\
&+ \left.(n - n_q)\nu^2\right) + \nu^2 + \frac{\sigma^2}{n} + \rho^2\frac{\sum_{p=1}^{k} n_p^2}{n^2} \\
&= \sigma^2\left(1 - \frac{1}{n}\right) + \rho^2\left(1 - \frac{2n_q}{n} + \frac{\sum_{p=1}^{k_0} n_p^2}{n^2}\right).
\end{aligned}$$

In the computation, we assumed that Y_i belongs to the q^{th} segment and we used Equations (3.29) and (3.30). Then, denoting with Y_i^q the i^{th} element of the q^{th} segment,

$$\begin{aligned}
\mathrm{E}[\widehat{\rho^2}] &= \frac{1}{n}\sum_{q=1}^{k_0}\sum_{i=1}^{n_q}\mathrm{E}[(Y_i^q - \overline{Y})^2] \\
&= \frac{1}{n}\sum_{q=1}^{k_0} n_q\left[\sigma^2\left(1 - \frac{1}{n}\right) + \rho^2\left(1 - \frac{2n_q}{n} + \frac{\sum_{p=1}^{k_0} n_p^2}{n^2}\right)\right]
\end{aligned}$$

$$= \sigma^2 \left(1 - \frac{1}{n}\right) + \rho^2 \left(1 - \frac{\sum_{p=1}^{k_0} n_p^2}{n^2}\right). \tag{3.34}$$

Notice that when $k_0 = 1$ (i.e. having only one segment), $\mathrm{E}[\widehat{\rho^2}] = \sigma^2 \left(1 - \frac{1}{n}\right)$. In this degenerate case, the variance of the segment levels ρ^2 should be estimated with zero, instead $\widehat{\rho^2}$ estimates it with the variance of the data points.

Moreover, since $\sum_{p=1}^{k_0} n_p^2 \geq n$ (the equality holds only when $k_0 = n$), we obtain that

$$\sigma^2 \left(1 - \frac{1}{n}\right) \leq \mathrm{E}[\widehat{\rho^2}] \leq \left(1 - \frac{1}{n}\right) \left(\sigma^2 + \rho^2\right). \tag{3.35}$$

Hence, if n is large the expected value is between σ^2 and $\sigma^2 + \rho^2$, so that, if $\rho^2 \ll \sigma^2$, the estimator is almost unbiased for σ^2 (instead of ρ^2).

Definition of alternative estimator of ρ^2: $\widehat{\rho_1^2}$

Since the covariance between data points belonging to the same segment is ρ^2, we could try to use a circular version of the estimator of the autocovariance of a stationary time series

$$\widehat{\rho_1^2} := \frac{1}{n} \sum_{i=1}^{n} (Y_i - \overline{Y})(Y_{i+1} - \overline{Y}), \tag{3.36}$$

where $Y_{n+1} := Y_1$. To compute the expected value for each term of the summation, we need to consider the two cases:

1. Y_i and Y_{i+1} belong to different segments,
2. Y_i and Y_{i+1} belong to the same segment.

In the first case, if we suppose that Y_i belongs to the q^{th} segment and Y_{i+1} belongs to the $(q+1)^{th}$ segment, Equations (3.29) and (3.30) imply that

$$\begin{aligned}\mathrm{E}[(Y_i-\overline{Y})(Y_{i+1}-\overline{Y})] &= \mathrm{E}[Y_iY_{i+1}]-\frac{1}{n}\sum_{j=1}^{n}\mathrm{E}[Y_iY_j]-\frac{1}{n}\sum_{j=1}^{n}\mathrm{E}[Y_{i+1}Y_j]+\mathrm{E}[\overline{Y}^2]\\ &= \nu^2-\frac{1}{n}(n\nu^2+n_q\rho^2+\sigma^2)-\frac{1}{n}(n\nu^2+n_{q+1}\rho^2+\sigma^2)\\ &+\nu^2+\frac{\sigma^2}{n}+\rho^2\frac{\sum_{p=1}^{k}n_p^2}{n^2}\\ &= \rho^2\left(\frac{\sum_{p=1}^{k_0}n_p^2}{n^2}-\frac{n_q+n_{q+1}}{n}\right)-\frac{\sigma^2}{n}.\end{aligned}$$

In the second case, we obtain that

$$\begin{aligned}\mathrm{E}[(Y_i-\overline{Y})(Y_{i+1}-\overline{Y})] &= \nu^2+\rho^2-\frac{2}{n}(n\nu^2+n_q\rho^2+\sigma^2)+\nu^2+\frac{\sigma^2}{n}\\ &+\rho^2\frac{\sum_{p=1}^{k}n_p^2}{n^2}=\rho^2\left(\frac{\sum_{p=1}^{k_0}n_p^2}{n^2}-\frac{2n_q}{n}+1\right)-\frac{\sigma^2}{n}.\end{aligned}$$

Therefore, by summing over all $i=1,\ldots,n$, we obtain

$$\begin{aligned}&\sum_{i=1}^{n}\mathrm{E}[(Y_i-\overline{Y})(Y_{i+1}-\overline{Y})]\\ &=\sum_{q=1}^{k_0}\sum_{i=1}^{n_q-1}\mathrm{E}[(Y_i^q-\overline{Y})(Y_{i+1}^q-\overline{Y})]+\mathbb{I}_{\{k_0\geq 2\}}\sum_{q=1}^{k_0-1}\mathrm{E}[(Y_{n_q}^q-\overline{Y})(Y_1^{q+1}-\overline{Y})]\\ &+\mathrm{E}[(Y_n-\overline{Y})(Y_1-\overline{Y})]\\ &=\sum_{q=1}^{k_0}(n_q-1)\left[\rho^2\left(\frac{\sum_{p=1}^{k_0}n_p^2}{n^2}-\frac{2n_q}{n}+1\right)-\frac{\sigma^2}{n}\right]\\ &+\mathbb{I}_{\{k_0\geq 2\}}\sum_{q=1}^{k_0-1}\left[\rho^2\left(\frac{\sum_{p=1}^{k_0}n_p^2}{n^2}-\frac{n_q+n_{q+1}}{n}\right)-\frac{\sigma^2}{n}\right]\\ &+\mathbb{I}_{\{k_0\geq 2\}}\left[\rho^2\left(\frac{\sum_{p=1}^{k_0}n_p^2}{n^2}-\frac{n_{k_0}+n_1}{n}\right)-\frac{\sigma^2}{n}\right]-\frac{\sigma^2}{n}\mathbb{I}_{\{k_0=1\}}\end{aligned}$$

$$= \rho^2 \left[n + 2 - k_0 - 2\mathbb{I}_{\{k_0 \geq 2\}} - \frac{\sum_{p=1}^{k_0} n_p^2}{n^2} \left(n + k_0 - k_0 \mathbb{I}_{\{k_0 \geq 2\}} \right) \right]$$
$$+ \sigma^2 \left(\frac{k_0}{n} - 1 - \frac{k_0}{n} \mathbb{I}_{\{k_0 \geq 2\}} + \frac{1}{n} \mathbb{I}_{\{k_0 = 1\}} \right)$$

$$\Rightarrow \mathrm{E}[\widehat{\rho_1^2}] = \begin{cases} -\frac{\sigma^2}{n} & \text{if } k_0 = 1 \\ \frac{\rho^2}{n} \left(n - k_0 - \frac{\sum_{p=1}^{k_0} n_p^2}{n} \right) - \frac{\sigma^2}{n} & \text{if } k_0 \geq 2. \end{cases} \tag{3.37}$$

In the computation we considered two cases: $k_0 = 1$ and $k_0 \geq 2$. When $k_0 = 1$, Y_1 and Y_n belong to the same segment and thus they are dependent; when $k_0 \geq 2$, we suppose that the first and the last segment do not have the same level value and thus Y_1 and Y_n are independent. If $k_0 \geq 2$ and the first and the last segment had the same level value (event with a very low probability), then the first and the last segments would be joined together and so Y_1 and Y_n would be dependent. In this case, the expected value of the estimator would have the same formula, but with $k_0 - 1$ instead of k_0.

We can observe that, when $k_0 = 1$, the expected value is negative, while, when $k_0 \geq 2$, it can be negative or positive. Moreover, the coefficient of σ^2 is $-\frac{1}{n}$ and thus this addendum does not contribute much to the unbiasedness of the estimator.

The negativity of the expected value happens because the estimator is a generic estimator of the covariance and, in general, this quantity can be negative. To prevent the negativity of the estimator, we can use its absolute value. In this way, when the quantity in (3.36) is negative, we use the same estimator but with the sign changed in one of the factors of each product, $\widehat{\rho_1^2} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \overline{Y})(\overline{Y} - Y_{i+1})$. Hence, the meaning of the estimator is the same. We are interested only in the absolute value of the estimate and not in its sign. In fact, we already know that the correlation is positive and the negativity of the estimate is due only to the property of the estimator. Our final definition of the estimator is then

$$\widehat{\rho_1^2} := \frac{1}{n} \left| \sum_{i=1}^{n} (Y_i - \overline{Y})(Y_{i+1} - \overline{Y}) \right|.$$

3.1.6 Comparison among the proposed estimators on simulated data

In this subsection, we experimentally compare the behavior of all the estimators proposed previously, on the basis of their results obtained on the artificial datasets. The comparisons were accomplished using both the true and the estimated values of the other parameters involved in the estimation.

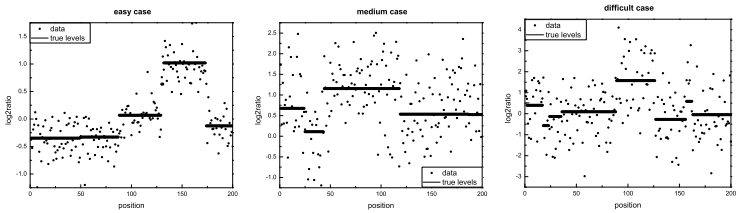

Fig. 3.1 The simulated data in the figure represent an easy, medium and difficult case, respectively. [Reprinted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

We used several types of artificial data. We call *sample* a sequence of data which represents the copy number data of a genomic region, we call *dataset* a set of samples, while *collection* a set of datasets.

In order to experimentally evaluate the behavior of all the estimators proposed, we used the artificial datasets sampled from the priors, defined in the hypotheses of the model. We always chose $\nu = 0.2$, while we changed the values of σ^2 and ρ^2 for each dataset, in order to study different situations of noise (some examples of data are in Figure 3.1 and the corresponding estimated profiles obtained by applying several methods are in Figures 3.7 and 3.20). The most problematic cases were the ones with $\rho^2 < \sigma^2$ (i.e. when the variance of the noise was higher than the variance of the segment levels), because in these cases it was hard to identify the true profile of the levels. We always used $n = 200$, similar to the mean number of probes of a small chromosome in the Affymetrix

GeneChip Mapping 10K Array (hence it represented a difficult case due to the small sample size), and $k_{\max} = 40$, in order to have at least 5 probes per segment on average.

Sometimes we needed datasets where all samples had the same true profile of the segment levels (i.e. K, T and M were sampled one time and only the noise varied in all samples). This type of dataset is called *dataset with replicates*. Otherwise, the dataset is called *without replicates* (i.e. each time we sampled K, T, M and added the noise to the profile). The number of samples per dataset was 100, for datasets with replicates, and 300 otherwise. We considered datasets with replicates in order to be able to compare the goodness of different types of estimations for a given profile.

We also compared the behavior of our boundary estimators using the artificial dataset already employed in [84], where three methods for copy number estimation were examined. This dataset contained 500 samples consisting of 20 chromosomes, each of 100 probes, which emulated the real copy number data. This dataset is referred to as *Simulated Chromosomes*.

Comparison among the hyper-parameter estimators

We applied the hyper-parameter estimators on eight datasets without replicates, considering different values for σ^2 and ρ^2 (examples of data are in Figure 3.2). To evaluate the behavior of the hyper-parameter estimators in all these cases, for each dataset we computed the (estimated) Mean Square Error (MSE) with respect to the true value of the parameter. The MSE is defined as the expected value of the square error between the estimator $\widehat{\Theta}$ and the parameter θ (to estimate),

$$\begin{aligned} \mathrm{MSE}_{\widehat{\Theta}} &= \mathrm{E}[(\widehat{\Theta} - \theta)^2] \\ &= \mathrm{Var}(\widehat{\Theta}) + (\mathrm{E}[\widehat{\Theta}] - \theta)^2 \\ &= \mathrm{Var}(\widehat{\Theta}) + \Big(\mathrm{Bias}(\widehat{\Theta})\Big)^2 . \end{aligned} \tag{3.38}$$

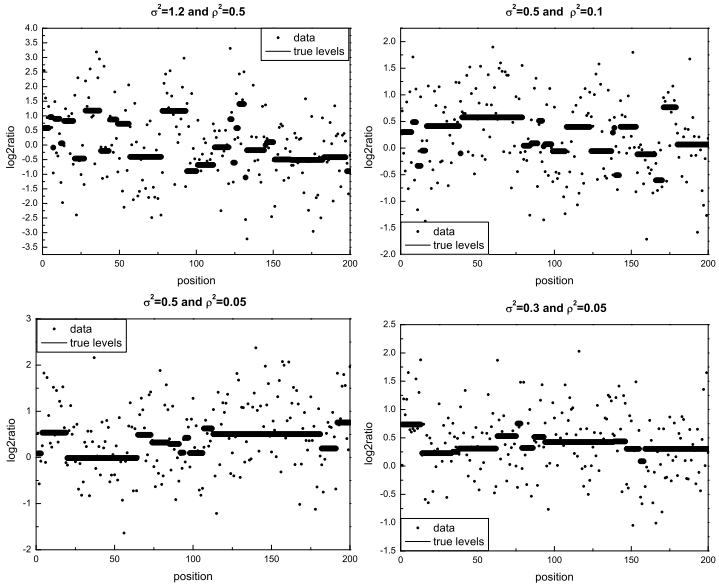

Fig. 3.2 True profiles of some datasets with replicates used in the comparisons. [Reprinted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

The $\mathrm{MSE}_{\widehat{\Theta}}$ is estimated with the corresponding arithmetic mean. Given N estimated values $\hat{\theta}_1, \ldots, \hat{\theta}_N$ of the parameter θ, the estimated Mean Square Error of $\widehat{\Theta}$ is defined as

$$\widehat{\mathrm{MSE}}_{\widehat{\Theta}} := \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta)^2.$$

If the estimator is unbiased $\hat{\mathrm{MSE}}_{\widehat{\Theta}}$ estimates the variance of the estimator (see Equation (3.38)).

Table 3.1 Estimated mean square error of the estimators $\hat{\nu}$, $\hat{\sigma}^2$, $\hat{\rho}^2$ and $\hat{\rho}^2_{\mathrm{I}}$ applied on datasets with different values of σ^2 and ρ^2. The table shows that $\mathrm{MSE}_{\hat{\nu}}$ increases with ρ^2 and $\mathrm{MSE}_{\hat{\sigma}^2}$ increases with σ^2. The estimator $\hat{\rho}^2_{\mathrm{I}}$ is generally better than $\hat{\rho}^2$ with respect to the MSE error measure. [Adapted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

error	$\sigma^2=0.1$ $\rho^2=0.5$	$\sigma^2=0.3$ $\rho^2=0.05$	$\sigma^2=0.5$ $\rho^2=0.02$	$\sigma^2=0.5$ $\rho^2=0.05$	$\sigma^2=0.5$ $\rho^2=0.1$	$\sigma^2=0.7$ $\rho^2=0.5$	$\sigma^2=1$ $\rho^2=0.05$	$\sigma^2=1.2$ $\rho^2=0.5$
$\widehat{\mathrm{MSE}}_{\hat{\nu}}$	0.0904	0.0091	0.0059	0.0094	0.021	0.067	0.0169	0.0729
$\widehat{\mathrm{MSE}}_{\hat{\sigma}^2}$	0.0042	0.0014	0.0041	0.0036	0.0037	0.0114	0.0123	0.0272
$\widehat{\mathrm{MSE}}_{\hat{\rho}^2}$	0.0633	0.0871	0.2508	0.2404	0.2426	0.4271	0.9921	1.3254
$\widehat{\mathrm{MSE}}_{\hat{\rho}^2_{\mathrm{I}}}$	0.068	0.0009	0.0008	0.0014	0.0047	0.0593	0.0024	0.0623

Looking at the errors of the estimations of ν, we observed that the error seemed to depend more on the ρ^2 value than on σ^2 value, increasing as the value of ρ^2 increased. This is a natural behavior due to the definition of this estimator. Since it is unbiased (see Subsection 3.1.5), $\widehat{\mathrm{MSE}}_{\hat{\nu}}$ is the average of the variance of $\hat{\nu}$ over the samples (notice that the variance is different in each sample because it depends on the segment number, see Equation (3.31)). Then, the behavior or $\widehat{\mathrm{MSE}}_{\hat{\nu}}$ is due to the fact that each variance is a linear combination of σ^2 and ρ^2, where the coefficient of the latter is greater than the one of the former (Equation (3.31)).

Regarding $\hat{\sigma}^2$, its error was always low, but we observed that it increased as the value of σ^2 increased, independently on ρ^2 value. In fact, although the estimator is biased and its bias is proportional to ρ^2 (Equation (3.33)), the value of the bias is small ($\rho^2 k_0/n$).

Comparing the behavior of the two different ρ^2 estimators, we saw that, apart from the first case in Table 3.1, $\hat{\rho}^2_{\mathrm{I}}$ always gave a lower error than $\hat{\rho}^2$ and we also noticed that the two errors had different order of magnitude. In the first case, the only one in which $\rho^2 > \sigma^2$, $\widehat{\mathrm{MSE}}_{\hat{\rho}^2}$ and $\widehat{\mathrm{MSE}}_{\hat{\rho}^2_{\mathrm{I}}}$ were comparable and $\widehat{\mathrm{MSE}}_{\hat{\rho}^2}$ was the lowest. We also observed that the error of

$\hat{\rho}^2$ did not depend on the value of ρ^2, while it increased as the value σ^2 increased. In fact, we know from Subsection 3.1.5 that $\hat{\rho}^2$ is biased and its bias highly depends on σ^2 (Equation (3.34)). On the contrary, the error of $\hat{\rho}_{\mathrm{I}}^2$ increased only as the value of ρ^2 increased, because its bias depends less on σ^2 (Equation (3.37)).

To summarize the accuracy of the estimators, we computed the estimated mean relative error over all datasets (Table 3.2),

$$\frac{\Delta\hat{\theta}}{\theta} := \frac{1}{MN}\sum_{j=1}^{M}\sum_{i=1}^{N}\frac{\left|\hat{\theta}_{ij}-\theta_j\right|}{\theta_j},$$

where M is the number of datasets, N the number of samples in each dataset and $\hat{\theta}_{ij}$ is the estimate of θ_j based on the i^{th} sample in the j^{th} dataset.

Table 3.2 Estimated mean relative error of the estimators $\hat{\nu}$, $\hat{\sigma}^2$, $\hat{\rho}^2$ and $\hat{\rho}_{\mathrm{I}}^2$ over all the datasets used. The results show that $\hat{\sigma}^2$ has the lowest error. The error of $\hat{\nu}$ is higher but acceptable. The estimator $\hat{\rho}_{\mathrm{I}}^2$ is better than $\hat{\rho}^2$ with respect to this error measure. [Reprinted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

type of error	value
$\Delta\hat{\nu}/\nu$	0.6376
$\Delta\hat{\sigma}^2/\sigma^2$	0.1524
$\Delta\hat{\rho}^2/\rho^2$	8.6217
$\Delta\hat{\rho}_{\mathrm{I}}^2/\rho_{\mathrm{I}}^2$	0.5840

From the results, we can deduce that $\hat{\sigma}^2$ is a good estimator because it was quite precise in all situations, while $\hat{\nu}$ was sometimes poor but in general acceptable. About the ρ^2 estimation, it is better to use $\hat{\rho}_{\mathrm{I}}^2$ than $\hat{\rho}^2$, when the variance of the noise is higher than the variance of the levels. Otherwise, it seems better to use $\hat{\rho}^2$ because it does not underestimate ρ^2 (see Subsection 3.1.5).

Comparison among the segment number estimators

We evaluated the quality of the estimators of the number of segments, using the eight datasets without replicates already employed for the comparison of the hyper-parameters estimators. The estimations were made using either the true values of the hyper-parameters or the estimated ones. In this way, we could also observe the behavior of the boundary estimators without the influence of the hyper-parameter estimation. The results are in Tables 3.3, 3.4 and 3.5.

Comparing the absolute, squared and 0-1 errors, we found that $\widehat{K}_2$ generally had the lowest upper bound (or its upper bound was very close to the lowest one) of the confidence interval at level 95%, for any type of error and any type of value of the parameter ρ^2 (see, for example, Figures 3.3 and 3.4). The estimator $\widehat{K}_1$ had a behavior similar to $\widehat{K}_2$, but it often had a larger confidence interval which contained the confidence interval of $\widehat{K}_2$. Moreover, all estimators always had a similar confidence interval of the 0-1 error, while using $\hat{\rho}^2$, in most cases the upper bound of the confidence interval of the absolute and the squared error was lower than using $\hat{\rho}_1^2$. All these results support the suggestion to use $\widehat{K}_2$ with $\hat{\rho}^2$.

We should also observe that in general $\widehat{K}_{01}$ underestimates k_0, while $\widehat{K}_1$ and $\widehat{K}_2$ overestimate it. In addition, the percentage of the underestimations increases using $\hat{\rho}^2$.

Comparison among the boundary estimators

We compared the boundary estimators on the same datasets, previously used for the estimators of the number of segments, with both the estimated and the true value of the parameters involved. The following errors were taken into account: the sum 0-1 error, the joint 0-1 error and the binary error, defined in Subsection 3.1.4, and the average square error,

$$\text{sum 0-1 error} = \sum_{p=1}^{k^0-1} \left(1 - \delta_{t_p,\widehat{T}_p}\right)$$

Table 3.3 Estimated mean errors ± standard deviation, of the estimators of k_0 applied to eight datasets without replicates, by using in the regression σ^2, ρ^2 and ν.

dataset	error	$\widehat{K}_{01}$	$\widehat{K}_1$	$\widehat{K}_2$
σ^2=0.1	0-1	0.82 ± 0.39	0.79 ± 0.41	0.84 ± 0.37
ρ^2=0.5	absolute	2.83 ± 2.94	2.87 ± 3.13	3 ± 3.17
	squared	16.65 ± 37.17	17.99 ± 42.61	19.02 ± 44.99
σ^2=0.3	0-1	0.94 ± 0.24	0.96 ± 0.2	0.98 ± 0.15
ρ^2=0.05	absolute	15.93 ± 19.86	16.97 ± 15.32	18.54 ± 13.92
	squared	646.88 ± 1512.8	521.87 ± 824.5	537.03 ± 719.3
σ^2=0.5	0-1	0.91 ± 0.28	0.96 ± 0.2	0.97 ± 0.17
ρ^2=0.1	absolute	18.09 ± 22.3	17.8 ± 16.1	18.4 ± 15.06
	squared	823.1 ± 1783.1	575.33 ± 895.8	564.7 ± 781.1
σ^2=0.5	0-1	0.98 ± 0.14	0.97 ± 0.17	0.98 ± 0.15
ρ^2=0.05	absolute	23.16 ± 25.62	21.73 ± 16.27	22.97 ± 14.9
	squared	1190.9 ± 2240.9	736.3 ± 942.4	748.9 ± 830.7
σ^2=0.5	0-1	0.95 ± 0.21	0.99 ± 0.11	1 ± 0.06
ρ^2=0.02	absolute	30.24 ± 30.14	25.24 ± 15.31	26.24 ± 13.82
	squared	1819.8 ± 2703	870.6 ± 877.4	878.8 ± 783.5
σ^2=0.7	0-1	0.94 ± 0.24	0.91 ± 0.28	0.93 ± 0.25
ρ^2=0.5	absolute	7.4 ± 8.35	7.89 ± 8.42	8.49 ± 8.6
	squared	124.21 ± 333	132.91 ± 302.73	145.82 ± 302.65
σ^2=1	0-1	0.95 ± 0.22	0.99 ± 0.11	0.99 ± 0.11
ρ^2=0.05	absolute	26.55 ± 28.34	24.46 ± 15.53	25.85 ± 14.15
	squared	1505.1 ± 2502.6	838.8 ± 893.8	867.5 ± 797
σ^2=1.2	0-1	0.94 ± 0.24	0.95 ± 0.23	0.97 ± 0.17
ρ^2=0.5	absolute	10.52 ± 14.89	10.71 ± 12.09	11.87 ± 11.73
	squared	331.71 ± 1061.8	260.55 ± 596.3	278.06 ± 537.4

$$\text{total 0-1 error} = 1 - \delta_{t,\widehat{T}}$$

$$\text{binary error} = k^0 - 1 - \sum_{i=1}^{n-1} \delta_{\tau_i^0,\widehat{\boldsymbol{\tau}}_i} = k^0 - 1 - \sum_{q=1}^{\hat{k}-1} \sum_{p=1}^{k^0-1} \delta_{t_p^0,\widehat{T}_q}$$

$$\text{average squared error} = \frac{1}{k^0-1} \sum_{p=1}^{k^0-1} \min_{q=1,\ldots,k-1} \left(t_q^0 - \widehat{T}_p\right)^2.$$

Table 3.4 Estimated mean errors ± standard deviation, of the estimators of k_0 applied to eight datasets without replicates, by using in the regression $\hat{v}$, $\hat{\sigma}^2$ and $\hat{\rho}^2$.

dataset	error	$\widehat{K}_{01}$	$\widehat{K}_1$	$\widehat{K}_2$
σ^2=0.1	0-1	0.86 ± 0.34	0.88 ± 0.33	0.9 ± 0.3
ρ^2=0.5	absolute	4.06 ± 3.65	3.74 ± 3.33	3.65 ± 3.18
	squared	29.78 ± 46.06	25.04 ± 39.8	23.45 ± 37.92
σ^2=0.3	0-1	0.97 ± 0.17	0.95 ± 0.21	0.97 ± 0.16
ρ^2=0.05	absolute	15.55 ± 11.61	13.44 ± 10.78	12.4 ± 10.27
	squared	376.16 ± 434.9	296.45 ± 381.7	258.91 ± 352.5
σ^2=0.5	0-1	0.95 ± 0.22	0.95 ± 0.23	0.92 ± 0.28
ρ^2=0.1	absolute	14.32 ± 10.92	12.37 ± 10.25	11.39 ± 9.78
	squared	323.78 ± 377	257.55 ± 334.4	224.93 ± 304.3
σ^2=0.5	0-1	0.97 ± 0.16	0.94 ± 0.23	0.97 ± 0.16
ρ^2=0.05	absolute	16.39 ± 10.65	14.58 ± 10.59	13.75 ± 10.22
	squared	381.9 ± 403.8	324.3 ± 377.7	293 ± 352.5
σ^2=0.5	0-1	0.97 ± 0.17	0.96 ± 0.2	0.97 ± 0.18
ρ^2=0.02	absolute	19.24 ± 11.84	17.88 ± 11.77	16.87 ± 11.41
	squared	509.9 ± 472.3	457.9 ± 450.4	414.3 ± 423.7
σ^2=0.7	0-1	0.94 ± 0.24	0.92 ± 0.28	0.95 ± 0.22
ρ^2=0.5	absolute	7.86 ± 6.54	6.79 ± 5.85	6.45 ± 5.52
	squared	104.46 ± 162.57	80.13 ± 131.64	71.96 ± 122.17
σ^2=1	0-1	0.96 ± 0.19	0.95 ± 0.23	0.92 ± 0.27
ρ^2=0.05	absolute	17.82 ± 12.01	16.86 ± 11.9	15.88 ± 11.63
	squared	461.4 ± 459.6	425.4 ± 440.6	387 ± 415.4
σ^2=1.2	0-1	0.94 ± 0.24	0.95 ± 0.23	0.95 ± 0.22
ρ^2=0.5	absolute	10.23 ± 8.53	8.49 ± 7.45	7.78 ± 6.74
	squared	177.17 ± 251.7	127.47 ± 195.4	105.74 ± 165.1

The first three errors were used in the definition of the different estimators, while the last one corresponds to the mean square error over the whole vector of estimated boundaries. As observed in Subsection 3.1.4, when the estimated segment number is used in the estimation of the boundaries, we are only able to compute the binary error, because it does not require that the vector of estimated boundaries has the same length as the vector of the true boundaries.

Table 3.5 Estimated mean errors ± standard deviation, of the estimators of k_0 applied to eight datasets without replicates, by using in the regression $\hat{v}$, $\hat{\sigma}^2$ and $\hat{\rho}_1^2$.

dataset	error	$\widehat{K}_{01}$	$\widehat{K}_1$	$\widehat{K}_2$
σ^2=0.1	0-1	0.9 ± 0.3	0.89 ± 0.32	0.88 ± 0.32
ρ^2=0.5	absolute	4.43 ± 5.68	4.17 ± 5.04	4.14 ± 4.96
	squared	51.71 ± 247.72	42.71 ± 197	41.65 ± 187.05
σ^2=0.3	0-1	0.92 ± 0.27	0.97 ± 0.17	0.97 ± 0.16
ρ^2=0.05	absolute	13.31 ± 14.11	13.7 ± 11.76	14.22 ± 11.57
	squared	375.62 ± 846.1	325.66 ± 563.1	335.82 ± 537.8
σ^2=0.5	0-1	0.96 ± 0.19	0.96 ± 0.2	0.98 ± 0.15
ρ^2=0.1	absolute	15.59 ± 15.56	15.4 ± 13.33	15.9 ± 12.94
	squared	484.11 ± 931.8	414.18 ± 642.3	419.68 ± 608.1
σ^2=0.5	0-1	0.96 ± 0.2	0.97 ± 0.16	0.98 ± 0.15
ρ^2=0.05	absolute	15.76 ± 16.22	15.06 ± 14.02	15.33 ± 13.69
	squared	510.6 ± 1079.2	422.9 ± 781.2	421.8 ± 742.8
σ^2=0.5	0-1	0.97 ± 0.16	0.97 ± 0.16	0.97 ± 0.16
ρ^2=0.02	absolute	15.45 ± 14.15	13.84 ± 11.86	13.85 ± 11.58
	squared	438.1 ± 830.1	331.8 ± 588.1	325.5 ± 559.3
σ^2=0.7	0-1	0.96 ± 0.2	0.95 ± 0.21	0.96 ± 0.2
ρ^2=0.5	absolute	8.33 ± 8.95	9.23 ± 9.08	10.05 ± 9.37
	squared	149.18 ± 360.27	167.28 ± 343.79	188.59 ± 350.74
σ^2=1	0-1	0.98 ± 0.15	0.97 ± 0.16	0.99 ± 0.1
ρ^2=0.05	absolute	14.41 ± 12.97	13.17 ± 10.66	13.14 ± 10.53
	squared	375.1 ± 705.5	286.7 ± 487.6	283.1 ± 470.6
σ^2=1.2	0-1	0.97 ± 0.16	0.95 ± 0.21	0.99 ± 0.11
ρ^2=0.5	absolute	9.69 ± 10.15	10.78 ± 9.67	11.78 ± 9.81
	squared	196.63 ± 501.2	209.46 ± 357.6	234.66 ± 354.7

Using the true values of v, σ^2, ρ^2 and k_0 (see Table 3.6), we observed that all the T estimators gave similar sum 0-1 errors and total 0-1 errors. With regard to the binary error, the results were similar, but they were more distinguishable with respect to the previous errors: $\widehat{T}_{joint}$, $\widehat{T}_{BinErr}$ and $\widehat{T}_{BinErrAk}$ gave the lowest errors and $\widehat{T}_{01}$ the highest. Considering the average squared error, we divided the datasets into two groups: the first one, where $\sigma^2 < \rho^2$, and the other ones, where $\sigma^2 > \rho^2$. In the first dataset,

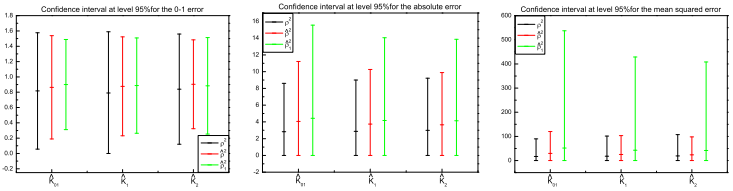

Fig. 3.3 Confidence intervals at 95% of the estimators of k_0 for the 0-1 error, the absolute error and the squared error, respectively, obtained on a dataset without replicates with $\sigma^2 = 0.1$ and $\rho^2 = 0.5$. The graphs show that, in each situation, $\widehat{K}_2$ always had the lowest upper bound of the interval. Using $\hat{\rho}^2$ the confidence intervals were shorter. [Reprinted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

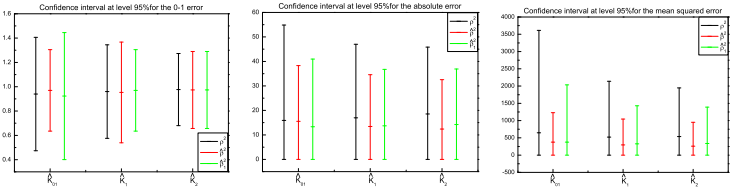

Fig. 3.4 Confidence intervals at 95% of the estimators of k_0 for the 0-1 error, the absolute error and the squared error, respectively, obtained on a dataset without replicates with $\sigma^2 = 0.3$ and $\rho^2 = 0.05$. The graphs show that, in each situation, $\widehat{K}_2$ always had the lowest upper bound of the interval. [Reprinted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

$\widehat{T}_{01}$ and $\widehat{T}_{BinErr}$ gave the lowest error and $\widehat{T}_{joint}$ the highest one, while, in the other datasets, $\widehat{T}_{01}$, $\widehat{T}_{joint}$ and $\widehat{T}_{BinErr}$ gave the lowest error and $\widehat{T}_{BinErrAk}$ the highest one.

Afterwards, we used in the regression estimated values of the parameters. We used $\widehat{K}_2$ to estimate the segment number and both $\widehat{\rho}^2$ and $\widehat{\rho}_1^2$ to estimate ρ^2 (Tables 3.7 and 3.8). In all the cases, looking at the binary er-

Table 3.6 Estimated mean errors ± standard deviation, of the estimators of t^0 applied to eight datasets without replicates, by using in the regression k_0, ν, σ^2 and ρ^2.

dataset	error	$\widehat{T}_{01}$	$\widehat{T}_{joint}$	$\widehat{T}_{BinErr}$	$\widehat{T}_{BinErrAk}$
	sum 0-1	10.29 ± 6.92	8.35 ± 5.24	7.69 ± 4.93	7.7 ± 4.93
σ^2=0.1	total 0-1	0.94 ± 0.23	0.94 ± 0.23	0.94 ± 0.23	0.94 ± 0.24
ρ^2=0.5	binary	10.29 ± 6.92	8.35 ± 5.24	7.69 ± 4.93	7.7 ± 4.93
	av. squared	7.13 ± 45.71	38.39 ± 133.66	7.45 ± 21.43	11.98 ± 70.63
	sum 0-1	16.9 ± 9.97	15.65 ± 8.87	15.15 ± 8.71	15.25 ± 8.71
σ^2=0.3	total 0-1	0.98 ± 0.13	0.98 ± 0.13	0.98 ± 0.13	0.98 ± 0.13
ρ^2=0.05	binary	16.9 ± 9.97	15.65 ± 8.87	15.15 ± 8.71	15.25 ± 8.71
	av. squared	196.23 ± 891.73	177.56 ± 811.37	150.35 ± 926.13	151.53 ± 591.5
	sum 0-1	17.1 ± 9.83	15.66 ± 8.6	15.25 ± 8.63	15.38 ± 8.59
σ^2=0.5	total 0-1	0.97 ± 0.16	0.97 ± 0.16	0.97 ± 0.16	0.97 ± 0.16
ρ^2=0.1	binary	17.1 ± 9.83	15.66 ± 8.6	15.25 ± 8.63	15.38 ± 8.59
	av. squared	204.94 ± 895.42	182.21 ± 687.28	133.18 ± 824.12	416.61 ± 2279.1
	sum 0-1	16.52 ± 9.1	15.64 ± 8.4	15.27 ± 8.23	15.35 ± 8.23
σ^2=0.5	total 0-1	0.97 ± 0.16	0.97 ± 0.16	0.97 ± 0.16	0.97 ± 0.16
ρ^2=0.05	binary	16.52 ± 9.1	15.64 ± 8.4	15.27 ± 8.23	15.35 ± 8.23
	av. squared	221.24 ± 1123.88	188.88 ± 821.38	152.02 ± 823.34	233.57 ± 1504.93
	sum 0-1	17.76 ± 10.13	17 ± 9.54	16.89 ± 9.41	16.95 ± 9.51
σ^2=0.5	total 0-1	0.98 ± 0.15	0.98 ± 0.15	0.98 ± 0.15	0.98 ± 0.15
ρ^2=0.02	binary	17.76 ± 10.13	17 ± 9.54	16.89 ± 9.41	16.95 ± 9.51
	av. squared	261.44 ± 891.24	208.92 ± 744.72	219.78 ± 906.24	351.09 ± 1889.43
	sum 0-1	15.87 ± 8.82	13.91 ± 7.41	13.19 ± 6.99	13.3 ± 7.07
σ^2=0.7	total 0-1	0.99 ± 0.11	0.99 ± 0.11	0.99 ± 0.1	0.99 ± 0.1
ρ^2=0.5	binary	15.87 ± 8.82	13.91 ± 7.41	13.19 ± 6.99	13.3 ± 7.07
	av. squared	150.34 ± 1370.74	129.08 ± 1235.01	100.17 ± 1232.16	102.42 ± 589.02
	sum 0-1	16.55 ± 10.39	15.68 ± 9.55	15.54 ± 9.68	15.63 ± 9.63
σ^2=1.0	total 0-1	0.97 ± 0.17	0.97 ± 0.17	0.97 ± 0.17	0.97 ± 0.17
ρ^2=0.05	binary	16.55 ± 10.39	15.68 ± 9.55	15.54 ± 9.68	15.63 ± 9.63
	av. squared	731.75 ± 3205.78	769.53 ± 3256.3	703.13 ± 3290.92	580.06 ± 2761.28
	sum 0-1	17.01 ± 9.49	15.34 ± 8.08	14.61 ± 7.85	14.73 ± 7.91
σ^2=1.2	total 0-1	0.98 ± 0.15	0.98 ± 0.14	0.98 ± 0.14	0.98 ± 0.14
ρ^2=0.5	binary	17.01 ± 9.49	15.34 ± 8.08	14.61 ± 7.85	14.73 ± 7.91
	av. squared	384.21 ± 2599.43	377.84 ± 2585.1	345.71 ± 2585.49	284.99 ± 2164.06

ror, we saw that $\widehat{T}_{01}$ gave the highest error, while the other estimators had the same behavior and $\widehat{T}_{BinErr}$ and $\widehat{T}_{BinErrAk}$ gave the lowest upper bound of the confidence interval. Moreover, in the cases in which $\sigma^2 > \rho^2$, we saw that there was a great difference between the errors obtained using $\widehat{\rho}^2$ and $\widehat{\rho}^2_1$: using $\widehat{\rho}^2_1$, the errors were significantly lower.

Table 3.7 Estimated mean binary error ± standard deviation, of the estimators of t^0 applied to eight datasets without replicates, by using in the regression $\widehat{K}_2$, $\hat{\nu}$, $\hat{\sigma}^2$ and $\hat{\rho}^2$. [Adapted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

dataset	$\widehat{T}_{01}$	$\widehat{T}_{joint}$	$\widehat{T}_{BinErr}$	$\widehat{T}_{BinErrAk}$
σ^2=0.1, ρ^2=0.5	11.6 ± 0.46	8.87 ± 0.33	8.68 ± 0.34	8.73 ± 0.34
σ^2=0.3, ρ^2=0.05	18.09 ± 0.64	17.63 ± 0.62	17.38 ± 0.62	17.39 ± 0.62
σ^2=0.5, ρ^2=0.1	18.2 ± 0.62	17.48 ± 0.59	17.25 ± 0.6	17.26 ± 0.6
σ^2=0.5, ρ^2=0.05	17.83 ± 0.59	17.58 ± 0.59	17.46 ± 0.59	17.46 ± 0.59
σ^2=0.5, ρ^2=0.02	19.47 ± 0.68	19.31 ± 0.68	19.33 ± 0.68	19.31 ± 0.68
σ^2=0.7, ρ^2=0.5	16.29 ± 0.54	14.72 ± 0.47	14 ± 0.46	14.01 ± 0.46
σ^2=1.0, ρ^2=0.05	17.77 ± 0.69	17.63 ± 0.69	17.64 ± 0.69	17.61 ± 0.69
σ^2=1.2, ρ^2=0.5	17.68 ± 0.58	16.62 ± 0.54	15.9 ± 0.53	15.91 ± 0.53

Table 3.8 Estimated mean binary error ± standard deviation, of the estimators of t^0 applied to eight datasets without replicates, by using in the regression $\widehat{K}_2$, $\hat{\nu}$, $\hat{\sigma}^2$ and $\hat{\rho}_1^2$. [Adapted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

dataset	$\widehat{T}_{01}$	$\widehat{T}_{joint}$	$\widehat{T}_{BinErr}$	$\widehat{T}_{BinErrAk}$
σ^2=0.1, ρ^2=0.5	11.48 ± 0.45	8.79 ± 0.33	8.47 ± 0.34	8.46 ± 0.34
σ^2=0.3, ρ^2=0.05	15.29 ± 0.53	13.73 ± 0.47	13.06 ± 0.46	13.05 ± 0.47
σ^2=0.5, ρ^2=0.1	15.69 ± 0.54	13.81 ± 0.47	13.18 ± 0.47	13.08 ± 0.47
σ^2=0.5, ρ^2=0.05	14.81 ± 0.49	13.31 ± 0.45	13.02 ± 0.44	12.97 ± 0.44
σ^2=0.5, ρ^2=0.02	15.84 ± 0.55	14.2 ± 0.51	14.03 ± 0.5	14.05 ± 0.5
σ^2=0.7, ρ^2=0.5	15.56 ± 0.5	13.29 ± 0.41	12.34 ± 0.4	12.34 ± 0.4
σ^2=1.0, ρ^2=0.05	14.78 ± 0.56	13.38 ± 0.5	13.12 ± 0.5	13.08 ± 0.5
σ^2=1.2, ρ^2=0.5	16.19 ± 0.53	14.17 ± 0.45	13.24 ± 0.43	13.22 ± 0.44

Therefore, the best estimators were $\widehat{T}_{BinErr}$, $\widehat{T}_{BinErrAk}$ and $\widehat{T}_{joint}$, but it seemed that $\widehat{T}_{BinErr}$, $\widehat{T}_{BinErrAk}$ were slightly better than $\widehat{T}_{joint}$, when we estimate the parameters. In fact, their definition of the error takes into account that the segment number must be estimated. Moreover, be-

tween these ones, we preferred the second one, because its computation depended less on the estimation of k_0 and thus it was more stable.

To choose between the estimators $\widehat{T}_{BinErrAk}$ and $\widehat{T}_{joint}$, we performed the segment level estimation, considering both ρ^2 estimators. We computed the $\widehat{\text{MSE}}$ of the estimated segment level per probe and its upper bound of the confidence interval at level 95% (the corresponding graphs regarding some datasets can be found in Figure 3.5). The results showed that, for a given estimator of ρ^2, in general $\widehat{T}_{BinErrAk}$ was better than $\widehat{T}_{joint}$ and the upper bound of the confidence interval of the $\widehat{\text{MSE}}$ of the former estimator was lower than that one of the latter. Moreover, using $\widehat{\rho}_1^2$ the error was generally lower.

In addition, we compared the behavior of our boundary estimators on dataset *Simulated Chromosomes*. Similarly to [84], to assess the goodness of the boundary estimation, we measured the sensitivity (proportion of true breakpoints detected) and the false discovery rate (FDR, i.e. proportion of false estimated breakpoints among the estimated ones), see Figure 3.6. The sensitivity and the FDR were computed not only looking at the exact position of the breakpoints (w = 0), but also accounting for a neighborhood of 1 or 2 probes around the true positions (w = 1, 2). To assess the influence of the boundary estimation on the profile estimation, we calculated the sum of squared distance (SSQ), the median absolute deviation (MAD) and the accuracy (proportion of probes correctly assigned to an altered or unaltered state). We also computed the accuracy inside and outside the aberrations separately, since the samples of dataset *Simulated Chromosomes* contained only few small copy number changes and thus the accuracy depended more on the correct estimation/classification of the probes in the "normal" regions. The results are in Table 3.9.

Since we estimated $\hat{\sigma}^2 = 0.026$ and $\hat{\rho}_1^2 = 0.031$ (using the whole dataset), we expected that we should obtain better results by using $\hat{\rho}^2$ because σ^2 was lower than ρ^2 and indeed the results obtained with $\hat{\rho}^2$ were slightly better than the others (see Table 3.9 and Figure 3.6). Moreover, we found in general that $\widehat{T}_{BinErrAk}$ had the highest sensitivity but also a higher FDR than $\widehat{T}_{01}$ (Figure 3.6). This was due to the fact that, although the estimated number of segments was the same, $\widehat{T}_{01}$ could estimate some

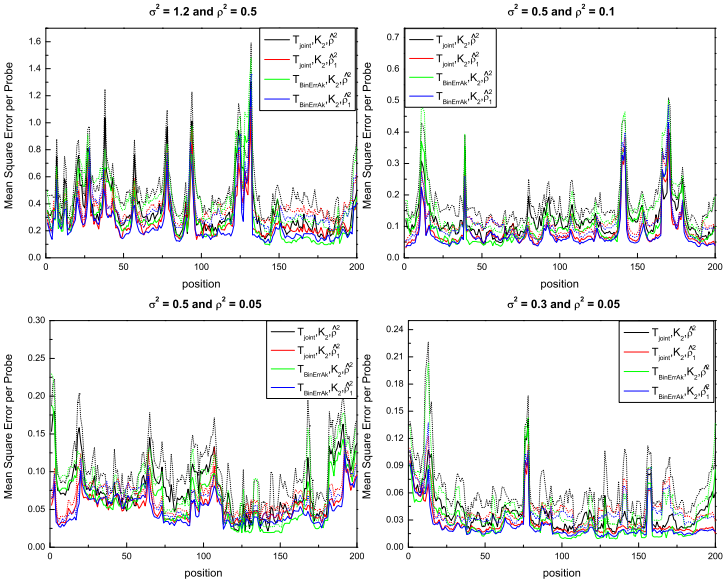

Fig. 3.5 Comparison of the segment level estimation by using $\widehat{T}_{joint}$ or $\widehat{T}_{BinErrAk}$ with the different ρ^2 estimators, on four datasets with replicates. The corresponding true profiles are in Figure 3.2. In general, using $\widehat{T}_{BinErrAk}$ we obtained a lower MSE per probe than using $\widehat{T}_{joint}$. For a fixed boundary estimator, often the error was lower by using $\hat{\rho}_1^2$ on the datasets with $\sigma^2 \gg \rho^2$. [Reprinted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

breakpoints with the same position, reducing the total number of breakpoints and thus reducing the FDR. We can see in Table 3.9 that the false estimated breakpoints did not negatively influence the profile estimation. In fact, the false breakpoints are often used by the algorithm in two ways: either to divide a long segment into two or more segments with close levels or, if it is difficult to determine the position of a breakpoint, to add, before

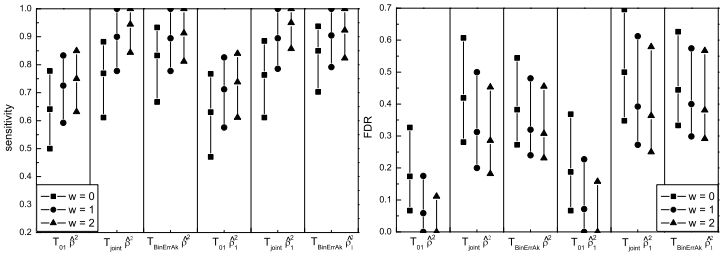

Fig. 3.6 Comparison of the sensitivity and FDR computed on the results obtained on dataset *Simulated Chromosomes* for all estimators of t^0 (apart $\widehat{\mathrm{T}}_{BinErr}$). We always used $\widehat{K}_2$, $\hat{\nu}$, $\hat{\sigma}^2$ and both ρ^2 estimators as estimators of the other parameters involved. Using $\widehat{\rho}^2$ instead of $\widehat{\rho}_1^2$, the FDR was lower. The estimator $\widehat{T}_{BinErrAk}$ had the highest sensitivity and the second lowest FDR. The FDR of $\widehat{T}_{01}$ was the lowest, but this is due to the fact that it reduces the number of the estimated breakpoints by assigning the same position to different breakpoints. [Adapted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

or after the aberration, a segment of one point. Overall, $\widehat{T}_{BinErrAk}$ with $\hat{\rho}^2$ performed best on this dataset.

In conclusion, we suggest to use $\widehat{T}_{BinErrAk}$, even if $\widehat{T}_{joint}$ is also a relatively good estimator at least in some cases. Regarding the estimation of ρ^2, it seems that it is better to use $\widehat{\rho}_1^2$ in presence of high noise.

3.1.7 Comparison with other methods

In this subsection we compare the original and modified versions of BPCR with other existing segmentation methods for genomic copy number estimation: CBS [59], CGHseg [63], GLAD [31], HMM [21], BioHMM [45] and Rendersome [57]. For thoroughness, in the modified versions of BPCR, we used both $\widehat{\rho}^2$ and $\widehat{\rho}_1^2$ as estimators of ρ^2, $\widehat{K}_2$ as estimator of

Table 3.9 Comparison among the error measures for profile estimation, obtained on dataset *Simulated Chromosomes*, for all estimators of t^0 (apart $\widehat{\mathrm{T}}_{BinErr}$). We always used $\widehat{K}_2$, $\hat{v}$, $\hat{\sigma}^2$ and both ρ^2 estimators as estimators of the other parameters involved. The estimator $\widehat{T}_{BinErrAk}$ always had the lowest SSQ and MAD errors and the highest accuracy both inside and outside the regions of aberration. Using $\hat{\rho}^2$ the performance was slightly better, because $\sigma^2 < \rho^2$ in this dataset. [Adapted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

ρ^2 estimator	$\widehat{\mathrm{T}}$ estimator	SSQ	MAD	accuracy	accuracy inside aberrations	accuracy outside aberrations
$\hat{\rho}^2$	$\widehat{\mathrm{T}}_{01}$	14.23	0.00877	0.889	0.961	0.883
$\hat{\rho}^2$	$\widehat{\mathrm{T}}_{joint}$	2.22	0.00840	0.904	0.992	0.892
$\hat{\rho}^2$	$\widehat{\mathrm{T}}_{BinErrAk}$	1.70	0.00733	0.936	0.992	0.932
$\hat{\rho}_1^2$	$\widehat{\mathrm{T}}_{01}$	9.74	0.00952	0.881	0.960	0.877
$\hat{\rho}_1^2$	$\widehat{\mathrm{T}}_{joint}$	2.67	0.00970	0.882	0.993	0.867
$\hat{\rho}_1^2$	$\widehat{\mathrm{T}}_{BinErrAk}$	1.85	0.00781	0.929	0.993	0.920

k_0 and both $\widehat{T}_{joint}$ and $\widehat{T}_{BinErrAk}$ as estimators of the boundaries. We used $\widehat{\rho}^2$ when the noise was low ($\sigma^2 < \rho^2$) and otherwise $\widehat{\rho}_1^2$.

To assess the performance of the several methods, we used three types of artificial datasets. The first type consisted of four datasets with replicates used in the comparison among the estimators. This collection of datasets is called *Cases*.

The second type consisted of datasets adapted from the datasets used in [41] to compare several methods for copy number estimation. In these datasets, each sample was an artificial chromosome of 100 probes, where the copy number value was zero apart from the central part where there was an aberration. The authors considered several widths of aberration: 40, 20, 10 and 5 probes. The noise was always distributed as $\mathcal{N}(0, 0.25^2)$, while the signal to noise ratio (SNR) was 4, 3, 2 or 1. The SNR was defined as the ratio between the height of the aberration and the standard deviation of the noise. The data of the paper consisted of datasets of 100 samples for each combination of width and SNR.

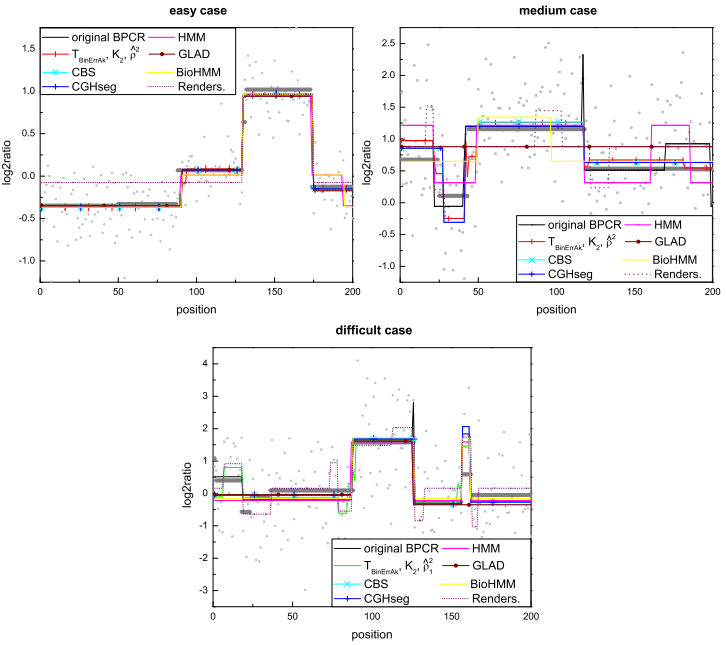

Fig. 3.7 Estimated profiles of the data shown in Figure 3.1, obtained by applying several piecewise constant methods. In each plot, the grey segments represent the true profile and the dots are the raw data points. [Adapted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

We defined our datasets in the following way. For a fixed SNR value, we constructed a chromosome with four aberrations of width of 40, 20, 10 and 5 probes, respectively, by joining the corresponding four types of chromosome of the previous datasets. This collection of datasets is called *Four aberrations*. In the following, we will consider only the datasets with SNR=3 (medium noise) and SNR=1 (high noise).

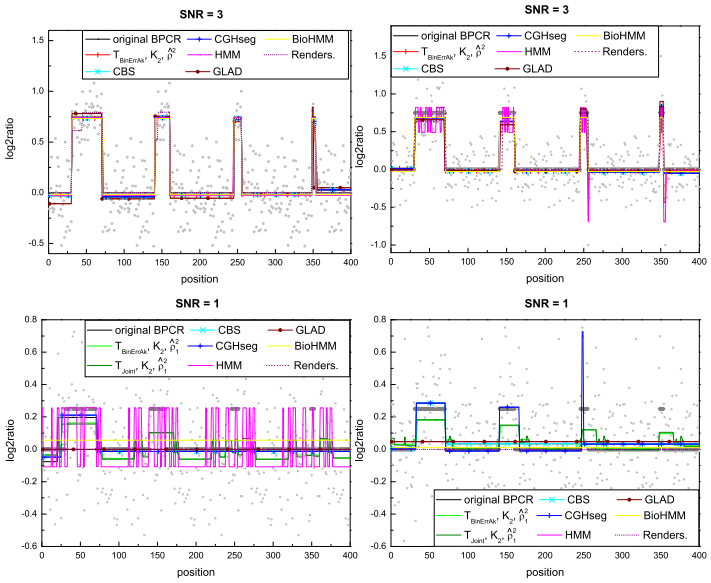

Fig. 3.8 The plots show the differences in the level estimation among the piecewise constant methods on samples with SNR = 3 and SNR = 1: some algorithms are unable to identify the small aberrations in presence of high noise. In each graph, the grey segments represent the true profile. [Reprinted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

The third type of dataset used was the *Simulated Chromosomes* dataset (described in Subsection 3.1.6).

Figure 3.7 shows the estimated profiles with these segmentation methods of three examples of data in collection *Cases* (the true profile are in Figure 3.1). Figure 3.8 displays examples of estimated profiles of data in collection *Four aberrations*.

Error measures used in the comparison

We used two different measures to examine the behavior of the different methods on the collections *Cases* and *Four aberrations*: the root mean square error (for both) and the ROC curve (only for the latter). Each point of the ROC curve has as coordinates the false positive rate (FPR) and the true positive rate (TPR) for a certain threshold. The TPR is defined as the fraction of probes in the true aberrations whose estimated value is above the threshold considered, while the FPR consists in the fraction of probes outside the true aberrations whose estimated value is above the threshold. Hence, the ROC curve measures the accuracy of the method in the detection of the true aberrations.

Instead, the evaluation of the several methods on dataset *Simulated Chromosomes* was accomplished using the error measures described in [84] already used in the study of the boundary estimators (see Subsection 3.1.6).

Results

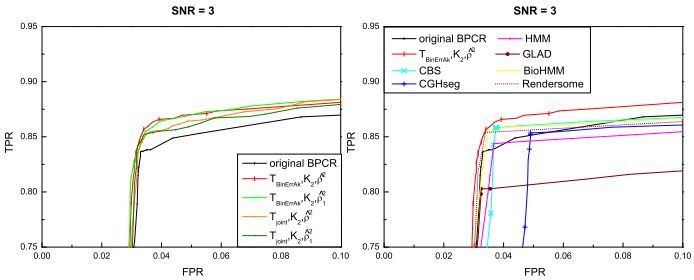

Fig. 3.9 Zoomed ROC curves of several piecewise constant methods applied to dataset with SNR = 3. On this easy type of data, all the methods (apart from GLAD) performed well, since their ROC curves were close to the top left corner of the plot. [Adapted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

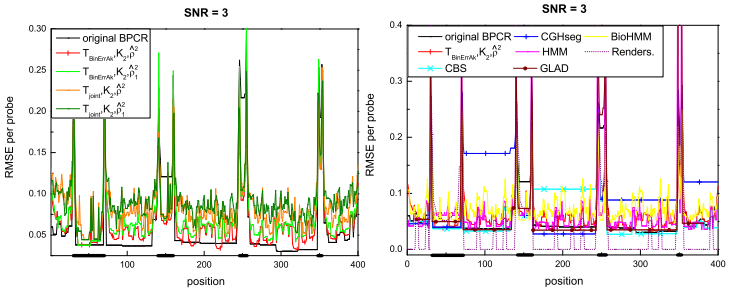

Fig. 3.10 RMSE of several piecewise constant methods applied to dataset with SNR = 3. The black segments on the horizontal axis correspond to the regions of aberration. On this dataset, HMM and the modified BPCR with $\hat{\rho}^2$, $\widehat{K}_2$ and $\widehat{T}_{BinErrAk}$ achieved a low RMSE both inside and outside the aberrations, compared to the other algorithms. Nevertheless, HMM had the highest error at the breakpoint positions. Hence, the modified BPCR with $\hat{\rho}^2$, $\widehat{K}_2$ and $\widehat{T}_{BinErrAk}$ performed better than all other methods with respect to the RMSE per probe measure. [Adapted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

In general, in presence of medium noise, the GLAD method performed worst, since it had a high error in the level estimation of the small peaks, while, for high noise, often both GLAD and Rendersome failed to detect the aberrations (Figures 3.11 and 3.12). The CGHseg method did not usually exhibit an appropriate level estimation except sometimes for segments of large width (for example in Figure 3.10). This is due to the fact that CGHseg estimates the level of a segment with the arithmetic mean of the data points in the segment and this estimator performs poorly if the segment contains few data points and the breakpoint estimation is not accurate. The CBS method, in general, performed quite well, but it was unable to detect aberrations of small width, especially when the noise was high (Figure 3.12).

On the collection *Cases* and the dataset of *Four aberrations* with SNR=3, the RMSE plots and the ROC curves of the HMM method showed

that it generally estimated the profile well, but sometimes it exhibited high errors near breakpoint positions (see, for example, Figure 3.10), likely because it was unable to determine the true position of the breakpoints precisely. Moreover, on the dataset with SNR=1, we recognized the true issues of the estimation with HMM. The RMSE plot showed that it had a high error outside the regions of the aberrations, while inside these regions the error was always more or less the same. Hence, it often failed also in the estimation of the largest aberration, the easiest one to detect (see the corresponding errors in Figure 3.12). The reason of this behavior of the RMSE is the following. The method estimated the true profile either with only one segment, or more often with a profile consisting of a lot of small segments, but all with the same level. Since in the latter case the estimated levels were close to that one of the true aberrations, the RMSE was low in the regions of the aberrations but high outside them. However, the estimated profiles were not similar to the true one. In presence of medium noise (SNR=3), the method BioHMM was more precise than HMM in the determination of the breakpoint positions and in the level estimation (Figure 3.10), while for high noise it behaved similarly to HMM (Figure 3.12).

On collection *Cases*, sometimes the original version of BPCR had a lower error than the modified versions, mostly in regions corresponding to large segments. Nevertheless, the version of BPCR with $\widehat{T}_{BinErrAk}$ and $\widehat{\rho}_1^2$ seemed to be globally the best performing method on these datasets.

Zooming the ROC curve of dataset with SNR=3 (see Figure 3.9) and comparing the original and the modified versions of BPCR, the versions which used the boundary estimator $\widehat{T}_{BinErrAk}$ were the best ones and they were almost indistinguishable. They were also the best performing methods in comparison with the other segmentation methods. We can understand better the different behavior of the methods by looking at the RMSE in Figure 3.10. Comparing the different versions of BPCR, the methods which used the estimator $\widehat{T}_{joint}$ had higher error in the regions outside the aberrations. In all the cases, using $\hat{\rho}_1$ we obtained a higher error than using $\hat{\rho}$. The original version of BPCR was very good at detecting the wider aberration, but it was the worst at detecting the other aberrations. It was also good in estimating the levels in the regions outside the aberrations.

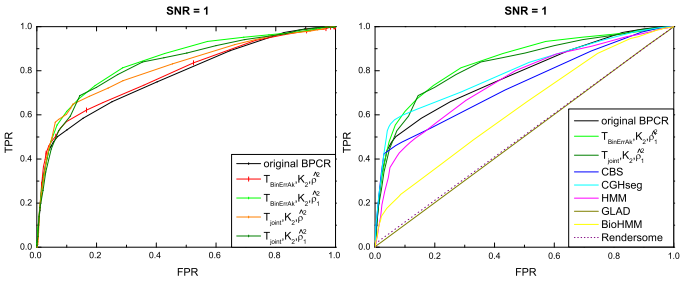

Fig. 3.11 ROC curves of several piecewise constant methods applied to dataset with SNR = 1. On this type of dataset with high noise, the modified BPCR with $\hat{\rho}_1^2$, $\widehat{K}_2$ and $\widehat{T}_{BinErrAk}$ was the best performing method with respect to the ROC curve measure, because its curve was globally the highest one at the top left corner of the plot. GLAD, Rendersome and BioHMM were the worse methods in detecting the aberrations on this dataset. [Adapted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

This means that the original BPCR is good only in the estimation of long segments. Hence, the version which uses $\widehat{T}_{BinErrAk}$ and $\hat{\rho}$ is preferred.

In conclusion, when $\sigma^2 < \rho^2$ (when $\sigma^2 > \rho^2$), the version of BPCR with $\widehat{T}_{BinErrAk}$ and $\widehat{\rho}^2$ ($\widehat{\rho}_1^2$) generally gave the best estimation compared to the other versions of BPCR and to the other methods (Figures 3.9, 3.10, 3.11 and 3.12). We will call this modified version of BPCR, mBPCR.

We could not choose the "best" performing method only on the dataset with SNR=1, because this case showed the limits of all the methods considered. The problem regarding the modified versions of BPCR was essentially represented by the estimation of the number of segments. The ROC curves (Figure 3.11) of the modified versions of BPCR with $\widehat{\rho}_1^2$ were the closest to the left and the top sides of the box, while the RMSE plot (Figure 3.12) showed that these methods were the best methods in the estimation of the levels inside the aberrations, but not outside them. In general, in case of very high noise, all the modified versions of BPCR well detected the aberrations, but had problems in the estimation of the profile outside

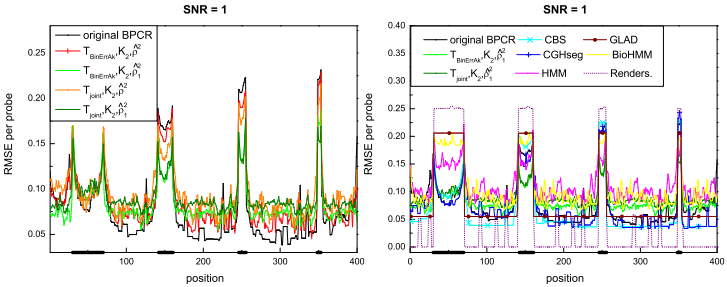

Fig. 3.12 RMSE of several piecewise constant methods applied to dataset with SNR = 1. The black segments on the horizontal axis correspond to the regions of aberration. On this very noisy dataset, the modified BPCR with $\hat{\rho}_1^2$, $\widehat{K}_2$ and $\widehat{T}_{BinErrAk}$ always had a low RMSE per probe, even though its error was not the lowest one outside the aberrations and inside the first one. On the contrary, CBS and CGHseg had the lowest error in these regions, but the highest error inside the small aberrations. Hence, globally the modified BPCR with $\hat{\rho}_1^2$, $\widehat{K}_2$ and $\widehat{T}_{BinErrAk}$ performed better than the other algorithms on this dataset, with respect to the RMSE measure. [Adapted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

them because of the poor estimation of the number of segments. In fact, $\widehat{K}_2$ tends to overestimate the number of segments and this problem worsens using $\widehat{\rho}_1^2$. In conclusion, in a situation with very high noise, using $\widehat{\rho}_1^2$ the BPCR methods detect better the small segments, but, at the same time, the large ones are divided in small segments. On the other hand, using $\widehat{\rho}^2$, smaller segments are not detected and are joined to the closest large segment.

Finally, the comparison performed on dataset *Simulated Chromosomes* showed that CBS and mBPCR better estimated the profiles (see Table 3.10). Regarding the breakpoint error measures (see Figure 3.13), we found that mBPCR had the highest sensitivity (hence, it was the best method in determining the exact position of the breakpoints), but also a higher FDR than CBS. We have already explained in the previous subsec-

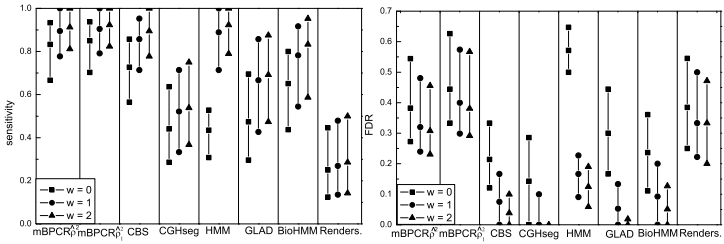

Fig. 3.13 Comparison of the sensitivity and FDR computed on the results obtained on dataset *Simulated Chromosomes*, using several piecewise constant methods. The mBPCR method (considering both ρ^2 estimators) had the highest sensitivity, hence it was the method that determined the breakpoints location with highest precision. Nevertheless, it had a higher FDR than CBS. [Adapted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

tion the possible reason of the high FDR of mBPCR and we can observe again that this fact did not influence negatively the profile estimation (see the SSQ error in Table 3.10). The GLAD method showed a low sensitivity and low FDR, apart from the case regarding the exact position of the breakpoints (w = 0). This means that it underestimated the segment number and the estimated breakpoints were not located exactly at their true positions. Also CGHseg underestimated the number of segments because of low sensitivity and FDR, while HMM had low sensitivity and high FDR when w = 0 and vice versa in the other cases, which means that it often detected the true segment number, but it was unable to put the breakpoints at their exact position. Instead, BioHMM solved the issue of HMM with w = 0, but had an overall lower sensitivity than HMM. Rendersome missed several true aberrations (lowest sensitivity) and detected some false aberrations (medium FDR).

Table 3.10 Comparison of the level estimations obtained by using several piecewise constant methods on dataset *Simulated Chromosomes*. In this comparison, the methods CBS and mBPCR exhibited the lowest SSQ error in the profile estimation and the highest accuracy inside the aberrated regions. On the other hand, HMM, BioHMM and Rendersome had the highest accuracy outside the aberrations, but a high SSQ error. Therefore, the former group of algorithms globally estimated a better profile than the latter. Because of its definition, the MAD error is less informative: it does not take into account if a small number of probes are wrongly estimated, but these probes could correspond to breakpoints or small aberrations. [Adapted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

method	SSQ	MAD	accuracy	accuracy inside aberrations	accuracy outside aberrations
mBPCR $\hat{\rho}^2$	1.70	0.00733	0.936	0.992	0.932
mBPCR $\hat{\rho}_1^2$	1.85	0.00781	0.929	0.993	0.920
CBS	1.56	0.00705	0.953	0.985	0.950
CGHseg	5.42	0.00795	0.925	0.885	0.956
HMM	4.47	0.00350	0.993	0.968	0.997
GLAD	4.15	0.00846	0.939	0.930	0.952
BioHMM	5.69	0.003647	0.990	0.949	0.999
Rendersome	19.13	0	0.920	0.289	1

3.1.8 Definition of the mBPCR algorithm

In Subsections 3.1.3, 3.1.4 and 3.1.5, we introduced new estimators for the parameters involved in BPCR. In Subsection 3.1.6 we selected the best performing ones on the basis of the empirical results obtained on artificial datasets. In particular, we found that the best way is to estimate the segment number with $\widehat{K}_2$ and the boundaries with $\widehat{T}_{BinErrAk}$ (or possibly $\widehat{T}_{joint}$).

Concerning the estimation of the variance of the segment levels, we found that the original estimator $\hat{\rho}^2$ overestimates ρ^2 (variance of the segment levels) by an addendum proportional to σ^2 (variance of the noise), see Equation (3.34). Hence, the estimation fails when $\sigma^2 > \rho^2$. The new

estimator $\hat{\rho}_1^2$ is more precise but slightly underestimates ρ^2, leading to an overestimation of the segment number. Applying both estimators on artificial datasets, we found that, in general, the best way is to use $\hat{\rho}^2$ when $\sigma^2 < \rho^2$ (low noise), but to use $\hat{\rho}_1^2$ when $\sigma^2 > \rho^2$ (high noise), even if it could lead to a slight overfitting. On real DNA copy number data, commonly $\sigma^2 > \rho^2$.

In Subsection 3.1.7, we compared the new versions of BPCR with other methods which also estimate the copy number as a piecewise constant function: CBS [59], HMM [21], CGHseg [63], GLAD [31], BioHMM [45] and Rendersome [57]. As a whole, the results showed that the version of BPCR which uses $\widehat{T}_{BinErrAk}$ gave the best estimation on the dataset used. However, when $\sigma^2 \gg \rho^2$ it is hard to understand which method is the most appropriate. Most of the other methods were not able to detect aberrations with a small width (5 and 10 probes) and the same was true for the version of BPCR which uses $\widehat{T}_{BinErrAk}$ and $\hat{\rho}^2$. On the other hand, the use of $\hat{\rho}_1^2$ led to the detection of the smaller segments, but the larger ones were often divided in small segments and sometimes the segments consisted of only one point.

Therefore, we define mBPCR as the BPCR version which uses $\widehat{K}_2$ and $\widehat{T}_{BinErrAk}$. As a general rule, we still suggest to use $\hat{\rho}^2$ when $\sigma^2 < \rho^2$ and to use $\hat{\rho}_1^2$ when $\sigma^2 > \rho^2$. However, especially on real data, it is preferable to estimate ρ^2 either on a whole dataset or on a "typical sample", which shows a sufficient number of segments (i.e. level changes). In the context of copy number estimation, such a "typical sample" could be a cell line with many copy number aberrations.

3.1.9 Application to real data

In this subsection, we show how mBPCR performed compared to other piecewise constant estimation methods on real data. We used samples from three mantle cell lymphoma cell lines (JEKO-1, GRANTA-519, REC-1) previously analyzed by us with the Affymetrix GeneChip Mapping 10K Array (Affymetrix, Santa Clara, CA), [69]. We also used the data obtained

on JEKO-1, by using the higher density Affymetrix GeneChip Mapping 250K Nsp Array (unpublished). We considered eight recurrent gene regions of aberration in lymphoma plus other two gene regions (*BIRC3* and *LAMP1*) and we compared the corresponding copy numbers obtained by the several piecewise constant methods with those obtained by the FISH technique in [69]. Lastly, we show a comparison among the estimated profiles of chromosome 11 of JEKO-1.

Gene copy number estimation

The knowledge of the true underling profile is required to properly evaluate the methods. In general, large aberrations on chromosomes can be detected with conventional karyotype analysis or with FISH and one could use this information for the evaluation procedure, but the width of these aberrations is so large that all the methods can detect them well, leading to a useless comparison. For this reason, we decided to take into account only genes to compare the piecewise constant methods.

In the comparison, as previously published [69], when two FISH copy numbers had been assigned to one gene, the first number should correspond to the copy number detected in the majority of the cells. We assigned two estimated copy numbers to one gene, when the position of the gene is between two SNPs and the method assigned two different values to these SNPs.

The results on REC-1 (Table 3.11) did not show any significant difference among the methods, instead those on GRANTA-519 (Table 3.12) showed that GLAD was unable to detect the true copy number in five cases, while HMM, BioHMM and Rendersome detected an amplification on *MALT1* greater than what detected by FISH analysis. All methods did not detect the true copy number of *ATM*, probably because the SNPs around *ATM* are far away from the corresponding FISH region (about 1Mb) and the deletion affects only this region. Only mBPCR with $\hat{\rho}_1^2$ and HMM detected a breakpoint between the two SNPs around *ATM* region, indicating a copy number change.

Table 3.11 Copy number estimation results obtained on sample REC-1. Globally, all methods behaved equally well on this data. Only Rendersome was unable to detect the correct copy number of *D13S319*. [Reprinted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

gene region	FISH CN	mBPCR $\hat{\rho}^2$	mBPCR $\hat{\rho}_1^2$	CBS	CGHseg	HMM	GLAD	BioHMM	Rendersome
BCL6	2/3	2.89	2.85	2.89	2.90	2.05	2.79	2.85	2.86
C-MYC	2	2.02	1.99	2.06	2.12	2.05	2.07	2.02	2.02
CCND1	2	2.01	1.92	2.01	2.05	2.05	2.07	2.02	2.02
BIRC3	2/3	2.01	2.22	2.01	2.05	2.05	2.07	2.02	2.02
ATM	2	2.01	1.80	2.01	2.05	2.05	2.07	2.02	2.02
D13S319	2	2.03	2.40	1.98	2.03	2.05	2.01	2.02	*2.89*
LAMP1	2	1.82	1.76	1.98	1.98	2.05	2.01	2.02	2.02
TP53	1/2	1.11	1.17	1.11	1.11	1.10	1.55	1.10	1.20
MALT1	2/3	2.12	2.25	2.02	2.12	2.05	2.09	2.02	2.02
BCL2	2	2.12	2.25	2.02	2.12	2.05	2.09	2.02	2.02

Table 3.12 Copy number estimation results obtained on sample GRANTA-519. On this data, the GLAD method often did not detect the correct gene copy number. The method mBPCR with $\hat{\rho}_1^2$ estimated the gene copy number always well, apart from *ATM* whose copy number was estimated different from the FISH copy number by all methods. [Reprinted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

gene region	FISH CN	mBPCR $\hat{\rho}^2$	mBPCR $\hat{\rho}_1^2$	CBS	CGHseg	HMM	GLAD	BioHMM	Rend.
BCL6	2	2.11	2.10	2.11	2.07	2.11	1.85	2.12	2.04
C-MYC	2	2.07	2.00	2.08	2.11	1.99	*6.22/1.37*	2.12	2.04
CCND1	2	2.06	2.03	2.06	2.34	2.20	2.4	2.12	2.04
BIRC3	2/3	2.06	1.76	2.06	*1.14/2.34*	2.11	2.4	2.12	2.04
ATM	1	*2.06*	*2.01/1.61*	*2.06*	*2.34*	*2.11/1.14*	*2.4*	*2.12*	*2.04*
D13S319	2	2.01	2.00	2.03	2.05	2.07	2.26	2.12	2.04
LAMP1	2	2.01	2.00	2.03	2.05	1.98	1.58	2.12	2.04
TP53	1	1.10	1.16	1.13	1.01	1.13	*1.85*	1.36	1.08
MALT1	3	3.36	3.05	3.17	3.04	*5.30*	*2.16*	*4.78*	*4.28*
BCL2	ampl	5.46	5.10	5.52	6.12	5.30	*2.16*	4.78	7.22

Table 3.13 Copy number estimation results obtained on 250K Array data of sample JEKO-1. All methods behaved equally good. The method HMM had problem in determining the right position of one breakpoint around the *C-MYC* amplification. All methods estimated the copy number of *CCND1* differently from FISH technique. [Reprinted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

gene region	FISH CN	mBPCR $\hat{\rho}^2$	mBPCR $\hat{\rho}_1^2$	CBS	CGHseg	HMM	GLAD	BioHMM	Rendersome
BCL6	3/2	3.06	3.06	3.02	3.06	2.96	2.98	3.04	2.98
C-MYC	ampl	7.12	7.10	7.14	6.87	*6.70/2.63*	7.28	6.51	7.72
CCND1	2	*3.51*	*3.51*	*3.51*	*3.51*	*3.44*	*3.51*	*3.52*	*3.60*
BIRC3	4/5	4.20	4.19	4.20	4.24	4.24	4.23	4.26	3.60
ATM	4	4.20	4.19	4.20	4.24	4.24	4.23	4.26	3.60
D13S319	4	3.72	3.72	3.81	3.82	3.72	3.81	3.73	3.64
LAMP1	4	3.67	3.67	3.82	3.67	3.72	3.69	3.73	3.64
TP53	2/3	2.57	2.69	2.22	2.76	2.34	2.76	2.83	2.90
MALT1	4	3.52	3.52	3.59	3.59	3.50	3.55	3.52	3.50
BCL2	4	3.52	3.52	3.59	3.59	3.50	3.55	3.52	3.50

Regarding the JEKO-1 data, since the cell line is triploid, to obtain more realistic copy number value, we centered the estimated $\log_2$ratio around $\log_2 3$. With the denser 250K Array data, all methods behaved equally good. Only HMM had a problem in the detection of the breakpoint corresponding to the *C-MYC* amplification (see Table 3.13). On both arrays, all methods identified a gain (copy number 3 or 4) at the *CCND1* position, while the copy number detected by FISH was 2. This fact cannot be explained as previously for *ATM*, because this region is well covered by SNPs. Instead, on the JEKO-1 10K Array data (Table 3.14), the noisiest among all samples, we can see several cases in which CBS, HMM and GLAD did not detect correctly the gene copy number (for example, *BCL2* and *MALT1*). This occurred more frequently to BioHMM and Rendersome, while only once to CGHseg (*LAMP1*). The method mBPCR with $\hat{\rho}_1^2$ always estimated gene copy numbers correctly, apart from *CCND1*.

Table 3.14 Copy number estimation results obtained on 10K Array data of sample JEKO-1. On this noisy data, BioHMM and Rendersome often estimated the gene copy number wrongly, while this occurred only sometimes to CBS, HMM and GLAD. The method mBPCR with $\hat{\rho}_1^2$ correctly estimated the gene copy numbers, apart from *CCND1* whose copy number was estimated by all methods differently from the FISH technique. [Adapted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

gene region	FISH CN	mBPCR $\hat{\rho}^2$	mBPCR $\hat{\rho}_1^2$	CBS	CGHseg	HMM	GLAD	BioHMM	Rendersome
BCL6	3/2	2.97	2.99	2.97	2.90	2.92	2.92	3.14	2.92
C-MYC	ampl	12.11	9.35	10.27	10.27	13.95	9.82	8.26	*13.10/3.11*
CCND1	2	*4.08*	*3.77*	*4.08*	*4.08*	*3.84*	*3.79*	*3.14*	*3.50*
BIRC3	4/5	4.08	4.29	4.08	4.08	3.84	3.79	*3.14*	3.50
ATM	4	4.08	4.29	4.08	4.08	3.84	3.79	*3.14*	*3.50/2.39*
D13S319	4	3.72	3.59	3.57	3.72	3.62	3.58	*3.14*	*3.43*
LAMP1	4	*3.41*	3.82	*3.41*	*3.41*	3.62	*2.49*	*3.14*	*3.43*
TP53	2/3	2.81	3.00	2.83	2.50	*3.52*	2.93	3.14	2.93
MALT1	4	3.63	3.62	*3.48*	3.64	*3.42*	*3.42*	*3.14*	*3.42*
BCL2	4	3.63	3.62	*3.48*	3.64	*3.42*	*3.42*	*3.14*	*3.42*

Profile estimation

To compare the profile estimations, we chose the sample JEKO-1 because, using the results obtained on both types of array, we could at least understand which regions were more realistically estimated. Up to now, validated whole chromosome profiles are not available. Among all chromosomes, we chose chromosome 11 since three of the previous genes belong to that: *CCND1* (around 69.17Mb), *BIRC3* (around 101.7Mb) and *ATM* (around 107.6Mb).

From the graphs in Figure 3.14 we can observe that, among all the piecewise constant methods, only mBPCR with $\hat{\rho}_1^2$ was able to detect the high amplification after position 110Mb on the 10K Array data, while it was recognized by all methods (apart from BioHMM) on the 250K Array data. Moreover, on the 10K Array data, almost all methods detected a false deletion around position 3Mb, due to the presence of a sequence of

outliers, and BioHMM did not find any copy number change in the chromosome. On the 250K Array data, HMM and Rendersome had problems in recognizing the last part of the chromosome as a flat region. Moreover, on the 10K Array data, Rendersome estimated several outliers as true aberrations and, on the 250K Array data, it was unable (contrary to all other algorithms) to identify the whole region from about 78Mb to 111Mb as gained.

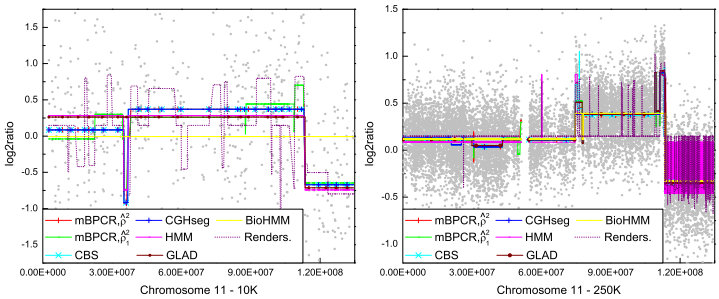

Fig. 3.14 Comparison among the piecewise constant estimated profiles of chromosome 11 of JEKO-1 using both 10K Array and 250K Array data. Only mBPCR with $\hat{\rho}_1^2$ was able to detect the high amplification after position 110Mb on the 10K Array data. On the other hand, all methods (apart from BioHMM) recognized it on the 250K Array data. [Reprinted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

3.2 Estimation with a continuous curve: the mBRC and BRCAk methods

Using the same hypotheses of the "piecewise constant setting" (see Section 2.1), Hutter in [32, 33] derived also a way to estimate the function

with a continuous/smoothing curve instead of a piecewise constant one. In fact, we can estimate the segment level $\widetilde{\mathrm{M}}_s$ at a generic position s, using the fact that it belongs to some segment p and in this segment $\widetilde{\mathrm{M}}_s = \mathrm{M}_p$. Then, summing over all the possible segments, we can compute its posterior distribution in the following way:

$$
\begin{aligned}
&p(\widetilde{\mu}_s \,|\, y, K = k_0) \\
&= \sum_{p=1}^{k_0} \sum_{i=0}^{s-1} \sum_{j=s}^{n} p(\mu_p, T_{p-1} = i, T_p = j \,|\, y, K = k_0) \\
&= \sum_{p=1}^{k_0} \sum_{i=0}^{s-1} \sum_{j=s}^{n} \frac{p(\mu_p, y \,|\, T_{p-1} = i, T_p = j, K = k_0) \mathrm{P}(T_{p-1} = i, T_p = j \,|\, K = k_0))}{p(y \,|\, K = k_0)} \\
&= \sum_{p=1}^{k_0} \sum_{i=0}^{s-1} \sum_{j=s}^{n} \left(\frac{p(y_{0i} \,|\, K_{0i} = p-1) p(\mu_p, y_{ij} \,|\, K_{ij} = 1) p(y_{jn} \,|\, K_{jn} = k_0 - p)}{p(y \,|\, K = k_0)} \right. \\
&\quad \left. \cdot \frac{\binom{i-1}{p-2} \binom{n-j}{k_0-1-p}}{\binom{n-1}{k_0-1}} \right) \\
&= \frac{1}{\binom{n-1}{k_0-1} p(y \,|\, K = k_0)} \sum_{p=1}^{k_0} \sum_{i=0}^{s-1} \binom{i-1}{p-2} p(y_{0i} \,|\, K_{0i} = p-1) \\
&\quad \cdot \sum_{j=s}^{n} p(\mu_p, y_{ij} \,|\, K_{ij} = 1) \binom{n-j}{k_0-1-p} p(y_{jn} \,|\, K_{jn} = k_0 - p). \qquad (3.39)
\end{aligned}
$$

To obtain (3.39), we used Bayes' rule, the independence of data points belonging to different segments and the uniform prior distribution of the boundaries.

Finally, the corresponding Bayesian estimate of $\widetilde{\mathrm{M}}_s$ given $\hat{k}$ is

$$
\begin{aligned}
\widehat{\widetilde{\mu}}_s &= \mathrm{E}[\widetilde{\mathrm{M}}_s \,|\, y, \hat{k}] \qquad (3.40) \\
&= \int_{\mathbb{R}} \widetilde{\mu}_s p(\widetilde{\mu}_s \,|\, y, K = \hat{k})
\end{aligned}
$$

$$= \frac{1}{\binom{n-1}{\hat{k}-1} p(y \mid K = \hat{k})} \sum_{p=1}^{\hat{k}} \sum_{i=0}^{s-1} \binom{i-1}{p-2} p(y_{0i} \mid K_{0i} = p-1)$$
$$\cdot \sum_{j=s}^{n} \binom{n-j}{\hat{k}-1-p} p(y_{jn} \mid K_{jn} = \hat{k}-p) \int_{\mathbb{R}} \mu_p p(y_{ij} \mid \mu_p, K_{ij} = 1) p(\mu_p) d\mu_p \quad (3.41)$$

for all $s = 1, \ldots, n$. The vector $\widehat{\widetilde{\mu}}$ is called *Bayesian Regression Curve* (*BRC*).

In Subsection 3.2.1, we define another type of Bayesian regression curve (called BRCAk) and, in Subsection 3.2.2, we compare the original BRC with BRCAk and BRCs which use a different estimator of k_0. Finally, in Subsection 3.2.3 we evaluate the best performing BRCs in comparison with existing smoothing methods.

3.2.1 Improved regression curve: the BRCAk

As we saw in Section 3.1, there are cases in which the estimation of a parameter of our interest can be made independently of other parameters by integration. The computation of the BRC (see Equations (3.39) and (3.40)) suggests to average also over the number of segments by considering the posterior probability of $\widetilde{\mathrm{M}}_s$, given only the sample point y,

$$\begin{aligned}
\widehat{\widetilde{\mu}}_s &:= \mathrm{E}[\widetilde{\mathrm{M}}_s \mid y] \\
&= \int_{\mathbb{R}} \widetilde{\mu}_s p(\widetilde{\mu}_s \mid y) d\widetilde{\mu}_s \\
&= \sum_{k=1}^{k_{\max}} p(k \mid y) \int_{\mathbb{R}} \widetilde{\mu}_s p(\widetilde{\mu}_s \mid y, k) d\widetilde{\mu}_s \\
&= \sum_{k=1}^{k_{\max}} p(k \mid y) \mathrm{E}[\widetilde{\mathrm{M}}_s \mid y, k] \qquad (3.42)
\end{aligned}$$

Unfortunately, the computation of this quantity requires time $O(n^2 k_{\max}^2)$ (see Section 3.3), hence it could be a problem with samples of big size.

This new type of $\widetilde{\mathrm{M}}_s$ estimation is referred to as *Bayesian Regression Curve Averaging over k* (*BRCAk*).

The same procedure cannot be applied for the estimation of the levels in the piecewise constant regression, because in that case we need to know the partition of the whole interval.

3.2.2 Comparison among the regression curves on simulated data

We compared the estimation of the levels of BRC with the one of BRCAk, also taking into account the influence of the different estimators of the parameters on the final results. To valuate the performance of the methods, we used the root mean square error (RMSE) per probe, computed with respect to the true profile of the levels. For this purpose, we necessarily needed datasets with replicates and we used collection *Cases* (described in Subsection 3.1.7).

Comparing the BRC using the true value of ρ^2 and the three types of segment number estimators, we saw that using $\widehat{K}_1$ and $\widehat{K}_2$ we obtained a similar behavior which was better than using $\widehat{K}_{01}$ (see Figure 3.15). When we compared the BRC using the true value of the segment number and the different estimators of ρ^2 (see Figure 3.16), it was not clear which estimator was better. The same happened also when we made the same comparison using the BRCAk (see Figure 3.17).

Finally, we fixed an estimator of ρ^2 and we made the comparison among the BRC with the different segment number estimators and the BRCAk. Using $\widehat{\rho}_1^2$ (see Figure 3.18), the behavior of the RMSE per probe was more or less the same for all the BRCs and the BRCAk. While using $\widehat{\rho}^2$ (see Figure 3.19), it seemed that it was better to use the BRCAk.

In conclusion, using BRCAk we generally obtained a better or equal result with respect to the BRC. Moreover, we observed that, using BRC, it was better to estimate the number of segments with $\widehat{K}_1$ or $\widehat{K}_2$. We define mBRC the BRC which uses $\widehat{K}_2$ as estimator of the number of the segments.

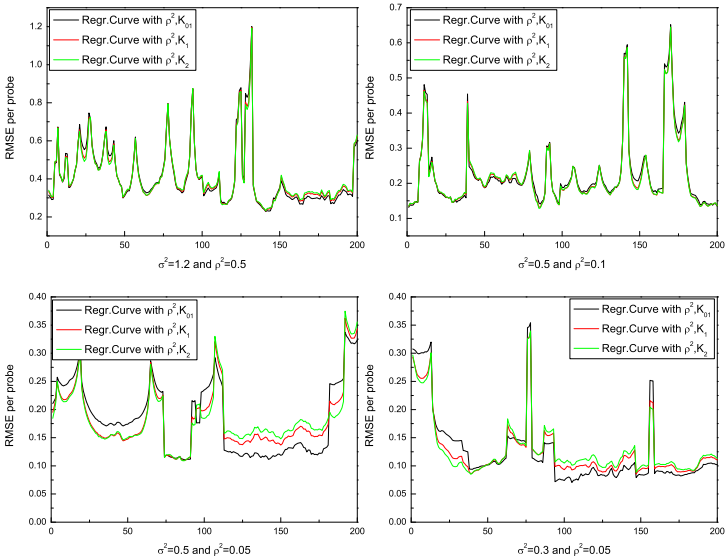

Fig. 3.15 RMSE per probe of the BRCs by using different estimators of k_0. The corresponding true profiles are in Figure 3.2. In general, using $\widehat{K}_1$ or $\widehat{K}_2$ the RMSE per probe error is lower, but when ρ^2 is closer to 0 (graphs at the bottom), sometimes using $\widehat{K}_{01}$ we obtain a better estimate.

Note that we still have to solve the problem to determine which is the best estimator of ρ^2. In most cases, the profile obtained by using $\widehat{\rho}_1^2$ was better than using $\hat{\rho}^2$ (for example, see the plots at the bottom of Figure 3.16). This is due to the fact that sometimes $\widehat{\rho}_1^2$ slightly underestimated ρ^2, leading to overfitting. Still we recommend to use $\widehat{\rho}_1^2$, even if it could lead to a slight overfitting especially in the case of few segments.

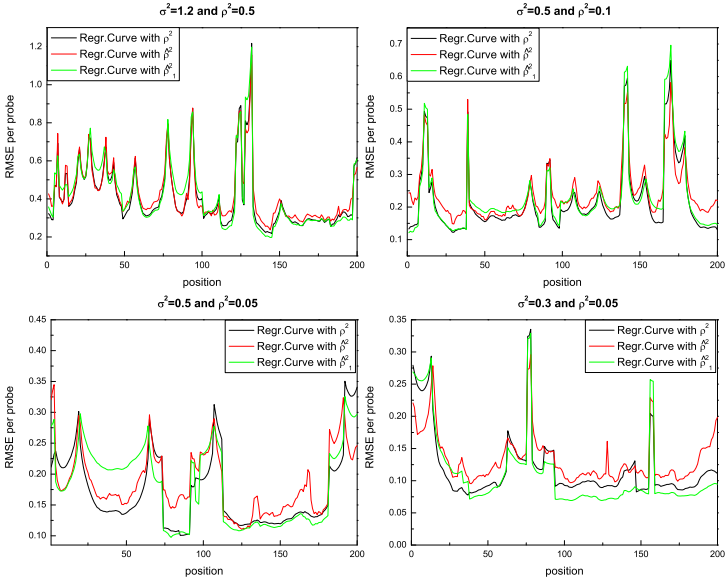

Fig. 3.16 RMSE per probe of the BRCs by using different estimators of ρ^2, on four datasets with replicates. The corresponding true profiles are in Figure 3.2. The graphs do not show clearly which ρ^2 estimator was better with respect to this error measure. Sometimes the error committed using $\hat{\rho}_1^2$ was lower than using $\hat{\rho}^2$, probably because $\hat{\rho}_1^2$ can lead to a slight overfitting. [Reprinted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

3.2.3 Comparison with other smoothing methods

We compared the several versions of the Bayesian regression curves with methods which estimate the copy number as a continuous curve: lowess, wavelet [28], quantreg [18] and smoothseg [29]. Lowess is the acronym of "Locally Weighted Smoothing" (implemented in the stats library of R) and

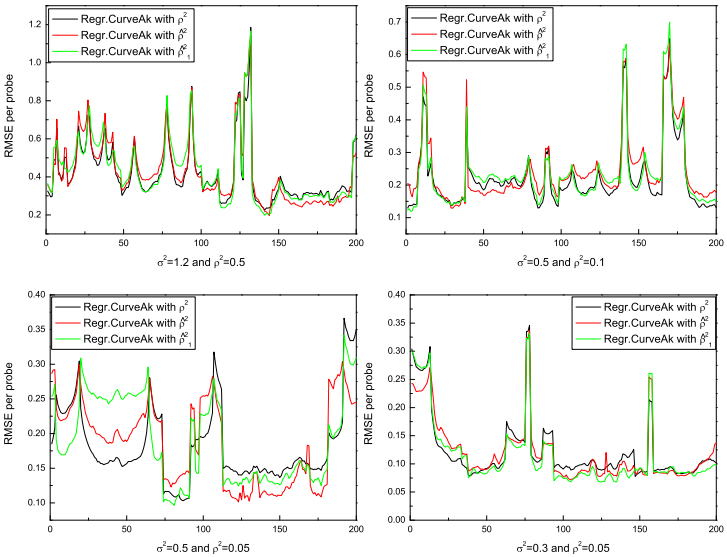

Fig. 3.17 RMSE per probe of BRCAk by using different estimators of ρ^2. The corresponding true profiles are in Figure 3.2. The graphs do not show clearly which ρ^2 estimator is better with respect to this error measure. As in Figure 3.16, sometimes the error committed using the estimated ρ^2 is lower than using the true value of ρ^2, probably because of overfitting.

it is one of the methods considered in the comparison performed in [41]. As we saw previously, both the mBRC and the BRCAk perform well, so we tested both versions with both estimators of ρ^2.

To asses the performance of the methods, we considered collections of artificial datasets already used in the comparison of the piecewise constant methods (see Subsection 3.1.7): *Cases* and *Four aberrations* (datasets SNR = 3 and SNR = 1). As previously, the error measure considered were: the RMSE (for both) and the ROC curve (for the latter). Instead, since

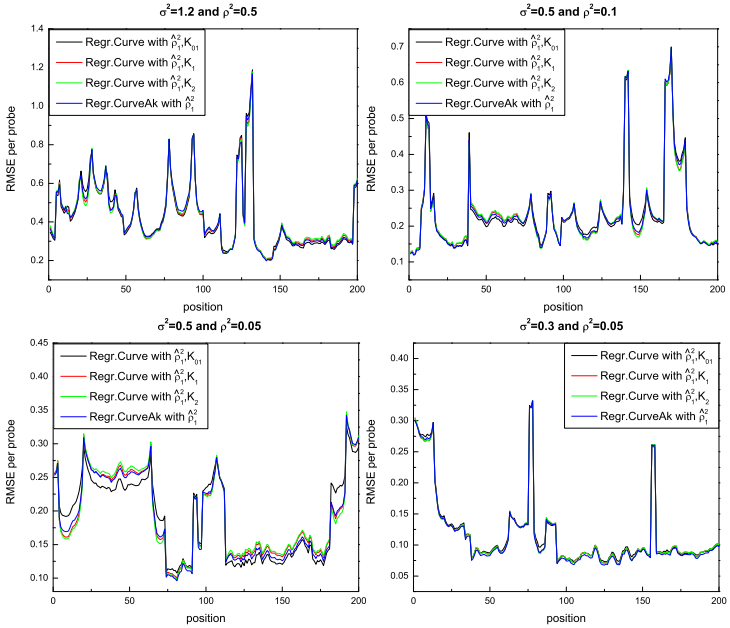

Fig. 3.18 RMSE per probe of the several Bayesian regression curves, using $\hat{\rho}^2_1$ as the estimator of ρ^2, on four datasets with replicates. The corresponding true profiles are in Figure 3.2. Using the estimator $\hat{\rho}^2_1$, all the regression curves gave similar RMSE per probe curve. [Reprinted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

some error measures of [84] suppose that the estimated profile is piecewise constant, we did not apply this group of methods to dataset *Simulated Chromosomes*. Figure 3.20 shows the profiles of three examples of data in collection *Cases* estimated by these smoothing methods (the true profile are in Figure 3.1). Figure 3.21 displays examples of estimated profiles of data in collection *Four aberrations*.

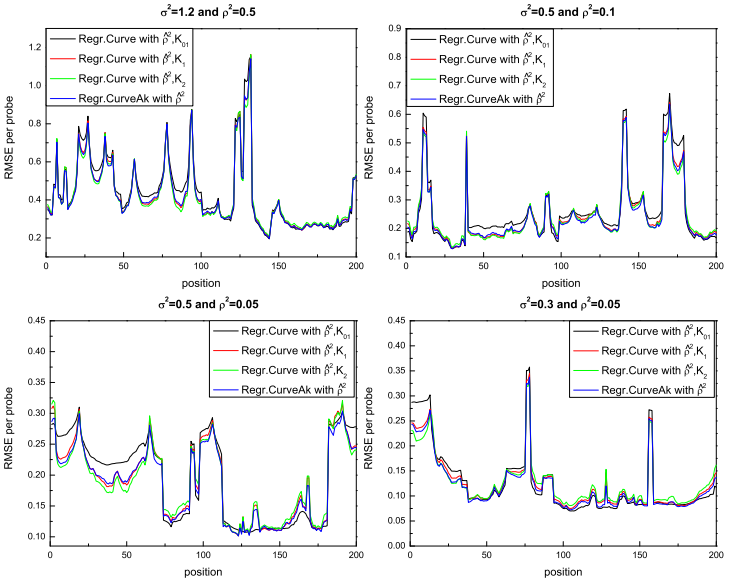

Fig. 3.19 RMSE per probe of the several Bayesian regression curves, using $\hat{\rho}^2$ as the estimator of ρ^2, on four datasets with replicates. The corresponding true profiles are in Figure 3.2. The graphs show that, using $\hat{\rho}^2$, BRCAk always had the lowest RMSE per probe and thus performed better than the other BRCs. [Reprinted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

Results

In general, we found that all methods detected the regions of aberration quite well (see, for example, Figures 3.23 and 3.25). The wavelet method showed a higher error in the level estimation of the aberrations in the datasets SNR = 3 and SNR = 1 (Figures 3.23 and 3.25). The methods lowess and quantreg had the highest RMSE in the collection *Cases*, while

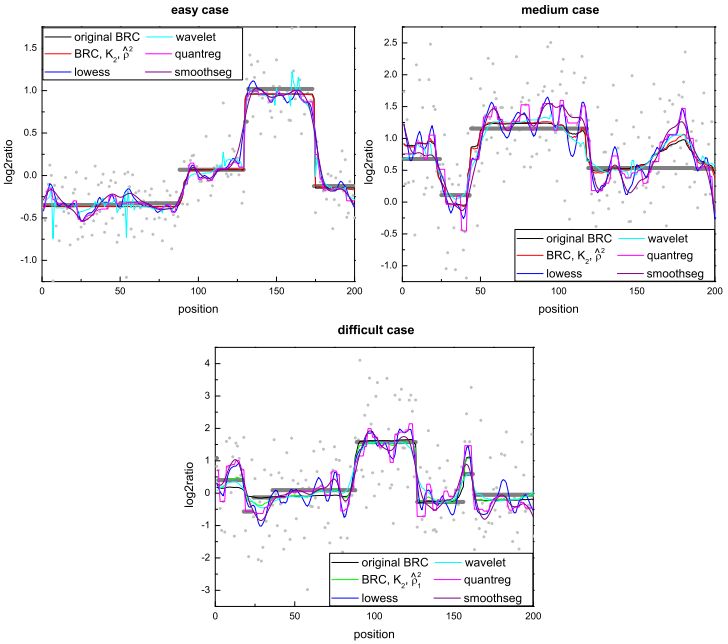

Fig. 3.20 Estimated profiles of the data shown in Figure 3.1, obtained by applying some smoothing methods. In each plot, the grey segments represent the true profile and the dots are the raw data points. [Adapted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

their error was not significantly different outside and inside the aberrations on datasets with SNR = 1, 3. Therefore, in the last cases the error was low inside the aberrations and high outside them in comparison with the other methods. The method smoothseg showed a similar behavior, but with a lower error.

Regarding the BRCs, all of them obtained a quite good estimation when applied to the datasets of collection *Cases*. On dataset SNR = 3, the ROC

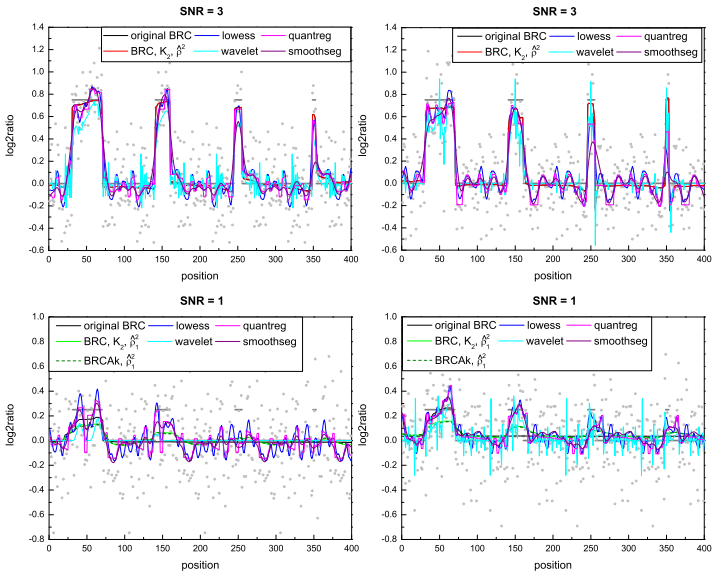

Fig. 3.21 The plots show the differences in the level estimation among the smoothing methods on a samples with SNR = 3 and SNR = 1: some oscillate more in the regions outside the aberrations. In cases of high noise, the more oscillating the profiles are, the harder it is to identify which regions correspond to the aberrations. In each graph, the grey segments represent the true profile. [Reprinted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

curves (see Figure 3.22) of the BRCs which use the estimator $\widehat{\rho}_1^2$ were slightly better than the other ones and, in general, all the modified versions were better than the original one. But the mBRC with $\widehat{\rho}^2$ and the BRCAk with $\widehat{\rho}^2$ obtained the best RMSE. All BRCs gave a similar ROC curve on datasets SNR = 1, the corresponding RMSE (Figure 3.25) shown that the BRCs which use the estimator $\widehat{\rho}_1^2$ had an error more stable than the other ones. In fact, they had also a low error in the aberrations with a small

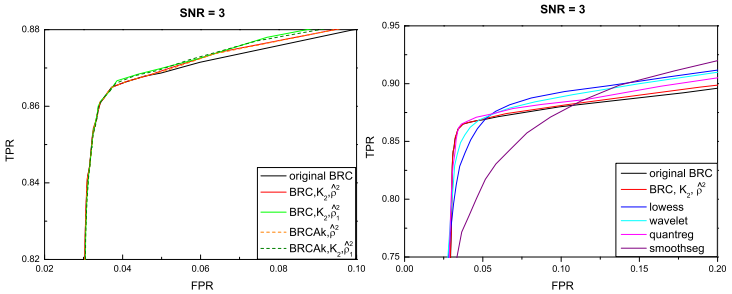

Fig. 3.22 Zoomed ROC curves of several smoothing methods applied to dataset with SNR = 3. The intersection among the ROC curves was due to the differences of the methods in the level estimation outside the aberrations. The more oscillating were the estimated curves in these regions, the closer were the corresponding ROC curves to the top side of the graph. In our case, an oscillating estimated profile is very different from the true one. [Adapted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

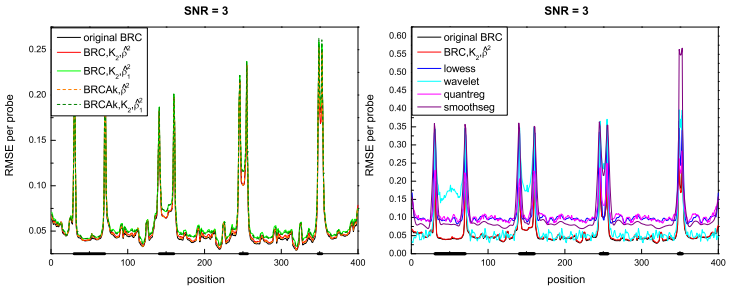

Fig. 3.23 RMSE of several smoothing methods applied to dataset with SNR = 3. The black segments on the horizontal axis correspond to the regions of aberration. On this dataset, both the original BRC and the version of BRC with $\widehat{K}_2$ and $\widehat{\rho}^2$ had everywhere the lowest error. [Adapted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

width. On this data the mBRC with $\widehat{\rho}_1^2$ seemed performing slightly better than the BRCAk with $\widehat{\rho}_1^2$.

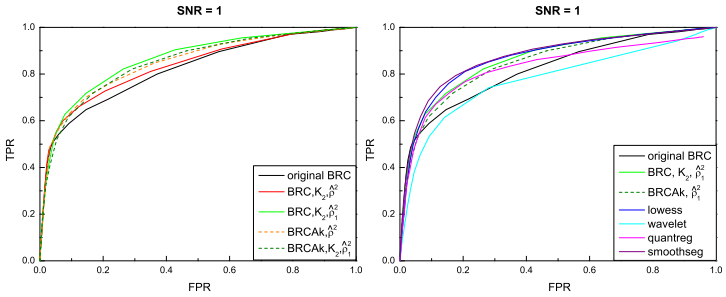

Fig. 3.24 ROC curves of several smoothing methods applied to dataset with SNR = 1. On this very noisy data, the methods smoothseg and lowess seemed to be the best ones, since their ROC curves were the highest at the top left corner of the plot. The third best method was BRC with $\widehat{K}_2$ and $\widehat{\rho}_1^2$. [Adapted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

We also found that the ROC measure was affected by oscillations in the estimated curve, which led to ROC curves intersected and difficult to be interpreted (Figure 3.22). This complex behavior is a consequence of the way in which lowess, wavelet, quantreg and smoothseg yielded oscillating curves with positive and negative values outside the aberrations; while BRCs estimated the true profile with a line almost flat and close to zero (see the examples in Figure 3.21). Hence, when the threshold (used for computing the ROC curve) is negative, the proportion of probes outside the aberrations which are above the threshold (FPR) of the BRCs is greater than the one of the other methods. At the same time, the TPR of wavelet and lowess increases, because the wrongly estimated levels of the probes inside the aberrations are above the threshold.

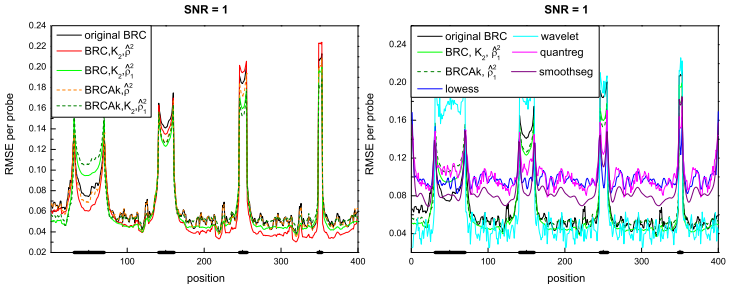

Fig. 3.25 RMSE of several smoothing methods applied to dataset with SNR = 1. The black segments on the horizontal axis correspond to the regions of aberration. The graphs show that the method lowess, quantreg and smoothseg had more or less the same error inside and outside the aberrations. Instead, the BRC version with $\widehat{K}_2$ and $\widehat{\rho}_1^2$ and BRCAk with $\widehat{\rho}_1^2$ had a very low error outside the aberrations and not the highest error inside them, thus globally they performed better than the other algorithms with respect to the RMSE measure. [Adapted from BioMed Central Ltd: *BMC Bioinformatics* [65], copyright (2009), available under Creative Commons Attribution 2.0 Generic]

In conclusion, mBRC and BRCAk gave in general a better estimation than the other BRCs and the other smoothing methods considered. Regarding the ρ^2 estimation, we found that it is better to use $\widehat{\rho}^2$, if $\sigma^2 < \rho^2$, and $\widehat{\rho}_1^2$, if $\sigma^2 > \rho^2$.

3.3 Dynamic programming

As we saw in Sections 3.1 and 3.2, the Bayesian estimation needs the computation of the posterior probabilities of the variables involved, but the complexity of the problem does not allow us to find them analytically. Fortunately the computation can be done by using dynamic programming (a more detailed explanation can be found in [32, 33]).

First of all, we introduce some notations:

$$
\begin{aligned}
\mathrm{M}_{l,m} &= (\mathrm{M}_{l+1},\ \ldots,\ \mathrm{M}_m)\\
T_{l,m} &= (T_l,\ \ldots,\ T_m)\\
Y_{i,j} &= (Y_{i+1},\ \ldots,\ Y_j)\\
K_{i,j} &= \text{number of the segments in } Y_{i,j}.
\end{aligned}
$$

The dynamic program is based on the fact that, for a fixed segment number and partition, the data points that belong to different segments are independent. Then, the joint probability distribution of the data points and the levels in $m-l$ segments can be decomposed in the product of the same joint probability distribution in the first $p-l$ segments and in the last $m-p$ segments, for any p such that $l < p \leq m$:

$$
\begin{aligned}
&p(y_{t_l,t_m},\, \mu_{l,m}\,|\,t_{l,m},\, K_{t_l,t_m} = m-l)\\
&= \prod_{i=l}^{m-1} p(y_{t_i,t_{i+1}}\,|\,\mu_{i+1},t_{i,i+1},K_{t_i,t_{i+1}} = 1)p(\mu_{i+1}\,|\,t_{i,i+1},K_{t_i,t_{i+1}} = 1)\\
&= p(y_{t_l,t_p},\mu_{l,p}\,|\,t_{l,p},K_{t_l,t_p} = p-l)p(y_{t_p,t_m},\mu_{p,m}\,|\,t_{p,m},K_{t_p,t_m} = m-p).
\end{aligned}
$$

We define the following quantity to perform the dynamic program

$$
\mathrm{A}^r_{i,j} := \int_{\mathbb{R}} \mu_p^r p(y_{i,j},\mu_p|\,K_{i,j} = 1)d\mu_p \qquad \text{for all } i < p \leq j,
$$

where $\widetilde{\mu}_{i+1} = \ldots = \widetilde{\mu}_j = \mu_p$, since all the data points considered belong to the same segment, which we call p. One should notice that, if $r = 0$,

$$
\mathrm{A}^0_{i,j} = p(y_{i,j}\,|\,K_{i,j} = 1),
$$

i.e., it is the density of the data $y_{i,j}$ given that they belong to the same segment, while if $r \geq 1$, $\mathrm{A}^r_{i,j}$ is related to the moments of M_p with respect to the conditional probability given y, t and k,

$$
\begin{aligned}
\mathrm{E}[\mathrm{M}_p^r\,|\,y,t,k] &= \mathrm{E}[\mathrm{M}_p^r\,|\,y_{T_{p-1},T_p}, T_{p-1} = i, T_p = j, K_{T_{p-1},T_p} = 1]\\
&= \frac{\int_{\mathbb{R}} \mu_p^r p(y_{i,j},\mu_p|\,K_{i,j} = 1)d\mu_p}{p(y_{T_{p-1},T_p}\,|\,T_{p-1} = i, T_p = j, K_{T_{p-1},T_p} = 1)} = \frac{\mathrm{A}^r_{i,j}}{\mathrm{A}^0_{i,j}} \quad (3.43)
\end{aligned}
$$

where $i = t_{p-1}$ and $j = t_p$, for all $p = 1, \ldots, k$.

The left and right recursion of the dynamic program are made on the following quantities

$$\begin{aligned} \mathrm{L}_{k+1,j} &:= \binom{j-1}{k} p(y_{0,j} | K_{0,j} = k+1) \\ \mathrm{R}_{k+1,i} &:= \binom{n-i-1}{k} p(y_{i,n} | K_{i,n} = k+1). \end{aligned} \tag{3.44}$$

Given the position of t_k, the probability distribution of the data in $k+1$ segments can be decomposed in the product between the joint distribution in the first k segments and the joint distribution in the last segment. Hence, we can compute the probability distribution of the data in $k+1$ segments summing this products for all possible positions of t_k,

$$\begin{aligned} & p(y_{0,j} | K_{0,j} = k+1) \\ &= \sum_{h=k}^{j-1} p(y_{0,j} | T_k = h, K_{0,j} = k+1) \mathrm{P}(T_k = h | K_{0,j} = k+1) \\ &= \sum_{h=k}^{j-1} p(y_{0,j} | T_k = h, K_{0,j} = k+1) \frac{\binom{h-1}{k-1}}{\binom{j-1}{k}} \\ &= \frac{1}{\binom{j-1}{k}} \sum_{h=k}^{j-1} \binom{h-1}{k-1} p(y_{0,h} | K_{0,h} = k) p(y_{h,j} | K_{h,j} = 1) \\ &\Rightarrow \mathrm{L}_{k+1,j} = \sum_{h=k}^{j-1} \mathrm{L}_{k,h} \mathrm{A}^0_{h,j}. \end{aligned} \tag{3.45}$$

Similarly, we can sum, over all possible position of t_1, the products between the joint distribution in the first segment and the joint distribution in the last k segments, obtaining

$$\mathrm{R}_{k+1,i} = \sum_{h=i+1}^{n-k} \mathrm{A}^0_{i,h} \mathrm{R}_{k,h}.$$

The recursion starts with $\mathrm{L}_{0,j} := \delta_{0,j}$ and $\mathrm{R}_{0,i} := \delta_{i,n}$, because there are zero segments only if the two endpoints are equal.

The Bayesian estimators of the original BPCR and BRC

For the definition of the Bayesian estimators, we need to compute the evidence and the posterior distributions of the parameters. Since

$$\mathrm{L}_{k,n} = \mathrm{R}_{k,0} = \tbinom{n-1}{k-1} p(y\,|\,k), \tag{3.46}$$

the *evidence* of the sample point y turns out to be

$$\mathrm{E} := p(y) = \sum_{k=1}^{k_{\max}} p(y\,|\,k)\,p_K(k) = \frac{1}{k_{\max}} \sum_{k=1}^{k_{\max}} \frac{\mathrm{L}_{k,n}}{\binom{n-1}{k-1}}.$$

Then, the posterior probabilities of the number of segments and the boundaries can be written in the following way

$$p(k\,|\,y) = \frac{p(y\,|\,k)p(k)}{p(y)} \tag{3.47}$$

$$= \frac{\mathrm{L}_{k,n}}{\mathrm{E}k_{\max}\binom{n-1}{k-1}} \qquad \forall k \in \mathbb{K} \tag{3.48}$$

$$\begin{aligned}
&\mathrm{P}(T_p = h\,|\,y,k) \\
&= \frac{p(y\,|\,T_p = h, K = k)\mathrm{P}(T_p = h\,|\,K = k)}{p(y\,|\,k)} \\
&= \frac{p(y_{0h}\,|\,K_{0h} = p)p(y_{hn}\,|\,K_{hn} = k-p)}{p(y\,|\,k)} \frac{\binom{h-1}{p-1}\binom{n-h}{k-1-p}}{\binom{n-1}{k-1}} \qquad (3.49) \\
&= \frac{\mathrm{L}_{p,h}\mathrm{R}_{k-p,h}}{\mathrm{L}_{k,n}} \qquad (3.50)
\end{aligned}$$

for each $p = 1, \ldots, k_0 - 1$ and for all $h \in \{p, \ldots, n-1\}$.

Now we can compute the estimators. Equation (3.48) implies that the MAP estimate of the segment number given the sample point y is

$$\widehat{k}_{01} := \arg\max_{k \in \mathbb{K}} p(k \,|\, y) = \arg\max_{k \in \mathbb{K}} \frac{\mathrm{L}_{k,n}}{\binom{n-1}{k-1}}.$$

As we can see in Equations (3.6) and (3.50), the estimate for the boundaries needs the knowledge of the true number of segments, thus we use $\hat{k}_{01}$ instead of k_0,

$$\widehat{t}_p := \arg\max_{h \in \{p,\, \ldots,\, n-(\widehat{k}-p)\}} \mathrm{P}(T_p = h \,|\, y, \hat{k}_{01}) = \arg\max_{h \in \{p,\, \ldots,\, n-(\widehat{k}_{01}-p)\}} \mathrm{L}_{p,h} \mathrm{R}_{\widehat{k}_{01}-p,h},$$

for $p = 1, \ldots, \hat{k} - 1$. The computation of the estimate of the r^{th} moment of the level of the p^{th} segment needs the knowledge of the segment number and the partition of the data (see Equation (3.43)), so we use the estimated ones

$$\widehat{\mu}_p^r := \mathrm{E}[\mathrm{M}_p^r \,|\, y, \widehat{t}, \hat{k}_{01}] = \frac{\mathrm{A}_{i,j}^r}{\mathrm{A}_{i,j}^0},$$

where $i = \widehat{t}_{p-1}$ and $j = \widehat{t}_p$, for all $p = 1, \ldots, \hat{k}_{01}$.

To estimate the segment level $\widetilde{\mathrm{M}}_s$ at a generic position s with BRC, it is sufficient to rewrite Equation (3.41) using the definitions of $\mathrm{A}_{i,j}^0$, $\mathrm{L}_{k+1,j}$ and $\mathrm{R}_{k+1,j}$:

$$\widehat{\tilde{\mu}}_s = \sum_{i=0}^{s-1} \sum_{j=s}^{n} \mathrm{F}_{i,j}^0(\hat{k}_{01}) \qquad \text{with } \mathrm{F}_{i,j}^0(\hat{k}_{01}) := \frac{1}{\mathrm{L}_{\widehat{k}_{01},n}} \sum_{p=1}^{\widehat{k}_{01}} \mathrm{L}_{p-1,i} \mathrm{A}_{i,j}^0 \mathrm{R}_{\widehat{k}_{01}-p,j}, \tag{3.51}$$

for all $s = 1, \ldots, n$.

The boundary estimator $\widehat{T}_{\text{joint}}$

The boundary estimator $\widehat{T}_{\text{joint}}$ is defined as

$$\widehat{T}_{\text{joint}} := \underset{t \in \mathbb{T}_{k_0,n}}{\arg\max}\, p(t \,|\, K = k_0, Y). \tag{3.52}$$

The explicit formula for the joint boundary distribution, given k_0 and the sample point y, is

$$\begin{aligned} p(t \,|\, y, K = k_0) &= \frac{\left[\prod_{p=0}^{k_0-1} p(y_{t_p,t_{p+1}} \,|\, t_p, t_{p+1}, K_{t_p,t_{p+1}} = 1)\right] p(t \,|\, K = k_0)}{p(y \,|\, K = k_0)} \\ &= \frac{\left[\prod_{p=0}^{k_0-1} \mathrm{A}^0_{t_p,t_{p+1}}\right]}{\mathrm{L}_{k_0,n}}, \end{aligned} \tag{3.53}$$

by using the definition of $\mathrm{A}^0_{i,j}$ and Equation (3.46). Thus, the estimated boundaries are

$$\widehat{t}_{\text{joint}} = \underset{t \in \mathbb{T}_{k_0,n}}{\arg\max} \prod_{p=0}^{k_0-1} \mathrm{A}^0_{t_p,t_{p+1}}. \tag{3.54}$$

If we look at Equation (3.54), the computation of $\widehat{T}_{\text{joint}}$ seems to be complex, because it needs to maximize $\prod_{p=0}^{k_0-1} \mathrm{A}^0_{t_p,t_{p+1}}$ for all possible combination of the boundaries. Actually, the computation can be done with a dynamic program in time $O(n k_{\max})$.

Equation (3.54) implies that if we know that there are two segments in $y_{i,j}$, we can estimate the inner boundary with

$$\underset{h \in \{i+2,\ldots,j-1\}}{\arg\max}\; \mathrm{A}^0_{i,h} \mathrm{A}^0_{h,j}.$$

From this observation, we can define the following recursion,

$$\mathrm{W}^1_{0,i} := \mathrm{A}^0_{0,i}$$

$$\mathrm{W}_{0,i}^{p+1} := \max_{h\in\{p,\ldots,i-1\}} \mathrm{W}_{0,h}^{p}\mathrm{A}_{h,i}^{0} \qquad \text{for } p=1,\ldots,\hat{k}-2, \tag{3.55}$$

where $\mathrm{W}_{0,i}^{p}$ represents the maximum probability that $y_{0,i}$ is divided into p segments over all combinations of the boundaries. As a consequence, the components of the vector of the boundary estimator turn out to be,

$$\widehat{t}_k = \operatorname*{arg\,max}_{h\in\{k,\ldots,\widehat{t}_{k+1}-1\}} \mathrm{W}_{0,h}^{k}\mathrm{A}_{h,\widehat{t}_{k+1}}^{0} \qquad \text{for } k=\hat{k}-1,\ldots,1, \tag{3.56}$$

with $\widehat{t}_{\hat{k}} := n$.

The boundary estimators $\widehat{T}_{\,\mathrm{BinErr}}$ and $\widehat{T}_{\,\mathrm{BinErrAk}}$

In Section3.1, we defined the boundary estimators $\widehat{T}_{\,\mathrm{BinErr}}$ and $\widehat{T}_{\,\mathrm{BinErrAk}}$ as the inverse image of $\widehat{\boldsymbol{\tau}}_{\,\mathrm{BinErr}}$ and $\widehat{\boldsymbol{\tau}}_{\,\mathrm{BinErrAk}}$, respectively, under function (3.19). The latter were specified as

$$\begin{aligned}
\widehat{\boldsymbol{\tau}}_{\,\mathrm{BinErr}} &= \mathrm{E}\left[\sum_{i=1}^{n-1}\boldsymbol{\tau}_i\tau_i' \,\middle|\, Y,\hat{k}\right] \\
&= \sum_{\{i=1,\ldots,n-1\,:\,\tau_i'=1\}} \mathrm{P}(\boldsymbol{\tau}_i=1\,|\,Y,\hat{k}) \\
\widehat{\boldsymbol{\tau}}_{\,\mathrm{BinErrAk}} &= \mathrm{E}\left[\sum_{i=1}^{n-1}\boldsymbol{\tau}_i\tau_i' \,\middle|\, Y\right] \\
&= \sum_{\{i=1,\ldots,n-1\,:\,\tau_i'=1\}} \mathrm{P}(\boldsymbol{\tau}_i=1\,|\,Y).
\end{aligned}$$

To compute these estimators, we first find the indices $i_1, \ldots, i_{\widehat{k}-1}$ corresponding to the $\widehat{k}-1$ highest $\{\mathrm{P}(\boldsymbol{\tau}_i=1\,|\,y,\hat{k})\}_{i=1}^{n-1}$ or $\{\mathrm{P}(\boldsymbol{\tau}_i=1\,|\,y)\}_{i=1}^{n-1}$, respectively, and then $\widehat{\tau}_{\,\mathrm{BinErr}}$ or $\widehat{\tau}_{\,\mathrm{BinErrAk}}$, respectively, is the vector such that $\widehat{\tau}_{i_p}=1$ for $p=1,\ldots,\widehat{k}-1$.

The calculation of $\{\mathrm{P}(\boldsymbol{\tau}_i = 1\,|\,y,\hat{k})\}_{i=1}^{n-1}$ and $\{\boldsymbol{\tau}_i = 1\,|\,y)\}_{i=1}^{n-1}$ is derived from Equations (3.24), (3.48) and (3.50) and the definitions of $\mathrm{A}^0_{i,j}$, $\mathrm{L}_{k+1,j}$ and $\mathrm{R}_{k+1,j}$:

$$\begin{aligned}
\mathrm{P}(\boldsymbol{\tau}_i = 1\,|\,y,\hat{k}) &= \sum_{p=1}^{\min(i,\hat{k}-1)} \mathrm{P}(T_p = i\,|\,y,\hat{k}) \\
&= \sum_{p=1}^{\min(i,\hat{k}-1)} \frac{\mathrm{L}_{p,i}\mathrm{R}_{\hat{k}-p,i}}{\mathrm{L}_{\hat{k},n}} \\
\mathrm{P}(\boldsymbol{\tau}_i = 1\,|\,y) &= \sum_{k=2}^{k_{\max}} \mathrm{P}(\boldsymbol{\tau}_i = 1\,|\,y,k)p(k\,|\,y) \\
&= \sum_{k=2}^{k_{\max}} \frac{\mathrm{L}_{p,i}\mathrm{R}_{k-p,i}}{\mathrm{L}_{k,n}} \frac{\mathrm{L}_{k,n}}{\binom{n-1}{k-1} k_{\max}\mathrm{E}}.
\end{aligned}$$

The BRCAk

The computation of the BRCAk is derived similarly to the one of BRC. Equations (3.42) and (3.51) imply that

$$\begin{aligned}
\widehat{\mathrm{M}}_s &= \sum_{k=1}^{k_{\max}} p(k\,|\,y) \sum_{i<t\le j} \mathrm{F}^0_{i,j}(k) \\
&= \sum_{k=1}^{k_{\max}} \frac{1}{Ek_{max}\binom{n-1}{k-1}} \sum_{p=1}^{k} \mathrm{L}_{p-1,i}\mathrm{A}^0_{i,j}\mathrm{R}_{k-p,j}.
\end{aligned}$$

From the last equation, we can notice that the calculation of this quantity required time $O(n^2k^2_{\max})$, hence it might represent a problem with samples of big size.

3.4 Change of the prior distribution of K

In Section 3.1, we often observed that the estimation of a profile with k_0=1 can be problematic. Namely, we explained in Subsection 3.1.1 that, in this case, the variance of the levels ρ^2 should be estimated with zero (since the levels can assume only one value, if we know only the data). Also for this reason, we proposed the estimator $\hat{\rho}_1^2$ in Subsection 3.1.5. Nevertheless, we found that $\widehat{K}_2$ wrongly estimated the number of segments in all samples with k_0=1 of the eight datasets without replicates used in Subsection 3.1.6. The mean value (over the 48 samples) of $\hat{K}_2$ was 15. Therefore, here we propose a modification of the model which avoid this issue.

In order to modify the model, we first study the reason of the failure of the model. As stated before, if k_0=1, then $p(\mu \,|\, \nu, \rho^2) = \delta_{\mu,\nu}$ and the distribution of the data, conditioned only on the hyper-parameters, is

$$\begin{aligned}
\mathrm{A}^0_{i,j} &= p(y_{ij} \,|\, K_{i,j} = 1, \nu, \rho^2, \sigma^2) \\
&= \int_{\mathbb{R}} p(\mu_p \,|\, \nu, \rho^2) \prod_{h=i+1}^{j} p(y_h \,|\, \mu_p, \sigma^2) d\mu_p \\
&= \prod_{h=i+1}^{j} p(y_h \,|\, \nu, \sigma^2) \\
&= \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{j-i}{2}} e^{-\frac{1}{2\sigma^2} \sum_{h=i+1}^{j} (y_h - \nu)^2}.
\end{aligned}$$

Consequently,

$$\mathrm{A}^0_{i,h} \mathrm{A}^0_{h,j} = \mathrm{A}^0_{i,j} \tag{3.57}$$

for all $0 \leq i \leq h \leq j \leq n$.

In this particular case, we can derive an explicit formula of the posterior distribution of K. By using the definition of $\mathrm{L}_{k,n}$ in Equation (3.44), the posterior turns out to be,

$$p(k \,|\, y) = \frac{p(y \,|\, k) p(k)}{\sum_{k'=1}^{k_{\max}} p(y \,|\, k') p(k')}$$

$$= \frac{\mathrm{L}_{k,n}\binom{n-1}{k-1}^{-1} p(k)}{\sum_{k'=1}^{k_{\max}} \mathrm{L}_{k',n}\binom{n-1}{k'-1}^{-1} p(k')}. \tag{3.58}$$

In order to completely calculate the posterior, we derive an explicit formula for $\mathrm{L}_{k,n}$. We want to show by induction that, for all $j = k, \ldots, n$,

$$\mathrm{L}_{k,j} = \binom{j-1}{k-1} \mathrm{A}^0_{0,j}, \qquad k \geq 1. \tag{3.59}$$

For $k = 1$, $\mathrm{L}_{1,j}$=$\mathrm{A}^0_{0,j}$ by definition (see Section 3.3) and the relation in Equation (3.59) is true for this value of k. We assume that Equation (3.59) holds for k, then we have to show that Equation (3.59) holds for $k+1$. By using Equations (3.45) and (3.57), and

$$\sum_{j=k}^{n} \binom{j}{k} = \binom{n+1}{k+1},$$

(from [7]), we obtain

$$\begin{aligned}
\mathrm{L}_{k+1,j} &= \sum_{h=k}^{j-1} \mathrm{L}_{k,h} \mathrm{A}^0_{h,j} \\
&= \sum_{h=k}^{j-1} \binom{h-1}{k-1} \mathrm{A}^0_{0,h} \mathrm{A}^0_{h,j} \\
&= \mathrm{A}^0_{0,j} \sum_{h=k}^{j-1} \binom{h-1}{k-1} \\
&= \mathrm{A}^0_{0,j} \sum_{h=k-1}^{j-2} \binom{h}{k-1} \\
&= \binom{j-1}{k} \mathrm{A}^0_{0,j}.
\end{aligned}$$

Finally, by substituting Equation (3.59) in (3.58), we obtain the explicit formula of the posterior distribution of K,

$$\begin{aligned} p(k\,|\,y) &= \frac{\mathrm{L}_{k,n}\binom{n-1}{k-1}^{-1} p(k)}{\sum_{k'=1}^{k_{\max}} \mathrm{L}_{k',n}\binom{n-1}{k'-1}^{-1} p(k')} \\ &= \frac{\binom{n-1}{k-1}\mathrm{A}^0_{0,n}\binom{n-1}{k-1}^{-1} p(k)}{\sum_{k'=1}^{k_{\max}} \binom{n-1}{k'-1}\mathrm{A}^0_{0,n}\binom{n-1}{k'-1}^{-1} p(k')} \\ &= \frac{p(k)}{\sum_{k'=1}^{k_{\max}} p(k')} = p(k). \end{aligned}$$

i.e. the posterior is equal to the prior. Also the prior and posterior of the boundaries and the levels are equal,

$$\begin{aligned} p(t\,|\,y,k) &= \frac{p(y\,|\,t,k)p(t\,|\,k)}{p(y\,|\,k)} \\ &= \frac{\prod_{p=1}^{k}\mathrm{A}^0_{t_{p-1},t_p}\binom{n-1}{k-1}^{-1}}{\binom{n-1}{k-1}^{-1}\mathrm{L}_{k,n}} \\ &= \frac{\prod_{p=1}^{k}\mathrm{A}^0_{t_{p-1},t_p}\binom{n-1}{k-1}^{-1}}{\binom{n-1}{k-1}^{-1}\binom{n-1}{k-1}\mathrm{A}^0_{0,n}} \\ &= \frac{1}{\binom{n-1}{k-1}} = p(t\,|\,k) \\ p(\mu_p\,|\,y,t,k) &= \frac{p(y\,|\,\mu_p,t,k)p(\mu_p)}{\int_{\mathbb{R}} p(y\,|\,\mu_p,t,k)p(\mu_p)d\mu_p} \\ &= \frac{p(y\,|\,\mu_p,t,k)\delta_{\mu_p,\nu}}{p(y\,|\,\nu,t,k)} \end{aligned}$$

$$= \delta_{\mu_p,\nu} = p(\mu_p), \qquad p = 1,\ldots,k.$$

We applied the definition of $A^0_{i,j}$ and Equations (3.46), (3.59) and (3.57), to derive the equality $p(t\,|\,y,k) = p(t\,|\,k)$. If the data belong to only one segment, they cannot supply deep information about the distribution of the levels and thus the posterior and the prior knowledge about the parameters are equal.

The estimator $\widehat{K}_2$ is defined as the argument which minimizes the posterior squared error and, since K assumes only discrete values, the estimator is equal to the closest integer to the posterior expected value. Therefore, with a uniform prior, $\widehat{K}_2$ is the closest integer to the midpoint of the interval [1,$k_{\max}$], and thus the estimate of k_0 is wrong. On the other hand, the estimator $\widehat{K}_{01}$ represents the argument which maximizes the posterior, but with a uniform prior we do not have a maximum.

The prior of K does not include our prior default assumption that the data consists of only one segment. The segments with levels different from ν are "deviations". To include this knowledge in the prior, we need a distribution which gives the highest probability to $\{K = 1\}$ (and non-zero probability to $\{K = k\}$, $k = 2, \ldots, k_{\max}$). With this type of prior, the correct estimator between $\widehat{K}_{01}$ and $\widehat{K}_2$ is $\widehat{K}_{01}$, because $k = 1$ maximizes the posterior.

We still need to define the form of the prior distribution. Since we previously used a uniform prior and an estimator which minimizes the posterior expected squared error, now we consider a prior similar to $1/k^2$ and an estimator which minimizes the 0-1 error. In order to obtain a "nice" (easy to compute) constant of normalization, we define the new prior as

$$p(k) = \frac{k_{\max}}{k_{\max}+1}\,\frac{1}{k(k+1)}, \qquad k \in \mathbb{K}. \tag{3.60}$$

The new version of mBPCR consists in this modification of both the prior and the estimator of K.

In order to verify empirically the correctness of our modification, we estimated again the profile of the 48 samples with $k_0 = 1$, using the new version of mBPCR and $\hat{\rho}_1$. Now, in 47 samples $\hat{k} = 1$ and only in one

sample $\hat{k} = 2$. In order to verify that the estimation did not worsen in the other cases (i.e. $k_0 \neq 1$), we re-estimated the samples of the collection *Simulated chromosomes*, because these artificial data are similar to real data (although with lower noise). To compare the estimation of the original and the new version of mBPCR, we used the same error measures as in the comparison of the boundary estimators Subsection 3.1.6. As explained in Subsection 3.1.6, on these data we expected to achieve a better estimation by using $\hat{\rho}^2$, since $\sigma^2 < \rho^2$. Therefore, we comment only the results obtained with $\hat{\rho}^2$.

The new version of mBPCR had a lower FDR in the breakpoint estimation and its sensitivity (only regarding $w = 0,1$) was only slightly lower than the corresponding one of the original mBPCR (Figure 3.26). The SSQ error slightly increased, but also the accuracy slightly increased because the accuracy in the detection of normal regions slightly increased (Table 3.15). Overall, the results of the two methods were not significantly different and thus we decided to use the new version of mBPCR.

Table 3.15 Comparison among the error measures for profile estimation, obtained on dataset *Simulated Chromosomes*, for the original and the new version of mBPCR. We used both ρ^2 estimators. The original mBPCR had the lowest SSQ, while the new version the highest accuracy of normal regions. Nevertheless, the results of both methods were overall similar.

method	SSQ	MAD	accuracy	accuracy inside aberrations	accuracy outside aberrations
mBPCR, $\hat{\rho}^2$	1.70	0.00733	0.936	0.992	0.932
mBPCR, $\hat{\rho}^2_1$	1.85	0.00781	0.929	0.993	0.920
new mBPCR, $\hat{\rho}^2$	1.76	0.00729	0.940	0.990	0.936
new mBPCR, $\hat{\rho}^2_1$	1.90	0.00772	0.934	0.991	0.927

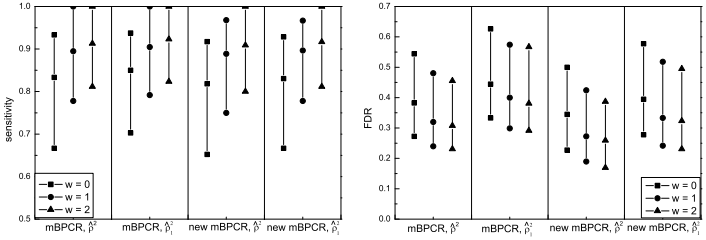

Fig. 3.26 Comparison of the sensitivity and FDR computed on the results obtained on dataset *Simulated Chromosomes* using the original and the new version of mBPCR. The new mBPCR has a lower FDR and, only for $w = 0, 1$, its sensitivity is slightly lower than the one of the original mBPCR.

Chapter 4
Statistical model for the integration of copy number and LOH data

Cancer and several human diseases are caused by genomic aberrations, which can affect the homozygous status and/or the DNA copy number (see Subsection 1.1.4). The former aberrations are often displayed by unusual long stretches of homozygous SNPs, called *loss of heterozygosity* (LOH) regions. The latter consist in genomic regions with DNA copy number different from two.

SNP microarrays are able to measure simultaneously both the DNA copy number and the genotype at each SNP position considered [30]. By integrating both types of data, we can better identify several types of lesions of the genome (regarding combinations of both DNA copy number and LOH aberrations). For example, when one copy of a chromosomal segment is deleted, we usually detect a long stretch of homozygous SNPs (since the microarray is unable to distinguish between the presence of only one copy and the presence of two equal copies), but the same homozygous status can also occur for other reasons, such as uniparental disomy (see Subsection 1.1.4). In this situation, the knowledge of both types of data can lead to the correct interpretation of the phenomenon, while with only one type of data it would not be possible. Another example is when an amplified genomic segment is present: if one of the two copies of the segment is highly amplified, then, even for heterozygous SNPs, all SNPs of

the region will be likely detected as homozygous, because the DNA quantity of one allele is much higher than the other one. In this case again, the integration of both types of data is able to better identify the dosage of the DNA aberration.

Many methods have been developed for the estimation of the copy number profile (see, for example, [21, 31, 59, 63, 65]) and others for the discovery of LOH regions, without distinguishing if they are caused by either the loss of one copy or other genomic events like uniparental disomy or autozygosity (see, for example, [2, 8, 56]). To the best of our knowledge, only one method integrates these two types of data for the estimation of both copy number and LOH aberrations and it uses HMM [72]. Other two methods use both copy number and LOH data to find copy number changes: QuantiSNP [13] and PennCNV [81]. The former employs an Objective Bayes HMM, instead the latter uses HMM and can be applied only to Illumina high-density SNP microarray.

In this chapter, we propose a method which estimates the copy number profile and the stretches of homozygous SNPs at the same time, using both LOH and DNA copy number data. The estimation procedure consists of a Bayesian piecewise constant regression, thus we call our algorithm *genomic Bayesian Piecewise Constant Regression* (gBPCR). Our model is more general than [72], since the latter cannot be applied to data, whose DNA sample come from a mixture of cell populations (which is usually the case for samples of patients affected by cancer). Moreover, the algorithm in [72] needs the specification of some parameters by the user and is sensitive to their values.

Because of the complexity of the biological model, we first describe (in Section 4.1) a preliminary simplified model (called Model 1), which estimates the copy number events exploiting the relationship between copy number and LOH data. Therefore, it does not detect the LOH regions without copy number changes (called *copy-neutral LOH*), which are due to events like uniparental disomy, and does not distinguish the normal regions from the gained one (because we suppose that the capability of detection of the homozygous status is the same in these two types of regions). Subsequently, in Sections 4.2 and 4.3, we add to the model the detection of copy-neutral LOH regions (Model 2) and of gained ones (Model 3). In

Section 4.4, we show how we estimated the parameters of the models and in Section 4.5 how we adjust the values of some of these parameters with respect to the noise of the sample that we want to estimate. In Section 4.6, we use artificial data to compare the estimators of the breakpoints proposed in Section 4.1, and to compare gBPCR with dChip and CNAT 4.01 (described in Section 2.3). Finally, in Section 4.7, we apply gBPCR to real data and in Section 4.8, we show how to modify the dynamic programming used for the computation of mBPCR (Section 3.3), to perform gBPCR.

A preliminary version of the method has been published in [66], while a complete version has been submitted [67].

4.1 Model 1: relationship between LOH and copy number data

Although in nature copy number are integers, the raw copy number detected by the microarray are usually continuous values, due to technical procedures. Also, the samples often contain a percentage of normal cells together with the neoplastic cells, which can contribute.

As we explained in Chapter 2, it is common practise to treat copy number data in a $\log_2$ratio scale (to assume that the errors are normally distributed) and to estimate the copy number profile with a piecewise constant function where the levels assume real values. For the purpose of our model, we estimate this profile by mBPCR (see Chapter 3), that we have shown to outperform well-known other methods on several datasets.

Commonly, in biomedical/cancer research, after estimating the $\log_2$ratio profile, the copy number aberrations are defined as those regions with values outside an interval around zero (we recall that, in the $\log_2$ratio scale, zero represents CN = 2, i.e. a normal copy number). Often, the interval is a statistical confidence interval computed on the basis of the samples of the whole dataset.

In Model 1, our aim is to better classify the copy number changes, trying to reduce the number of false positive, by exploiting the relationship between copy number and LOH data.

4.1.1 Mathematical model of the biology mechanism

The aim of Model 1 is to obtain a better estimation of the true underlying copy number events, using both the information given by copy number and LOH data. In a genomic region, a copy number event is defined as a particular class of copy number values. The definition of the categories, in which the copy number values are divided will follow from the description of the LOH data.

For the purpose of better identifying the copy number events, we can consider two classes of SNP conditions: *Heterozygosity* (Het) and *Homozygosity* (Hom). The microarray is unable to distinguish among a homozygosity due to the presence of two equal nucleotides or the one due to the loss or high amplification of one of them. Hence, the presence of heterozygosity can ensure that the copy number is normal or gained with a high probability, while the homozygosity can be due to different events. It follows that there are only four relevant classes of copy number events, that can be distinguished by looking at the LOH data. Therefore, if we call $\widetilde{Z}_i$ the random variable which represents a copy number event at SNP i, it can assume only the following values:

- $\widetilde{Z}_i = 2$ when $\mathrm{CN} > 4$ ("amplification")
- $\widetilde{Z}_i = 0$ when $1 < \mathrm{CN} \leq 4$ ("normal or gain")
- $\widetilde{Z}_i = -1$ when $\mathrm{CN} = 1$ ("loss")
- $\widetilde{Z}_i = -2$ when $\mathrm{CN} = 0$ ("homozygous deletion").

The homozygous deletion corresponds to the loss of both copies of a genomic region. Ideally, the microarray should detect a "NoCall" genotype at the corresponding SNP position (i.e. it should not be able to identify the genotype of the SNP). Although not common since cancer DNA samples usually contain a mixture of tumor and normal cells, the information given

by the "NoCall" genotype can be useful to better distinguish between a mono-allelic deletion and a bi-allelic (homozygous) deletion.

Therefore, three different LOH variables are present in the model: the true homozygous status in normal cells (X^N), the homozygous status in "cancer" cells (X), which is the consequence of copy number changes (now we do not consider other biological events), and the homozygous status detected by the microarray (Y). The components of the first two random vectors can assume only values in $\mathbb{X} = \{Het, Hom\}$ and $\mathbb{X}^* = \{\emptyset, Het, Hom\}$, respectively, and we suppose that they are independently distributed as Bernoulli random variable. Instead, the components of Y can assume values in $\mathbb{Y} = \{NoCall, Het, NHet\}$ ($NHet$ stands for "not heterozygous", since the microarray cannot distinguish between two equal nucleotides, i.e. homozygosity, and a loss of one copy).

Figure 4.1 shows a summary of the model. Ideally, at each SNP i, the homozygous status in "cancer" cells X_i is completely determined by the corresponding value in normal cells X_i^N and the copy number event occurred $\widetilde{Z}_i$, by the following relations:

$$\begin{aligned}
\mathrm{P}(X_i = x \,|\, X_i^N = x, \widetilde{Z}_i = 2) &= 1, \qquad x \in \mathbb{X}, \\
\mathrm{P}(X_i = x \,|\, X_i^N = x, \widetilde{Z}_i = 0) &= 1, \qquad x \in \mathbb{X}, \\
\mathrm{P}(X_i = Hom \,|\, X_i^N = x, \widetilde{Z}_i = -1) &= 1, \qquad x \in \mathbb{X}, \\
\mathrm{P}(X_i = \emptyset \,|\, X_i^N = x, \widetilde{Z}_i = -2) &= 1, \qquad x \in \mathbb{X}.
\end{aligned}$$

Nevertheless, the homozygous status of "cancer" cells measured by the microarray (Y_i) is affected by several sources of errors (that will be explained in Subsection 4.1.2).

4.1.2 Hypothesis of the model

The genome of cancer cells can be divided in subregions where the copy number is constant. Since we divided the copy number values in four

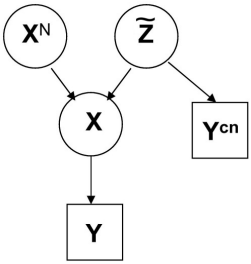

Fig. 4.1 Scheme of Model 1. The vector X of the homozygous status of all SNPs in cancer cells is completely determined, given the vector X^N of their homozygous status in normal cells and the vector $\widetilde{Z}$ of their corresponding copy number events. Using this relationship among X, X^N and $\widetilde{Z}$, we can estimate $\widetilde{Z}$, given the observations Y^{cn} and Y (respectively, the raw $\log_2$ratio of the copy number and the homozygous status in cancer cells detected by the microarray) and by specifying the prior distribution of X^N. The observations Y^{cn} are used to defined the prior distribution of $\widetilde{Z}$ in the Bayesian model.

classes (i.e. the copy number events), we can also consider regions with the same copy number event.

Let us consider a genomic region where the microarray measures the DNA copy number and the genotype at n SNP loci. Then, from the previous discussion, the vector of the copy number events at all positions $\widetilde{Z} = \left(\widetilde{Z}_1, \ldots, \widetilde{Z}_n\right)$ can be seen as a piecewise constant function. This function consists of k_0 intervals with the same copy number event and with boundaries $0 = t_0^0, < t_1^0, \ldots < t_{k_0-1}^0 < t_{k_0}^0 = n$, so that $\widetilde{Z}_{t_{p-1}^0+1} = \ldots = \widetilde{Z}_{t_p^0} = Z_p$ for all $p = 0, \ldots, k_0$. Therefore, to estimate this function we use a Bayesian piecewise constant regression, which determines the number of segments k_0, the boundaries t^0 and the copy number events $Z = \left(Z_1, \ldots, Z_{k_0}\right)$.

For any sample, we assume to have the LOH data given by the microarray (Y) and the mBPCR estimated profile of the $\log_2$ ratio of the copy number. The estimated $\log_2$ratio profile consists of $\hat{k}^{cn}$ intervals with boundaries $\hat{t}^{cn}$=$(0 = \hat{t}_0^{cn}, \hat{t}_1^{cn}, \ldots, \hat{t}_{k_0}^{cn} = n)$ and levels of the segments $\widehat{\mu} \in \mathbb{R}^{\hat{k}^{cn}}$. This estimated profile is used only to define the prior distribution of the random

vector Z (see Subsection 4.1.3), while the genotyping data are used to infer Z. Notice that we do not suppose to know X^N, i.e. the homozygous status in normal cells. Moreover, we assume that, given the true value of the homozygous status in normal cells X^N and the copy number event Z at each position, the LOH data points $\{Y_i\}_{i=1}^n$ are independent, since their values depend only on both noise and genotyping detection errors.

The model implies that, given k_0 and t^0, the posterior distribution of $\widetilde{Z}$ is

$$p(\widetilde{z}\,|\,y,t^0,k_0) \propto \prod_{p=1}^{k_0} \prod_{i=t_{p-1}^0+1}^{t_p^0} \sum_{x\in\mathbb{X}} p(y_i\,|\,X_i^N=x,Z_p=z_p)\mathrm{P}(X_i^N=x)\mathrm{P}(Z_p=z_p),$$

and thus, only conditioning with respect to the LOH data points y, the posterior becomes

$$p(\widetilde{z}\,|\,y) \propto \sum_{k\in\mathbb{K}} \sum_{t\in\mathbb{T}_{k,n}} p(\widetilde{z}\,|\,y,t,k)\mathrm{P}(T=t\,|\,K=k)\mathrm{P}(K=k),$$

where $\mathbb{K}$ and $\mathbb{T}_{k,n}$ are the domains of k and t, respectively (defined as in Subsection 3.1.1).

To specify the model (see Figure 4.1), we need to define the likelihood, i.e. the conditional distribution of Y, given $\widetilde{Z}$ and X^N. To model it, we take into account all the variability that can affect the genotype detection, such as the polymerase chain reaction (PCR) amplification, the presence of different cancer cell subpopulations or normal cells and the amplification of only one copy. For example, the probabilities $\mathrm{P}(Y_i = NHet\,|\,X_i^N = Het,\widetilde{Z}_i=0)$ and $\mathrm{P}(Y_i = Het\,|\,X_i^N = Hom,\widetilde{Z}_i=0)$ are not zero, because of the error in the genotyping detection even in case of a normal DNA sample. The probabilities $\mathrm{P}(Y_i = Het\,|\,X_i^N = Het,\widetilde{Z}_i=-2)$ and $\mathrm{P}(Y_i = NHet\,|\,X_i^N = Hom,\widetilde{Z}_i=-2)$ are related to detection errors due to the presence of normal cells and/or different types of cancer cell subpopulations, or to PCR amplification errors, while $\mathrm{P}(Y_i = NHet\,|\,X_i^N = Het,\widetilde{Z}_i=2)$ is related to errors that can be due to amplification of only one allele. Also

$\mathrm{P}(Y_i = Het \,|\, X_i^N = Het, \widetilde{Z}_i = -1)$ and $\mathrm{P}(Y_i = NHet \,|\, X_i^N = Het, \widetilde{Z}_i = -2)$ account for the errors that can be due to the presence of subpopulations.

To complete the Bayesian model, we need to define the prior distributions of the other random variables. For the parameters K and T, we consider the same distributions of the last version of mBPCR (see Section 3.4):

$$P(K = k) = \frac{k_{\max}+1}{k_{\max}} \frac{1}{k(k+1)}, \qquad k \in \mathbb{K},$$
$$P(T = t \,|\, K = k) = \frac{1}{\binom{n-1}{k-1}}, \qquad t \in \mathbb{T}_{k,n},$$

where $\mathbb{K} = \{1, \ldots, k_{\max}\}$ and $\mathbb{T}_{k,n}$ is a subspace of $\mathbb{N}_0^{k+1}$ such that $t_0 = 0$, $t_k = n$ and $t_q \in \{1, \ldots, n-1\}$ for all $q = 1, \ldots, k-1$, in an ordered way and without repetitions.

The $\{X_i^N\}_{i=1}^n$ are assumed to be independent and Bernoulli distributed, but with a different parameter $q_i := \mathrm{P}(X_i^N = Het)$. This set of parameters $\{q_i\}_{i=1}^n$ does not need to be estimated, because the documentation related to the microarray already provides the probabilities of heterozygosity of all SNPs in the array.

4.1.3 Definition of the prior distribution of Z

The only prior that we have not yet defined is the one of Z. While the estimated levels of the $\log_2$ratio profile are continuous variables, Z classifies the copy number as discrete events. Then, the major problem consists in mapping the continuous values into the discrete values of Z, i.e. in defining a partition of the $\log_2$ratio values such that each interval corresponds to a particular copy number event.

In literature, most methods determine a confidence interval around zero and then consider all the $\log_2$ratio values above this interval "gain" and all values below "loss" (see, for example, [55, 79]). This method is not suit-

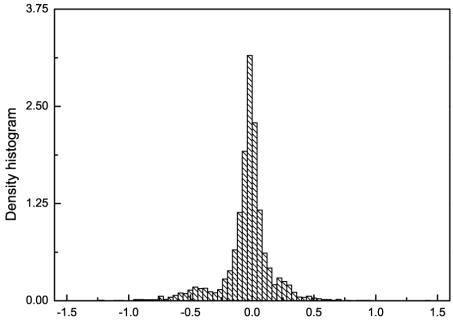

Fig. 4.2 Example of a density histogram of estimated log2ratio levels. The data come from the mBPCR estimated log2ratio levels of the profiles of 14 HIV lymphoma cell lines in [10].

able in our case, since we want to classify also the events $\{\text{CN=0}\}$ and $\{\text{CN>4}\}$. Looking at the density histogram of the estimated $\log_2$ratio values (see, for example, Figure 4.2), we can see that they have a multimodal density with peaks corresponding to $\{\text{CN=1}\}$, $\{\text{CN=2}\}$ and $\{\text{CN=3,4}\}$. Sometimes, as in Figure 4.2, we can separate the peaks of $\{\text{CN=3}\}$ and $\{\text{CN=4}\}$. Similarly to [27], we model this density as a mixture of normal distributions. Once the parameters of the density are estimated, we can define a function to map the $\log_2$ratio values into the copy number event values:

$$f_{LOGtoZ}(x) = \begin{cases} 2 \text{ if } x > \widehat{m}_4 + 3\widehat{s}_4 \\ 0 \text{ if } \widehat{m}_2 - 3\widehat{s}_2 < x \leq \widehat{m}_4 + 3\widehat{s}_4 \\ -1 \text{ if } \widehat{m}_1 - 3\widehat{s}_1 < x \leq \widehat{m}_2 - 3\widehat{s}_2 \\ -2 \text{ if } x < \widehat{m}_1 - 3\widehat{s}_1, \end{cases} \tag{4.1}$$

where $(\widehat{m}_{cn}, \widehat{s}^2_{cn})$ are, respectively, the estimated mean and variance of the normal distribution corresponding to $\{\text{CN=}cn\}$.

From the definition of f_{LOGtoZ}, for all $p = 1, \ldots, \hat{k}_{cn}$, we define the prior distribution of Z_p as:

$$\begin{aligned} \text{P}(Z_p = 2) &= \text{P}\left(M_p \geq \widehat{m}_4 + 3\widehat{s}_4 \,\middle|\, cn\right) \\ \text{P}(Z_p = 0) &= \text{P}\left(\widehat{m}_2 - 3\widehat{s}_2 < M_p \leq \widehat{m}_4 + 3\widehat{s}_4 \,\middle|\, cn\right) \\ \text{P}(Z_p = -1) &= \text{P}\left(\widehat{m}_1 - 3\widehat{s}_1 < M_p \leq \widehat{m}_2 - 3\widehat{s}_2 \,\middle|\, cn\right) \end{aligned}$$

$$\mathrm{P}(Z_p = -2) = \mathrm{P}\left(M_p \leq \widehat{m}_1 - 3\widehat{s}_1 \,\middle|\, cn\right)$$

where we *cn* represents all copy number information (both raw data and estimated profile by mBPCR) and M_p is the random variable representing the $\log_2$ratio value in the p^{th} segment. From the mBPCR model, given *cn*, the conditional posterior distribution of any M_p is $\mathcal{N}(\hat{\mu}_p, \hat{V}_p)$, where $(\hat{\mu}_p, \hat{V}_p)$ are the posterior mean and variance of M_p estimated by mBPCR.

Remark 4.1. We notice that this definition of the prior distribution of Z takes into account that often the DNA sample comes from a mixture of cell populations and thus the measured copy numbers are not integers. There are also other technical reasons in the procedure that lead to this change of domain of the copy number values. It follows that it is not possible to partition the $\log_2$ratio values in intervals around the $\log_2$ratio of the corresponding true copy number, such as:

$$f_{LOGtoZ}(x) = \begin{cases} 2 \text{ if } x > \log_2 \frac{4}{2} + \frac{1}{2}\left(\log_2 \frac{4}{2} + \log_2 \frac{5}{2}\right) \\ 0 \text{ if } \log_2 \frac{1}{2} + \frac{1}{2}\left|\log_2 \frac{1}{2}\right| < x \leq \log_2 \frac{4}{2} + \frac{1}{2}\left(\log_2 \frac{4}{2} + \log_2 \frac{5}{2}\right) \\ -1 \text{ if } \log_2 \frac{1}{2} - \frac{1}{2}\left|\log_2 \frac{1}{2}\right| < x \leq \log_2 \frac{1}{2} + \frac{1}{2}\left|\log_2 \frac{1}{2}\right| \\ -2 \text{ if } x < \log_2 \frac{1}{2} - \frac{1}{2}\left|\log_2 \frac{1}{2}\right|, \end{cases}$$

i.e.

$$f_{LOGtoZ}(x) = \begin{cases} 2 \text{ if } x > 2.16 \\ 0 \text{ if } -0.5 < x \leq 2.16 \\ -1 \text{ if } -1.5 < x \leq -0.5 \\ -2 \text{ if } x < -1.5. \end{cases}$$

In fact, as example, in Figure 4.2, we can observe how poorly the above function can partition the $\log_2$ratio values of 14 HIV lymphoma cell lines in [10].

4.1.4 The estimation

To estimate the piecewise constant profile of the copy number events, first we use the estimators for k_0 and t^0 of the last version of the mBPCR method (see Chapter 3):

$$\widehat{K}_{01} = \arg\max_{k\in\mathbb{K}} p(k\,|\,Y, cn), \tag{4.2}$$

$$\widehat{T}_{BinErrAk} = \arg\max_{t'\in\mathbb{T}_{\hat{k},n}} \mathrm{E}\left[\sum_{q=1}^{\hat{k}-1}\sum_{p=1}^{k_0-1}\delta_{t'_q,t^0_p}\,\middle|\,Y, cn\right]. \tag{4.3}$$

As we previously saw, $\widehat{T}_{BinErrAk}$ corresponds to the $\hat{k}_{01}$ positions which have the highest posterior probability to be a breakpoint. A difference with respect to mBPCR consists in the level estimation. While in the copy number model the levels were continuous random variables, now they assume categorical values. Hence, they are estimated separately (as before) with the MAP estimator instead of the posterior expected value,

$$\widehat{Z}_p = \arg\max_{z=-2,-1,0,2} \mathrm{P}(Z_p = z\,|\,Y, \hat{t}, \hat{k}, cn), \tag{4.4}$$

where $\hat{t}$ and $\hat{k}$ are any estimate of t^0 and k_0, respectively.

Let us define $y_{ij} = (y_{i+1}, \ldots, y_j)$, representing the LOH data points in the interval $[i+1, j]$, and K_{ij} as the random variable which represents the number of segments in the interval $[i+1, j]$. Using Bayes Theorem and the independence of the LOH data points belonging to different segments, the probability in Equation (4.4), given the LOH data y, can be written as,

$$\begin{aligned}
&\mathrm{P}(Z_p = z\,|\,y, \hat{t}, \hat{k}, cn)\\
&= \mathrm{P}(Z_p = z\,|\,y_{\widehat{t}_{p-1},\widehat{t}_p}, \widehat{t}_{p-1}, \widehat{t}_p, \widehat{K}_{\widehat{t}_{p-1},\widehat{t}_p} = 1, cn)\\
&= \frac{\mathrm{P}(y_{\widehat{t}_{p-1},\widehat{t}_p}\,|\,Z_p = z, \widehat{t}_{p-1}, \widehat{t}_p, \widehat{K}_{\widehat{t}_{p-1},\widehat{t}_p} = 1)\mathrm{P}(Z_p = z\,|\,\widehat{t}_{p-1}, \widehat{t}_p, \widehat{K}_{\widehat{t}_{p-1},\widehat{t}_p} = 1, cn)}{\mathrm{P}(y_{\widehat{t}_{p-1},\widehat{t}_p}\,|\,\widehat{t}_{p-1}, \widehat{t}_p, \widehat{K}_{\widehat{t}_{p-1},\widehat{t}_p} = 1, cn)}
\end{aligned} \tag{4.5}$$

Therefore, if the boundary estimator misses a clear boundary between $\widehat{t}_{p-1}$ and $\widehat{t}_p$, then the probability at the denominator of Equation (4.5) could be zero and thus the level would not be estimated. The only way to prevent this event consists in using a good estimator for the boundaries.

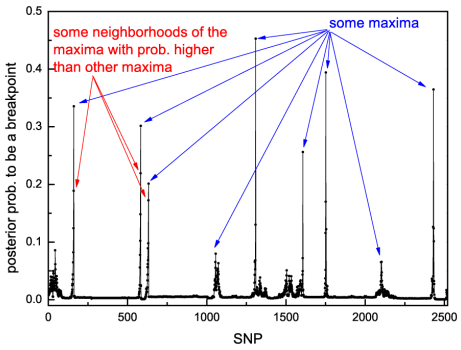

Fig. 4.3 Example of estimated posterior probabilities to be a breakpoint. The graph shows, for each probe, the estimated posterior probability to be a breakpoint on a sample of dataset *B*.

Previously, in Subsection 3.1.6, we found that the boundary estimator $\widehat{T}_{BinErrAk}$ is an estimator with a high sensitivity, but medium FDR. The problem of this estimator is the following. The vector p of the posterior probabilities to be a breakpoint at each point of the sample usually represents a multimodal function with maxima at the breakpoint positions, but often in a neighborhood of each maximum there are other points with high probability because of the uncertainty (see Figure 4.3). If we take the first k_0 points with the highest probability (according to the definition of $\widehat{T}_{BinErrAk}$), we could take points in the neighborhood of the higher maxima and not some maxima with a lower probability (see Figure 4.3). Thus, if k_0 was estimated with its exact value then the sensitivity of the $\widehat{T}_{BinErrAk}$ would be lower. In this case, we could lose important breakpoints so that the denominator in Equation (4.5) would become zero. In practice, $\widehat{K}_{01}$ often slightly overestimates k_0, because of the high noise of the data, and thus this phenomenon should not happen, but to prevent even this rare case we searched for a way to improve the estimation of the boundaries.

Since the vector of the posterior probabilities usually shows the position of the breakpoints clearly in correspondence to the maxima, we estimate the number of the segments and the breakpoints with the number of peaks and the locations of their maxima, respectively (see the next subsubsection). After applying a kernel method to reduce the noise of the function, the algorithm for the determination of the peaks uses two thresholds: for the determination of the peaks (thr_1) and for the definition of the values close to zero (thr_2). We will denote the corresponding estimators by $\widehat{K}_{Peaks,thr_1,thr_2}$ and $\widehat{T}_{Peaks,thr_1,thr_2}$.

In Subsection 4.6.1, we will consider several pairs of thresholds and we will apply the corresponding estimators to simulated data, in order to determine the best paired thresholds and to compare their performance with $\widehat{T}_{BinErrAk}$. We will also compare $\widehat{T}_{BinErrAk}$ with $\widehat{T}_{Joint}$, another boundary estimator described in Subsection 3.1.4.

Algorithm to determine maxima of a multimodal function

We have just introduced the paired estimators $(\widehat{K}_{Peaks,thr_1,thr_2}, \widehat{T}_{Peaks,thr_1,thr_2})$ for the number of segments and the breakpoints. They correspond to the number and the locations of the peaks of the vector p of the posterior probabilities to be a breakpoint at each SNP location. We derived an algorithm for the determination of the maxima in a multimodal function to compute them.

Let us assume that we have to determine the positions of the maxima of a multimodal function f and we know its values at positions $\{1, \ldots, n\}$ (called $f = \{f_1, \ldots, f_n\}$). Moreover, the values f are affected by noise (in fact, in our case f is the posterior probability to be a breakpoint at each position, which depends on the estimates of parameters).

In this framework, we have derived an algorithm to determine the positions of the maxima of f:

1. **Denoising of f.** In order to denoise the function, we use a regression method with kernel basis, obtaining $\hat{f}$.

2. **Selection of only one position per peak.** We identify the positions which belong to the same peak through a threshold thr_1 (i.e. an interval A corresponds to a peak if all elements of $\hat{f}_A$ are greater than thr_1). Then, among the positions belonging to the same peak, we select the one with the highest value of $\hat{f}$. The vector of guess locations is called q^0.
3. **Final selection of the peak locations.** Lastly, we choose all locations $i \in q^0$ such that $\hat{f}_i > thr_2$. The new vector of locations is denoted by q^1. The use of a second threshold is necessary, because the function f (i.e. the estimated posterior probabilities, in our case) can have small peaks also when it assumes values very close to zero (due to the noise).
   Moreover, since we cannot estimate more than k_{max} breakpoints (because of the definition of the prior of K), if more than k_{max} peaks are selected, then the algorithm chooses the ones corresponding to the k_{max} highest values of the set $\{\hat{f}_i \,|\, i \in q^1\}$.

The described algorithm depends on the value of the thresholds thr_1 and thr_2. In the simulations in Section 4.6, we will try several pairs of the following types of thresholds:

- $thr_{005} = max(0.005, \text{quantile of } \hat{p} \text{ at } 0.95)$
- $thr_{01} = max(0.01, \text{quantile of } \hat{p} \text{ at } 0.95)$
- $thr_{01\_90} = max(0.01, \text{quantile of } \hat{p} \text{ at } 0.90)$
- $thr_{mad} = median(\hat{p}) + 3 * mad(\hat{p})$

where *mad* is the median absolute deviation. All these thresholds derive from different definitions of which probability values are to be considered significant.

4.2 Model 2: addition of the IBD/UPD region detection

LOH data are used in biology not only to better identify regions of loss and amplifications, but, especially, to detect regions of copy-neutral LOH, i.e.

long genomic segments with both copies equal. These regions can be identified by unusual long stretches of homozygous SNPs, with normal copy number. In the past, copy-neutral LOH regions were usually explained as a consequence of an uniparental disomy event (UPD) (see [40]). Recently, long homozygous segments have also been detected in genomes of normal individuals, supporting the hypothesis that some copy-neutral LOH segments might represent autozygosity (see, for example, [9, 48, 43]). In literature, it has been shown a relationship between some tumors and both types of aberrant event (see, for example, [4, 5, 22]).

As explained in Subsection 1.1.4, the uniparental isodisomy (iUPD) occurs when two copies of a part of a chromosome are two replicates of one homologue of one parent, while the uniparental heterodisomy (hUPD) occurs when both homologues are inherited from the same parent. In cancer cells, iUPD can also occur when an homologue of a part of a chromosome is lost and the remaining homologue is duplicated. Instead, the autozygosity describes a situation where the homologues are identical by descendent (IBD), because they are inherited from a common ancestor. Consequently, the iUPD or IBD can be detected because they appear as a long sequence of homozygous SNPs with a low probability to occur, while the hUPD consists in a sequence of both homozygous and heterozygous SNPs as in a normal condition. It follows that, without the genotypes of the parents, from SNP data we can only detect the uniparental isodisomy or IBD segments. In the following, we will consider only these two events, referring to them as IBD/UPD events.

Since an IBD/UPD event, by definition, only exists in regions of normal copy number, the only probabilities which are affected by the presence of this event are those involving $\{Z = 0\}$. Therefore, we define the following sets of conditional probabilities $\{\mathrm{P}(Y_i = y \,|\, X_i^N = x, \widetilde{Z}_i = 0, \widetilde{U}_i = 0), y \in \mathbb{Y}, x \in \mathbb{X}\}$ and $\{\mathrm{P}(Y_i = y \,|\, \widetilde{Z}_i = 0, \widetilde{U}_i = 1), y \in \mathbb{Y}\}$, where the variable $\widetilde{U}_i$ indicates if an IBD/UPD event occurred at SNP i. We can notice that, given $\{\widetilde{U}_i = 0, \widetilde{Z}_i = 0\}$, the distribution of Y_i is equal to the conditional distribution with respect to only $\{\widetilde{Z}_i = 0\}$ in Model 1, since the latter was modeled with no possibility of an IBD/UPD event. Instead, in case of an IBD/UPD event, we do not need to condition with respect to X_i^N, since,

in case of a somatic iUPD event, the genotype of an iUPD region is independent of the homozygosity or heterozygosity of same region in a normal cell. Otherwise, in case of autozygosity or germ line iUPD, the genotypes of normal and cancer cells are the same and it has no sense to condition one to the other.

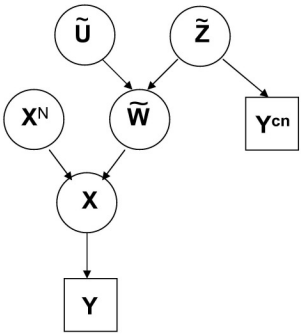

Fig. 4.4 Scheme of Model 2 and 3. The vector $\widetilde{W}$ of aberration events represents the lesions derived from both IBD/UPD events ($\widetilde{U}$) and copy number event ($\widetilde{Z}$), at each SNP position. The vector X of the homozygous status of all SNPs in cancer cells is completely determined, given the vector X^N of their homozygous status in normal cells and the vector $\widetilde{W}$ of their corresponding aberration events. Using this relationship among X, X^N and $\widetilde{W}$, we can estimate $\widetilde{W}$, given the observations Y^{cn} and Y (respectively, the raw $\log_2$ratio of the copy number and the homozygous status in cancer cells detected by the microarray) and by specifying the prior distributions of $\widetilde{U}$ and X^N. The observations Y^{cn} are used to defined the prior distribution of $\widetilde{Z}$ in the Bayesian model.

In the new framework, we define the vector of the aberration events at n SNP loci with $\widetilde{W}=(\widetilde{W}_1, \ldots, \widetilde{W}_n)$. Each component i of the vector assumes values: -3 ($\widetilde{Z}_i$=0 and $\widetilde{U}_i$=1), -2 ($\widetilde{Z}_i$=-2), -1 ($\widetilde{Z}_i$=-1), 0 ($\widetilde{Z}_i$=0 and $\widetilde{U}_i$=0), 2 ($\widetilde{Z}_i$=2); a graphical representation of the model is given in Figure 4.4. As previously, we can divide the genome in intervals corresponding to the same aberration event, i.e the profile of the aberrations consists of k_0 intervals, with boundaries $0 = t_0^0 < t_1^0 < \ldots < t_{k_0-1}^0 < t_{k_0}^0 = n$, so that

$\widetilde{W}_{t^0_{p-1}+1} = \ldots = \widetilde{W}_{t^0_p} =: W_p$, for all $p = 1, \ldots, k_0$. The estimation procedure is similar to the one of Model 1. The estimators of k_0 and t^0 are the same and, given $\hat{k}$ and $\hat{t}$ (any estimate of k_0 and t^0, respectively), we estimate the aberration events in each interval with their MAP estimators,

$$\widehat{W}_p = \underset{w=-3,-2,-1,0,2}{\arg\max}\ \mathrm{P}(W_p = w \,|\, Y, \hat{t}, \hat{k}, cn). \tag{4.6}$$

Notice that, for $w = -2, -1, 2$, the posterior probability $\mathrm{P}(W_p = w \,|\, Y, \hat{t}, \hat{k}, cn)$ is equal to $\mathrm{P}(Z_p = w \,|\, Y, \hat{t}, \hat{k}, cn)$, while for $w = -3, 0$ we have,

$$\mathrm{P}(W_p = -3 \,|\, Y, \hat{t}, \hat{k}, cn) = \mathrm{P}(Z_p = 0 \,|\, U_p = 1, Y, \hat{t}, \hat{k}, cn)\mathrm{P}(U_p = 1) \tag{4.7}$$

$$\mathrm{P}(W_p = 0 \,|\, Y, \hat{t}, \hat{k}, cn) = \mathrm{P}(Z_p = 0 \,|\, U_p = 0, Y, \hat{t}, \hat{k}, cn)\mathrm{P}(U_p = 0) \tag{4.8}$$

and we assume that $p_{upd} := \mathrm{P}(U_p = 1)$, for any $p = 1, \ldots, \hat{k}$.

Remark 4.2. The addition of the IBD/UPD detection to the model is necessary for the analysis of real data, in case of either a normal or a "cancer" sample. As we explain before, the IBD regions can be present also in the genome of normal sample, and both IBD and UPD regions can be potentially related to the disease of a patient and thus present in his genome. Therefore, in any sample we can expect to find IBD/UPD regions. Without considering the IBD/UPD aberration in the model, many regions with this alteration would be seen as mono-allelic or bi-allelic deletion, due to the high percentage of homozygous SNPs in them, leading to a poor estimation of the aberration event profile of the patient.

4.3 Model 3: addition of the gained region detection

In the description of Model 1, we explained our assumption that there is no difference in the genotyping detection between a normal or gained region. Therefore, in Model 1 (and in Model 2), we defined a single class for the normal or gained regions. But, for the biological studies, it is relevant

to distinguish these two copy number events and this distinction is based essentially on the estimated copy number (since there is no difference in the distribution of the detected genotypes, due to the previous discussion). As a consequence, the probability of Y_i given a normal (i.e. $\{\widetilde{Z}_i = 0\}$) or gained copy number (i.e. $\{\widetilde{Z}_i = 1\} = \{\widetilde{W}_i = 1\}$) is the same,

$$\begin{aligned}\mathrm{P}(Y_i = y \,|\, X_i^N = x, \widetilde{Z}_i = 1) &= \mathrm{P}(Y_i = y \,|\, X_i^N = x, \widetilde{Z}_i = 0) \\ &= \mathrm{P}(Y_i = y \,|\, \widetilde{Z}_i = 0, \widetilde{U}_i = 1) p_{upd} \\ &+ \mathrm{P}(Y_i = y \,|\, X_i^N = x, \widetilde{Z}_i = 0, \widetilde{U}_i = 0)(1 - p_{upd}).\end{aligned}$$

We need also to define two distinct prior probabilities for the normal copy number and the gain event. Similarly to its previous definition in Subsection 4.1.3, for all $p = 1, \ldots, \hat{k}^{cn}$, the new prior of Z_p is simply given by,

$$\begin{aligned}\mathrm{P}(Z_p = 2) &= \mathrm{P}\left(M_p \geq \widehat{m}_4 + 3\widehat{s}_4 \,\middle|\, cn\right) \\ \mathrm{P}(Z_p = 1) &= \mathrm{P}\left(\widehat{m}_2 + 3\widehat{s}_2 < M_p \leq \widehat{m}_4 + 3\widehat{s}_4 \,\middle|\, cn\right) \\ \mathrm{P}(Z_p = 0) &= \mathrm{P}\left(\widehat{m}_2 - 3\widehat{s}_2 < M_p \leq \widehat{m}_2 + 3\widehat{s}_2 \,\middle|\, cn\right) \\ \mathrm{P}(Z_p = -1) &= \mathrm{P}\left(\widehat{m}_1 - 3\widehat{s}_1 < M_p \leq \widehat{m}_2 - 3\widehat{s}_2 \,\middle|\, cn\right) \\ \mathrm{P}(Z_p = -2) &= \mathrm{P}\left(M_p \leq \widehat{m}_1 - 3\widehat{s}_1 \,\middle|\, cn\right).\end{aligned}$$

In the following, Model 3 (which is the complete model) will be called *genomic Bayesian Piecewise Constant Regression* (gBPCR).

4.4 Estimation of the parameters of the model

In the previous sections we have described the gBPCR model, which depends on the specification of several parameters regarding both the likelihood and the priors. Here, we described how we estimated these parameters, by using mainly published datasets.

4.4.1 Estimation of the parameters of the likelihood

The set of conditional probabilities $\{\mathrm{P}(Y_i = y \,|\, X_i^N = x, \widetilde{W}_i = w)$, $y \in \mathbb{Y}$, $x \in \mathbb{X}$, w=-2, -1, 0, 2$\}$ are considered as parameters of the model. To quantify them, we needed paired normal-cancer samples, since they are related to the probability of detecting a certain homozygous status in a cancer cell, given the corresponding one in a normal cell of the same patient and under some copy number event. Therefore, to compute maximum likelihood estimates of these parameters, we used some breast cancer cell line samples of [42, 87], suitable for our purpose. The genotyping calling algorithm used was BRLMM [1].

The following two probabilities are related to a normal copy number event and they represent errors due to the detection of *NoCall* instead of *NHet* and *Het*, respectively,

$$\begin{aligned}\mathrm{P}(Y_i = NoCall \,|\, X_i^N = Hom, \widetilde{W}_i = 0) &= \delta_1\\ \mathrm{P}(Y_i = NoCall \,|\, X_i^N = Het, \widetilde{W}_i = 0) &= \delta_2.\end{aligned}$$

To estimate them, we used chromosome 1 of two replicates of the normal cell line HCC38 BL. In this chromosome we did not find any SNP with homozygous call in one sample and heterozygous call in the other one, thus we assumed that the detected genotypes (different from *NoCall*) were all correct. Instead, we found 27 *NoCall* SNPs in both samples, so that we eliminated them from the analysis. The estimated parameters were:

$$\begin{aligned}\hat{\delta}_1 &= \frac{\#\{\text{ SNPs homozygous in one sample and } NoCall \text{ in the other}\}}{\#\{\text{homozygous SNPs in at least one sample}\}}\\ \hat{\delta}_2 &= \frac{\#\{\text{ SNPs heterozygous in one sample and } NoCall \text{ in the other}\}}{\#\{\text{heterozygous SNPs in at least one sample}\}}\end{aligned}$$

Regarding the other probabilities related to a normal copy number $\mathrm{P}(Y_i = Het \,|\, X_i^N = Hom, \widetilde{W}_i = 0)$ and $\mathrm{P}(Y_i = NHet \,|\, X_i^N = Het, \widetilde{W}_i = 0)$, we set them as the genotyping detection error.

$\{\mathrm{P}(Y_i = y \,|\, X_i^N = x, \widetilde{W}_i = w),\ y \in \mathbb{Y},\ x \in \mathbb{X},\ w$=-2, -1, 2$\}$ are related to errors due also to the presence of a subpopulation of cells in the tumor sample (normal cells or tumor cells in another stage of the disease). Hence, to estimate them, we used the human breast carcinoma cell lines HCC1143 and HCC38, because we had samples containing 0, 60, 70, 80, 90, 100% of tumor cells for each cell line [42, 87]. We defined some regions of amplification, loss and homozygous deletion on the basis of the regions of copy number changes indicated by [42, 87]. For each region, we looked at the copy number value of the SNPs to better identify the start/end SNP of the aberrant region, since, in [87, 42], they were only denoted by the corresponding cytobands. Finally, the estimations of the probabilities were performed using the maximum likelihood estimators averaging over all the samples of the same cell line with some percentage of tumor cells, for w = -2, -1, 2,

$$\hat{\mathrm{P}}(Y_i = Het \,|\, X_i^N = x, \widetilde{W}_i = w) = \frac{1}{5\sum_c n_c} \sum_{\substack{c \in \{\text{HCC38,} \\ \text{HCC1143}\}}} \sum_{h=1}^{5} \sum_{j=1}^{n_c} (SNP_{j,c,h}^2 + SNP_{j,c,h})$$

$$\hat{\mathrm{P}}(Y_i = NoCall \,|\, X_i^N = x, \widetilde{W}_i = w) = \frac{1}{5\sum_c n_c} \sum_{\substack{c \in \{\text{HCC38,} \\ \text{HCC1143}\}}} \sum_{h=1}^{5} \sum_{j=1}^{n_c} (SNP_{j,c,h}^2 - SNP_{j,c,h})$$

where n_c is the total number of the SNPs in the regions with $\{X_i^N = x, \widetilde{W}_i = w\}$ of cell line c, and $SNP_{j,c,h}$ is a value assign to j^{th} SNP of sample h of cell line c, based on its homozygous status $Y_{j,c,h}$:

$$SNP_{j,c,h} = \begin{cases} -1 \text{ if } Y_{j,c,h} = NoCall \\ 0 \text{ if } Y_{j,c,h} = NHet \\ 1 \text{ if } Y_{j,c,h} = Het. \end{cases}$$

It remains to estimate the following probabilities related to the IBD/UPD events,

$$\mathrm{P}(Y_i = NHet \,|\, \widetilde{W}_i = -3) = \delta_3$$
$$\mathrm{P}(Y_i = NoCall \,|\, \widetilde{W}_i = -3) = \delta_4.$$

For their estimation, we used 11 IBD/UPD regions previously found by us on 5 samples of patients with hairy cell leukemia [19] and on the B-

cell lymphoma cell line KARPAS-422 (unpublished). All regions were detected by dChip [8]. Their width is between 3Mb and 100Mb (covering from 300 to 9800 SNPs), so that they are large enough to be really considered IBD/UPD regions.

We computed the estimators for δ_3 and δ_4 simply using the frequency of $NHet$ and $NoCall$ in the selected IBD/UPD regions, respectively. We found $\hat{\delta}_4$ equal to the arithmetic mean of $\hat{\delta}_1$ and $\hat{\delta}_2$ (the errors of detecting a $NoCall$ in a normal region instead of $NHet$ or Het, respectively), which is a realistic result since we can have both homozygous and heterozygous SNPs in a UPD region.

4.4.2 Estimation of the parameter p_{upd}

We expect the prior probability of an IBD/UPD event to be low. In order to estimate the order of magnitude of this parameter, we considered two studies on IBD regions: [4] and [48]. In the former, they considered as IBD regions only stretches of at least 50 homozygous SNPs (with at maximum 2% of heterozygous) longer than 4Mb and the platform used was the Affymetrix GeneChip Human Mapping 50K Array. In the latter, a denser microarray was used and the stretches considered were longer than 1Mb (with at least 50 probes) or longer than 3Mb. Hence, using the data of the former paper (only the normal samples), we estimated $p_{upd} \approx 1.7 \cdot 10^{-3}$. Instead, with the data of the latter, we estimated $p_{upd} \approx 1.5 \cdot 10^{-3}$, considering all regions greater than 1Mb, while $p_{upd} \approx 1.46 \cdot 10^{-4}$, considering only the regions greater than 3Mb. The differences in the estimated values are due to the different resolution of the technologies used (in fact, in the former the number of SNPs used was 58,960, while in the latter was 3,107,620). Moreover, the probability depends also on the minimum length allowed for these regions. The wider the regions are, the higher is the probability that the regions represent "abnormalities" and the lower becomes the probability of their occurrence (so that p_{upd} is lower). Therefore, in the following applications in Sections 4.6 and 4.7, we will use two values: $p_{upd} = 10^{-3}$ and $p_{upd} = 10^{-4}$.

Another possible way to solve the problem could be to assign a prior probability to p_{upd} (for example, we could know its range and use a uniform distribution in this range) and integrate it out in the equations of the model. In this way, the equations would depend only on the expected value of p_{upd}.

4.5 Adjustment of the parameters related to *NoCall*

The probabilities $\{\mathrm{P}(Y_i = NoCall \,|\, X_i^N = x, \widetilde{W}_i = w), x \in \mathbb{X}, w - 3, -2, -1, 0, 2\}$ are related to the detection of *NoCall*s under some conditions. Generally, the presence of *NoCall*s is not only due to difficulties of the microarray in the detection of the genotype (technical noise), but also to the noise of the sample because of the differences in quality of extracted DNA. Therefore, we need to adjust the estimated values of these parameters on the basis of the sample noise.

Since usually the *NoCall* rate (i.e. percentage of *NoCall*s in the sample) increases with the noise of the sample, we assume that, given $\{X_i^N = x, \widetilde{W}_i = z\}$, the probability of detecting a *NoCall* at SNP i in sample s is proportional to a parameter $p_{x,z}$ (which depends on the technical noise) by a factor θ_s (which depends on the sample noise),

$$\mathrm{P}(Y_i = NoCall \,|\, X_i^N = x, \widetilde{W}_i = z, s) \approx p_{x,z}\theta_s. \tag{4.9}$$

By conditioning over the values of X_i^N and estimating $\mathrm{P}(X_i^N = Het) = 1/2$ for a generic SNP i, we compute the *NoCall* rate in regions with copy number event z,

$$\begin{aligned} &\mathrm{P}(Y_i = NoCall \,|\, \widetilde{W}_i = z, s) && (4.10)\\ &= \mathrm{P}(Y_i = NoCall \,|\, X_i^N = Hom, \widetilde{W}_i = z, s)\mathrm{P}(X_i^N = Hom) \\ &+ \mathrm{P}(Y_i = NoCall \,|\, X_i^N = Het, \widetilde{W}_i = z, s)\mathrm{P}(X_i^N = Het) && (4.11)\\ &\approx p_{Het,z}\theta_s\mathrm{P}(X_i^N = Het) + p_{Hom,z}\theta_s\mathrm{P}(X_i^N = Hom) \end{aligned}$$

$$\approx \theta_s \frac{p_{Het,z} + p_{Hom,z}}{2}. \tag{4.12}$$

Therefore, for any pair of samples (sample 1 and 2), we can write the conditional probability of *NoCall*, given $\{X_i^N = x,\ \widetilde{W}_i = z\}$, in sample 1 in terms of the corresponding probability in sample 2,

$$\begin{aligned} &\mathrm{P}(Y_i = NoCall \,|\, X_i^N = x, \widetilde{W}_i = z, s = 1) \\ &\approx \mathrm{P}(Y_i = NoCall \,|\, X_i^N = x, \widetilde{W}_i = z, s = 2) \frac{\mathrm{P}(Y_i = NoCall \,|\, \widetilde{W}_i = z, s = 1)}{\mathrm{P}(Y_i = NoCall \,|\, \widetilde{W}_i = z, s = 2)} \end{aligned} \tag{4.13}$$

because, applying Equations (4.9) and (4.12),

$$\begin{aligned} \mathrm{P}(Y_i = NoCall \,|\, X_i^N = x, \widetilde{W}_i = z, s = 1) &\approx \frac{\theta_1 \frac{1}{2}(p_{Het,z} + p_{Hom,z})}{\theta_2 \frac{1}{2}(p_{Het,z} + p_{Hom,z})} p_{x,z}\theta_2 \\ &= p_{x,z}\theta_1. \end{aligned}$$

In the following, we will denote the sample to estimate with $s = 1$ and the reference sample with $s = 2$.

By using Equation (4.13), the values of the parameters related to *NoCall* detection are adjusted for sample 1,

$$\widehat{\mathrm{P}}(Y_i = NoCall \,|\, X_i^N = x, \widetilde{W}_i = z, s = 1) = \frac{r_1(z)}{r_2(z)} \widehat{\mathrm{P}}(Y_i = NoCall \,|\, X_i^N = x, \widetilde{W}_i = z, s = 2),$$

for z=-2,-1,0,2, where $r_1(z)$ and $r_2(z)$ are an estimate of the *NoCall* rate, in regions with copy number event z, for sample 1 and 2, respectively. By applying Equation (4.11) with $\mathrm{P}(X_i^N = Het)$ = 1/2, $r_2(z)$ can be computed from the estimated values of $\mathrm{P}(Y_i = NoCall \,|\, X_i^N = Het, \widetilde{W}_i = z)$ and $\mathrm{P}(Y_i = NoCall \,|\, X_i^N = Hom, \widetilde{W}_i = z)$,

$$\begin{aligned} r_2(z) := \widehat{\mathrm{P}}(Y_i = NoCall \,|\, \widetilde{W}_i = z, s = 2) \\ = \frac{1}{2}\widehat{\mathrm{P}}(Y_i = NoCall \,|\, X_i^N = Het, \widetilde{W}_i = z, s = 2) \\ + \frac{1}{2}\widehat{\mathrm{P}}(Y_i = NoCall \,|\, X_i^N = Hom, \widetilde{W}_i = z, s = 2), \end{aligned}$$

for $z = -2,-1,0,2$. Instead, $r_1(z)$ is the frequency of *NoCall* in regions with copy number event z of sample 1, for $z = -2,-1,0,2$.

The estimated value of the probability $\mathrm{P}(Y_i = NoCall \,|\, \widetilde{W}_i = -3)$ is adjusted in a different way. As expected, on the reference samples we found that

$$\begin{aligned}\widehat{\mathrm{P}}(Y_i = NoCall \,|\, \widetilde{W}_i = -3, s) &= \frac{1}{2}\widehat{\mathrm{P}}(Y_i = NoCall \,|\, X_i^N = Het, \widetilde{W}_i = z, s) \\ &+ \frac{1}{2}\widehat{\mathrm{P}}(Y_i = NoCall \,|\, X_i^N = Hom, \widetilde{W}_i = z, s) \\ &\approx \widehat{\mathrm{P}}(Y_i = NoCall \,|\, \widetilde{W}_i = 0, s),\end{aligned}$$

that is the *NoCall* rate in IBD/UPD regions is approximately equal to the *NoCall* rate in normal regions. Therefore,

$$\widehat{\mathrm{P}}(Y_i = NoCall \,|\, \widetilde{W}_i = -3, s = 1) = r_1(0).$$

In Subsection 4.6.2, we will compare the estimations resulting from gBPCR with and without the adjustment of these parameters.

4.6 Simulations

In this section, we apply gBPCR to artificial data. First, we compare the boundary estimators (described in Subsection 4.1.4) on data simulated using Model 1. Then, we evaluate the detection of IBD/UPD regions on the artificial dataset of [85], in comparison with two well-known methods for LOH estimation. Using the same data, we also show the difference in the estimation by using the adjustment of the parameters.

4.6.1 Comparison among the breakpoint estimators on simulated data

In Subsection 4.1.4, we have described several possible boundary estimators: $\widehat{T}_{BinErrAk}$, $\widehat{T}_{Joint}$ and $\widehat{T}_{Peaks,thr_1,thr_2}$. The last one actually defines a class of estimators which depend on the values of the thresholds thr_1 and thr_2. In the comparisons, we tried several pairs of the thresholds defined in Subsection 4.1.4.

We assessed the quality of all the estimators of k_0 and t^0 considered, by applying them on two artificial datasets (called datasets A and B), each of 100 samples. We used as prior probabilities of heterozygosity the ones given by the annotation file of Affymetrix for the SNPs of chromosome 22 in the Affymetrix GeneChip Mapping 250K NspI microarray (Affymetrix, Santa Clara, CA, USA), hence the number of data points in each sample is $n = 2520$. Since our complete model (Model 3) does not provide a realistic way to simulate IBD/UPD regions and the identification of gained regions depends mainly on copy number data, the samples were generated using Model 1.

Simulation description

Since the method assumes to know the estimated copy number profile given by mBPCR, for both datasets we fixed the estimated segment number $\hat{k}^{cn} = 15$, the estimated boundaries $\hat{t}^{cn} = (0, 27, 31, 161, 273, 585, 633, 1006, 1050, 1054, 1309, 1607, 1754, 2100, 2432, 2520)$ (generated uniformly random given $\hat{k}^{cn} = 15$) and the prior probabilities of Z (in Table 4.1, for dataset A, and Table 4.2, for dataset B). The profiles of the samples in dataset A should be estimated easily, since in each segment the prior distribution of Z is more peaked with respect to dataset B.

Given the previous parameters and the estimated values of the other parameters of the model, we used the following steps to generate each LOH sample:

Table 4.1 Prior distribution of Z in the simulated dataset A.

	segment														
prior	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV
$P(Z_p = 2)$	0	0	0	0	0.8	0	0	0	0	0.8	0	0	0.8	0	0
$P(Z_p = 0)$	0.1	0.8	0.1	0.8	0.2	0.8	0.1	0	0.8	0.2	0.1	0.8	0.2	0.8	0.1
$P(Z_p = -1)$	0.8	0.2	0.8	0.2	0	0.2	0.8	0.2	0.2	0	0.8	0.2	0	0.2	0.8
$P(Z_p = -2)$	0.1	0	0.1	0	0	0	0.1	0.8	0	0	0.1	0	0	0	0.1

Table 4.2 Prior distribution of Z in the simulated dataset B.

	segment														
prior	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV
$P(Z_p = 2)$	0	0.1	0	0.1	0.5	0.1	0	0	0.1	0.5	0	0.1	0.5	0.1	0
$P(Z_p = 0)$	0.1	0.6	0.1	0.6	0.4	0.6	0.1	0.1	0.6	0.4	0.1	0.6	0.4	0.6	0.1
$P(Z_p = -1)$	0.6	0.3	0.6	0.3	0.1	0.3	0.6	0.4	0.3	0.1	0.6	0.3	0.1	0.3	0.6
$P(Z_p = -2)$	0.3	0	0.3	0	0	0	0.3	0.5	0	0	0.3	0	0	0	0.3

1. we generated a true profile of the homozygous status X^N, by using the prior probabilities of heterozygosity, described previously,
2. we generated a true copy number event profile $\widetilde{Z}$, by using the prior distribution of Z (notice that in some cases the final profile can have less than 15 segments, since if consecutive segments have the same copy number event they are joined together),
3. given the true copy number event profile and the profile of the homozygous status, we generated Y (the profile of the homozygous status in cancer cells detected by the microarray), by using the conditional probability distributions of Model 1.

Description of the error measures

To evaluate the performance of the estimators, we used several error measures, that we have already employed for the comparisons in Chapter 3.

For the estimation of the number of segments, we considered the following errors:

$$\begin{aligned} \text{0-1 error} &= 1 - \delta_{\hat{k},k_0} \\ \text{absolute error} &= |\hat{k} - k_0| \\ \text{squared error} &= (\hat{k} - k_0)^2. \end{aligned}$$

For the evaluation of the boundary estimation, we computed the binary error, i.e.

$$k_0 - 1 - \sum_{q=1}^{\hat{k}-1} \sum_{p=1}^{k_0-1} \delta_{\hat{t}_q, t_p^0},$$

the sensitivity (proportion of true breakpoints detected) and the false discovery rate (FDR, i.e. proportion of false estimated breakpoints among the estimated ones). The last two measures were calculated not only looking at the exact position of the breakpoints (w = 0), but also accounting for a neighborhood of up to 6 SNPs around the true position (w = 1, ..., 6). Finally, to asses the influence of the boundary estimation on the profile estimation, we calculated the sum 0-1 error and the sum of squared distance (SSQ), which are defined as

$$\begin{aligned} \text{sum 0-1 error} &= \sum_{i=1}^{n} \left(1 - \delta_{\widehat{\widetilde{Z}}_i, \widetilde{Z}_i^0}\right) \\ \text{SSQ} &= \sum_{i=1}^{n} (\widehat{\widetilde{Z}}_i - \widetilde{Z}_i^0)^2. \end{aligned}$$

We also measured the sensitivity and the FDR for all copy number events.

To compare the estimators, we considered not only the error measures computed on their estimates but also on their "final" estimates (denoted by "final" or F). In fact, since the levels are categorical variables and they are estimated separately (see Equation (4.4)), if the estimated levels of contiguous segments are the same, then they are joined together ("merging" step). Therefore, after the "merging", the number of the segments can be lower than the estimated one.

Table 4.3 Binary error of the boundary estimations obtained with several boundary estimators on datasets *A* and *B*. On the former dataset $(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$ outperforms all other methods, while on the latter $(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$, $(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$, $(\widehat{K}_{Peaks,01,01}, \widehat{T}_{Peaks,01,01})$ and $(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$ give the lower binary errors.

dataset	method	type	binary error
A	$(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$	estimated	4.19
		final	4.43
	$(\widehat{K}_{01}, \widehat{T}_{Joint})$	estimated	7.23
		final	7.23
	$(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$	estimated	6.09
		final	6.16
B	$(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$	estimated	6.53
		final	7.37
	$(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$	estimated	6.97
		final	7.33
	$(\widehat{K}_{Peaks,01,01}, \widehat{T}_{Peaks,01,01})$	estimated	6.97
		final	7.33
	$(\widehat{K}_{Peaks,01\_90,01\_90}, \widehat{T}_{Peaks,01\_90,01\_90})$	estimated	7.21
		final	7.46
	$(\widehat{K}_{Peaks,mad,mad}, \widehat{T}_{Peaks,mad,mad})$	estimated	7.40
		final	7.51
	$(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$	estimated	6.59
		final	7.00
	$(\widehat{K}_{Peaks,mad,01}, \widehat{T}_{Peaks,mad,01})$	estimated	7.53
		final	7.64

Results of the comparisons

We applied the following pairs of estimators to dataset *A*: $(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$, $(\widehat{K}_{01}, \widehat{T}_{Joint})$ and $(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$. We found that $(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$ and $(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$ were the best performing methods. In particular, the former had the lowest binary error, regarding both the estimated boundaries and the "final" ones (see Table 4.3) and the lowest "final" FDR (see Figure 4.5), while the errors regarding the "final" estimation

of the number of segments were similar (see Table 4.4). As a consequence, regarding the level estimation, $(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$ had the lowest errors (see Table 4.5) and almost always the highest sensitivity and lowest FDR (see Tables 4.6 and 4.7).

Table 4.4 Error measures regarding the estimation of the number of segments on both datasets *A* and *B*. The estimations were obtained using several types of estimators. On the former dataset $(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$ and $(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$ perform equally good, on the latter the best performing methods are $(\widehat{K}_{Peaks,mad,01}, \widehat{T}_{Peaks,mad,01})$ and $(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$, followed by $(\widehat{K}_{Peaks,01,01}, \widehat{T}_{Peaks,01,01})$ and $(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$.

dataset	method	type	err 0-1	err 1	err 2	$\#(\hat{k} > k_0)$
A	$(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$	estimated	1	15.74	256.2	100
		final	0.85	1.73	4.35	20
	$(\widehat{K}_{01}, \widehat{T}_{Joint})$	estimated	1	15.74	256.2	100
		final	1	15.69	254.51	100
	$(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$	estimated	0.99	7.43	65.55	99
		final	0.84	2.33	9.55	65
B	$(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$	estimated	0.98	6.61	57.29	96
		final	0.89	2.46	9.22	16
	$(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$	estimated	1	12.72	177.22	100
		final	0.92	3.75	20.31	82
	$(\widehat{K}_{Peaks,01,01}, \widehat{T}_{Peaks,01,01})$	estimated	1	12.62	175.98	100
		final	0.92	3.74	20.28	81
	$(\widehat{K}_{Peaks,01\_90,01\_90}, \widehat{T}_{Peaks,01\_90,01\_90})$	estimated	1	15.12	246.38	100
		final	0.98	5.75	40.61	94
	$(\widehat{K}_{Peaks,mad,mad}, \widehat{T}_{Peaks,mad,mad})$	estimated	1	17.08	300.16	100
		final	0.99	7.03	58.09	99
	$(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$	estimated	1	12.62	175.98	100
		final	0.98	5.47	38.25	91
	$(\widehat{K}_{Peaks,mad,01}, \widehat{T}_{Peaks,mad,01})$	estimated	0.9	2.98	14.3	75
		final	0.83	1.73	4.95	40

From these results, we decided to not apply the estimators $(\widehat{K}_{01}, \widehat{T}_{Joint})$ on dataset *B* and we also decided to try other paired thresholds for

Table 4.5 The table shows some error measures regarding the copy number event estimation obtained with several methods on datasets *A* and *B*. While $(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$ outperforms the other methos on the former dataset, on the latter it obtains a poor estimation of the copy number events in comparison with the other methods. On dataset *B*, the methods which achieve the lowest errors are: $(\widehat{K}_{Peaks,01,01}, \widehat{T}_{Peaks,01,01})$, $(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$, $(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$ and $(\widehat{K}_{Peaks,mad,01}, \widehat{T}_{Peaks,mad,01})$.

dataset	method	sum 0-1 err	SSQ	$\sqrt{SSQ/n}$
	$(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$	51.53	86.08	0.19
A	$(\widehat{K}_{01}, \widehat{T}_{Joint})$	146.91	596.78	0.49
	$(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$	91.99	345.64	0.37
	$(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$	421.79	1226.59	0.70
	$(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$	110.39	287.21	0.34
	$(\widehat{K}_{Peaks,01,01}, \widehat{T}_{Peaks,01,01})$	109.39	286.15	0.34
B	$(\widehat{K}_{Peaks,01\_90,01\_90}, \widehat{T}_{Peaks,01\_90,01\_90})$	141.65	370.78	0.38
	$(\widehat{K}_{Peaks,mad,mad}, \widehat{T}_{Peaks,mad,mad})$	154.56	424.2	0.41
	$(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$	109.39	286.15	0.34
	$(\widehat{K}_{Peaks,mad,01}, \widehat{T}_{Peaks,mad,01})$	111.75	283.77	0.34

$\widehat{T}_{Peaks,thr_1,thr_2}$, in order to reduce the FDR of the boundary estimation. The results showed that the methods which obtained a better estimation of the number of segments were, in order: $(\widehat{K}_{Peaks,mad,01}, \widehat{T}_{Peaks,mad,01})$, $(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$, $(\widehat{K}_{Peaks,01,01}, \widehat{T}_{Peaks,01,01})$ and $(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$; see Table 4.4. Instead, regarding the boundary estimation, the methods with the lowest binary error were: $(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$, $(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$, $(\widehat{K}_{Peaks,01,01}, \widehat{T}_{Peaks,01,01})$ and $(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$; see Table 4.3.

In general, the methods $(\widehat{K}_{Peaks,01,01}, \widehat{T}_{Peaks,01,01})$ and $(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$ always obtained similar results and the latter perform slightly worse than the former (e.g. in the estimation of k_0). Moreover, the methods $(\widehat{K}_{Peaks,01\_90,01\_90}, \widehat{T}_{Peaks,01\_90,01\_90})$ and $(\widehat{K}_{Peaks,mad,mad}, \widehat{T}_{Peaks,mad,mad})$ had always all the error measures higher than $(\widehat{K}_{Peaks,01,01}, \widehat{T}_{Peaks,01,01})$. Therefore, the pair of estimators $(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$,

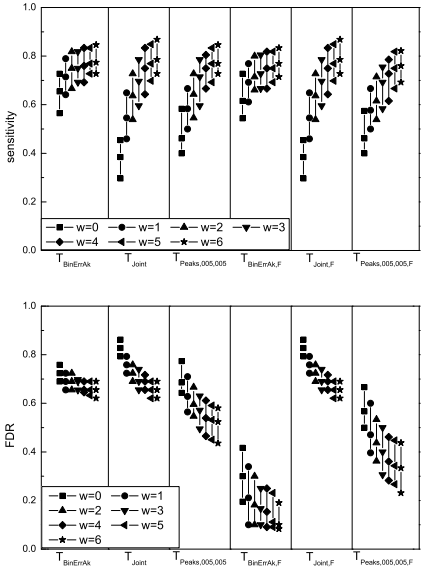

Fig. 4.5 Sensitivity (on the top) and FDR (at the bottom) of all boundary estimators applied to dataset *A*. The estimator $\widehat{T}_{BinErrAk}$ achieves the lowest "final" FDR and $\widehat{T}_{Joint}$ has the highest FDR. For $w = 6$, the sensitivity of all estimators look similar.

$(\widehat{K}_{Peaks,01\_90,01\_90}, \widehat{T}_{Peaks,01\_90,01\_90})$ and $(\widehat{K}_{Peaks,mad,mad}, \widehat{T}_{Peaks,mad,mad})$ will not be considered in the following discussions.

The lowest errors in the estimation of the number of the segments were achieved by $(\widehat{K}_{Peaks,mad,01}, \widehat{T}_{Peaks,mad,01})$; see Table 4.4. Moreover, using this procedure, k_0 was underestimated in about half of the cases and thus also the FDR regarding the boundary estimation was the lowest one (see Figure 4.6). As a consequence, all the error measures regarding the level estimation were among the best ones (see Table 4.5, and Tables 4.6 and 4.7). Instead, using $(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$, the number of seg-

Table 4.6 Sensitivity in the detection of each type of copy number event on datasets A and B. On dataset A, $(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$ and $(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$ seem to have globally the highest sensitivity, while, on the latter dataset, $(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$, $(\widehat{K}_{Peaks,01,01}, \widehat{T}_{Peaks,01,01})$ and $(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$ outperform all other methods.

dataset	method	sensitivity			
		$Z=2$	$Z=0$	$Z=-1$	$Z=-2$
A	$(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$	0.803	0.987	0.984	0.995
	$(\widehat{K}_{01}, \widehat{T}_{Joint})$	0.912	0.977	0.926	0.931
	$(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$	0.849	0.985	0.963	0.961
B	$(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$	0.681	0.932	0.968	0.555
	$(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$	0.894	0.983	0.961	0.946
	$(\widehat{K}_{Peaks,01,01}, \widehat{T}_{Peaks,01,01})$	0.896	0.983	0.961	0.946
	$(\widehat{K}_{Peaks,01\_90,01\_90}, \widehat{T}_{Peaks,01\_90,01\_90})$	0.884	0.981	0.940	0.930
	$(\widehat{K}_{Peaks,mad,mad}, \widehat{T}_{Peaks,mad,mad})$	0.893	0.979	0.928	0.923
	$(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$	0.896	0.983	0.961	0.946
	$(\widehat{K}_{Peaks,mad,01}, \widehat{T}_{Peaks,mad,01})$	0.889	0.984	0.963	0.942

ments was almost always overestimated (see Table 4.4) and thus the algorithm detected the highest number of true breakpoints (in fact, it had the highest sensitivity, in Figure 4.7, and a low binary error, in Table 4.3). But due to the higher number of segments, the algorithm found also a higher number of false breakpoints than $(\widehat{K}_{Peaks,mad,01}, \widehat{T}_{Peaks,mad,01})$ (see Figure 4.6).

From the study of the behavior of $(\widehat{K}_{Peaks,mad,01}, \widehat{T}_{Peaks,mad,01})$ and $(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$, we can understand the role of the two thresholds in our algorithm for the determination of the maxima in a multimodal function (see Subsection 4.1.4). The threshold thr_1 is used to decide which points belong to the same peak: all the points, between two regions of points below thr_1, are considered in the same peak. Hence, with a low threshold, more points are considered belonging to the same peak and thus we can eliminate lot of false breakpoints (like in $(\widehat{K}_{Peaks,mad,01}, \widehat{T}_{Peaks,mad,01})$). But, at the same time, if two true peaks are close, then it is

Table 4.7 FDR in the detection of each type of copy number event on datasets *A* and *B*. On dataset *A*, $(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$ and $(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$ seem to have globally the lowest FDR, while, on the latter dataset, $(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$, $(\widehat{K}_{Peaks,01,01}, \widehat{T}_{Peaks,01,01})$ and $(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$ outperform all other methods.

dataset	method	FDR			
		$Z=2$	$Z=0$	$Z=-1$	$Z=-2$
A	$(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$	0.039	0.020	0.036	0.000
	$(\widehat{K}_{01}, \widehat{T}_{Joint})$	0.232	0.110	0.027	0.002
	$(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$	0.141	0.064	0.029	0.001
B	$(\widehat{K}_{01}, \widehat{T}_{BinErrAk})$	0.017	0.047	0.306	0.025
	$(\widehat{K}_{Peaks,005,005}, \widehat{T}_{Peaks,005,005})$	0.044	0.031	0.069	0.020
	$(\widehat{K}_{Peaks,01,01}, \widehat{T}_{Peaks,01,01})$	0.043	0.031	0.068	0.020
	$(\widehat{K}_{Peaks,01\_90}, \widehat{T}_{Peaks,01\_90})$	0.085	0.036	0.081	0.028
	$(\widehat{K}_{Peaks,mad,mad}, \widehat{T}_{Peaks,mad,mad})$	0.106	0.041	0.079	0.034
	$(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$	0.043	0.031	0.068	0.020
	$(\widehat{K}_{Peaks,mad,01}, \widehat{T}_{Peaks,mad,01})$	0.038	0.026	0.075	0.023

possible that they are considered as only one peak, losing a true breakpoint (low sensitivity). Instead, the threshold thr_2 is used to choose which estimated breakpoints are significant for the regression, i.e. if their posterior probabilities are to be considered different from zero. Therefore, using a lower value of thr_2, we select a higher number of breakpoints obtaining a higher percentage of both false ones (high FDR) and true ones (high sensitivity, as in $(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$).

In conclusion, from these results we suggest the use of the following pairs of estimators: $(\widehat{K}_{Peaks,01,01}, \widehat{T}_{Peaks,01,01})$, $(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$ or $(\widehat{K}_{Peaks,mad,01}, \widehat{T}_{Peaks,mad,01})$.

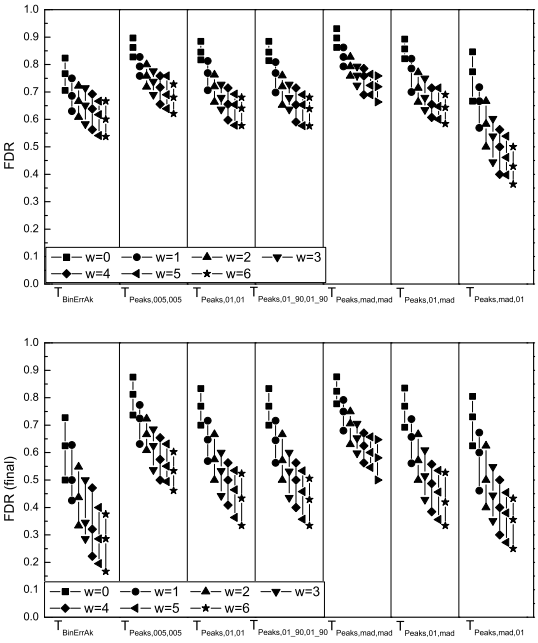

Fig. 4.6 FDR of all boundary estimators applied to dataset *B* computed on both the original estimates (on the top) and the final ones (at the bottom). The estimator $\hat{T}_{Peaks,mad,01}$ has the lowest FDR. Instead, $\widetilde{T}_{BinErrAk}$ achieves the lowest final FDR, followed by $\hat{T}_{Peaks,mad,01}$.

4.6.2 Comparisons on simulated data with LOH regions

In order to evaluate the IBD/UPD detection of gBPCR, we applied it to simulated data of [85]. These data are based on three real samples of the HapMap dataset [47], obtained with the Affymetrix GeneChip Mapping 250K NspI. For each sample and signal to noise ratio (SNR) value, they

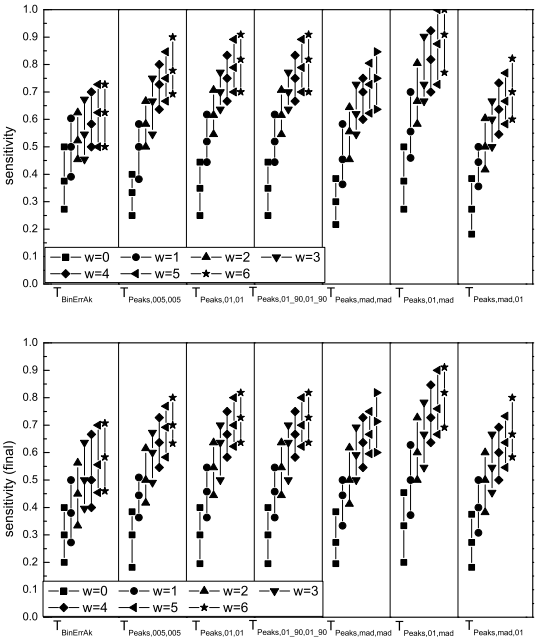

Fig. 4.7 Sensitivity of all boundary estimators applied to dataset *B* computed on both the original estimates (on the top) and the final ones (at the bottom). The estimator $\widehat{T}_{Peaks,01,mad}$ has the highest sensitivity (and also the highest final one).

simulated two profiles: one with regions of copy-neutral LOH and one with regions of loss. In both cases the number of regions was 50 and their width ranged from 20 SNPs to a whole chromosome. The values of SNR considered were: 5, 2 and 1.25. The simulated samples were in .CEL file format, thus we used BRLMM [1] to extract the genotyping data and CNAT 4.01 [2] for the raw copy number data.

Similar to [85], we compared the estimation of gBPCR with the ones given by two well-known methods in the field: dChip [8] and CNAT 4.01 [2], that we have both described in Section 2.3. The evaluation has been done computing the true positive rate (TPR) and the false positive rate (FPR), i.e. the proportion of SNPs inside the LOH regions that are correctly identified (as belonging to a LOH region) and the proportion of SNPs outside these segments that are wrongly identified (as belonging to them), respectively. We also calculated the average 0-1 error over the SNPs (the 0-1 error is zero, if the SNP is correctly classified, and one, otherwise). We used $(\widehat{K}_{Peaks,01,01}, \widehat{T}_{Peaks,01,01})$, $(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$ or $(\widehat{K}_{Peaks,mad,01}, \widehat{T}_{Peaks,mad,01})$ as paired estimators of the number of segments and the boundaries, and either $p_{upd} = 10^{-3}$ or $p_{upd} = 10^{-4}$ as the prior probability of IBD/UPD.

Since CNAT does not consider the *NoCall* SNPs (called *non-informative* SNPs) for the estimation of the LOH profile, first we compare the TPR and FPR computed using only the informative SNPs.

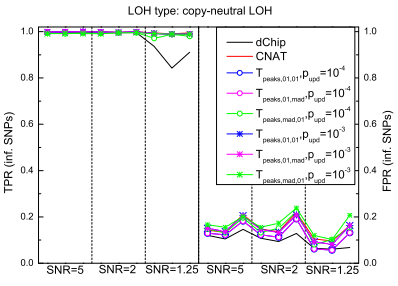

Fig. 4.8 TPR and FPR (computed only on informative SNP) of all methods applied to the samples, with regions of copy-neutral LOH, of the dataset in [85]. The FPR is almost always below 0.2, for all methods. All methods, apart from dChip, always have a TPR close to 1. The three points per SNR correspond to the three samples used.

Regarding the IBD/UPD detection (see Figure 4.8), all methods maintained a similar and low FPR (usually below 0.2), for all samples and all SNRs. The TPR of CNAT and all versions of gBPCR was always closed to one, while dChip achieved a lower TPR (about 0.9) in the samples with

SNR = 1.25. The average 0-1 error (over the informative SNPs) of all methods was similar (see Figure 4.10).

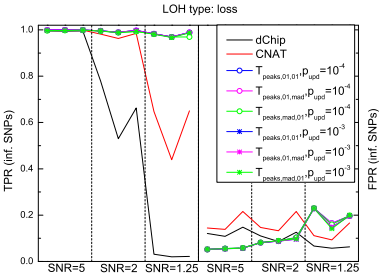

Fig. 4.9 TPR and FPR (computed only on informative SNP) of all methods applied to the samples, with regions of loss, of the dataset in [85]. The FPR is almost always below 0.2, for all methods. The TPR of dChip and CNAT decreases as the noise increases. Instead, all versions of gBPCR always maintain a TPR close to 1. The three points per SNR correspond to the three samples used.

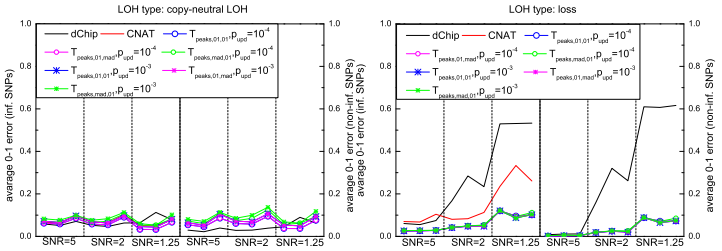

Fig. 4.10 Average 0-1 error (computed on both informative and non-informative SNPs) of all methods applied to the samples of the dataset in [85]. All methods have a similar error in the samples with regions of copy-neutral LOH. In the samples with regions of loss, the error of gBPCR is always small and, when SNR=1.25, the errors of CNAT and dChip are more than twice the one of gBPCR.

In the estimation of losses (see Figure 4.9), we observed that, again, the FPR was always below or close to 0.2. In case of dChip and CNAT, we saw that the FPR decreased as the noise increased, while the opposite occurred for the versions of gBPCR. It is natural to observe an increasing of the FPR with the noise, because the higher the noise, the more difficult it is to perform the estimation. Therefore, the unnatural behavior of the FPR of dChip and CNAT is related to the fact that they lose in "power of detection" in presence of high noise. In fact, also their TPR decreased as the noise increased and dChip even achieved a TPR close to zero in the samples with SNR = 1.25. Only gBPCR maintained a TPR always close to one. We also observe that the average 0-1 error (over the informative SNPs) of CNAT was at least twice the one of gBPCR in the samples with SNR = 1.25 (see Figure 4.10).

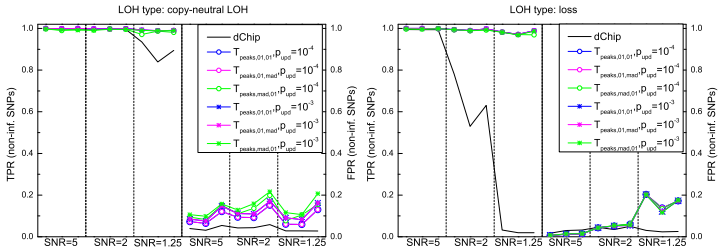

Fig. 4.11 TPR and FPR (computed only on non-informative SNP) of all methods applied to the samples of the dataset in [85]. All versions of gBPCR have a higher TPR than dChip in samples with high noise (especially in the samples with regions of loss), while the FPR of dChip is usually lower than gBPCR. Among the versions of gBPCR, $(\widehat{K}_{Peaks,01,01}, \widehat{T}_{Peaks,01,01})$ and $(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$ with $p_{upd} = 10^{-4}$ more often achieve a lower FDR than the others. The three points per SNR correspond to the three samples used.

Using only the non-informative SNPs, we obtained that the TPRs, regarding the estimation of both the copy-neutral LOH and the loss, were similar to the ones previously described, for both dChip and all versions of

gBPCR (see Figures 4.11). The FPR of dChip never exceeded 0.06 and the one of gBPCR was usually lower than 0.2. Moreover, among the versions of gBPCR, $(\widehat{K}_{Peaks,01,01}, \widehat{T}_{Peaks,01,01})$ and $(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$ with $p_{upd} = 10^{-4}$ more often achieved a lower FDR than the others.

Globally, all versions of gBPCR behaved similarly on these data and they outperformed CNAT and dChip. We also observed that dChip failed to give a good estimation in presence of high noise. Due to the results obtained on the non-informative SNPs, we suggest to use $(\widehat{K}_{Peaks,01,01}, \widehat{T}_{Peaks,01,01})$ or $(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$ with $p_{upd} = 10^{-4}$.

Finally, on these data we observed the effect of the adjustment of the model parameters related to the *NoCall* detection (see Section 4.5). At low or medium noise, we could not see any significant differences in the goodness of the estimation (see, for example, Figure 4.12). Instead, in presence of high noise, the FPR regarding the IBD/UPD detection without the adjustment of the model parameters was close to one. In fact, in this situation a segment with normal copy number is more often classified as IBD/UPD, since the *NoCall* rate is higher and, without the correction, the IBD/UPD segments are allowed to contain a higher percentage of *NoCall*s with respect to the normal ones. Instead, with the adjustment, we allow all types of regions to have a higher number of *NoCall*s in proportion to the noise, obtaining a less biased estimation.

4.7 Application to real data

In this section, we show the behavior of gBPCR on real data. We will use $(\hat{K}_{Peaks,01}, \hat{T}_{Peaks,01})$, $(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$ or $(\hat{K}_{Peaks,mad,01}, \hat{T}_{Peaks,mad,01})$ as paired estimators of the number of segments and the boundaries, and either $p_{upd} = 10^{-3}$ or $p_{upd} = 10^{-4}$ as prior probability of IBD/UPD.

One of the two datasets we used consisted of paired samples of patients affected by chronic lymphocytic leukemia (CLL), which then progressed in diffuse large B-cell lymphoma (DLBCL), see [70, 71]. For two patients

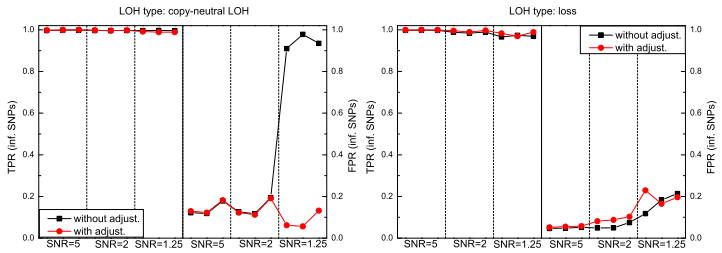

Fig. 4.12 TPR and FPR (computed only on informative SNP) of the versions of gBPCR, which use $(\hat{K}_{Peaks,01,01}, \hat{T}_{Peaks,01,01})$, $p_{upd} = 10^{-4}$ and with or without the adjustment of the parameters, applied to the samples of the dataset in [85]. In case of high noise, the version without the correction has a FPR close to 1, in the detection of copy-neutral LOH regions. The three points per SNR correspond to the three samples used.

we had also a third sample. In general, samples coming from the same patient should present the same copy-neutral LOH regions (the germ line ones) for the majority of the genome. Hence, we used them to evaluate the IBD/UPD detection of our method. The second dataset consisted of 18 patients affected by CLL, see [20].

In [20, 70, 71], the copy number of some genomic regions was also measured with FISH. Therefore, on these regions we compared the copy number event estimation of our procedure.

Results regarding the copy number event identification

We recall that a sample of a patient can contain also normal cells and/or other subpopulations of tumor cells in a subsequent stage of the disease. Therefore, some aberrations may be present in just a small percentage of the cells. In fact, we can see in Figure 4.13 that the $\log_2$ratio values corresponding to normal, gain, loss regions are sufficiently well separated only

when the copy number changes are borne in at least 60% of the cells in the sample. As a consequence, we aim to detect the copy number changes borne in at least the 60% of the cells, otherwise we cannot ensure that the identified aberrations are true and not due to the noise of microarray data (the noise is so high that aberrations borne in only a small percentage of cells can be seen as noise and viceversa). To detect aberrations in even less cell content, it is sufficient to change the prior of Z with thresholds closer to zero. In practice, the prior of Z influences stronger the discovery of the gains than the one of the other copy number events, because the determination of gains depends mainly on the estimated $\log_2$ratio values (instead of the LOH data).

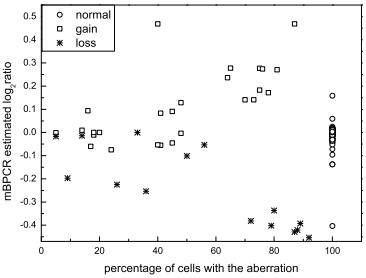

Fig. 4.13 Example of CN event classification. The plot shows the estimated $\log_2$ratio values (given by mBPCR), as function of the estimated percentage of cells bearing the aberration (given by FISH). The CN changes are classified as loss, gain or normal, using the results given by the FISH. For the normal regions, we set the percentage as 100%. We can observe that only the aberrations borne in at least 60% of the cells are clearly separated. For the graph, we used the data of the patients in [70, 71].

In the samples considered for the comparison, we had a total of 169 regions measured by FISH (which provides also an estimate of the percentage of cells bearing the aberration): 38 regions were gains or losses in at least 60% of the cells (called *detectable aberrations*), 33 were gains or losses in less than 60% and 98 were identified as normal segments. Regarding the detectable aberrations, only two copy number events were not

identified by all versions of the method. One loss was not found, because the estimated $\log_2$ratio was very close to zero, and the other, because of a different percentage of *Het* SNPs from what expected by the algorithm.

All versions found 11 of the 33 less detectable copy number changes and other 2 were discovered by $(\hat{K}_{Peaks,01,01}, \hat{T}_{Peaks,01,01})$ and $(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$ (with both values of p_{upd}). In two of the 98 normal segments, all our estimators discovered an aberration, but one of these copy number changes was equal to the one discovered in the same region of the paired sample, thus it was likely to be true.

Instead, by simply using the thresholds of the prior of Z for the classification of the copy number events (similarly to what is usually done), we detected one alteration less than what found by $(\hat{K}_{Peaks,01,01}, \hat{T}_{Peaks,01,01})$ and $(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$, and other 5 normal regions were seen as aberrations.

For the analysis of the results, we have to consider that the samples used for FISH came from peripheral blood, for the CLL samples, and from paraffin embedded tissues or lymph node, for the DLBCLs. Because of the consequently different cell content, in the former case, the results are better estimated. Moreover, the samples used for microarray and FISH might not be exactly the same, hence the percentage of cells which carry the aberrations can be different and a discordance between the two techniques is possible.

In conclusion, all the versions behaved similarly and equally good in estimating the copy number changes on these samples. The best performing estimators were $(\hat{K}_{Peaks,01,01}, \hat{T}_{Peaks,01,01})$ and $(\widehat{K}_{Peaks,01,mad}, \widehat{T}_{Peaks,01,mad})$ (with both values of p_{upd}).

Results regarding IBD/UPD region detection

For the evaluation of the IBD/UPD region detection, we considered the two patients with three samples. For the first patient (called Patient 1), we had: one sample from normal cells in peripheral blood (called sample 1.1),

one from neoplastic cells in peripheral blood at CLL phase (called sample 1.2) and the last one from neoplastic cells in lymph node at DLBCL phase (called sample 1.3). For the second patient (Patient 2), we had: one sample from neoplastic cells in peripheral blood at CLL phase (called sample 2.1), one from neoplastic cells in lymph node at DLBCL phase (called sample 2.2) and the last one from neoplastic cells in peripheral blood at a further progression of the DLBCL (called sample 2.3).

Applying the six versions of the method to the three samples of Patient 1, we found that the number of aberrations in each sample increased with the progression of the disease. The lower number of segments discovered in sample 1.1 could be also due to a higher *NoCall* rate in comparison to the other samples. The same happened for sample 2.3 of Patient 2.

We compared the IBD/UPD segments found in the three samples of each patient and we divided them into three classes (see Table 4.8):

- equal regions: segments that are exactly the same in two or three samples;
- overlapping regions: segments that are common in at least two samples but do not have the same boundaries;
- single sample regions: the remaining segments.

Then, we defined the number of distinct regions as the sum of all these regions and the number of validated ones as the sum of all types of regions except the single sample regions. For both patients, the lowest number of distinct regions was found by all estimators with $p_{upd} = 10^{-4}$, while the highest by $(\hat{K}_{Peaks,mad,01}, \hat{T}_{Peaks,mad,01})$ with $p_{upd} = 10^{-3}$. In general, the estimator $(\hat{K}_{Peaks,mad,01}, \hat{T}_{Peaks,mad,01})$ (with both values of p_{upd}) gave the highest proportion of equal regions (24-30%) and, consequently, the lowest proportion for the overlapping regions (49-52%). For all the methods, the single sample regions were about the 20% of the distinct regions in Patient 2, but the majority of them had length less than 50 SNPs. Instead, since the samples of Patient 1 belonged to different stages of the disease, in this patient we found a higher number of single sample regions and most of them were wider than 50 SNPs. In fact, the majority of these regions was detected in sample 1.3, thus they were likely to be somatic. In gen-

Table 4.8 Results regarding the IBD/UPD region detection, obtained on two patients using the three pairs of estimators $(\hat{K}_{Peaks,01,01}, \hat{T}_{Peaks,01,01})$, $(\hat{K}_{Peaks,01,mad}, \hat{T}_{Peaks,01,mad})$ and $(\hat{K}_{Peaks,mad,01}, \hat{T}_{Peaks,mad,01})$ and, as probability of IBD/UPD, either $p_{upd} = 10^{-4}$ or $p_{upd} = 10^{-3}$.

Patient 1:						
types of regions	$p_{upd} = 10^{-4}$			$p_{upd} = 10^{-3}$		
	01,01	01,*mad*	*mad*,01	01,01	01,*mad*	*mad*,01
in sample 1.1 (total)	213	213	212	330	330	351
in sample 1.2 (total)	337	337	347	443	443	480
in sample 1.3 (total)	391	391	412	468	468	511
distinct (total)	438	438	456	567	567	604
equal (%)	22	22	24	21	21	25
overlapping (%)	52	52	49	53	53	52
validated (%)	73	73	73	74	74	77
single sample (%)	27	27	27	26	26	23
% of single sample $<$ 50 SNPs	28	28	30	53	53	48
Patient 2:						
in sample 2.1 (total)	376	376	400	470	470	511
in sample 2.2 (total)	384	384	401	400	470	514
in sample 2.3 (total)	177	177	177	292	292	306
distinct (total)	438	438	461	555	555	603
equal (%)	27	27	30	26	26	28
overlapping (%)	52	52	49	54	54	52
validated (%)	79	79	79	80	80	80
single sample (%)	21	21	21	20	20	20
% of single sample $<$ 50 SNPs	62	62	58	68	68	67

eral, we validated about 73-77% of the regions detected in Patient 1 and $\sim$79-80% of the regions in Patient 2.

Finally, we observed that in both patients the number of identified regions was higher using $p_{upd} = 10^{-3}$ than $p_{upd} = 10^{-4}$. For the first patient, the average length of the regions was $\sim$77 SNPs with $p_{upd} = 10^{-3}$ and $\sim$87 SNPs with $p_{upd} = 10^{-4}$, while for the second $\sim$74 SNPs with $p_{upd} = 10^{-3}$ and $\sim$82 SNPs with $p_{upd} = 10^{-4}$. The estimators $(\hat{K}_{Peaks,01},$

$\hat{T}_{Peaks,01}$) and ($\widehat{K}_{Peaks,01,mad}$, $\widehat{T}_{Peaks,01,mad}$) gave always the maximum average length (measured in SNPs). Although the average length of this kind of DNA lesions has not been defined yet, it is better to be more conservative (i.e. find a lower number of regions, but with a higher number of SNPs) to avoid false positives. Considering the latter and taking into account the previous results, we suggest to use $p_{upd} = 10^{-4}$ and preferably ($\hat{K}_{Peaks,01,01}$, $\hat{T}_{Peaks,01,01}$) or ($\widehat{K}_{Peaks,01,mad}$, $\widehat{T}_{Peaks,01,mad}$).

4.8 Computation of the posteriors and dynamic programming

Since we assume a uniform prior distribution for the boundaries (see Subsection 4.1.2), we can use the same recursion employed in mBPCR to calculate the posterior distributions (see Section 3.3). Then, for computational purpose, we only need to make explicit the formula of A_{ij}^0 (the conditional density of Y_{ij} at y_{ij}, given *cn* and knowing that Y_{ij} belongs to only one segment called *p*).

First, let us consider Model 1. Conditioning with respect to Z_p and using the independence of the data points Y_s ($s = i+1, \ldots, j$) given Z_p, we obtain

$$\begin{aligned}
\mathrm{A}_{ij}^0 &= p(y_{ij} \,|\, K_{ij} = 1, T_{p-1} = i, T_p = j, cn) \\
&= \sum_{z \in \{-2,-1,0,2\}} p(y_{ij}, Z_p = z \,|\, K_{ij} = 1, T_{p-1} = i, T_p = j, cn) \\
&= \sum_{z \in \{-2,-1,0,2\}} p(y_{ij} \,|\, Z_p = z, K_{ij} = 1, T_{p-1} = i, T_p = j) \mathrm{G}_{i,j}(z) \\
&= \sum_{z \in \{-2,-1,0,2\}} \prod_{s=i+1}^{j} p(y_s \,|\, \widetilde{Z}_s = z) \mathrm{G}_{i,j}(z), \qquad (4.14)
\end{aligned}$$

where $\mathrm{G}_{i,j}(z) = \mathrm{P}(Z_p = z \,|\, K_{ij} = 1, T_{p-1} = i, T_p = j, cn)$. The computation of $\mathrm{G}_{i,j}(z)$, given *cn*, can be done by using Bayes Theorem and the equiva-

lence between the events $\{K_{ij}=1\}$ and $\{\widetilde{Z}_s = Z_p = z \text{ for all } s = i+1, \ldots, j, \text{ for some } z \in \{-2,-1,0,2\}\}$,

$$\begin{aligned}
\mathrm{G}_{i,j}(z) &= \mathrm{P}(Z_p = z \,|\, K_{ij}=1, T_{p-1}=i, T_p=j, cn) \\
&= \mathrm{P}\left(\bigcap_{s=i+1}^{j} \{\widetilde{Z}_s = z\} \,|\, K_{ij}=1, T_{p-1}=i, T_p=j, cn \right) \\
&= \frac{\mathrm{P}(\bigcap_{s=i+1}^{j}\{\widetilde{Z}_s = z\}, K_{ij}=1 \,|\, T_{p-1}=i, T_p=j, cn)}{\mathrm{P}(K_{ij}=1 \,|\, T_{p-1}=i, T_p=j, cn)} \\
&= \frac{\mathrm{P}(\bigcap_{s=i+1}^{j}\{\widetilde{Z}_s = z\} \,|\, cn)}{\sum_{z\in\{-2,-1,0,2\}} \mathrm{P}(\bigcap_{s=i+1}^{j}\{\widetilde{Z}_s = z\} \,|\, cn)},
\end{aligned} \tag{4.15}$$

where,

$$\mathrm{P}\left(\bigcap_{s=i+1}^{j} \{\widetilde{Z}_s = z\} \,\middle|\, cn \right) = \prod_{\hat{t}^{cn}_{\tilde{q}} \in \mathscr{T}_{i,j}} \mathrm{P}(Z_{\hat{t}^{cn}_{\tilde{q}}} = z \,|\, cn), \tag{4.16}$$

with $\mathscr{T}_{i,j} = \{\hat{t}^{cn}_{\tilde{q}} \mid \hat{t}^{cn}_{\tilde{q}} \in [i+1, j),\ \tilde{q} = 1, \ldots, \hat{k}^{cn}\} \cup \min\{\hat{t}^{cn}_{\tilde{q}} \mid \hat{t}^{cn}_{\tilde{q}} \geq j,\ \tilde{q} = 1, \ldots, \hat{k}^{cn}\}$ (we denote the cardinality of $\mathscr{T}_{i,j}$ with $|\mathscr{T}_{i,j}| =: N_{\mathscr{T},i,j}$ and the indices of the boundaries in $\mathscr{T}_{i,j}$ with $\tilde{q}_1, \ldots, \tilde{q}_{N_{\mathscr{T},i,j}}$). Notice that the prior probability of $\widetilde{Z}_{ij}$ is based on the copy number estimation and that only the $\widetilde{Z}_s$ belonging to different segments are independent. Therefore, in Equation (4.16) we partitioned the interval $[i+1, j]$ in subintervals $\{I_p\}_{p=1}^{N_{\mathscr{T},i,j}}$, by using the boundaries $\{\hat{t}^{cn}_{\tilde{q}} \mid \hat{t}^{cn}_{\tilde{q}} \in [i+1, j], \tilde{q} = 1, \ldots, \hat{k}^{cn}\}$:

$$\begin{aligned}
I_1 &= [i+1, \hat{t}^{cn}_{\tilde{q}_1}] \subseteq (\hat{t}^{cn}_{\tilde{q}_0}, \hat{t}^{cn}_{\tilde{q}_1}] \\
I_p &= (\hat{t}^{cn}_{\tilde{q}_{p-1}}, \hat{t}^{cn}_{\tilde{q}_p}], \qquad p = 2, \ldots, N_{\mathscr{T},i,j} - 1, \\
I_{N_{\mathscr{T},i,j}} &= \left(\hat{t}^{cn}_{\tilde{q}_{N_{\mathscr{T},i,j}-1}}, j\right] \subseteq \left(\hat{t}^{cn}_{\tilde{q}_{N_{\mathscr{T},i,j}-1}}, \hat{t}^{cn}_{\tilde{q}_{N_{\mathscr{T},i,j}}}\right],
\end{aligned}$$

where $\hat{t}^{cn}_{\tilde{q}_0} = \hat{t}^{cn}_{\tilde{q}_1 - 1}$. We can notice that, by definition, any interval I_p of $[i+1, j]$ is either equal or contained in a segment of the partition generated by the estimated $\log_2$ratio profile (called $I_{\tilde{q}_p} = (\hat{t}^{cn}_{\tilde{q}_{p-1}}, \hat{t}^{cn}_{\tilde{q}_p}]$). Consequently,

$$\mathrm{P}\left(\bigcap_{s\in I_p}\{\widetilde{Z}_s=z\}\,\middle|\,cn\right)=\mathrm{P}\left(\bigcap_{s\in I_{\hat{q}p}}\{\widetilde{Z}_s=z\}\,\middle|\,cn\right)=\mathrm{P}\left(Z_{\hat{t}^{cn}_{\hat{q}p}}=z\,|\,cn\right),$$

and thus,

$$\begin{aligned}\mathrm{P}\left(\bigcap_{s=i+1}^{j}\{\widetilde{Z}_s=z\}\,\middle|\,cn\right)&=\prod_{p=1}^{N_{\mathscr{T},i,j}}\mathrm{P}\left(\bigcap_{s\in I_p}\{\widetilde{Z}_s=z\}\,\middle|\,cn\right)\\&=\prod_{\hat{t}^{cn}_{\hat{q}}\in\mathscr{T}_{i,j}}\mathrm{P}(Z_{\hat{t}^{cn}_{\hat{q}}}=z\,|\,cn).\end{aligned}$$

Using the dynamic programming defined in Section 3.3, we can compute both $p(k\,|\,Y,cn)$ and $p(t\,|\,Y,cn)$ and thus estimate k_0 and t^0 as described in Subsection 4.1.4. Finally, from Equations (4.5) and (4.14), it follows that the posterior distribution of Z_p can be written as

$$\mathrm{P}(Z_p=z\,|\,y,\hat{t},\hat{k},cn)=\frac{\prod_{i=\hat{t}_{p-1}+1}^{\hat{t}_p}\mathrm{P}(y_i\,|\,\widetilde{Z}_i=z)\mathrm{G}_{\hat{t}_{p-1},\hat{t}_p}(z)}{\mathrm{A}^0_{\hat{t}_{p-1},\hat{t}_p}}\tag{4.17}$$

for $z=-2,-1,0,2$, and we can derive the MAP estimate of Z_p, for each $p=1,\ldots,\hat{k}$.

If we consider Model 2, Equation (4.14) for the computation of the quantity A^0_{ij} becomes,

$$\begin{aligned}\mathrm{A}^0_{ij}&=\mathrm{P}(\underline{y}_{ij}\,|\,K_{ij}=1,T_{p-1}=i,T_p=j,cn)\\&=\sum_{z\in\{-2,-1,2\}}\prod_{s=i+1}^{j}\mathrm{P}(y_s\,|\,\widetilde{Z}_s=z)\mathrm{G}_{i,j}(z)+\Bigg[\prod_{s=i+1}^{j}\mathrm{P}(y_s\,|\,\widetilde{Z}_s=0,\widetilde{U}_s=1)\\&\quad\cdot\ p_{upd}+\prod_{s=i+1}^{j}\mathrm{P}(y_s\,|\,\widetilde{Z}_s=0,\widetilde{U}_s=0)(1-p_{upd})\Bigg]\mathrm{G}_{i,j}(0).\end{aligned}\tag{4.18}$$

Moreover, for any $p=1,\ldots,\hat{k}$, the posterior probability of W_p is

$$\mathrm{P}(W_p = w \,|\, \underline{y}, \underline{\hat{t}}, \hat{k}, cn) = \mathrm{P}(Z_p = w \,|\, \underline{y}, \underline{\hat{t}}, \hat{k}, cn),$$

for w = -2, -1, 2, and

$$\begin{aligned}&\mathrm{P}(W_p = w \,|\, \underline{y}, \underline{\hat{t}}, \hat{k}, cn)\\ &= p_{upd}^{-w/3}(1-p_{upd})^{(3+w)/3}\frac{\prod_{i=\hat{t}_{p-1}+1}^{\hat{t}_p}\mathrm{P}(y_i \,|\, \widetilde{W}_i = w)\mathrm{G}_{\hat{t}_{p-1},\hat{t}_p}(0)}{\mathrm{A}^0_{\hat{t}_{p-1},\hat{t}_p}},\end{aligned}$$

for $w = -3, 0$, by using a derivation similar to the one of Equation (4.17).

Since we assume that there is no difference in the genotype detection between normal and gained regions, in Model 3 the probabilities $p(y \,|\, Z_p = 0, \hat{t}, \hat{k}, cn)$ and $p(y \,|\, Z_p = 1, \hat{t}, \hat{k}, cn)$ are equal. Therefore, the computation of A^0_{ij} in Equation (4.18) becomes

$$\begin{aligned}\mathrm{A}^0_{ij} &= \mathrm{P}(y_{ij} \,|\, K_{ij} = 1, T_{p-1} = i, T_p = j, cn)\\ &= \sum_{z\in\{-2,-1,2\}}\prod_{s=i+1}^{j}\mathrm{P}(y_s \,|\, \widetilde{Z}_s = z)\mathrm{G}_{i,j}(z) + \Bigg[\prod_{s=i+1}^{j}\mathrm{P}(y_s \,|\, \widetilde{Z}_s = 0, \widetilde{U}_s = 1)\\ &\cdot\ p_{upd} + \prod_{s=i+1}^{j}\mathrm{P}(y_s \,|\, \widetilde{Z}_s = 0, \widetilde{U}_s = 0)(1-p_{upd})\Bigg]\left[\mathrm{G}_{i,j}(0) + \mathrm{G}_{i,j}(1)\right].\end{aligned}$$

where now $G_{ij}(z)$ is calculated taking into account five classes of copy number events, instead of four as in Equation (4.15),

$$\mathrm{G}_{i,j}(z) = \frac{\mathrm{P}(\bigcap_{s=i+1}^{j}\{\widetilde{Z}_s = z\} \,|\, cn)}{\sum_{z=-2}^{2}\mathrm{P}(\bigcap_{s=i+1}^{j}\{\widetilde{Z}_s = z\} \,|\, cn)}.$$

Moreover, the posterior probabilities of $\{W_p = w\}$ are the same as before for $w = -2$, -1, 0, 2, while for $w = 1$,

$$\mathrm{P}(W_p = 1 \,|\, y, \hat{t}, \hat{k}, cn) = \mathrm{P}(Z_p = 1 \,|\, y, \hat{t}, \hat{k}, cn)$$

$$
\begin{aligned}
&= \frac{p(y_{\hat{t}_{p-1},\hat{t}_p} \,|\, Z_p = 1, \hat{t}, \hat{k}, cn)\mathrm{P}(Z_p = 1 \,|\, \hat{t}, \hat{k}, cn)}{p(y_{\hat{t}_{p-1},\hat{t}_p} \,|\, \hat{t}, \hat{k}, cn)} \\
&= \frac{\mathrm{P}(y_{\hat{t}_{p-1},\hat{t}_p} \,|\, Z_p = 0, \hat{t}, \hat{k}, cn)\mathrm{P}(Z_p = 1 \,|\, \hat{t}, \hat{k}, cn)}{\mathrm{A}^0_{\hat{t}_{p-1}\hat{t}_p}} \\
&= \Big[p(y_{\hat{t}_{p-1},\hat{t}_p} \,|\, Z_p = 0, U_p = 1, \hat{t}, \hat{k}, cn)\mathrm{P}(U_p = 1) \\
&+ \; p(y_{\hat{t}_{p-1},\hat{t}_p} \,|\, Z_p = 0, U_p = 0, \hat{t}, \hat{k}, cn)\mathrm{P}(U_p = 0)\Big] \\
&\cdot \; \frac{\mathrm{P}(Z_p = 1 \,|\, \hat{t}, \hat{k}, cn)}{\mathrm{A}^0_{\hat{t}_{p-1}\hat{t}_p}},
\end{aligned}
$$

by using Equation (4.5), the definition of A^0_{ij}, the equality between $p(y \,|\, Z_p = 0, \hat{t}, \hat{k}, cn)$ and $p(y \,|\, Z_p = 1, \hat{t}, \hat{k}, cn)$ (given by the previous discussion) and the conditioning with respect to U_p.

Conclusion

In this first part of the Thesis, I developed four statistical methods for the analysis of genomic data: mBPCR, mBRC, BRCAk and gBPCR. mBPCR, mBRC and BRCAk are three Bayesian regression algorithms for the estimation of the DNA copy number profile either as a piecewise constant function (mBPCR) or as a continuous curve (mBRC and BRCAk). These algorithms represent an improvement of the corresponding algorithms (BPCR and BRC) presented by Hutter in [32, 33], by changing the definition of most of the parameter estimators involved in the statistical procedure. The main changes regarded the estimation of the variance of the segment levels ρ^2 and the estimation of the breakpoints T because of the following issues:

- the original estimator of ρ^2 ($\hat{\rho}^2$) was biased for ρ^2 and almost asymptotically unbiased for the variance of the noise σ^2 when $\rho^2 \ll \sigma^2$ (which is usually the case in real data);
- the original estimator of T ($\widehat{T}_{01}$) did not take into account the dependency among the boundaries and could estimate multiple breakpoints at the same position, loosing segments.

Since ρ^2 represents also the covariance between any two data points belonging to the same segment, I defined a new estimator $\hat{\rho}^2_1$, which is similar to the estimator of the autocovariance of a stationary time series. $\hat{\rho}^2_1$ is

asymptotically unbiased for ρ^2. Regarding the boundaries, after applying a binary transformation on the vector of breakpoints, I defined a binary error (i.e. a measure of dissimilarities between binary vectors). The new estimator $\widehat{T}_{BinErrAk}$ is the inverse image (through the binary transformation) of the argument which minimizes the posterior expected value of the binary error.

I tested the performance of mBPCR versus BPCR and other well-known methods, for the estimation of the copy number profile as a piecewise constant function, and I showed that mBPCR outperformed all other algorithms on both simulated and real data. Moreover, I compared mBRC and BRCAk with BRC and other well-known smoothing methods for the analysis of copy number data, by using simulated datasets, and I obtained that the both proposed algorithms gave a better estimation of the profiles than the others. Finally, some results obtained with mBPCR on real data have been validated by using FISH technique.

Since several relationships can be found between the homozygous status measured by the microarray and the presence of an altered copy number, I developed a statistical method (gBPCR) for the estimation of both copy number and LOH aberrations by integrating the information given by both copy number and LOH data. To the best of my knowledge, only another algorithm exists in literature which uses the same input data for the same purpose [72]. But the latter is not able to handle with data whose DNA sample come from a mixture of cell populations (which is usually the case in cancer data). Instead, the Bayesian statistical model employed in gBPCR to describe the relationship among the random variables involved allows the identification of the aberrations present in only a percentage of the cells in the DNA sample.

The estimation procedure of gBPCR is a Bayesian regression method similar to mBPCR, with an improvement of the breakpoint estimator to obtain an highest sensitivity. I compared gBPCR with other two well-known algorithms for LOH estimation, by using simulated datasets. The results showed that gBPCR is comparable with the others on data with low and medium noise, but it outperformed them on data with high noise. Moreover, I applied gBPCR on samples coming from cancer patients, also di-

rectly providing useful data for cancer research, and I validated some of the results found.

In the last years, new algorithms have been developed for the preprocessing of SNP microarray data. Since now they allow also the estimation of the allelic copy number, in the future I would like to extend the model of gBPCR to account also this type of information which should be less noisy than the LOH data.

Part II
Estimators of the intensity of stationary fibre processes applied to angiogenesis

Chapter 5
Statistics of fibre processes

Fibre processes are random geometric objects that can be used in medicine, biology, material science, to model structures like capillaries, radices and nervature of fibrous material. For stationary processes (i.e. with distribution invariant under translations), a quantity that characterize the process is the density of its length (called *intensity*). Therefore, statistical methods for the estimation of the intensity of a fibre process may offer relevant tools for applications.

In [53, 64], we introduced some estimators of the intensity of a stationary two-dimensional fibre process, obtained by intersecting the fibre process under study with another (simulated) test fibre process in $\mathbb{R}^2$, satisfying sufficient regularity properties, and considering the associated counting measure of the intersection points. Asymptotic properties of the estimators and, especially, strong consistency are thus retrieved, based on the regularity of the test process (in particular, on ergodicity). The proposed estimators were a generalization of some estimators presented in literature in [58, 78], which had no asymptotic properties, since were obtained via the intersection with a finite deterministic fibre system.

In Section 5.1 we recall from the literature some basic definitions and notations related to fibre processes; in Section 5.2 we introduce the estimators present in literature which have inspired our estimators; in Section 5.3

we describe our estimators and their properties and in Section 5.4 we consider some particular examples of fibre processes of which the ergodicity is easily proven, and which thus can be used as test processes in the estimation procedure.

5.1 Preliminaries

In this section we will briefly introduce the basic definitions and concepts of the theory of fibre processes. A further description of the general theory of fibre processes can be found, for example, in [78], while we suggest [73] for the theory of weighted fibre processes and [15] for the general theory of random measures and for the Palm theory.

Definition 5.1. A **fibre** is a subset of $\mathbb{R}^2$ which can be represented as the image of a curve $\gamma(t) = (\gamma_1(t), \gamma_2(t))$, with $t \in [0,1]$, such that $\gamma : [0,1] \to \mathbb{R}^2$ is a one-to-one mapping with continuous derivative.

From its definition, a fibre has finite length and it cannot intersect itself (because γ is one-to-one). Thus, we can define the measure of a fibre as its length:

$$\mu_\gamma(B) = \int_0^1 \mathbb{I}_B(\gamma(t)) \sqrt{(\gamma_1'(t))^2 + (\gamma_2'(t))^2}\, dt \qquad \forall B \in \mathfrak{B}^2,$$

(here and in the following, we will denote by $\mathfrak{B}^k$ the Borel σ-algebra in $\mathbb{R}^k$).

Definition 5.2. A **fibre system** is a set $\varphi \subset \mathbb{R}^2$ which can be represented as a union of at most countably many fibres $\gamma^{(i)}$, with the property that any compact set is intersected by only a finite number of fibres and such that distinct fibres have only endpoints in common, i.e. $\gamma^{(i)}((0,1)) \cap \gamma^{(j)}((0,1)) = \emptyset$ for $i \neq j$.

This definition provides computational advantages, since from real (2D) images, in general, it would be difficult to assign in a unique way

different branches of an intersection of two lines to two different fibres. Moreover, the definition implies that we can define the measure corresponding to a fibre system φ as the sum of the measures of its fibres $\gamma^{(i)}$,

$$\mu_\varphi(B) = \sum_{\gamma^{(i)} \in \varphi} \mu_{\gamma^{(i)}}(B) \qquad \forall B \in \mathfrak{B}^2.$$

It can be shown that the measure μ_φ is σ-finite, i.e. is finite for every bounded Borel set (see [78]).

Let $\mathscr{S}$ be the set of all possible fibre systems and $\mathfrak{S}$ be the σ-algebra of subsets of $\mathscr{S}$ generated by the sets $\left\{\varphi \in \mathscr{S} : \mu_\varphi(B) \in C\right\}$, with $B \in \mathfrak{B}^2, C \in \mathfrak{B}^1$.

Definition 5.3. A **fibre process** Φ is a random variable over a probability space $(\Omega, \mathfrak{F}, \mathscr{P})$ assuming values in $(\mathscr{S}, \mathfrak{S})$. The distribution of Φ is the probability measure P induced on $(\mathscr{S}, \mathfrak{S})$. We will denote by μ_Φ the random measure associated to Φ.

Remark 5.1. From the previous definition of a fibre process, we can identify the fibre process with its random measure and study the distribution of μ_Φ instead of P to characterize the process.

Another approach for defining and studying a fibre process is the one introduced by Zähle [86]. The previous definitions of a fibre and a fibre system allow to use basic tools of differential geometry to define and compute their lengths. Alternatively, we can model a system of fibres as a Hausdorff ($\mathscr{H}^1$) rectifiable set of dimension 1 [6, 86] and a fibre process can be considered a random set. This definition allows to consider more general geometric objects and other estimation procedures, than the ones proposed in the following sections, could be deduced (see for example [64]). Anyway, since the purpose of this thesis is to concentrate on the asymptotic properties of three particular estimators (see Sections 5.2 and 5.3), we will stick to our definitions to model a fibre process, because they are the most suitable for our study.

The most important properties that a fibre process may have are stationarity, isotropy and ergodicity. To define these properties, we need to

introduce the definition of translation and rotation of a fibre process, and of the corresponding measures, as follows,

$$\begin{aligned}
&\mathrm{T}_x\phi = \{y-x : y \in \phi\} \Rightarrow \mu_{\mathrm{T}_x\phi}(B) = \mu_\phi(\mathrm{T}_x B)\\
&\mathrm{R}_\beta\phi = \{\mathrm{R}_\beta y : y \in \phi\} \Rightarrow \mu_{\mathrm{R}_\beta\phi}(A) = \mu_\phi(\mathrm{R}_{-\beta}A)\\
&\qquad \text{where } \mathrm{R}_\beta(x_1,x_2) = (x_1\cos(\beta) - x_1\sin(\beta), x_1\sin(\beta) + x_2\cos(\beta)).
\end{aligned}$$

Definition 5.4.

1. A fibre process is **stationary** if its distribution P is stationary, i.e. $\mathrm{P}(\mathrm{T}_x A) = \mathrm{P}(A)$ for each $x \in \mathbb{R}^2, A \in \mathfrak{S}$.
2. A fibre process is **isotropic** if its distribution P is isotropic, i.e. $\mathrm{P}(\mathrm{R}_\beta A) = \mathrm{P}(A)$ for each $\beta \in \mathbb{R}, A \in \mathfrak{S}$.
3. A fibre process is **ergodic** if its distribution P is ergodic, i.e. for any $A \in \mathfrak{S}$, such that $\mathscr{P}(\mathrm{T}_x A \cap A) = \mathscr{P}(A)$ for each $x \in \mathbb{R}^2$, then $\mathscr{P}(A) = 0$ or 1.

Any random measure can be characterized by its moments. In the following we will use only the first and the second moment measure, called M_1 and M_2, respectively, which are defined as follows,

$$\begin{aligned}
\mathrm{E}\left[\int_{\mathbb{R}^2} f(x)\mu_\Phi(dx)\right] &=: \int_{\mathbb{R}^2} f(x)\,\mathrm{M}_1(dx)\\
\mathrm{E}\left[\int_{\mathbb{R}^2}\int_{\mathbb{R}^2} g(x,y)\mu_\Phi(dx)\mu_\Phi(dy)\right] &=: \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} g(x,y)\,\mathrm{M}_2(dx\times dy),
\end{aligned}$$

for all non-negative measurable functions f on $\mathbb{R}^2$ and g on $\mathbb{R}^2\times\mathbb{R}^2$.

In particular, when $f(x) = \mathbb{I}_B(x)$ and $g(x,y) = \mathbb{I}_B(x)\mathbb{I}_B(y)$, where $B \in \mathfrak{B}^2$ and $\mathbb{I}_B$ is its corresponding indicator function, we obtain that $\mathrm{M}_1(B) = \mathrm{E}[\mu_\Phi(B)]$ and $\mathrm{M}_2(B\times B) = \mathrm{E}\left[(\mu_\Phi(B))^2\right]$, i.e. the first moment measure is the expected length of the process in the Borel set B and the second moment measure is the second moment of the random variable $\mu_\Phi(B)$.

In case of stationary fibre processes, the Radon-Nikodym derivative of M_1 with respect to the Lebesgue measure is constant and is called *intensity* (see Theorem 3.2 in [51]). We will denote it by L_A. Moreover, the second

moment measure can be factorized in the following way (see [77])

$$\mathrm{M}_2(dx_1 \times dx_2) = \mathrm{L}_A^2\, dx \times \mathscr{K}(dh) \qquad \text{with } h = x_2 - x_1 \text{ and } x = x_1,$$

where $\mathscr{K}(dh)$ is called *reduced second moment measure*. As observed in [77], this measure can be described by the following function

$$\mathrm{K}(r, \alpha) := \mathscr{K}(s(0, r, \alpha)), \qquad \text{with } r \geq 0,\ \alpha \in [0, 2\pi), \tag{5.1}$$

where $s(0, r, \alpha)$ denotes a circular sector with radius r and angle α centered at the origin. By applying the Palm theory, it can be easily seen that the quantity $\mathrm{L}_A\, \mathrm{K}(r, \alpha)$ is the mean total length of the fibres in a sector with radius r and angle α centered at a "typical" fibre point (see [77]). If the process is also isotropic, K does not depend on α and in this case we define $\mathrm{K}(r) := \mathrm{K}(r, 2\pi)$.

If the second moment measure of a fibre process is finite, we can also define its variance in the following way: given $B \in \mathfrak{B}^2$, $\mathrm{Var}(\Phi(B)) = \mathrm{M}_2(B \times B) - (\mathrm{M}_1(B))^2$. Moreover, if the process is stationary,

$$\begin{aligned} \mathrm{Var}(\Phi(B)) &= \mathrm{L}_A^2 \left(\int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \mathbb{I}_B(x)\, \mathbb{I}_B(x+h)\, dx\, \mathscr{K}(dh) - (\nu_2(B))^2 \right) \\ &= \mathrm{L}_A^2 \left(\int_{\mathbb{R}^2} \nu_2(B \cap \mathrm{T}_h B)\, \mathscr{K}(dh) - (\nu_2(B))^2 \right), \end{aligned} \tag{5.2}$$

and if it is also isotropic

$$\mathrm{Var}(\Phi(B)) = \mathrm{L}_A^2 \left(\int_0^\infty \int_0^{2\pi} \nu_2(B \cap B_{(r,\alpha)})\, \frac{d\alpha}{2\pi}\, \mathrm{K}(dr) - (\nu_2(B))^2 \right).$$

For a stationary fibre process, it is possible to define another important quantity: the *angle distribution* ϑ_{P}. Let $\mathscr{S}_0$ be the set of all $\varphi \in \mathscr{S}$ which contain the origin and whose tangent at the origin is uniquely determined, then we define the direction of the fibre at the origin 0 via the following function $w : \mathscr{S} \to [0, \pi]$:

$$w(\varphi) = \begin{cases} \text{the angle in } [0,\pi) \text{ formed in the} & \text{if } \varphi \in \mathscr{S}_0 \\ \text{upper half-plane by the tangent of } \varphi \text{ at } 0 & \\ \text{and the positive abscissa axis} & \\ \pi & \text{if } \varphi \in \mathscr{S} \setminus \mathscr{S}_0. \end{cases}$$

As a consequence, $w(\mathrm{T}_x\varphi)$ denotes the direction of the fibre system φ at point $x \in \varphi$. The distribution ϑ_{P} can be seen as the distribution of the direction in a "typical" fibre point and its definition comes from the following theorem:

Theorem 5.1 (Theorem 3.3 in [51]). *Let Φ be a stationary fibre process with intensity L_A. Then there exists a unique probability measure ϑ_{P} on $[0,\pi)$ such that for all measurable functions $f : \mathbb{R}^2 \times [0,\pi] \to [0,\infty)$,*

$$\int_{\mathscr{S}} \int_{\mathbb{R}^2} f(x, w(\mathrm{T}_x\varphi))\, \mu_\varphi(dx)\, \mathrm{P}(d\varphi) = \mathrm{L}_A \int_{\mathbb{R}^2} \int_{[0,\ \pi)} f(x,\ \alpha)\, \vartheta_{\mathrm{P}}(d\alpha)\, dx.$$

When the fibre process is motion invariant (i.e. stationary and isotropic), the angle distribution is uniform on $[0,\pi)$ (see Theorem 5.3 in [51]).

A quantity that characterizes a fibre system ψ is its projection on a straight line with direction $\beta^\perp$,

Definition 5.5. Let r_β be a line forming an angle β with the abscissa axis, $r_\beta^\perp$ be a line orthogonal to r_β and $\psi \in \mathscr{S}$ be a fibre system. Then we define the total length (computed with multiplicity) of the projection of ψ in the direction $\beta^\perp$ as

$$\rho_\psi(\beta) = \int_{r_\beta^\perp} \chi(\psi \cap \mathrm{T}_y r_\beta)\, dy = \int_\psi \left|\sin(w(\mathrm{T}_y\psi) - \beta)\right| \mu_\psi(dy),\ \ \beta \in [0,\pi),$$

where χ is the Euler-Poincaré characteristic.

Analogously to the case of marked point process, we can consider the direction of a fibre in any of its points as a weight. Hence, we can generalize the theory of fibre processes defining the weighted fibre processes with a general weight (see [73, 78]), of which the unweighted fibre processes are a particular case.

5.2 Intensity estimators in literature

Let us consider a compact window of observation W for the fibre process Φ. If we want to estimate the intensity of the process, the simplest and unbiased estimator is the ratio between the length of the fibres in W and the area of W

$$\widehat{\mathrm{L}}_A^{measure}(W) := \frac{\mu_\Phi(W)}{\nu_2(W)}. \tag{5.3}$$

Note that this estimator has variance $\mathrm{Var}(\Phi(W))/\nu_2(W)^2$. If Φ is an ergodic process, a direct consequence of Corollary 10.2.V in [15], is that the estimator is also strongly consistent, i.e.

$$\widehat{\mathrm{L}}_A^{measure}(A_n) \underset{n\to\infty}{\to} \mathrm{L}_A \qquad \text{a.s. and in } L^1 \text{ norm},$$

where $\{A_n\}_{n\in\mathbb{N}}$ is a convex averaging sequence of sets in $\mathfrak{B}^2$, defined as follows,

Definition 5.6 (Definition 10.2.I in [15]). The sequence $\{A_n\}_{n\in\mathbb{N}}$ of bounded Borel sets in $\mathbb{R}^n$ is a **convex averaging sequence** if

1. each A_n is convex;
2. $A_n \subseteq A_{n+1}$, for $n \geq 1$;
3. $r(A_n) \to \infty$ as $n \to \infty$, where $r(A) = \sup\{r\colon A \text{ contains a ball of radius } r\}$.

Note that the sequence $\{A_n\}_{n\in\mathbb{N}}$ can be revisited as a sequence of enlarging windows of observation, converging to $\mathbb{R}^2$. Since the fibre process under study is assumed to be stationary, this is a straightforward way to "enrich the sample".

Although the estimator $\widehat{\mathrm{L}}_A^{measure}$ has good theoretic properties, it gives poor results in practice because it is not easy to measure the length of fibres from digital images of the process. In fact, the easiest way to measure it is to count the pixels belonging to the fibres, obtaining thus only a rough measure. In $\mathbb{R}^2$ a pixel is two-dimensional, while a fibre is one-dimensional, so it is necessary to correct the counting of the pixels with

a factor that represents the mean length of fibres contained in one pixel. This factor usually depends on the specific geometry of the fibre and the resolution of the image and is not easy to be retrieved.

Because of this problem, we need to find alternative estimators, based on suitable counting measures. A first way to define some estimators of this type is through the intersection between the fibre process and a deterministic fibre system (called *test fibre system*).

Let Φ be a stationary fibre process with intensity L_A and angle distribution ϑ_P and let ψ be a deterministic fibre system with finite total length l_ψ. Mecke [50] and Ohser [58] studied the properties of the intersection point process $\Phi \cap \psi$ (subsequently considered also in [78]). This point process is finite, because ψ has finite total length, and can be seen as a marked point process, if we consider $w(T_y\Phi)$ as the mark in each intersection point y.

In [58], Ohser derived the following unbiased estimator of the quantity $L_A\,\vartheta_P([\gamma_1, \gamma_2))$, through the intersection between Φ and ψ,

$$\sum_{y \in \Phi \cap \psi} \frac{\mathbb{I}_{[\gamma_1,\,\gamma_2)}(w(T_y\Phi))}{\rho_\psi(w(T_y\Phi))}. \tag{5.4}$$

Its unbiasedness is a consequence of Lemma 2.3 in [50] and Theorem 2.1 in [58]. Moreover, setting $\gamma_1 = 0$ and $\gamma_2 = \pi$ in (5.4), we can obtain an unbiased estimator for L_A,

$$\sum_{y \in \Phi \cap \psi} \frac{1}{\rho_\psi(w(T_y\Phi))}. \tag{5.5}$$

Examples of test systems ψ considered in literature for this type of estimation are:

1. a set of segments of total length l_ψ with direction β
2. two systems of segments with, respectively, direction β and $\beta + \frac{\pi}{2}$ and total length l_1 and l_2
3. N circles of radius R.

Their corresponding estimators are:

$$\widehat{\mathrm{L}}_A^{segms} = \sum_{y\in\Phi\cap\psi} \frac{1}{l_\psi \left|\sin(w(\mathrm{T}_y\Phi)-\beta)\right|},$$
$$\widehat{\mathrm{L}}_A^{orthog\_segms} = \sum_{y\in\Phi\cap\psi} \frac{1}{l_1 \left|\sin(\beta-w(\mathrm{T}_y\Phi))\right| + l_2 \left|\cos(\beta-w(\mathrm{T}_y\Phi))\right|}$$
$$\widehat{\mathrm{L}}_A^{circles} = \frac{\#(\Phi\cap\psi)}{4NR}. \tag{5.6}$$

As Osher observed, the first estimator is not so good in practice because the denominator is close to zero, when $w(\mathrm{T}_y\Phi)$ is close to β. This can not happen to the second estimator, but, in any case, its computation can be affected by high computational errors, because it requires the approximation of the angle of the tangent in each of the intersection points. The last estimator is very good in practice, because we need only to count the number of intersections between the system and the process.

Osher's estimators are Crofton-like estimators and, in literature, other estimators of the length of a curve, which derives from Cauchy-Crofton Theorem or Crofton Theorem [6, 11], can be found. These estimators are especially emploied in sterology.

5.3 Intensity estimators due to the intersection with another fibre process

In general, any estimator of type (5.5) has no asymptotic properties. In fact, as we said before, the point process generated by the intersection between the fibre process and the test fibre system is finite, since the test fibre system has finite length. As a consequence the estimate does not change when we enlarge the window of observation. In [64] our aim was to recover an estimator based on a counting measure (thus easily computable) and strongly consistent.

In order to obtain an intersection point process defined on the whole space $\mathbb{R}^2$ (instead of a finite one), we can intersect our fibre process with another fibre process (called *test process*), independent from the first one.

We will see that the test fibre process must satisfy some regularity conditions, in order to obtain a strongly consistent estimator.

We will first introduce some properties of the point process generated by the intersection between two independent and stationary fibre processes and then we will use them to define estimators with good asymptotic properties. In general, we suppose that the two processes are independent, because we generate the test process independently of the fibre process that we want to estimate. We suppose also that both processes are stationary, since we can choose a stationary test process.

Let us recall the following proposition (see [78]),

Proposition 5.1. *Let Φ_1 and Φ_2 be two independent and stationary fibre processes, then the point process $\Phi_1 \cap \Phi_2$ is stationary.*

Let Φ_1 be the fibre process under study, having (unknown) intensity $L_{A,1}$, and let Φ_2 be a test fibre process of (known) intensity $L_{A,2}$. Then, from Lemma 3.2 in [50], the intensity P_A of the intersection point process is

$$P_A = L_{A,1} L_{A,2} \int_0^\pi \int_0^\pi |\sin(\alpha_2 - \alpha_1)| \, \vartheta_{P_1}(d\alpha_1) \, \vartheta_{P_2}(d\alpha_2).$$

If Φ_2 is also isotropic, then the intensity P_A becomes

$$P_A = L_{A,1} L_{A,2} \frac{2}{\pi}. \tag{5.7}$$

In this case, $L_{A,1}$ and P_A are proportional, thus a good estimator for P_A is also a good estimator for $L_{A,1}$. Since Φ_2 represents the test process, we can easily choose it isotropic.

Given a bounded window of observation $W \in \mathfrak{B}^2$, an unbiased estimator for P_A is

$$\widehat{P}_A(W) = \frac{N_{\Phi_1 \cap \Phi_2}(W)}{\nu_2(W)},$$

where $N_{\Phi_1 \cap \Phi_2}$ denotes the counting measure associated to the point process $\Phi_1 \cap \Phi_2$. Therefore, if the test process Φ_2 is isotropic, by using Equation (5.7), the corresponding unbiased estimator for $L_{A,1}$ is

$$\widehat{\mathrm{L}}_{A,1}(W) = \frac{N_{\Phi_1\cap\Phi_2}(W)}{\nu_2(W)}\,\frac{\pi}{2\,\mathrm{L}_{A,2}}. \tag{5.8}$$

Proposition 5.2. *Let Φ_1 and Φ_2 be independent and stationary fibre processes with intensity $\mathrm{L}_{A,1}$ and $\mathrm{L}_{A,2}$, respectively. Moreover, let us assume that Φ_2 is also isotropic. If the point process $\Phi_1 \cap \Phi_2$ is ergodic, then the estimator $\widehat{\mathrm{L}}_{A,1}(A_n)$ is strongly consistent, for any convex averaging sequence $\{A_n\}_{n\in\mathbb{N}}$ of Borel sets in $\mathbb{R}^2$.*

Proof. The proof follows from Corollary 10.2.V in [15], from which we obtain that

$$\frac{N_{\Phi_1\cap\Phi_2}(A_n)}{\nu_2(A_n)} \underset{n\to\infty}{\longrightarrow} \mathrm{P}_A \qquad \text{a.s. and in norm } L^1, \tag{5.9}$$

where $\{A_n\}_{n\in\mathbb{N}}$ is any convex averaging sequence of Borel sets in $\mathbb{R}^2$. As a consequence,

$$\widehat{\mathrm{L}}_{A,1}(A_n) \underset{n\to\infty}{\longrightarrow} \mathrm{P}_A\,\frac{\pi}{2\,\mathrm{L}_{A,\,2}} = \mathrm{L}_{A,1} \qquad \text{a.s. and in norm } L^1.$$

□

We can obtain another estimator for $\mathrm{L}_{A,1}$, if we mimic $\widehat{\mathrm{L}}_A^{segms}$, described in Section 5.2. Thus, for any given bounded window $W \in \mathfrak{B}^2$,

$$\widehat{\widehat{\mathrm{L}}}_{A,1}(W) = \frac{1}{\nu_2(B)\,\mathrm{L}_{A,2}} \sum_{y\in\Phi_1\cap\Phi_2} \frac{\mathbb{I}_W(y)}{\left|\sin(w(\mathrm{T}_y\Phi_2) - w(\mathrm{T}_y\Phi_1))\right|}. \tag{5.10}$$

In order to show the unbiasedness of this estimator, we need to generalize Lemma 3.1 in [50] for a point process of intersection with two weights, as follows.

Lemma 5.1. *Let Φ_1 and Φ_2 be two stationary and independent fibre processes with, respectively, distribution P_1 and P_2, intensity $\mathrm{L}_{A,1}$ and $\mathrm{L}_{A,2}$ and angle distribution ϑ_{P_1} and ϑ_{P_2}. Let $h : \mathbb{R}^2 \times [0,\pi) \times [0,\pi) \to [0,\infty)$ be a measurable function, then*

$$\mathrm{E}\left[\sum_{y\in\Phi_1\cap\Phi_2} h(y,\, w(\mathrm{T}_y\Phi_1),\, w(\mathrm{T}_y\Phi_2))\right]$$
$$= \mathrm{L}_{A,1}\,\mathrm{L}_{A,2}\int_{\mathbb{R}^2}\int_0^\pi\int_0^\pi h(y,\,\alpha_1,\,\alpha_2)\,|\sin(\alpha_2-\alpha_1)|\;\vartheta_{\mathrm{P}_1}(d\alpha_1)\,\vartheta_{\mathrm{P}_2}(d\alpha_2)\,dy.$$

We omit the proof because it is quite similar to the proof of Lemma 3.1 in [50] (anyway it can be found in [64]).

The unbiasedness of $\widehat{\widehat{\mathrm{L}}}_{A,\,1}(B)$ follows from the previous lemma using

$$h(y,\, w(\mathrm{T}_y\Phi_1),\, w(\mathrm{T}_y\Phi_2)) = \frac{\mathbb{I}_B(y)}{\left|\sin(w(\mathrm{T}_y\Phi_2)-w(\mathrm{T}_y\Phi_1))\right|}.$$

This result does not need any hypothesis of isotropy of Φ_2, hence this estimator can also be used in presence of an anisotropic test process.

Like for the previous estimator, we are interested in showing also the strong consistency of $\widehat{\widehat{\mathrm{L}}}_{A,\,1}(B)$. In order to do that, it is necessary to define a different set of marks for the point process. We now use the pair of tangent angles with respect to both Φ_1 and Φ_2 in each intersection point. In order to define the distribution of this mark, analogously to the definition of the angle distribution (see Section 5.1), we set $h(y, w(\mathrm{T}_y\Phi_1), w(\mathrm{T}_y\Phi_2)) = \mathbb{I}_{\mathbb{U}^2}(y)\,f(\alpha,\beta)$, where $\mathbb{U}^2 = [0,1]\times[0,1]$ and $f : [0,\pi)\times[0,\pi) \to [0,\infty)$ is a measurable function. The expected value of h, with respect to the point process, can be computed by applying Theorem 1 in [73], or the previous Lemma 5.1. Then, the distribution of the marks of the point process is such that

$$\int_0^\pi\int_0^\pi f(\alpha,\beta)\,\widetilde{\Theta}(d\alpha,\,d\beta)$$
$$= \frac{\mathrm{L}_{A,1}\mathrm{L}_{A,\,2}}{\mathrm{P}_A}\int_0^\pi\int_0^\pi f(\alpha_1,\,\alpha_2)\,|\sin(\alpha_2-\alpha_1)|\;\vartheta_{\mathrm{P}_1}(d\alpha_1)\,\vartheta_{\mathrm{P}_2}(d\alpha_2), \quad (5.11)$$

for each measurable function $f : [0,\pi)\times[0,\pi) \to [0,\infty)$.

Proposition 5.3. *Let Φ_1 and Φ_2 be two stationary and independent fibre processes with intensity $\mathrm{L}_{A,1}$ and $\mathrm{L}_{A,2}$, respectively. If the point process*

$\Phi_1 \cap \Phi_2$ is ergodic, then estimator $\widehat{\widehat{\mathrm{L}}}_{A,1}(A_n)$ is strongly consistent, for any convex averaging sequence $\{A_n\}_{n\in\mathbb{N}}$ of Borel sets in $\mathbb{R}^2$.

Proof. The proof follows from Corollary 10.2.VII in [15], which gives the following asymptotic result for any convex averaging sequence $\{A_n\}_{n\in\mathbb{N}}$ of Borel sets in $\mathbb{R}^2$

$$\frac{1}{\nu_2(A_n)} \sum_{y\in\Phi_1\cap\Phi_2} \frac{\mathbb{I}_{A_n}(y)}{|\sin(w(\mathrm{T}_y\Phi_2) - w(\mathrm{T}_y\Phi_1))|} \xrightarrow[n\to\infty]{} \mathrm{P}_A \int_0^\pi \int_0^\pi \frac{1}{|\sin(\beta-\alpha)|} \widetilde{\Theta}(d\alpha, d\beta)$$

a.s. and in norm L^1. Thus, by using the previous limit and Equation (5.11), we obtain

$$\widehat{\widehat{\mathrm{L}}}_{A,1}(A_n) \xrightarrow[n\to\infty]{} \frac{\mathrm{P}_A}{\mathrm{L}_{A,\,2}} \int_0^\pi \int_0^\pi \frac{1}{|\sin(\beta-\alpha)|} \widetilde{\Theta}(d\alpha, d\beta) = \mathrm{L}_{A,1} \qquad \text{a.s. } \square$$

Remark 5.2. We can observe that the definition of estimator $\widehat{\mathrm{L}}_A^{measure}$ derives from the theory of random measures: the fibre process is identified with its random measure of the length. On the other hand, Osher's estimators, $\widehat{\mathrm{L}}_{A,1}$ and $\widehat{\widehat{\mathrm{L}}}_{A,1}$ are based on the theory of fibre processes, defined as we did in Section 5.1.

5.4 Ergodicity and choice of the test process

An example of fibre process is the stationary Boolean fibre process,

$$\Phi = \sum_n (\Gamma_n \oplus X_n),$$

where $\{X_n\}_{n\in\mathbb{N}}$ is a homogeneous Poisson point process in $\mathbb{R}^2$ and $\{\Gamma_n\}_{n\in\mathbb{N}}$ is a sequence of i.i.d. random fibres (which have a.s. finite length), independent of the point process. We denote with $\oplus$ the *Minkowski sum*, defined as $A \oplus B = \{x_1 + x_2 \,|\, x_1 \in A, x_2 \in B\}$ with $A, B \in \mathfrak{B}^2$.

Examples of this type of process are: the stationary isotropic Poisson segment process and the stationary isotropic Poisson circle process. In the

first example the process $\{\Gamma_n\}_{n\in\mathbb{N}}$ consists in segments having midpoint at the origin, fixed length and random uniform direction, while in the second example $\{\Gamma_n\}_{n\in\mathbb{N}}$ consists in deterministic circles centered at the origin, having fixed radius.

Some stationary Boolean processes are also ergodic. In [64], we showed the ergodicity of the processes which satisfy the following conditions:

Conditions 5.1.

- for any Borel set B, it is possible to determine a region K_B such that all the points X_n for which $\Gamma_n + X_n$ intersects B, belong to K_B, i.e.

$$\mu_\Phi(B) = \sum_{n=0}^{N(K_B)} \mu_{\Gamma_n \oplus X_n}(B), \tag{5.12}$$

  where $N(A)$ denotes the number of points of the Poisson process that belong to the Borel set A;
- the fibres $\{\Gamma_n\}_{n\in\mathbb{N}}$ have maximum length $l < \infty$.

Proposition 5.4. *Let Φ be a stationary Boolean process, which satisfies Conditions 5.1. Then, Φ is an ergodic process.*

The two examples of Boolean processes considered above satisfy Conditions 5.1, so they are ergodic.

We note that, given $B \in \mathfrak{B}^2$, K_B is defined so that all fibres belonging to K_B^c do not intersect B. As a consequence, if $K_B \cap K_{\mathrm{T}_x B} = \emptyset$, then the fibres that intersect B cannot intersect $\mathrm{T}_x B$ and viceversa. To give an example of the set K_B, let us consider the stationary isotropic Poisson segment process with segment length $l < 1$. For a fixed fibre direction α, if $B = \mathbb{U}^2$, the region K_B^α is depicted in Figure (5.1), so that K_B is the union of all K_B^α with $\alpha \in [0, \pi)$.

We have defined intensity estimators based on the intersection between two fibre processes and we have proven that they are strongly consistent if the intersection point process is ergodic. In [64], we also proved the following proposition.

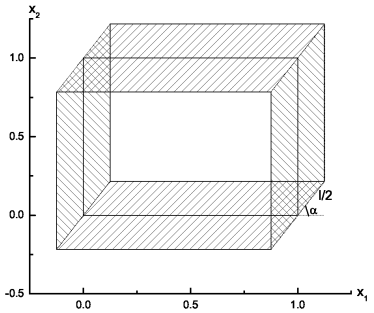

Fig. 5.1 K_B^α in case of a stationary isotropic Poisson segment process with segment length $l < 1$, when $B = \mathbb{U}^2$.

Proposition 5.5. *Let Φ_1 and Φ_2 be two independent stationary Boolean processes, such that both Φ_1 and Φ_2 satisfy Conditions 5.1. Then, the intersection process $\Phi_1 \cap \Phi_2$ is ergodic.*

This last proposition ensures that if the fibre process that we are analyzing is well approximated by a stationary Boolean fibre process satisfying Conditions 5.1, then it is sufficient to choose as test process another Boolean fibre process satisfying Conditions 5.1, in order to obtain consistent estimators of the intensity.

Chapter 6
A central limit theorem for functionals of point processes

In Chapter 5, we showed that a technique to estimate the intensity of a stationary planar fibre process Φ_1 is to intersect the process with another fibre process Φ_2 [53, 64]. The counting measure corresponding to the point process of the intersection $\Phi_1 \cap \Phi_2$ is related (by a Crofton-type Formula) to the intensity of the fibre process under study and, if a suitable test fibre process Φ_2 is used, the obtained estimators are unbiased and strongly consistent. The main difficulty is now to prove the asymptotic normality, that is a Central Limit Theorem (CLT), for such estimators.

In [34], Ivanoff established central limit theorems (CLTs) for the counting measure of a spatially homogeneous point process in $\mathbb{R}^n$. Penrose and Yukich (for example, in [61]) proved several CLTs for functionals of two types of point processes in $\mathbb{R}^n$: the Poisson and the binomial point process. In their proofs it is crucial the independence of the points of the process which fall in disjoint Borel sets (that is having independent increments). The spatial point process of intersections of fibres on a plane has in general not independent increments, since the (not null) length of fibres implies that points located at a distance lower than the maximum length of a fibre (if it exists) are correlated. But if we assume that the fibres have a.s. finite length, lower than $l \in \mathbb{R}_+$, then intersection points at a distance greater than l are independent (if those fibres are generated independently). Fur-

thermore the point process of intersections is in general not isotropic, if the fibres have a non-uniform random orientation.

Based on the previous motivations, in this chapter we prove a CLT for functionals of point processes in $\mathbb{R}^2$, which are stationary and independent in Borel sets at distance greater than l. The point process does not need to be isotropic, differently from the assumptions in the theorems in [34, 61]. Like in [61], our CLT considers increasing windows of observation and, for simplicity, we will consider only rectangular windows, but extensions to windows of more general shape can be easily proven.

The proof of our CLT, that is stated in Section 6.4, is based on the application of the Central Limit Theorem for Martingale Differences (CLTMD) [49]; thus in Sections 6.2 and 6.3 we need to show that the three main conditions needed to apply the CLTMD are satisfied under our assumptions. In particular, the first two conditions needed for the CLTMD are proven assuming particular mixing conditions, which are milder than the independence at distance l, in the view of a possible future extension of our CLT to point processes derived by the intersection of two fibre processes whose fibres do not have (a.s.) finite length.

In Section 6.5, we retrieve the asymptotic normality of the estimators of the intensity described in Section 5.3, by applying our CLT. For further details on the proof of our CLT see [52].

6.1 Notations and basic assumptions

To prove our CLT in $\mathbb{R}^2$, we will try to mimic the $\mathbb{R}^2$-version of the proof of Theorem 3.1 in [61]. Therefore, we will now explain all objects involved in that theorem and the ingredients used for its proof.

In [61], the CLT (in its $\mathbb{R}^2$-version) is defined on a sequence of sets $\{B_n\}_{n\geq 1} \subseteq \mathbb{R}^2$, which satisfies the following conditions,

Conditions 6.1.

1. B_n is a bounded Borel set in $\mathbb{R}^2$, for all $n \geq 1$,
2. $\nu_2(B_n) = \frac{n}{\lambda}$, for some constant $\lambda > 0$ and for all $n \geq 1$,

3. $\lim_{n\to\infty} B_n = \mathbb{R}^2$ (in the sense that $\bigcup_{n\geq 1}\bigcap_{m\geq n} B_m = \mathbb{R}^2$),
4. $\lim_{n\to\infty} \frac{\nu_2(\partial_r B_n)}{n} = 0$, for all $r > 0$,
5. there exists a constant β_1 such that $\mathrm{diam}(B_n) \leq \beta_1 n^{\beta_1}$ for all $n \geq 1$,

where $\partial_r A = \bigcup_{\mathbf{x}\in\partial A} Q_r(\mathbf{x})$, with $Q_r(\mathbf{x}) = [-r, r]^2 + \mathbf{x}$, for any $r > 0$, and ∂A is the boundary of A, for any Borel set $A \subset \mathbb{R}^2$.

For example, Conditions 6.1 hold for a sequence of squares centered at the origin with side $\sqrt{\frac{n}{\lambda}}$ (i.e. $B_n = Q_{\frac{1}{2}\sqrt{\frac{n}{\lambda}}}(\mathbf{0})$), for $n \geq 1$, or a sequence of Euclidean balls centered at the origin with ray $\sqrt{\frac{n}{\pi\lambda}}$, for $n \geq 1$.

In Theorem 3.1 of [61], Penrose and Yukich consider a generic functional H on a Poisson point process $\mathscr{P}$, which satisfies the following conditions (in $\mathbb{R}^2$):

Conditions 6.2.

1. H is a real-valued functional defined for all subsets of $\mathbb{R}^2$,
2. H is translation-invariant,
3. H is **weakly stabilizing on $\mathscr{B}$ with respect to the point process $\mathscr{P}$**, i.e. there exists a random variable $\Delta(\infty)$ such that for any $\mathscr{B}$-valued sequence $\{A_n\}_{n\geq 1}$ that tends to $\mathbb{R}^2$ (i.e. $\bigcup_{n\geq 1}\bigcap_{m\geq n} A_m = \mathbb{R}^2$),

$$\Delta(\mathscr{P}\cap A_n) \underset{n\to\infty}{\longrightarrow} \Delta(\infty), \qquad \text{a.s.},$$

4. H satisfies the **bounded moments condition on $\mathscr{B}$ with respect to the point process $\mathscr{P}$**, i.e.

$$\sup_{A\in\mathscr{B}:\,\mathbf{0}\in A} \left\{\mathrm{E}[\Delta(\mathscr{P}\cap A)^4]\right\} < \infty,$$

where $\Delta(A) = H(A\cup\{\mathbf{0}\}) - H(A)$, for all $A \subset \mathbb{R}^2$, and $\mathscr{B}$ is the collection of all regions $A \subset \mathbb{R}^2$ of the form $A = \{B_n + \mathbf{x} : \mathbf{x} \in \mathbb{R}^2, n \geq 1\}$, for a given sequence of sets $\{B_n\}_{n\geq 1} \subseteq \mathbb{R}^2$ satisfying Conditions 6.1.

Given a Poisson point process $\mathscr{P}$ of intensity λ and a functional H which satisfies Conditions 6.2, in Theorem 3.1 in [61] it is established that there exists $\sigma^2 \geq 0$ such that

$$n^{-1}\mathrm{Var}(H(\mathscr{P}_n)) \underset{n\to\infty}{\longrightarrow} \sigma^2$$

and

$$n^{-1/2}(H(\mathscr{P}_n) - \mathrm{E}[H(\mathscr{P}_n)]) \underset{n\to\infty}{\xrightarrow{\mathscr{D}}} \mathscr{N}(0,\sigma^2), \tag{6.1}$$

where $\mathscr{P}_n = \mathscr{P} \cap B_n$ and $\{B_n\}_{n\geq 1} \subseteq \mathbb{R}^2$ is a sequence of sets satisfying Conditions 6.1. Therefore, in the following, we will show a convergence similar to (6.1) for a point process $\mathscr{P}$ satisfying lower regularity conditions, and a functional H such that:

Conditions 6.3.

1. H is a positive real-valued functional defined for all subset of $\mathbb{R}^2$,
2. H is translation-invariant,
3. H is additive,
4. $H(\{\mathbf{0}\}) < \infty$.

We observe that the additivity of the functional and the fact that $H(\{\mathbf{0}\}) < \infty$ ensure both Property 3 (with $\Delta(\infty) = H(\{\mathbf{0}\})\mathbb{I}_{\{\mathbf{0}\notin\mathscr{P}\}}$) and Property 4 in Conditions 6.2. Moreover, we will require that $H(\mathscr{P}\cap .)$ has finite second or fourth moment measure.

Proposition 6.1. *Let H be a functional for which Conditions 6.3 hold and $\{B_n\}_{n\geq 1}$ be a sequence of sets in $\mathbb{R}^2$, satisfying Conditions 6.1. Then, H is weakly stabilizing on $\mathscr{B}$ and it satisfies the bounded moments condition on $\mathscr{B}$, for every point process $\mathscr{P}$.*

Proof. First, we notice that, since H is additive,

$$\Delta(\mathscr{P}\cap A) = H(\mathscr{P}\cap A\cup\{\mathbf{0}\}) - H(\mathscr{P}\cap A) = H(\{\mathbf{0}\})\mathbb{I}_{\mathbf{0}\notin\mathscr{P}\cap A}, \tag{6.2}$$

for all $A\subset\mathbb{R}^2$. Let $\{A_n\}_{n\geq 1}$ be any $\mathscr{B}$-valued sequence that tends to $\mathbb{R}^2$, then there exists $\bar{n}$ such that for each $n\geq\bar{n}$, $\mathbf{0}\in A_n$. Thus, by setting $\Delta(\infty) = H(\{\mathbf{0}\})\mathbb{I}_{\{\mathbf{0}\notin\mathscr{P}\}}$, it turns out that, for each $n\geq\bar{n}$,

$$\begin{aligned}|\Delta(\mathscr{P}\cap A_n) - \Delta(\infty)| &= H(\{\mathbf{0}\})\left|\mathbb{I}_{\mathbf{0}\notin\mathscr{P}\cap A_n} - \mathbb{I}_{0\notin\mathscr{P}}\right| \\ &= H(\{\mathbf{0}\})\left|\mathbb{I}_{0\notin\mathscr{P}} - \mathbb{I}_{0\notin\mathscr{P}}\right| = 0,\end{aligned}$$

i.e. $\lim_{n\to\infty} |\Delta(A_n) - \Delta(\infty)| = 0$ a.s.. Since this limit is independent of the chosen $\mathscr{B}$-valued sequence of sets $\{A_n\}_{n\geq 1}$, H is weakly stabilizing on $\mathscr{B}$ with $\Delta(\infty) = H(\{\mathbf{0}\})\mathbb{I}_{\{\mathbf{0}\notin\mathscr{P}\}}$. Moreover, Equation (6.2) and Property 4 of Conditions 6.3 imply that

$$\sup_{A\in\mathscr{B}:0\in A}\left\{\mathrm{E}[\Delta(\mathscr{P}\cap A)^4]\right\} = (H(\{\mathbf{0}\}))^4 \sup_{A\in\mathscr{B}:0\in A} \mathrm{P}(0\notin\mathscr{P}\cap A) \leq (H(\{\mathbf{0}\}))^4 < \infty,$$

i.e. H satisfies the bounded moments condition on $\mathscr{B}$. □

Since we will relax the regularity of the point process, we need to change the convergency properties of the sequence of Borel sets $\{B_n\}_{n\geq 1}$. We suppose that the sequence of Borel sets $\{B_n\}_{n\geq 1}$ satisfies the following properties:

Conditions 6.4.

1. $B_n \subseteq B_{n+1}$, for all $n \geq 1$;
2. for each $n \geq 1$, B_n is a rectangle with horizontal side $L\cdot o_n$ and vertical side $L\cdot v_n$, o_n, $v_n \in \mathbb{N}$ and $L > 0$;
3. there exists $n_1 > 0$ such that $v_n \leq o_n$ for all $n \geq n_1$;
4. there exists $\alpha > 0$ such that

$$\frac{o_n}{n^\alpha} \underset{n\to\infty}{\longrightarrow} c_1; \tag{6.3}$$

5. there exists $\beta > 0$ such that $\{v_n/o_n^\beta\}_{n\in\mathbb{N}}$ is a monotone sequence and

$$\frac{v_n}{o_n^\beta} \underset{n\to\infty}{\longrightarrow} c_2. \tag{6.4}$$

For each rectangle B_n, we define a grid of squares of side L that covers it. If we call $Q_{L/2}(\mathbf{x}_{i,j})$ the square of side L centered at the point $\mathbf{x}_{i,j}$ having abscissa i and ordinate j of the grid, then we can write

$$B_n = \bigcup_{i=1}^{o_n}\bigcup_{j=1}^{v_n} Q_{L/2}(\mathbf{x}_{i,j}).$$

The choice of the constant L will depend on the characteristics of the point process. For example, if we consider a *point process independent at distance* l ($l > 0$), then the choice $L \geq l$ guarantees the independence of $H(\mathscr{P} \cap .)$, when computed in non contiguous squares. Let us define rigorously the property of independence at distance l.

Definition 6.1. Let $\mathscr{P}$ be a point process in $\mathbb{R}^2$. We say that $\mathscr{P}$ is a **point process independent at distance** l, if there exist $0 < l < \infty$ such that $\mathscr{P} \cap A$ and $\mathscr{P} \cap B$ are independent for each $A, B \in \mathfrak{B}^2$ with $d(A,B) > l$, where $d(.,.)$ is the distance defined by

$$d(A,B) = \inf_{\mathbf{x} \in A, \mathbf{y} \in B} \|\mathbf{x} - \mathbf{y}\|_2, \qquad A, B \subseteq \mathbb{R}^2,$$

and $\|.\|_2$ denotes the Euclidean norm.

Note that if a point process is independent at distance l then it is also independent at any distance $l' \geq l$.

The proof of Theorem 3.1 in [61] requires the definition of a martingale difference in order to apply the CLT for martingale differences (Theorem 2.3 of [49]). We will now define the martingale difference suitable for our setting.

Since the point process is stationary, without loss of generality, we can suppose that the centers of the squares of the grid belong to $\mathbb{Z}_L^2 = \{\mathbf{x}L : \mathbf{x} \in \mathbb{Z}^2\}$ and, for each $\mathbf{x} = (x_1, x_2) \in \mathbb{Z}_L^2$, we define the σ-algebra $\mathfrak{F}_{\mathbf{x}} = \sigma(\{\mathscr{P} \cap Q_{L/2}(\mathbf{y}) \,|\, \mathbf{y} = (y_1, y_2) \in \mathbb{Z}_L^2, y_1 \leq x_1\})$. We can observe that, for any $\mathbf{x}, \mathbf{y} \in \mathbb{Z}_L^2$, $\mathfrak{F}_{\mathbf{x}} = \mathfrak{F}_{\mathbf{y}}$, if $x_1 = y_1$, and $\mathfrak{F}_{\mathbf{x}} \subset \mathfrak{F}_{\mathbf{y}}$, if $x_1 < y_1$.

For a fixed $n \in \mathbb{N}$, we define the filtration $\{\mathfrak{G}_0, \ldots, \mathfrak{G}_{o_n}\}$, where $\mathfrak{G}_0$ is the trivial σ-algebra and $\mathfrak{G}_i = \mathfrak{F}_{\mathbf{x}_{i,1}}$, for $i = 1, \ldots, o_n$. Then, we can write

$$H(\mathscr{P}_n) - \mathrm{E}[H(\mathscr{P}_n)] = \sum_{i=1}^{o_n} [\mathrm{E}[H(\mathscr{P}_n) \,|\, \mathfrak{G}_i] - \mathrm{E}[H(\mathscr{P}_n) \,|\, \mathfrak{G}_{i-1}]] =: \sum_{i=1}^{o_n} D_i,$$

where $D_i = \mathrm{E}[H(\mathscr{P}_n) \,|\, \mathfrak{G}_i] - \mathrm{E}[H(\mathscr{P}_n) \,|\, \mathfrak{G}_{i-1}]$, for $i = 1, \ldots, o_n$. We observe that $\{D_i\}_{i=1}^{o_n}$ is a martingale difference, since $\{H(\mathscr{P}_n) \,|\, \mathfrak{G}_i\}_{i=0}^{o_n}$ is a martingale.

Note that our filtration is different from the one used in [61], since the definition of $\{\mathfrak{G}_i\}_{i=0}^{o_n}$ allows us to use the independence at distance $l = L$ (or milder mixing conditions; see Conditions 6.5). In fact, in [61], since the authors considered point processes $\mathscr{P}$ having independent increments (note that a process with independent increments is independent at distance $l = 0$), any choice for the side $L > 0$ of the squares of the grid led to the definition of a suitable filtration. Thus they used a grid of squares of side $L = 1$ (to cover each set B_n) and, for each $\mathbf{x} \in \mathbb{Z}^2$, they defined the σ-algebra $\mathfrak{F}_{\mathbf{x}} = \sigma(\{\mathscr{P} \cap Q_{L/2}(\mathbf{y}) \,|\, \mathbf{y} \in \mathbb{Z}^2, \mathbf{y} \preceq \mathbf{x}\})$, where $\preceq$ denotes the lexicographic order in $\mathbb{Z}^2$.

For the next Propositions and computations, we need to decompose D_i, $i = 1, \ldots, o_n$, into the following sums:

$$\begin{aligned} D_i \;=\; & \sum_{j=1}^{v_n} \left[\sum_{h=1}^{i-1} H(\mathscr{P} \cap Q_{L/2}(\mathbf{x}_{h,j})) - H(\mathscr{P} \cap Q_{L/2}(\mathbf{x}_{h,j})) \right] \\ & + \sum_{j=1}^{v_n} \sum_{h=i}^{o_n} \mathrm{E}[H(\mathscr{P} \cap Q_{L/2}(\mathbf{x}_{h,j})) \,|\, \mathfrak{G}_i] - \mathrm{E}[H(\mathscr{P} \cap Q_{L/2}(\mathbf{x}_{h,j})) \,|\, \mathfrak{G}_{i-1}] \\ =: & \sum_{j=1}^{v_n} \sum_{h=i}^{o_n} \Delta_{\mathbf{x}_{h,j},\mathbf{x}_{i,1}}, \end{aligned} \tag{6.5}$$

where $\Delta_{\mathbf{x}_{h,j},\mathbf{x}_{i,1}} = \mathrm{E}[H(\mathscr{P} \cap Q_{L/2}(\mathbf{x}_{h,j})) \,|\, \mathfrak{G}_i]$-$\mathrm{E}[H(\mathscr{P} \cap Q_{L/2}(\mathbf{x}_{h,j})) \,|\, \mathfrak{G}_{i-1}]$, for each triple of indices (h, j, i). For the stationarity of the process, we can easily see that, fixed $h \geq i$ and $i > 1$, the elements of the family $\{\Delta_{\mathbf{x}_{h,j},\mathbf{x}_{i,1}}\}_{j=1}^{v_n}$ have all the same distribution and thus also $\{\sum_{h=i}^{o_n} \Delta_{\mathbf{x}_{h,j},\mathbf{x}_{i,1}}\}_{j=1}^{v_n}$ for $i > 1$. For $i = 1$,

$$\begin{aligned} \Delta_{\mathbf{x}_{1,j},\mathbf{x}_{1,1}} &= H(\mathscr{P} \cap Q_{L/2}(\mathbf{x}_{1,j})) - \mathrm{E}[H(\mathscr{P} \cap Q_{L/2}(\mathbf{x}_{h,j}))] \\ \Delta_{\mathbf{x}_{h,j},\mathbf{x}_{1,1}} &= \mathrm{E}[H(\mathscr{P} \cap Q_{L/2}(\mathbf{x}_{h,j})) \,|\, \mathfrak{G}_1] - \mathrm{E}[H(\mathscr{P} \cap Q_{L/2}(\mathbf{x}_{h,j}))], \qquad h > 1, \end{aligned}$$

because of the definition of the filtration. In general, for each $\mathbf{x}, \mathbf{y} \in \mathbb{Z}_L^2$, we define

$$\Delta_{\mathbf{y}}^0 = H(\mathscr{P} \cap Q_{L/2}(\mathbf{y})) - \mathrm{E}[H(\mathscr{P} \cap Q_{L/2}(\mathbf{y}))]$$

$$\Delta^0_{\mathbf{y},\mathbf{x}} = \mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{y}))\,|\,\mathfrak{F}_{\mathbf{x}}] - \mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{y}))]$$
$$\Delta_{\mathbf{y},\mathbf{x}} = \mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{y}))\,|\,\mathfrak{F}_{\mathbf{x}}] - \mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{y}))\,|\,\mathfrak{F}_{\mathbf{x}-L\mathbf{e}_1}],$$

where $\mathbf{e}_1$=(1,0). We can observe that if $H(\mathscr{P}\cap.)$ has finite fourth moment measure, then we can prove that $\mathrm{E}[(\Delta^0_{\mathbf{y}})^4]$, $\mathrm{E}[(\Delta^0_{\mathbf{y},\mathbf{x}})^4]$, $\mathrm{E}[(\Delta_{\mathbf{y},\mathbf{x}})^4] < \infty$, by using Jensen's inequality. In fact, given any two σ-algebras $\mathfrak{F}$ and $\mathfrak{G}$, the following inequalities hold

$$\begin{aligned}
&\mathrm{E}[(\mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{y}))\,|\,\mathfrak{F}] - \mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{y}))\,|\,\mathfrak{G}])^4]\\
&\quad\le \mathrm{E}[(\mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{y}))\,|\,\mathfrak{F}]^2 + \mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{y}))\,|\,\mathfrak{G}]^2)^2]\\
&\quad\le 2\mathrm{E}[\mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{y}))^4\,|\,\mathfrak{F}] + \mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{y}))^4\,|\,\mathfrak{G}]]\\
&\quad= 4\mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{y}))^4] =: 4E_{\mathbf{0},4} < \infty \qquad (6.6)\\
&\mathrm{E}[(\mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{y}))\,|\,\mathfrak{F}] - \mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{y}))\,|\,\mathfrak{G}])^2]\\
&\quad\le \mathrm{E}[\mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{y}))^2\,|\,\mathfrak{F}] + \mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{y}))^2\,|\,\mathfrak{G}]]\\
&\quad= 2\mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{y}))^2] =: 2E_{\mathbf{0},2} < \infty \qquad (6.7)
\end{aligned}$$

where

$$E_{\mathbf{0},2} = \mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{0}))^2] \qquad (6.8)$$
$$E_{\mathbf{0},4} = \mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{0}))^4]. \qquad (6.9)$$

Moreover, if the point process is independent at distance $l = L$, the following lemma holds.

Lemma 6.1. *If $\mathscr{P}$ is a point process independent at distance $l = L$, then $\Delta^0_{\mathbf{x},\mathbf{x}-L\mathbf{e}_1}$ and $\Delta^0_{(x_1,x_2+Ld),\mathbf{x}-L\mathbf{e}_1}$ are independent, for any $\mathbf{x} = (x_1,x_2) \in \mathbb{Z}^2_L$ and $d \in \mathbb{N}$ with $d \ge 4$.*

Proof. Since the point process $\mathscr{P}$ is independent in Borel sets with distance greater than L, we have that, for each $\mathbf{x} = (x_1,x_2) \in \mathbb{Z}^2_L$, $H(\mathscr{P}\cap Q_{L/2}(\mathbf{x}))$ can depend only on the events in

$$\sigma\left(\left\{\mathscr{P}\cap Q_{L/2}(x_1+d_1L, x_2+d_2L)\,|\,d_1,d_2 = -1,0,1\right\}\right).$$

As a consequence,

$$\begin{aligned}
&\Delta^0_{\mathbf{x},\mathbf{x}-L\mathbf{e}_1}\\
&= E[H(\mathscr{P}\cap Q_{L/2}(\mathbf{x}))\,|\,\mathfrak{F}_{\mathbf{x}-L\mathbf{e}_1}] - \mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{x}))]\\
&= \mathrm{E}\left[H(\mathscr{P}\cap Q_{L/2}(\mathbf{x}))\,|\,\sigma\left(\left\{\mathscr{P}\cap Q_{L/2}(x_1-L,x_2+d_2L)\,|\,d_2=-1,0,1\right\}\right)\right]\\
&- \mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{x}))],
\end{aligned}$$

and

$$\begin{aligned}
&\Delta^0_{(x_1,x_2+Ld),\mathbf{x}-L\mathbf{e}_1}\\
&= \mathrm{E}\left[H(\mathscr{P}\cap Q_{L/2}((x_1,x_2+Ld)))\,|\,\sigma\left(\left\{\mathscr{P}\cap Q_{L/2}(x_1-L,x_2+(d+d_2)L)\,|\right.\right.\right.\\
&\qquad \left.\left.\left. d_2=-1,0,1\right\}\right)\right] - \mathrm{E}[H(\mathscr{P}\cap Q_{L/2}((x_1,x_2+Ld)))].
\end{aligned}$$

Then, $\Delta^0_{\mathbf{x},\mathbf{x}-L\mathbf{e}_1}$ and $\Delta^0_{(x_1,x_2+Ld),\mathbf{x}-L\mathbf{e}_1}$ are independent if $d\geq 4$, because the σ-algebras involved in the conditional expectations in $\Delta^0_{\mathbf{x},\mathbf{x}-L\mathbf{e}_1}$ and $\Delta^0_{(x_1,x_2+Ld),\mathbf{x}-L\mathbf{e}_1}$ are independent. □

In case of a point process independent at distance $l=L$, Equation (6.5) becomes,

$$\begin{aligned}
D_i &= \sum_{j=1}^{v_n}\sum_{h=i}^{i+1}\mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{x}_{h,j}))\,|\,\mathfrak{G}_i] - \mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{x}_{h,j}))\,|\,\mathfrak{G}_{i-1}]\\
&\quad + \sum_{j=1}^{v_n}\sum_{h=i+2}^{o_n}\left[\mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{x}_{h,j}))] - \mathrm{E}[H(\mathscr{P}\cap Q_{L/2}(\mathbf{x}_{h,j}))]\right]\\
&= \sum_{j=1}^{v_n}\sum_{h=i}^{i+1}\Delta_{\mathbf{x}_{h,j},\mathbf{x}_{i,1}} =: \sum_{j=1}^{v_n} M_{i,j}, \qquad i=1,\ldots,o_n-1, \qquad (6.10)\\
D_{o_n} &= \sum_{j=1}^{v_n}\Delta_{\mathbf{x}_{o_n,j},\mathbf{x}_{o_n,1}}, \qquad (6.11)
\end{aligned}$$

where $M_{i,j}{=}\sum_{h=i}^{i+1}\Delta_{\mathbf{x}_{h,j},\mathbf{x}_{i,1}}$ and, in general, for each $\mathbf{x},\mathbf{y}\in\mathbb{Z}^2_L$, we define

$$M_{\mathbf{y},\mathbf{x}} = \sum_{d=0}^{1} \Delta_{\mathbf{y}+dL\mathbf{e}_1,\mathbf{x}}.$$

From the previous discussion, $\{M_{i,j}\}_{i>1,j}$ have all the same distribution. Moreover, if $H(\mathscr{P} \cap .)$ has finite second or fourth moment measure, then Jensen's inequality and Inequalities (6.6) and (6.7) imply that, respectively,

$$\mathrm{E}[M_{i,j}^2] = \mathrm{E}[(\Delta_{\mathbf{x}_{i,j},\mathbf{x}_{i,1}} + \Delta_{\mathbf{x}_{i+1,j},\mathbf{x}_{i,1}})^2] < 8E_{\mathbf{0},2} < \infty, \tag{6.12}$$

$$\mathrm{E}[M_{i,j}^4] = \mathrm{E}[(\Delta_{\mathbf{x}_{i,j},\mathbf{x}_{i,1}} + \Delta_{\mathbf{x}_{i+1,j},\mathbf{x}_{i,1}})^4] < 64E_{\mathbf{0},4} < \infty, \tag{6.13}$$

for any $i = 1, \ldots, o_n$ and $j = 1, \ldots, v_n$.

Similarly to the proof of Theorem 3.1 in [61], in Section 6.4 we will show that, to obtain the asymptotic normality of $H(\mathscr{P}_n)$, it is sufficient to prove that the hypotheses of the CLT for martingale differences are satisfied for the martingale array $\{o_n^{-\gamma} D_i : i = 1, \ldots, o_n\}_{n\geq 1}$, i.e.

$$\sup_{n\geq 1} \mathrm{E}\left[\max_{1\leq i\leq o_n} (o_n^{-\gamma} D_i)^2\right] < \infty, \tag{6.14}$$

$$o_n^{-\gamma} \max_{1\leq i\leq o_n} |D_i| \xrightarrow[n\to\infty]{P} 0 \tag{6.15}$$

$$o_n^{-2\gamma} \sum_{i=1}^{o_n} D_i^2 \xrightarrow[n\to\infty]{L^1} \tau^2, \tag{6.16}$$

hold for some $\tau^2 \geq 0$. In Sections 6.2 and 6.3, we will prove relations (6.14), (6.15) and (6.16) and, in Section 6.4, we will derive the asymptotic normality of $H(\mathscr{P}_n)$. Note that the requirement of independence at distance $l = L$ is needed only to prove (6.16), but not for the proof of (6.14) and (6.15), where milder conditions can be assumed.

6.2 Proof of the first two conditions of the CLT for martingale differences for a stationary point process

Before proving relations (6.14) and (6.15), we define a set of conditions that the point process must satisfy for our purpose. These conditions can be seen as mixing conditions, which may substitute the independence at distance $l = L$ and are implied by it.

Conditions 6.5. The point process $\mathscr{P}$ is stationary, $H(\mathscr{P} \cap .)$ has finite fourth moment measure and there exist two non-negative functions f_1 and f_2 and two nonnegative constants δ_1 and δ_2 such that

$$\mathrm{E}\left[\left(\mathrm{E}[H(\mathscr{P}\cap A)\,|\,\mathfrak{F}_{\mathbf{x}}] - \mathrm{E}[H(\mathscr{P}\cap A)]\right)^2\right] \le f_1(\nu_2(A)) \text{ with}$$
$$\frac{f_1(\nu_2(A))}{\nu_2(A)^{\delta_1}} \underset{\nu_2(A)\to\infty}{\longrightarrow} \text{constant}, \tag{6.17}$$
$$\mathrm{E}\left[\left(\mathrm{E}[H(\mathscr{P}\cap A)\,|\,\mathfrak{F}_{\mathbf{x}}] - \mathrm{E}[H(\mathscr{P}\cap A)]\right)^4\right] \le f_2(\nu_2(A)) \text{ with}$$
$$\frac{f_2(\nu_2(A))}{\nu_2(A)^{\delta_2}} \underset{\nu_2(A)\to\infty}{\longrightarrow} \text{constant}, \tag{6.18}$$

for every $\mathbf{x} \in \mathbb{Z}_L^2$ and every convex set $A \in \mathfrak{B}^2$ such that $d(A, R_{\mathbf{x}}) > L$, $R_x = \{Q_{L/2}(\mathbf{y}) \,|\, \mathbf{y} \in \mathbb{Z}_L^2, y_1 \le x_1\}$. Moreover, there exist two non-negative functions f_3 and f_4 and two nonnegative constants δ_3 and δ_4 such that

$$\mathrm{E}\left[\left(\mathrm{E}[H(\mathscr{P}\cap B)\,|\,\mathfrak{F}_{\mathbf{x}}] - \mathrm{E}[H(\mathscr{P}\cap B)]\right)^2\right] \le f_3(h(B)) \text{ with}$$
$$\frac{f_3(h(B))}{h(B)^{\delta_3}} \underset{h(B)\to\infty}{\longrightarrow} \text{constant}, \tag{6.19}$$
$$\mathrm{E}\left[\left(\mathrm{E}[H(\mathscr{P}\cap B)\,|\,\mathfrak{F}_{\mathbf{x}}] - \mathrm{E}[H(\mathscr{P}\cap B)]\right)^4\right] \le f_4(h(B)) \text{ with}$$
$$\frac{f_4(h(B))}{h(B)^{\delta_4}} \underset{h(B)\to\infty}{\longrightarrow} \text{constant}, \tag{6.20}$$

for every $\mathbf{x} \in \mathbb{Z}_L^2$ and every convex set $B \in \mathfrak{B}^2$ such that $B \cap R_{\mathbf{x}} = \emptyset$ and $e(B, R_{\mathbf{x}}) \le L$, where

$$h(B) = \max_{\ell \in \mathscr{L}} \nu_1(\ell \cap B), \qquad \text{with } \mathscr{L} = \{\text{lines in } \mathbb{R}^2 \text{ parallel to the vertical axis}\},$$

and for each $A, B \subseteq \mathbb{R}^2$,

$$e(A,B) = \sup_{\mathbf{x} \in A, \mathbf{y} \in B} \|\mathbf{x} - \mathbf{y}\|_2.$$

Finally, there exists a nonnegative constant δ_5 such that for each $B \in \mathfrak{B}^2$,

$$\frac{\mathrm{E}\left[(H(\mathscr{P} \cap B) - \mathrm{E}[H(\mathscr{P} \cap B)])^4\right]}{\nu_2(B)^{\delta_5}} \underset{\nu_2(B) \to \infty}{\longrightarrow} \text{constant.} \tag{6.21}$$

If the point process is independent at distance $l = L$, then Conditions 6.5 are satisfied when $\delta_5 = 2$, f_1, $f_2 \equiv 0$, $f_3(h(B))$=$2E_{\mathbf{0},2}\,(9h(B)/L - 11)$,

$$f_4(h(B)) = \begin{cases} 82944 E_{\mathbf{0},4} & \text{if } h(B) < 11L \\ E_{\mathbf{0},4}\left(972\left(\frac{h(B)}{L}\right)^2 - 6744\frac{h(B)}{L} - 41924\right) & \text{if } h(B) \ge 11L, \end{cases}$$

where $E_{\mathbf{0},2}$ and $E_{\mathbf{0},4}$ are defined by (6.8) and (6.9); thus $\delta_1 = \delta_2 = 0$, $\delta_3 = 1$ and $\delta_4 = 2$. The proof of these results can be found in Subsection 6.2.1.

Remark 6.1. If a point process satisfies Conditions 6.5, then Relations (6.17) and (6.18) hold also for $(\mathrm{E}[H(\mathscr{P} \cap A) \,|\, \mathfrak{F}_{\mathbf{x}}] - \mathrm{E}[H(\mathscr{P} \cap A) \,|\, \mathfrak{F}_{\mathbf{x}-L\mathbf{e}_1}])$ with a function proportional to the corresponding f_j function. In fact, if $\mathbf{x} \in \mathbb{Z}_L^2$ and $A \in \mathfrak{B}^2$ is a convex set such that $d(A, R_{\mathbf{x}}) > L$, then

$$\begin{aligned} &\mathrm{E}\left[(\mathrm{E}[H(\mathscr{P} \cap A) \,|\, \mathfrak{F}_{\mathbf{x}}] - \mathrm{E}[H(\mathscr{P} \cap A) \,|\, \mathfrak{F}_{\mathbf{x}-L\mathbf{e}_1}])^p\right] \\ &\le 2^{p-1}\mathrm{E}\left[(\mathrm{E}[H(\mathscr{P} \cap A) \,|\, \mathfrak{F}_{\mathbf{x}}] - \mathrm{E}[H(\mathscr{P} \cap A)])^p\right] \\ &+ 2^{p-1}\mathrm{E}\left[(\mathrm{E}[H(\mathscr{P} \cap A) \,|\, \mathfrak{F}_{\mathbf{x}-Le_1}] - \mathrm{E}[H(\mathscr{P} \cap A)])^p\right] \\ &= 2^{p-1}\mathrm{E}\left[(\mathrm{E}[H(\mathscr{P} \cap A) \,|\, \mathfrak{F}_{\mathbf{x}}] - \mathrm{E}[H(\mathscr{P} \cap A)])^p\right] \\ &+ 2^{p-1}\mathrm{E}\left[(\mathrm{E}[H(\mathscr{P} \cap \mathrm{T}_{L\mathbf{e}_1} A) \,|\, \mathfrak{F}_{\mathbf{x}}] - \mathrm{E}[H(\mathscr{P} \cap A)])^p\right] \end{aligned}$$

$$\leq 2^p f_j(\nu_2(A)), \tag{6.22}$$

where $j = 1$, if $p = 2$, and $j = 2$, if $p = 4$. Analogously, for every $\mathbf{x} \in \mathbb{Z}_L^2$ and every convex set $B \in \mathfrak{B}^2$ such that $B \cap R_{\mathbf{x}} = \emptyset$ and $e(B, R_{\mathbf{x}}) \leq L$,

$$\mathrm{E}\left[(\mathrm{E}[H(\mathscr{P} \cap B) \,|\, \mathfrak{F}_{\mathbf{x}}] - \mathrm{E}[H(\mathscr{P} \cap \mathrm{T}_{-L\mathbf{e}_1} B) \,|\, \mathfrak{F}_{\mathbf{x}-L\mathbf{e}_1}])^p\right] \leq 2^p f_j(h(B)), \tag{6.23}$$

where $j = 3$, if $p = 2$, and $j = 4$, if $p = 4$. Moreover, from Equation (6.21), for all $B \in \mathfrak{B}^2$,

$$\frac{\mathrm{Var}\,(H(\mathscr{P} \cap B))}{\nu_2(B)^{\delta_5/2}} \underset{\nu_2(B)\to\infty}{\longrightarrow} \text{constant}. \tag{6.24}$$

Following the proof of Theorem 3.1 in [61], sufficient conditions to prove hypotheses (6.14) and (6.15) of the CLT for martingale differences are

$$o_n^{-2\gamma} \sum_{i=1}^{o_n} \mathrm{E}[D_i^2] \;<\; C < \infty, \qquad \text{for all } n \geq 1 \tag{6.25}$$

$$o_n^{-4\gamma} \sum_{i=1}^{o_n} \mathrm{E}[D_i^4] \to 0, \qquad n \to \infty. \tag{6.26}$$

Therefore, in the following, we will show that these two conditions hold for point processes that satisfy Conditions 6.5. In Section 6.4, we will prove that (6.25) and (6.26) are sufficient conditions for (6.14) and (6.15).

Proposition 6.2. *Let $\mathscr{P}$ be a stationary point process such that Conditions 6.5 hold. Let $\{B_n\}_{n\geq 1}$ be a sequence of rectangles which satisfies Conditions 6.4. Then,*

$$o_n^{-2\gamma} \sum_{i=1}^{o_n} \mathrm{E}[D_i^2] \;<\; C < \infty, \qquad \textit{for all } n \geq 1 \tag{6.27}$$

$$o_n^{-4\gamma} \sum_{i=1}^{o_n} \mathrm{E}[D_i^4] \to 0, \qquad n \to \infty, \tag{6.28}$$

hold for γ such that $\gamma \geq \max((\delta_3\beta+1)/2, (1+\delta_1(1+\beta))/2, (\delta_5\beta+2)/4)$ *and* $\gamma > \max((\delta_4\beta+1)/4, (1+\delta_2(1+\beta))/4)$.

Proof. Let us observe that, for $i \geq 2$,

$$
\begin{aligned}
&\mathrm{E}\left[\left(\sum_{j=1}^{v_n}\sum_{h=i}^{i+1}\Delta_{\mathbf{x}_{h,j},\mathbf{x}_{i,1}}\right)^p\right] \\
&\leq 2^{p-1}\mathrm{E}\left[\left(H\left(\mathscr{P}\cap\bigcup_{j=1}^{v_n}Q_{L/2}(\mathbf{x}_{2,j})\right)-\mathrm{E}\left[H\left(\mathscr{P}\cap\bigcup_{j=1}^{v_n}Q_{L/2}(\mathbf{x}_{2,j})\right)\right]\right)^p\right] \\
&+ 2^{p-1}\mathrm{E}\left[\left(\mathrm{E}\left[H\left(\mathscr{P}\cap\bigcup_{j=1}^{v_n}Q_{L/2}(\mathbf{x}_{3,j})\right)|\mathfrak{F}_{\mathbf{x}_{2,1}}\right]\right.\right. \\
&\left.\left.- \mathrm{E}\left[H\left(\mathscr{P}\cap\bigcup_{j=1}^{v_n}Q_{L/2}(\mathbf{x}_{2,j})\right)|\mathfrak{F}_{\mathbf{x}_{1,1}}\right]\right)^p\right], \qquad (6.29)
\end{aligned}
$$

$$
\begin{aligned}
&\mathrm{E}\left[\left(\sum_{j=1}^{v_n}\Delta_{\mathbf{x}_{i,j},\mathbf{x}_{i,1}}\right)^p\right] \\
&\leq 2^{p-1}\mathrm{E}\left[\left(H\left(\mathscr{P}\cap\bigcup_{j=1}^{v_n}Q_{L/2}(\mathbf{x}_{2,j})\right)-\mathrm{E}\left[H\left(\mathscr{P}\cap\bigcup_{j=1}^{v_n}Q_{L/2}(\mathbf{x}_{2,j})\right)\right]\right)^p\right] \\
&+ 2^{p-1}\mathrm{E}\left[\left(\mathrm{E}\left[H\left(\mathscr{P}\cap\bigcup_{j=1}^{v_n}Q_{L/2}(\mathbf{x}_{2,j})\right)|\mathfrak{F}_{\mathbf{x}_{1,1}}\right]\right.\right. \\
&\left.\left.- \mathrm{E}\left[H\left(\mathscr{P}\cap\bigcup_{j=1}^{v_n}Q_{L/2}(\mathbf{x}_{2,j})\right)\right]\right)^p\right], \qquad (6.30)
\end{aligned}
$$

by using Jensen's inequality and the stationarity of the process.

To show Inequality (6.27), first we use the definition of D_i in Equation (6.5) and then we apply Jensen's inequality,

$$o_n^{-2\gamma}\sum_{i=1}^{o_n}\mathrm{E}[D_i^2]$$

$$
\begin{aligned}
&\leq 3o_n^{-2\gamma}\mathrm{E}\left[\left(\sum_{j=1}^{v_n}\Delta_{\mathbf{x}_{1,j},\mathbf{x}_{1,1}}\right)^2+\left(\sum_{j=1}^{v_n}\Delta_{\mathbf{x}_{2,j},\mathbf{x}_{1,1}}\right)^2+\left(\sum_{j=1}^{v_n}\sum_{h=3}^{o_n}\Delta_{\mathbf{x}_{h,j},\mathbf{x}_{1,1}}\right)^2\right]\\
&+2o_n^{-2\gamma}\sum_{i=2}^{o_n-2}\mathrm{E}\left[\left(\sum_{j=1}^{v_n}\sum_{h=i}^{i+1}\Delta_{\mathbf{x}_{h,j},\mathbf{x}_{i,1}}\right)^2+\left(\sum_{j=1}^{v_n}\sum_{h=i+2}^{o_n}\Delta_{\mathbf{x}_{h,j},\mathbf{x}_{i,1}}\right)^2\right]\\
&+o_n^{-2\gamma}\mathrm{E}\left[\left(\sum_{j=1}^{v_n}\sum_{h=o_n-1}^{o_n}\Delta_{\mathbf{x}_{h,j},\mathbf{x}_{o_n-1,1}}\right)^2\right]+o_n^{-2\gamma}\mathrm{E}\left[\left(\sum_{j=1}^{v_n}\Delta_{\mathbf{x}_{o_n,j},\mathbf{x}_{o_n,1}}\right)^2\right].
\end{aligned}
$$

Now, to the right-hand side of the previous inequality, we apply Inequalities (6.29) and (6.30), Properties (6.17) and (6.19) of Conditions 6.5, and Properties (6.22) and (6.23),

$$
\begin{aligned}
&o_n^{-2\gamma}\sum_{i=1}^{o_n}\mathrm{E}[D_i^2]\\
&\leq o_n^{-2\gamma}(3+4(o_n-3)+2+2)\mathrm{Var}\left(H\left(\mathscr{P}\cap\bigcup_{j=1}^{v_n}Q_{L/2}(\mathbf{x}_{1,j})\right)\right)\\
&+o_n^{-2\gamma}(3+16(o_n-3)+8+2)f_3\left(h\left(\bigcup_{j=1}^{v_n}Q_{L/2}(\mathbf{x}_{2,j})\right)\right)\\
&+3o_n^{-2\gamma}f_1\left(\nu_2\left(\bigcup_{j=1}^{v_n}\bigcup_{h=3}^{o_n}Q_{L/2}(\mathbf{x}_{h,j})\right)\right)\\
&+8o_n^{-2\gamma}(o_n-3)\max_{i=2,\ldots,o_n-2}f_1\left(\nu_2\left(\bigcup_{j=1}^{v_n}\bigcup_{h=i+2}^{o_n}Q_{L/2}(\mathbf{x}_{h,j})\right)\right)\\
&=o_n^{-2\gamma}\left[3f_1\left(L^2\nu_n(o_n-2)\right)+8(o_n-3)f_1\left(L^2\nu_n(o_n-3)\right)\right]\\
&+o_n^{-2\gamma}\left[(16o_n-35)f_3\left(\nu_n\right)+(4o_n-5)\mathrm{Var}\left(H\left(\mathscr{P}\cap\bigcup_{j=1}^{v_n}Q_{L/2}(\mathbf{x}_{1,j})\right)\right)\right].
\end{aligned}
$$

Finally, using Property (6.4) of $\{o_n\}_{n\geq 1}$ and $\{v_n\}_{n\geq 1}$, the properties of f_1 and f_3 stated, respectively, in (6.17) and in (6.19), and Limit (6.24), we obtain Inequality (6.27), for $\gamma \geq \max((\delta_3\beta+1)/2, (1+\delta_1(1+\beta))/2, (\delta_5\beta+2)/4)$.

Analogously, we prove now Limit (6.28). The definition of D_i in Equation (6.5) and Jensen's inequality imply that

$$
\begin{aligned}
& o_n^{-4\gamma} \sum_{i=1}^{o_n} \mathrm{E}[D_i^4] \\
& \leq 3^3 o_n^{-4\gamma} \mathrm{E}\left[\left(\sum_{j=1}^{v_n} \Delta_{\mathbf{x}_{1,j},\mathbf{x}_{1,1}}\right)^4 + \left(\sum_{j=1}^{v_n} \Delta_{\mathbf{x}_{2,j},\mathbf{x}_{1,1}}\right)^4 + \left(\sum_{j=1}^{v_n}\sum_{h=3}^{o_n} \Delta_{\mathbf{x}_{h,j},\mathbf{x}_{1,1}}\right)^4\right] \\
& + 2^3 o_n^{-4\gamma} \sum_{i=2}^{o_n-2} \mathrm{E}\left[\left(\sum_{j=1}^{v_n}\sum_{h=i}^{i+1} \Delta_{\mathbf{x}_{h,j},\mathbf{x}_{i,1}}\right)^4 + \left(\sum_{j=1}^{v_n}\sum_{h=i+2}^{o_n} \Delta_{\mathbf{x}_{h,j},\mathbf{x}_{i,1}}\right)^4\right] \\
& + o_n^{-4\gamma} \mathrm{E}\left[\left(\sum_{j=1}^{v_n}\sum_{h=o_n-1}^{o_n} \Delta_{\mathbf{x}_{h,j},\mathbf{x}_{o_n-1,1}}\right)^4\right] + o_n^{-4\gamma} \mathrm{E}\left[\left(\sum_{j=1}^{v_n} \Delta_{\mathbf{x}_{o_n,j},\mathbf{x}_{o_n,1}}\right)^4\right].
\end{aligned}
$$

As before, to the right-hand side of the previous inequality, we apply Inequalities (6.29) and (6.30), Properties (6.18) and (6.20) of Conditions 6.5, and Properties (6.22) and (6.23),

$$
\begin{aligned}
& o_n^{-4\gamma} \sum_{i=1}^{o_n} \mathrm{E}[D_i^4] \\
& \leq o_n^{-4\gamma} \mathrm{E}\left[\left(H\left(\mathscr{P} \cap \bigcup_{j=1}^{v_n} Q_{L/2}(\mathbf{x}_{1,j})\right) - \mathrm{E}\left[H\left(\mathscr{P} \cap \bigcup_{j=1}^{v_n} Q_{L/2}(\mathbf{x}_{1,j})\right)\right]\right)^4\right] \\
& \cdot \ (3^3 + 2^6(o_n-3) + 2^3 + 2^3) + 3^3 o_n^{-4\gamma} f_2\left(v_2\left(\bigcup_{j=1}^{v_n}\bigcup_{h=3}^{o_n} Q_{L/2}(\mathbf{x}_{j,h})\right)\right) \\
& + o_n^{-4\gamma}(3^3 + 2^{10}(o_n-3) + 2^7 + 2^3) f_4\left(h\left(\bigcup_{j=1}^{v_n} Q_{L/2}(\mathbf{x}_{2,j})\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
&+ 2^7 o_n^{-4\gamma}(o_n-3) \max_{i=2,\ldots,o_n-2} f_2\left(v_2\left(\bigcup_{j=1}^{v_n}\bigcup_{h=i+2}^{o_n} Q_{L/2}(\mathbf{x}_{j,h})\right)\right)\\
&\le o_n^{-4\gamma}\left[(1024o_n-2909)f_4(v_n)+27f_2\left(L^2 v_n(o_n-2)\right)\right]\\
&+ 128(o_n-3)o_n^{-4\gamma}f_2\left(L^2 v_n(o_n-3)\right)+o_n^{-4\gamma}(64o_n-149)\cdot\\
&\cdot\ \mathrm{E}\left[\left(H\left(\mathscr{P}\cap\bigcup_{j=1}^{v_n} Q_{L/2}(\mathbf{x}_{1,j})\right)-\mathrm{E}\left[H\left(\mathscr{P}\cap\bigcup_{j=1}^{v_n} Q_{L/2}(\mathbf{x}_{1,j})\right)\right]\right)^4\right].
\end{aligned}
$$

Using Property (6.4) of $\{o_n\}_{n\ge1}$ and $\{v_n\}_{n\ge1}$, the properties of f_2 and f_4 stated, respectively, in (6.18) and in (6.20), and Limit (6.21), we obtain Inequality (6.28), for $\gamma > \max((\delta_4\beta+1)/4, (1+\delta_2(1+\beta))/4, (\delta_5\beta+1)/4)$.

As a consequence, Relations (6.27) e (6.28) hold if $\gamma \ge \max((\delta_3\beta+1)/2, (1+\delta_1(1+\beta))/2, (\delta_5\beta+2)/4)$ and $\gamma > \max((\delta_4\beta+1)/4, (1+\delta_2(1+\beta))/4)$. □

6.2.1 *Proof that Conditions 6.5 hold for a stationary point process independent at distance $l = L$*

Previously, we claimed that a suitable choice of the quantities f_i, δ_i, $i = 1, \ldots, 4$, exists such that Conditions 6.5 hold for a stationary point process independent at distance $l = L$. Now, we will prove it.

Proposition 6.3. *Let $\mathscr{P}$ be a stationary point process independent at distance $l = L < \infty$ and let H be a functional which satisfies Conditions 6.3. If $H(\mathscr{P}\cap .)$ has finite fourth moment measure, then Conditions 6.5 are satisfied for: $\delta_1 = \delta_2 = 0$, $\delta_3 = 1$, $\delta_4 = \delta_5 = 2$, f_1, $f_2 \equiv 0$, $f_3(h(B))$=$2E_{\mathbf{0},2}(9h(B)/L-11)$ and*

$$
f_4(h(B)) = \begin{cases} E_{\mathbf{0},4}82944 & \text{if } h(B) < 11L \\ E_{\mathbf{0},4}\left(972\left(\frac{h(B)}{L}\right)^2 - 6744\frac{h(B)}{L} - 41924\right) & \text{if } h(B) \ge 11L. \end{cases}
$$

Proof. If $A \in \mathfrak{B}^2$ is such that $d(A, R_{\mathbf{x}}) > L$, $\mathbf{x} \in \mathbb{Z}_L^2$, then, by using the independence of the point process at distance L, we obtain that, for any $n \in \mathbb{N}$,

$$\mathrm{E}\left[(\mathrm{E}[H(\mathscr{P} \cap A) \,|\, \mathfrak{F}_{\mathbf{x}}] - \mathrm{E}[H(\mathscr{P} \cap A)])^n\right] = (\mathrm{E}[H(\mathscr{P} \cap A)] - \mathrm{E}[H(\mathscr{P} \cap A)])^n = 0$$

and thus f_1, $f_2 \equiv 0$ and δ_1, δ_2=0.

Let $B \in \mathfrak{B}^2$ be such that $B \cap R_{\mathbf{x}} = \emptyset$ and $e(B, R_{\mathbf{x}}) \leq L$, $\mathbf{x} \in \mathbb{Z}_L^2$. Then, we can consider the smallest rectangle $R \supseteq B$ with horizontal side L and vertical side nL, $n \in \mathbb{N}$. Due to the properties of B, $n \leq \lfloor h(B) \rfloor / L + 1$ (here $\lfloor a \rfloor$ denotes the $\sup\{b \in \mathbb{N} \,|\, b \leq a\}$). Let us partition R into squares of side L and denote by Q_j the j^{th} square from the bottom. If we define

$$\begin{aligned} g(A) &= \mathrm{E}[H(\mathscr{P} \cap A) \,|\, \mathfrak{F}_{\mathbf{x}}] - \mathrm{E}[H(\mathscr{P} \cap A)], \qquad \forall A \in \mathfrak{B}^2, \\ g_j(B) &= g(B \cap Q_j), \end{aligned}$$

we obtain

$$\begin{aligned} \mathrm{E}\left[(g(B))^2\right] &= \mathrm{E}\left[\left(\sum_{j=1}^{n} g(B \cap Q_j)\right)^2\right] \\ &= \sum_{j=1}^{n} \mathrm{E}\left[(g_j(B))^2\right] + 2 \sum_{j=1}^{n-1} \sum_{j'=j+1}^{\max(n, j+4)} \mathrm{E}\left[g_j(B) g_{j'}(B)\right] \\ &= \sum_{j=1}^{n} \mathrm{E}\left[(g_j(B))^2\right] + 2 \sum_{j=n-3}^{n-1} \sum_{j'=j+1}^{n} \mathrm{E}\left[g_j(B) g_{j'}(B)\right] \\ &\quad + 2 \sum_{j=1}^{n-4} \sum_{j'=j+1}^{j+4} \mathrm{E}\left[g_j(B) g_{j'}(B)\right], \end{aligned} \tag{6.31}$$

where in the second step we used the fact that $g_j(B)$ and $g_{j'}(B)$ are independent if $j' > j + 4$ (see Lemma 6.1) and $\mathrm{E}[g_j(B)] = 0$, for every j.

Let us derive now some upper bounds for the quantities which appear in Equation (6.31). Regarding the first term, by using the Jensen's inequality, the stationarity of the point process and the additivity and positivity of

$H(\mathscr{P} \cap .)$, we obtain that

$$\begin{aligned}
\mathrm{E}\left[(g_j(B))^2\right] &\leq \mathrm{E}\left[\mathrm{E}[H(\mathscr{P} \cap B \cap Q_j) \mid \mathfrak{F}_{\mathbf{x}}]^2 + \mathrm{E}[H(\mathscr{P} \cap B \cap Q_j)]^2\right] \\
&\leq \mathrm{E}\left[\mathrm{E}\left[H(\mathscr{P} \cap B \cap Q_j)^2 \mid \mathfrak{F}_{\mathbf{x}}\right] + \mathrm{E}\left[H(\mathscr{P} \cap B \cap Q_j)^2\right]\right] \\
&= 2\mathrm{E}\left[H(\mathscr{P} \cap B \cap Q_j)^2\right] \leq 2E_{\mathbf{0},2}. \qquad (6.32)
\end{aligned}$$

For the expected value of the product, by using Holder's inequality and Inequality (6.32), we obtain

$$\begin{aligned}
\mathrm{E}\left[g_j(B) g_{j'}(B)\right] &\leq \mathrm{E}\left[\left|g_j(B) g_{j'}(B)\right|\right] \\
&\leq \mathrm{E}\left[(g_j(B))^2\right]^{\frac{1}{2}} \mathrm{E}\left[\left(g_{j'}(B)\right)^2\right]^{\frac{1}{2}} \\
&\leq 2E_{\mathbf{0},2}. \qquad (6.33)
\end{aligned}$$

Finally, by using the upper bounds in (6.32) and (6.33) and the condition $n \leq \lfloor h(B) \rfloor / L + 1$, Equation (6.31) becomes

$$\begin{aligned}
\mathrm{E}\left[(g(B))^2\right] &\leq 2E_{\mathbf{0},2}(9n - 20) \\
&\leq 2E_{\mathbf{0},2}\left(\frac{9h(B)}{L} - 11\right).
\end{aligned}$$

Hence, Relation (6.19) is satisfied for $f_3(h(B)) = 2E_{\mathbf{0},2}(9h(B)/L - 11)$ and $\delta_3 = 1$.

In a similar way, we can derive δ_4 and the form of function f_4 in Relation (6.20). By using the independence of $g_j(B)$ and $g_{j'}(B)$ if $j' > j + 4$ (see Lemma 6.1) and that $\mathrm{E}[g_j(B)] = 0$ for all j,

$$\begin{aligned}
&\mathrm{E}\left[(g(B))^4\right] \\
&= \mathrm{E}\left[\left(\sum_{j=1}^{n} g(B \cap Q_j)\right)^4\right] \qquad (6.34)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{n} \mathrm{E}\left[(g_j(B))^4\right] + 6\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n} \mathrm{E}\left[g_j(B)^2 g_{j'}(B)^2\right] \\
&+ 4\sum_{j=1}^{n-1}\sum_{j'=j+1}^{\min(j+4,n)} \mathrm{E}\left[g_j(B) g_{j'}(B)^3\right] + 4\sum_{j=1}^{n-1}\sum_{j'=j+1}^{\min(j+4,n)} \mathrm{E}\left[g_j(B)^3 g_{j'}(B)\right] \\
&+ 12\sum_{j=1}^{n-2}\sum_{j'=j+1}^{\min(j+4,n-1)}\sum_{h=j'+1}^{n} \mathrm{E}\left[g_j(B) g_{j'}(B) g_h(B)^2\right] \\
&+ 12\sum_{j=1}^{n-2}\sum_{j'=j+1}^{n-1}\sum_{h=j'+1}^{\min(j'+4,n)} \mathrm{E}\left[g_j(B)^2 g_{j'}(B) g_h(B)\right] \\
&+ 12\sum_{j=1}^{n-2}\sum_{j'=j+1}^{\min(j+4,n-1)}\sum_{h=j'+1}^{\min(j'+4,n)} \mathrm{E}\left[g_j(B) g_{j'}(B)^2 g_h(B)\right] \\
&+ 24\sum_{j=1}^{n-3}\sum_{j'=j+1}^{\min(j+4,n-2)}\sum_{h=j'+1}^{n-1}\sum_{h'=h+1}^{\min(h+4,n)} \mathrm{E}\left[g_j(B) g_{j'}(B) g_h(B) g_{h'}(B)\right]. \quad (6.35)
\end{aligned}
$$

We can derive an upper bound of the first term in Equation (6.35), by using the Jensen's inequality, the stationarity of the point process and the additivity and positivity of $H(\mathscr{P} \cap .)$,

$$
\begin{aligned}
\mathrm{E}\left[(g_j(B))^4\right] &\leq \mathrm{E}\left[\left(\mathrm{E}[H(\mathscr{P} \cap B \cap Q_j) \mid \mathfrak{F}_{\mathbf{x}}]^2 + \mathrm{E}[H(\mathscr{P} \cap B \cap Q_j)]^2\right)^2\right] \\
&\leq 2\mathrm{E}\left[\mathrm{E}[H(\mathscr{P} \cap B \cap Q_j) \mid \mathfrak{F}_{\mathbf{x}}]^4 + \mathrm{E}[H(\mathscr{P} \cap B \cap Q_j)]^4\right] \\
&\leq 2\mathrm{E}\left[\mathrm{E}\left[H(\mathscr{P} \cap B \cap Q_j)^4 \mid \mathfrak{F}_{\mathbf{x}}\right] + \mathrm{E}\left[H(\mathscr{P} \cap B \cap Q_j)^4\right]\right] \\
&= 4\mathrm{E}\left[H(\mathscr{P} \cap B \cap Q_j)^4\right] \leq 4E_{\mathbf{0},4}. \quad (6.36)
\end{aligned}
$$

Moreover, by applying Holder's inequality, its generalization and Inequality (6.36), the following inequalities hold

$$
\mathrm{E}\left[g_j^2(B) g_{j'}^2(B)\right] \leq \mathrm{E}[g_j^4(B)]^{\frac{1}{2}} \mathrm{E}[g_{j'}^4(B)]^{\frac{1}{2}} \leq 4E_{\mathbf{0},4}
$$

$$\begin{aligned}
\mathrm{E}\left[g_j(B)g_{j'}^3(B)\right] &\leq \mathrm{E}[g_j^4(B)]^{\frac{1}{4}}\mathrm{E}[g_{j'}^4(B)]^{\frac{3}{4}} \leq 4E_{\mathbf{0},4}\\
\mathrm{E}\left[g_j(B)g_{j'}(B)g_h^2(B)\right] &\leq \mathrm{E}[g_j^4(B)]^{\frac{1}{4}}\mathrm{E}[g_{j'}^4(B)]^{\frac{1}{4}}\mathrm{E}[g_h^4(B)]^{\frac{1}{2}} \leq 4E_{\mathbf{0},4}\\
\mathrm{E}\left[g_j(B)g_{j'}(B)g_h(B)g_{h'}(B)\right] &\leq \mathrm{E}[g_j^4(B)]^{\frac{1}{4}}\mathrm{E}[g_{j'}^4(B)]^{\frac{1}{4}}\mathrm{E}[g_h^4(B)]^{\frac{1}{4}}\mathrm{E}[g_{h'}^4(B)]^{\frac{1}{4}}\\
&\leq 4E_{\mathbf{0},4},
\end{aligned}$$

for any j, j', h and h'. Putting these results in Equation (6.35), for $n \geq 12$, we obtain

$$\mathrm{E}\left[(g(B))^4\right] \leq 4E_{\mathbf{0},4}(243n^2 - 2172n - 8552),$$

and thus for $h(B) > 11L$ (remembering that $n \leq \lfloor h(B)\rfloor/L + 1$),

$$\mathrm{E}\left[(g(B))^4\right] \leq E_{\mathbf{0},4}\left(972\left(\frac{h(B)}{L}\right)^2 - 6744\frac{h(B)}{L} - 41924\right).$$

For $h(B) \leq 11L$ (i.e. $n < 12$), we can simply use Jensen's inequality to the right-hand side of Equation (6.34) and Inequality (6.36),

$$\begin{aligned}
\mathrm{E}\left[(g(B))^4\right] &\leq n^3 \sum_{j=1}^{n} \mathrm{E}\left[(g\left(B \cap Q_j\right))^4\right]\\
&\leq n^4 4E_{\mathbf{0},4}\\
&< 82944E_{\mathbf{0},4}.
\end{aligned}$$

As a consequence, Relation (6.20) is satisfied for

$$f_4(h(B)) = \begin{cases} 82944E_{\mathbf{0},4} & \text{if } h(B) \leq 11L\\ E_{\mathbf{0},4}\left(972\left(\frac{h(B)}{L}\right)^2 - 6744\frac{h(B)}{L} - 41924\right) & \text{if } h(B) > 11L, \end{cases}$$

and $\delta_4 = 2$.

Analogously, using

$$g(A) = H(\mathscr{P} \cap A) - \mathrm{E}[H(\mathscr{P} \cap A)], \qquad \forall A \in \mathfrak{B}^2,$$

and the independence at distance $l = L$, we can prove Limit (6.21) for $\delta_5 = 2$. □

6.3 Proof of the third condition of the CLT for martingale differences for a stationary point process independent at distance l

In this section, we show the third hypothesis of the CLT for martingale differences (see Limit (6.16)) assuming the independence at distance l. Moreover, we consider two kinds of increasing sequences of rectangles $\{B_n\}_{n\geq 1}$: in the former all rectangles will have fixed vertical side (Proposition 6.5), in the the latter they will satisfy Conditions 6.4 with $L = l$ (Proposition 6.6).

Proposition 6.5 can be applied to point processes belonging either to $\mathbb{R} \times [a,b]$ or $\mathbb{R}_+ \times [a,b]$ (for some a, $b \in \mathbb{R}$). For example, a line process can be seen as a point process in $\mathbb{R} \times (0, 2\pi]$, if we parameterize each line ℓ with the signed distance between ℓ and the origin $\mathbf{0}$ and the angle between the abscissa axis and the line orthogonal to ℓ passing through $\mathbf{0}$ [78].

Proposition 6.4. *Let $\mathscr{P}$ be a stationary point process such that $H(\mathscr{P} \cap .)$ has finite second moment measure. Let us assume that there exist $0 < l < \infty$ such that $\mathscr{P} \cap A$ and $\mathscr{P} \cap B$ are independent for each A, $B \in \mathfrak{B}^2$ with $d(A,B) > l$. Let $\{B_n\}_{n\geq 1}$ be an increasing sequence of rectangles such that, for each n, B_n has vertical side $L \cdot V$ ($V \in \mathbb{N}$) and horizontal side $L \cdot o_n$, with $L = l$ ($o_n \in \mathbb{N}$ and $\{o_n\}_{n\geq 1}$ satisfies Limit (6.3)). Then*

$$\frac{1}{o_n - 2} \sum_{i=1}^{o_n} D_i^2 \xrightarrow[n\to\infty]{L^1} \mathrm{E}\left[\left(\sum_{j=1}^{V} M_{(L,jL),L\mathbf{u}} \right)^2 \right],$$

where $\mathbf{u} = (1,1)$.

Proof. Since the point process is independent at distance $l = L$, $D_i = \sum_{j=1}^{V} M_{i,j}$ if $i < o_n$ (see Equation (6.10)) and $D_{o_n} = \sum_{j=1}^{V} \Delta_{\mathbf{x}_{o_n,j},\mathbf{x}_{o_n,1}}$.

Let us notice that, by using Jensen's inequality and Equation (6.7),

$$\begin{aligned}
\mathrm{E}\left[\left|\frac{1}{o_n-2}D_1^2\right|\right] &\leq \frac{2V}{o_n-2}\sum_{j=1}^{V}\sum_{h=1}^{2}\mathrm{E}\left[\Delta^2_{\mathbf{x}_{h,j},\mathbf{x}_{1,1}}\right] \\
&\leq \frac{8V^2E_{\mathbf{0},2}}{o_n-2}\xrightarrow[n\to\infty]{}0 \\
\mathrm{E}\left[\left|\frac{1}{o_n-2}D_{o_n}^2\right|\right] &\leq \frac{V}{o_n-2}\sum_{j=1}^{V}\mathrm{E}\left[\Delta^2_{\mathbf{x}_{o_n,j},\mathbf{x}_{o_n,1}}\right] \\
&\leq \frac{2V^2E_{\mathbf{0},2}}{o_n-2}\xrightarrow[n\to\infty]{}0.
\end{aligned}$$

Therefore, to prove the thesis it is sufficient to show that

$$\frac{1}{o_n-2}\sum_{i=2}^{o_n-1}D_i^2\xrightarrow[n\to\infty]{L^1}\mathrm{E}\left[\left(\sum_{j=1}^{V}M_{(L,jL),L\mathbf{u}}\right)^2\right].$$

Let us define

$$F_{iL\mathbf{e}_1}=\sum_{j=1}^{V}M_{(iL,jL),iL\mathbf{e}_1},$$

for all $i\in\mathbb{N}$. Due to the properties of the point process, $\{F_{iL\mathbf{e}_1}\}_{i\in\mathbb{N}}$ is a stationary and ergodic sequence. Since $(.)^2$ is a continuous function, we can use the Birkhoff Ergodic Theorem (Theorem 2.3 in [62]) and thus, for any $\varepsilon>0$, there exists $M>0$ such that for all $m>M$,

$$\mathrm{E}\left[\left|\frac{1}{m}\sum_{i=1}^{m}F^2_{iL\mathbf{e}_1}-\mathrm{E}[F^2_{L\mathbf{e}_1}]\right|\right]<\varepsilon.$$

As a consequence, due to the stationarity of the point process, for any $\varepsilon>0$, we can take $N>0$ such that, for all $n>N$, $o_n-2>M$, and it holds

$$\mathrm{E}\left[\left|\frac{1}{o_n-2}\sum_{i=2}^{o_n-1} D_i^2 - \mathrm{E}[F^2_{L\mathbf{e}_1}]\right|\right] < \varepsilon.$$

Since $\mathrm{E}\left[\left(\sum_{j=1}^{V} M_{(L,jL),L\mathbf{u}}\right)^2\right] = \mathrm{E}[F^2_{L\mathbf{e}_1}]$, we have shown the thesis. □

We can observe that the proof applies also to rectangles $\{B_n\}_{n\geq 1}$ with a vertical side of any fixed length. Therefore, the following result holds.

Proposition 6.5. *Let $\mathscr{P}$ be a point process such that $H(\mathscr{P}\cap .)$ has finite second moment measure. Let us assume that there exist $0<l<\infty$ such that $\mathscr{P}\cap A$ and $\mathscr{P}\cap B$ are independent for each A, $B\in\mathfrak{B}^2$ with $d(A,B)>l$. Let $\{B_n\}_{n\geq 1}$ be an increasing sequence of rectangles such that, for each n, B_n has vertical side V ($V\in\mathbb{R}_+$) and horizontal side $L\cdot o_n$, with $L=l$ ($o_n\in\mathbb{N}$ and $\{o_n\}_{n\geq 1}$ satisfies Limit (6.3)). Then,*

$$\frac{1}{o_n-2}\sum_{i=1}^{o_n} D_i^2 \xrightarrow[n\to\infty]{L^1} \mathrm{E}\left[M^2\right],$$

where

$$\begin{aligned} M = {} & H(\mathscr{P}\cap\mathscr{R}_{L/2,V/2}(\mathbf{0})) - \mathrm{E}[H(\mathscr{P}\cap\mathscr{R}_{L/2,V/2}(\mathbf{0}))\,|\,\mathfrak{F}_{-L\mathbf{e}_1}] \\ & + \mathrm{E}[H(\mathscr{P}\cap\mathscr{R}_{L/2,V/2}(L\mathbf{e}_1))\,|\,\mathfrak{F}_{\mathbf{0}}] - \mathrm{E}[H(\mathscr{P}\cap R_{L/2,V/2}(\mathbf{0}))] \end{aligned}$$

and $\mathscr{R}_{l_1/2,l_2/2}(\mathbf{x})$ is a rectangle with horizontal side l_1, vertical side l_2 and centered at $\mathbf{x}$.

In case of a sequence $\{B_n\}_{n\geq 1}$ without fixed vertical side, first we need to prove the following two lemmas.

Lemma 6.2. *Let $\mathscr{P}$ be a stationary point process such that $H(\mathscr{P}\cap .)$ has finite second moment measure. Moreover, let $\{v_n\}_{n\in\mathbb{N}}$ and $\{o_n\}_{n\in\mathbb{N}}$ be two sequences of integers, which satisfy the properties stated in Conditions 6.4. If there exists $d_{max}\in\mathbb{N}$ such that $d_{max}<\infty$ and*

$$\mathrm{E}[M_{L\mathbf{u},L\mathbf{u}} M_{(L,jL),L\mathbf{u}}] = 0, \qquad \forall j > d_{max}, \tag{6.37}$$

and

$$\sup_{\substack{d\in\mathbb{N}\\ d\geq 2}} \mathrm{E}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\sum_{j=2}^{d} M_{(iL,L),(iL,L)}M_{(iL,jL),(iL,L)} - \sum_{j=2}^{d}\mathrm{E}[M_{L\mathbf{u},L\mathbf{u}}M_{(L,jL),L\mathbf{u}}]\right|\right] \underset{m\to\infty}{\longrightarrow} 0, \tag{6.38}$$

then

$$\frac{1}{(o_n-2)(v_n-1)}\sum_{i=1}^{o_n-2}\sum_{j=1}^{v_n-1}\sum_{j'=j+1}^{v_n} M_{(iL,jL),(iL,L)}M_{(iL,j'L),(iL,L)} \underset{n\to\infty}{\longrightarrow} \sum_{j=2}^{d_{max}}\mathrm{E}[M_{L\mathbf{u},L\mathbf{u}}M_{(L,jL),L\mathbf{u}}], \tag{6.39}$$

in L^1 norm.

Proof. In order to prove Limit (6.39), it is sufficient to show that the following two limits hold,

$$\begin{aligned} I_{1n} &:= \mathrm{E}\left[\left|\frac{1}{(o_n-2)(v_n-1)}\sum_{i=1}^{o_n-2}\sum_{j=1}^{v_n-1}\sum_{j'=j+1}^{v_n} M_{(iL,jL),(iL,L)}M_{(iL,j'L),(iL,L)}\right.\right.\\ &\quad \left.\left. - \frac{1}{v_n-1}\sum_{j=1}^{v_n-1}\mathrm{E}\left[\sum_{j'=2}^{v_n-j+1} M_{L\mathbf{u},L\mathbf{u}}M_{(L,j'L),L\mathbf{u}}\right]\right|\right] \underset{n\to\infty}{\longrightarrow} 0,\\ I_{2n} &:= \left|\mathrm{E}\left[\frac{1}{v_n-1}\sum_{j=1}^{v_n-1}\sum_{j'=2}^{v_n-j+1} M_{L\mathbf{u},L\mathbf{u}}M_{(L,j'L),L\mathbf{u}}\right] - \sum_{j=2}^{d_{max}}\mathrm{E}[M_{L\mathbf{u},L\mathbf{u}}M_{(L,jL),L\mathbf{u}}]\right| \underset{n\to\infty}{\longrightarrow} 0. \end{aligned}$$

First, we prove that $\lim_{n\to\infty} I_{1n} = 0$. Due to Hypothesis (6.38), for any $\varepsilon > 0$, there exists $M > 0$ such that for each $m > M$ and $d \geq 2$,

$$\mathrm{E}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\sum_{j=2}^{d} M_{(iL,L),(iL,L)}M_{(iL,jL),(iL,L)} - \sum_{j=2}^{d}\mathrm{E}[M_{L\mathbf{u},L\mathbf{u}}M_{(L,jL),L\mathbf{u}}]\right|\right] < \varepsilon.$$

Now we can choose $N > 0$ such that, for each $n > N$, $o_n - 2 > M$ and thus

$$I_{1n} = \mathrm{E}\left[\left|\frac{1}{(v_n-1)}\sum_{j=1}^{v_n-1}\left(\frac{1}{(o_n-2)}\sum_{i=1}^{o_n-2}\sum_{j'=j+1}^{v_n} M_{(iL,jL),(iL,L)}M_{(iL,j'L),(iL,L)}\right.\right.\right.$$

$$
\begin{aligned}
&- \mathrm{E}\left[\sum_{j'=2}^{v_n-j+1} M_{L\mathbf{u},L\mathbf{u}} M_{(L,j'L),L\mathbf{u}}\right]\Bigg)\Bigg| \Bigg] \\
&\leq \frac{1}{(v_n-1)} \sum_{j=1}^{v_n-1} \mathrm{E}\left[\left| \frac{1}{(o_n-2)} \sum_{i=1}^{o_n-2} \sum_{j'=j+1}^{v_n} M_{(iL,jL),(iL,L)} M_{(iL,j'L),(iL,L)} \right.\right. \\
&- \left.\left. \mathrm{E}\left[\sum_{j'=2}^{v_n-j+1} M_{L\mathbf{u},L\mathbf{u}} M_{(L,j'L),L\mathbf{u}}\right]\right|\right] \\
&\leq \frac{1}{(v_n-1)} \sum_{j=1}^{v_n-1} \varepsilon = \varepsilon.
\end{aligned}
$$

As a consequence $\lim_{n\to\infty} I_{1n} = 0$.

To prove $\lim_{n\to\infty} I_{2n} = 0$, first we use the linearity of the expected value and Property (6.37)

$$
\begin{aligned}
I_{2n} &= \left| \frac{1}{(v_n-1)} \sum_{j=2}^{v_n} \sum_{j'=1}^{v_n-j+1} \mathrm{E}\left[M_{L\mathbf{u},L\mathbf{u}} M_{(L,jL),L\mathbf{u}}\right] - \sum_{j=2}^{d_{max}} \mathrm{E}[M_{L\mathbf{u},L\mathbf{u}} M_{(L,jL),L\mathbf{u}}] \right| \\
&= \frac{1}{(v_n-1)} \left| \sum_{j=2}^{d_{max}} (2-j) \mathrm{E}[M_{L\mathbf{u},L\mathbf{u}} M_{(L,jL),L\mathbf{u}}] \right| \\
&\leq \frac{1}{(v_n-1)} \sum_{j=2}^{d_{max}} (j-2) \mathrm{E}\left[\left| M_{L\mathbf{u},L\mathbf{u}} M_{(L,jL),L\mathbf{u}} \right|\right].
\end{aligned}
$$

Finally, to the right-hand side of the previous inequality, we apply Holder's inequality and that $H(\mathscr{P} \cap .)$ has finite second moment measure, obtaining the thesis,

$$
\begin{aligned}
I_{2n} &\leq \frac{1}{(v_n-1)} \sum_{j=2}^{d_{max}} (j-2) \mathrm{E}[M^2_{L\mathbf{u},L\mathbf{u}}]^{\frac{1}{2}} \mathrm{E}[M^2_{(L,jL),L\mathbf{u}}]^{\frac{1}{2}} \\
&= \frac{(d_{max}-1)(d_{max}-2)}{2(v_n-1)} \mathrm{E}[M^2_{L\mathbf{u},L\mathbf{u}}] \to 0, \qquad n \to \infty.
\end{aligned}
$$

□

Lemma 6.3. *Let $\mathscr{P}$ be a stationary point process such that $H(\mathscr{P}\cap .)$ has finite second moment measure. Moreover, let $\{v_n\}_{n\in\mathbb{N}}$ and $\{o_n\}_{n\in\mathbb{N}}$ be two sequences of integers, which satisfy the properties stated in Conditions 6.4. If*

$$\sum_{d=1}^{\infty} \mathrm{E}[\Delta_{L\mathbf{u},L\mathbf{u}}\Delta_{(L,(1+d)L),L\mathbf{u}}] < \infty, \tag{6.40}$$

and

$$\sup_{d\in\mathbb{N}} \mathrm{E}\left[\left|\frac{1}{m}\sum_{j=1}^{m} \Delta_{(L,jL),L\mathbf{u}}\Delta_{(L,(j+d)L),L\mathbf{u}} - \mathrm{E}[\Delta_{L\mathbf{u},L\mathbf{u}}\Delta_{(L,(1+d)L),L\mathbf{u}}]\right|\right] \underset{m\to\infty}{\longrightarrow} 0, \tag{6.41}$$

then

$$\frac{1}{(o_n-2)(v_n-1)} \sum_{j=1}^{v_n-1}\sum_{j'=j+1}^{v_n} \Delta_{(L,jL),L\mathbf{u}}\Delta_{(L,j'L),L\mathbf{u}} \xrightarrow[n\to\infty]{L^1} 0. \tag{6.42}$$

Moreover, set $S_{j,d} = (\Delta^0_{(L,jL)} + \Delta^0_{(2L,jL),L\mathbf{u}})(\Delta^0_{(L,(j+d)L)} + \Delta^0_{(2L,(j+d)L),L\mathbf{u}})$, if

$$\sum_{d=1}^{\infty} \mathrm{E}[S_{1,d}] < \infty, \tag{6.43}$$

and

$$\sup_{d\in\mathbb{N}} \mathrm{E}\left[\left|\frac{1}{m}\sum_{j=1}^{m} S_{j,d} - \mathrm{E}[S_{1,d}]\right|\right] \underset{m\to\infty}{\longrightarrow} 0, \tag{6.44}$$

then,

$$\frac{1}{(o_n-2)(v_n-1)} \sum_{j=1}^{v_n-1}\sum_{j'=j+1}^{v_n} (\Delta^0_{(L,jL)} + \Delta^0_{(2L,jL),L\mathbf{u}})(\Delta^0_{(L,j'L)} + \Delta^0_{(2L,j'L),L\mathbf{u}}) \underset{n\to\infty}{\longrightarrow} 0, \tag{6.45}$$

in L^1 norm.

Proof. First, we prove Limit (6.42). Limit (6.41) implies that, for any $\varepsilon > 0$, there exists $M > 0$ such that for each $m > M$,

$$\mathrm{E}\left[\left|\frac{1}{m}\sum_{j=1}^{m}\Delta_{(L,jL),L\mathbf{u}}\Delta_{(L,(j+d)L),L\mathbf{u}}-\mathrm{E}[\Delta_{L\mathbf{u},L\mathbf{u}}\Delta_{(L,(1+d)L),L\mathbf{u}}]\right|\right]<\varepsilon,$$

for any $d\in\mathbb{N}$. Then, fixed $\varepsilon>0$, for each n such that $v_n>M$, we define $m_n=v_n-1-M$ and we can rewrite the sum in Limit (6.42) in the following way,

$$\begin{aligned}
&\frac{2}{(o_n-2)(v_n-1)}\sum_{d=1}^{v_n-1}\sum_{j=1}^{v_n-d}\Delta_{(L,jL),L\mathbf{u}}\Delta_{(L,(j+d)L),L\mathbf{u}}\\
&=\frac{2}{(o_n-2)(v_n-1)}\sum_{d=1}^{m_n}\sum_{j=1}^{v_n-d}\Delta_{(L,jL),L\mathbf{u}}\Delta_{(L,(j+d)L),L\mathbf{u}}\qquad(=:J_{1n})\\
&+\frac{2}{(o_n-2)(v_n-1)}\sum_{d=m_n+1}^{v_n-1}\sum_{j=1}^{v_n-d}\Delta_{(L,jL),L\mathbf{u}}\Delta_{(L,(j+d)L),L\mathbf{u}}\qquad(=:J_{2n}).
\end{aligned}$$

To prove that $\lim_{n\to\infty}J_{1n}=0$ in L^1 norm, first we note that

$$\begin{aligned}
&\mathrm{E}[|J_{1n}|]\\
&\le\mathrm{E}\left[\left|\frac{2}{o_n-2}\sum_{d=1}^{m_n}\left(\frac{1}{v_n-1}\sum_{j=1}^{v_n-d}\Delta_{(L,jL),L\mathbf{u}}\Delta_{(L,(j+d)L),L\mathbf{u}}-\mathrm{E}[\Delta_{L\mathbf{u},L\mathbf{u}}\Delta_{(L,(1+d)L),L\mathbf{u}}]\right)\right|\right]\\
&+\left|\frac{2}{o_n-2}\sum_{d=1}^{m_n}\mathrm{E}[\Delta_{L\mathbf{u},L\mathbf{u}}\Delta_{(L,(1+d)L),L\mathbf{u}}]\right|.
\end{aligned}$$

From the definition of m_n, it turns out that

$$\begin{aligned}
&\mathrm{E}\left[\left|\frac{1}{o_n-2}\sum_{d=1}^{m_n}\left(\frac{1}{v_n-1}\sum_{j=1}^{v_n-d}\Delta_{(L,jL),L\mathbf{u}}\Delta_{(L,(j+d)L),L\mathbf{u}}-\mathrm{E}[\Delta_{L\mathbf{u},L\mathbf{u}}\Delta_{(L,(1+d)L),L\mathbf{u}}]\right)\right|\right]\\
&\le\frac{1}{o_n-2}\sum_{d=1}^{m_n}\mathrm{E}\left[\left|\frac{1}{v_n-1}\sum_{j=1}^{v_n-d}\Delta_{(L,jL),L\mathbf{u}}\Delta_{(L,(j+d)L),L\mathbf{u}}-\mathrm{E}[\Delta_{L\mathbf{u},L\mathbf{u}}\Delta_{(L,(1+d)L),L\mathbf{u}}]\right|\right]\\
&\le\frac{1}{o_n-2}m_n\varepsilon\\
&\le\frac{v_n-1}{o_n-2}\varepsilon
\end{aligned}$$

$\leq 2\varepsilon$.

Moreover, Inequality (6.40) implies that

$$\left| \frac{1}{o_n - 2} \sum_{d=1}^{m_n} \mathrm{E}[\Delta_{L\mathbf{u},L\mathbf{u}} \Delta_{(L,(1+d)L),L\mathbf{u}}] \right| \underset{n\to\infty}{\longrightarrow} 0$$

and, consequently, $\lim_{n\to\infty} J_{1n} = 0$ in norm L^1. Regarding J_{2n},

$$\begin{aligned}
\mathrm{E}[|J_{2n}|] &\leq \frac{2}{(o_n-2)(v_n-1)} \sum_{d=m_n+1}^{v_n-1} \sum_{j=1}^{v_n-d} \mathrm{E}[|\Delta_{(L,jL),L\mathbf{u}} \Delta_{(L,(j+d)L),L\mathbf{u}}|] \\
&\leq \frac{2}{(o_n-2)(v_n-1)} \sum_{d=m_n+1}^{v_n-1} (v_n-d)\mathrm{E}[|\Delta_{L\mathbf{u},L\mathbf{u}} \Delta_{(L,(1+d)L),L\mathbf{u}}|] \\
&= \frac{2}{(o_n-2)(v_n-1)} \sum_{d_1=1}^{M} d_1 \mathrm{E}[|\Delta_{L\mathbf{u},L\mathbf{u}} \Delta_{(L,(1+v_n-d_1)L),L\mathbf{u}}|] \\
&\leq \frac{2}{(o_n-2)(v_n-1)} \sum_{d_1=1}^{M} d_1 \mathrm{E}[\Delta^2_{L\mathbf{u},L\mathbf{u}}]^{1/2} \mathrm{E}[\Delta^2_{(L,(1+v_n-d_1)L),L\mathbf{u}}]^{1/2} \\
&= \frac{2\mathrm{E}[\Delta^2_{L\mathbf{u},L\mathbf{u}}]}{(o_n-2)(v_n-1)} \frac{M(M+1)}{2} \underset{m\to\infty}{\longrightarrow} 0,
\end{aligned}$$

where we used the definition of M, Holder's inequality and the stationarity of the process.

The proof of Limit (6.45) needs the same steps used to show Limit (6.42). □

Now we prove that all the assumptions of Lemmas 6.2 and 6.3, apart from Limit (6.38), are satisfied by any stationary point process independent at distance l.

Lemma 6.4. *Let $\mathscr{P}$ be a stationary point process such that $H(\mathscr{P} \cap .)$ has finite second moment measure. Let us assume that there exist $0 < l < \infty$, such that $\mathscr{P} \cap A$ and $\mathscr{P} \cap B$ are independent for each A, $B \in \mathfrak{B}^2$ with*

$d(A,B) > l$. Then, assuming $L = l$, the assumptions of Lemma 6.3 hold and

$$\mathrm{E}\left[M_{L\mathbf{u},L\mathbf{u}} M_{(L,jL),L\mathbf{u}}\right] = 0, \qquad \forall j > 4.$$

Proof. From the definition of $M_{\mathbf{y},\mathbf{x}}$ and Lemma 6.1, it turns out that $\{M_{(L,jL),L\mathbf{u}}\}_{j\in\mathbb{N}}$ have all the same distribution with $\mathrm{E}[M_{L\mathbf{u},L\mathbf{u}}] = 0$ and, if $d \geq 4$, $M_{(L,jL),L\mathbf{u}}$ and $M_{(L,(j+d)L),L\mathbf{u}}$ are independent for any j. Thus,

$$\mathrm{E}[M_{L\mathbf{u},L\mathbf{u}} M_{(L,jL),L\mathbf{u}}] = \mathrm{E}[M_{L\mathbf{u},L\mathbf{u}}]\mathrm{E}[M_{(L,jL),L\mathbf{u}}] = 0, \qquad \forall j > 4.$$

Similarly,

$$\begin{aligned}
&\sum_{d=1}^{\infty} \mathrm{E}[\Delta_{L\mathbf{u},L\mathbf{u}} \Delta_{(L,(1+d)L),L\mathbf{u}}] \\
&= \sum_{j=2}^{4} \mathrm{E}[\Delta_{L\mathbf{u},L\mathbf{u}} \Delta_{(L,jL),L\mathbf{u}}] + \sum_{j=5}^{\infty} \mathrm{E}[\Delta_{L\mathbf{u},L\mathbf{u}}]\mathrm{E}[\Delta_{(L,jL),L\mathbf{u}}] \\
&= \sum_{d=1}^{3} \mathrm{E}[\Delta_{L\mathbf{u},L\mathbf{u}} \Delta_{(L,(1+d)L),L\mathbf{u}}]
\end{aligned}$$

and, using Holder's inequality,

$$\begin{aligned}
\sum_{d=1}^{3} \mathrm{E}[\Delta_{L\mathbf{u},L\mathbf{u}} \Delta_{(L,(1+d)L),L\mathbf{u}}] &\leq \sum_{d=1}^{3} \mathrm{E}[\Delta^2_{L\mathbf{u},L\mathbf{u}}]^{\frac{1}{2}} \mathrm{E}[\Delta^2_{(L,(1+d)L),L\mathbf{u}}]^{\frac{1}{2}} \\
&= 3\mathrm{E}[\Delta^2_{L\mathbf{u},L\mathbf{u}}] < \infty,
\end{aligned}$$

because $H(\mathscr{P} \cap .)$ has finite second moment measure. Using the same steps, we can show that Inequality (6.43) holds.

To show Limit (6.41), we observe that, due to the independence of the process at distance greater than L, the sequence $\{\Delta_{(L,jL),L\mathbf{u}} \Delta_{(L,(j+d)L),L\mathbf{u}}\}_{j\in\mathbb{N}}$ is ergodic for each $d \in \mathbb{N}$. Then, for the Birkhoff ergodic theorem

$$\mathrm{E}\left[\left|\frac{1}{m}\sum_{j=1}^{m} \Delta_{(L,jL),L\mathbf{u}} \Delta_{(L,(j+d)L),L\mathbf{u}} - \mathrm{E}[\Delta_{L\mathbf{u},L\mathbf{u}} \Delta_{(L,(1+d)L),L\mathbf{u}}]\right|\right] \underset{m\to\infty}{\longrightarrow} 0,$$

for any $d \in \mathbb{N}$. Moreover, since $\Delta_{(L,jL),L\mathbf{u}}$ and $\Delta_{(L,(j+d)L),L\mathbf{u}}$ are independent for any $j \in \mathbb{N}$ when $d \geq 4$, the product $\Delta_{(L,jL),L\mathbf{u}}\Delta_{(L,(j+d)L),L\mathbf{u}}$ has always the same distribution for any $j \in \mathbb{N}$ and $d \geq 4$. Thus,

$$\sup_{d\in\mathbb{N}} \mathrm{E}\left[\left|\frac{1}{m}\sum_{j=1}^{m} \Delta_{(L,jL),L\mathbf{u}}\Delta_{(L,(j+d)L),L\mathbf{u}} - \mathrm{E}[\Delta_{L\mathbf{u},L\mathbf{u}}\Delta_{(L,(1+d)L),L\mathbf{u}}]\right|\right]$$
$$= \sup_{d\leq 4} \mathrm{E}\left[\left|\frac{1}{m}\sum_{j=1}^{m} \Delta_{(L,jL),L\mathbf{u}}\Delta_{(L,(j+d)L),L\mathbf{u}} - \mathrm{E}[\Delta_{L\mathbf{u},L\mathbf{u}}\Delta_{(L,(1+d)L),L\mathbf{u}}]\right|\right] \underset{m\to\infty}{\longrightarrow} 0.$$

Limit (6.44) can be proven analogously. □

Remark 6.2. For any fixed $d \geq 2$, the hypotheses of stationarity and independence at distance $l = L$ of the point process ensures that

$$\mathrm{E}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\sum_{j=2}^{d} M_{(iL,L),(iL,L)}M_{(iL,jL),(iL,L)} - \sum_{j=2}^{d}\mathrm{E}[M_{L\mathbf{u},L\mathbf{u}}M_{(L,jL),L\mathbf{u}}]\right|\right] \underset{m\to\infty}{\longrightarrow} 0,$$

but they are not sufficient to ensure the uniform convergence with respect to d.

Now we have all the tools to prove Limit (6.16) for a sequence of increasing rectangles without fixed vertical side.

Proposition 6.6. *Let $\mathscr{P}$ be a stationary point process such that $H(\mathscr{P}\cap .)$ has finite second moment measure. Let us suppose that there exist $0 < l < \infty$ such that $\mathscr{P}\cap A$ and $\mathscr{P}\cap B$ are independent for each $A, B \in \mathfrak{B}^2$ with $d(A,B) > l$ and, set $L = l$, Limit (6.38) in Lemma 6.2 is satisfied. Let $\{B_n\}_{n\geq 1}$ be a sequence of rectangles which satisfies Conditions 6.4. Then,*

$$\frac{1}{(o_n-2)(v_n-1)}\sum_{i=1}^{o_n} D_i^2 \underset{n\to\infty}{\longrightarrow} \mathrm{E}[M_{L\mathbf{u},L\mathbf{u}}^2] + 2\mathrm{E}\left[M_{L\mathbf{u},L\mathbf{u}}\sum_{j=2}^{4} M_{(L,jL),L\mathbf{u}}\right],$$

in L^1 norm.

Proof. Since the point process is independent at distance L, $D_i = \sum_{j=1}^{v_n} M_{i,j}$ if $i < o_n$ (Equation (6.10)) and $D_{o_n} = \sum_{j=1}^{v_n} \Delta_{\mathbf{x}_{o_n,j},\mathbf{x}_{o_n,1}}$. Therefore,

$$\begin{aligned}
\frac{1}{(o_n-2)(v_n-1)}\sum_{i=1}^{o_n} D_i^2 &= \frac{1}{(o_n-2)(v_n-1)}\sum_{i=2}^{o_n-1}\sum_{j=1}^{v_n} M_{i,j}^2 \quad (=: I_{1n}) \\
&+ \frac{2}{(o_n-2)(v_n-1)}\sum_{i=2}^{o_n-1}\sum_{j=1}^{v_n-1}\sum_{j'=j+1}^{v_n} M_{i,j}M_{i,j'} (=: I_{2n}) \\
&+ \frac{1}{(o_n-2)(v_n-1)}\left(\sum_{j=1}^{v_n}\sum_{h=1}^{2}\Delta_{\mathbf{x}_{h,j},\mathbf{x}_{1,1}}\right)^2 \quad (=: I_{3n}) \\
&+ \frac{1}{(o_n-2)(v_n-1)}\left(\sum_{j=1}^{v_n}\Delta_{\mathbf{x}_{o_n,j},\mathbf{x}_{o_n,1}}\right)^2 \quad (=: I_{4n})
\end{aligned}$$

and we will study the convergence of I_{1n}, I_{2n}, I_{3n} and I_{4n}.

First, we show that $\lim_{n\to\infty} I_{1n} = \mathrm{E}[M^2_{L\mathbf{u},L\mathbf{u}}]$ in norm L^1. Since $\{M_{i,j}\}_{i>1,j}$ are stationary and the point process has finite second moment measure, $\mathrm{E}[M_{i,j}^2] = \mathrm{E}[M^2_{L\mathbf{u},L\mathbf{u}}] < \infty$, for each $i = 2,\ldots,o_n-1$ and $j = 1,\ldots,v_n$. Moreover, due to the independence at distance L, for any fixed j, $\{M_{i,j}\}_{i>1}$ is an ergodic sequence. We define,

$$F_{iL\mathbf{e}_1} = M_{iL\mathbf{e}_1,iL\mathbf{e}_1},$$

for each $i \in \mathbb{N}$. Since $\{F_{iL\mathbf{e}_1}\}_{n\geq 1}$ is a stationary and ergodic sequence (due to the properties $\{M_{i,j}\}_{i>1}$) and $(.)^2$ is a continuous function, we can use the Birkhoff ergodic theorem (Theorem 2.3 in [62]). Then,

$$\frac{1}{m}\sum_{i=1}^{m} F^2_{iL\mathbf{e}_1} \underset{m\to\infty}{\longrightarrow} \mathrm{E}[F^2_{L\mathbf{e}_1}],$$

in norm L^1, where $\mathrm{E}[M^2_{L\mathbf{u},L\mathbf{u}}] = \mathrm{E}[F^2_{L\mathbf{e}_1}]$. Using this result and the stationarity of the point process, we obtain

$$\mathrm{E}\left[\left|I_{1n} - \mathrm{E}[M^2_{L\mathbf{u},L\mathbf{u}}]\right|\right]$$

$$\begin{aligned}
&\leq \frac{1}{(v_n-1)}\sum_{j=1}^{v_n}\mathrm{E}\left[\left|\frac{1}{(o_n-2)}\sum_{i=2}^{o_n-1}M_{i,j}^2-\mathrm{E}[M_{L\mathbf{u},L\mathbf{u}}^2]\right|\right]+\frac{\mathrm{E}[M_{L\mathbf{u},L\mathbf{u}}^2]}{(v_n-1)}\\
&= \frac{v_n}{(v_n-1)}\mathrm{E}\left[\left|\frac{1}{(o_n-2)}\sum_{i=2}^{o_n-1}M_{i,1}^2-\mathrm{E}[M_{L\mathbf{u},L\mathbf{u}}^2]\right|\right]+\frac{\mathrm{E}[M_{L\mathbf{u},L\mathbf{u}}^2]}{(v_n-1)}\\
&\to 0, \qquad n\to\infty.
\end{aligned}$$

In order to show that $\lim_{n\to\infty} I_{2n} = 2\mathrm{E}[M_{L\mathbf{u},L\mathbf{u}}\sum_{j=2}^{4}M_{(L,jL),L\mathbf{u}}]$ in norm L^1, it is sufficient to prove that the hypotheses of Lemma 6.2 hold. Limit (6.38) is guaranteed by our assumptions, while Lemma 6.4 ensures that

$$\mathrm{E}\left[M_{L\mathbf{u},L\mathbf{u}}M_{(L,jL),L\mathbf{u}}\right]=0, \qquad \forall j>4.$$

Therefore, the limit holds.

Regarding I_{3n} and I_{4n}, we can rewrite them in the following way:

$$\begin{aligned}
I_{3n} &= \frac{1}{(o_n-2)(v_n-1)}\sum_{j=1}^{v_n}\left(\sum_{h=1}^{2}\Delta_{\mathbf{x}_{h,j},\mathbf{x}_{1,1}}\right)^2 \qquad (=: J_{1n})\\
&\quad+\frac{2}{(o_n-2)(v_n-1)}\sum_{j=1}^{v_n-1}\sum_{j'=j+1}^{v_n}\left(\sum_{h=1}^{2}\Delta_{\mathbf{x}_{h,j},\mathbf{x}_{1,1}}\right)\left(\sum_{h=1}^{2}\Delta_{\mathbf{x}_{h,j'},\mathbf{x}_{1,1}}\right) (=: J_{2n})\\
I_{4n} &= \frac{1}{(o_n-2)(v_n-1)}\sum_{j=1}^{v_n}\Delta^2_{\mathbf{x}_{o_n,j},\mathbf{x}_{o_n,1}} \qquad (=: K_{1n})\\
&\quad+\frac{2}{(o_n-2)(v_n-1)}\sum_{j=1}^{v_n-1}\sum_{j'=j+1}^{v_n}\Delta_{\mathbf{x}_{o_n,j},\mathbf{x}_{o_n,1}}\Delta_{\mathbf{x}_{o_n,j'},\mathbf{x}_{o_n,1}} \qquad (=: K_{2n})
\end{aligned}$$

As before, using the Birkhoff ergodic theorem,

$$\begin{aligned}
&\frac{1}{v_n}\sum_{j=1}^{v_n}\left(\sum_{h=1}^{2}\Delta_{\mathbf{x}_{h,j},\mathbf{x}_{1,1}}\right)^2\xrightarrow[n\to\infty]{L^1}\mathrm{E}\left[\left(\Delta^0_{L\mathbf{u}}+\Delta^0_{(2L,L),L\mathbf{u}}\right)^2\right]\\
&\frac{1}{v_n}\sum_{j=1}^{v_n}\Delta^2_{\mathbf{x}_{o_n,j},\mathbf{x}_{o_n,1}}\xrightarrow[n\to\infty]{L^1}\mathrm{E}\left[\Delta^2_{L\mathbf{u},L\mathbf{u}}\right]
\end{aligned}$$

and thus

$$J_{1n} = \frac{1}{o_n - 1} \frac{v_n}{v_n - 1} \frac{1}{v_n} \sum_{j=1}^{v_n} \left(\sum_{h=1}^{2} \Delta_{\mathbf{x}_{h,j}, \mathbf{x}_{1,1}} \right)^2 \xrightarrow[n\to\infty]{L^1} 0$$

$$K_{1n} = \frac{1}{o_n - 1} \frac{v_n}{v_n - 1} \frac{1}{v_n} \sum_{j=1}^{v_n} \Delta^2_{\mathbf{x}_{o_n,j}, \mathbf{x}_{o_n,1}} \xrightarrow[n\to\infty]{L^1} 0.$$

Finally, thanks to Lemma 6.4, we can apply Lemma 6.3 and then $\lim_{n\to\infty} J_{2n} = 0$ and $\lim_{n\to\infty} K_{2n} = 0$ in norm L^1. □

6.4 CLT for a positive functional of a stationary point process independent at distance l

We are now ready to formulate and prove our main result, by applying Propositions 6.2 and 6.6.

Theorem 6.1. *Let $\mathscr{P}$ be a stationary point process such that $\mathscr{H}(\mathscr{P} \cap .)$ has finite fourth moment measure. Let us suppose that there exist $0 < l < \infty$ such that $\mathscr{P} \cap A$ and $\mathscr{P} \cap B$ are independent for each A, $B \in \mathfrak{B}^2$ with $d(A,B) > l$ and, set $L = l$, Limit (6.38) in Lemma 6.2 is satisfied. Let $\{B_n\}_{n\geq 1}$ be a sequence of rectangles which satisfies Conditions 6.4. Then,*

$$n^{-\alpha\gamma}(H(\mathscr{P}_n) - \mathrm{E}[H(\mathscr{P}_n)]) \xrightarrow[n\to\infty]{\mathscr{D}} \mathscr{N}(0, \sigma^2),$$

where $\mathscr{P}_n = \mathscr{P} \cap B_n$, $\sigma^2 = (c_1^{\beta+1} c_2)\left(\mathrm{E}[M^2_{L\mathbf{u},L\mathbf{u}}] + 2\mathrm{E}\left[M_{L\mathbf{u},L\mathbf{u}} \Sigma_{j=2}^4 M_{(L,jL),L\mathbf{u}}\right]\right)$ and $\gamma = 1/2(\beta+1)$.

Proof. From the definition of the sequence of rectangles $\{B_n\}_{n\geq 1}$, for each n, we can define a lattice of points $\{\mathbf{x}_{i,j}\}_{i,j} \in \mathbb{R}^2$ such that

$$B_n = \bigcup_{i=1}^{o_n} \bigcup_{j=1}^{v_n} Q_{L/2}(\mathbf{x}_{i,j}).$$

Since our point process is supposed to be stationary, we can think to translate the sets $\{B_n\}_{n\geq 1}$ so that $\mathbf{x}_{i,j} \in \mathbb{Z}_L^2$ for all $i = 1, \ldots, o_n$ and $j = 1, \ldots, v_n$. Moreover, for each $\mathbf{x} \in \mathbb{Z}_L^2 = \{\mathbf{y}L : \mathbf{y} \in \mathbb{Z}^2\}$, we define the σ-algebra $\mathfrak{F}_{\mathbf{x}}=\sigma(\{\mathscr{P} \cap Q_{L/2}(\mathbf{y}) \,|\, \mathbf{y} \in \mathbb{Z}_L^2, y_1 \leq x_1\})$.

Similarly to the proof of Theorem 3.1 in [61], for any fixed n, we define a filtration $\{\mathfrak{G}_0, \ldots, \mathfrak{G}_{o_n}\}$, where $\mathfrak{G}_0$ is the trivial σ-algebra and $\mathfrak{G}_i = \mathfrak{F}_{\mathbf{x}_{i,1}}$, for $i = 1, \ldots, o_n$. As a consequence,

$$H(\mathscr{P}_n) - \mathrm{E}[H(\mathscr{P}_n)] = \sum_{i=1}^{o_n} [E[H(\mathscr{P}_n) \,|\, \mathfrak{G}_i] - E[H(\mathscr{P}_n) \,|\, \mathfrak{G}_{i-1}]] =: \sum_{i=1}^{o_n} D_i.$$

From its definition, fixed n, $\{D_i\}_{i=1}^{o_n}$ is a martingale difference and thus if we multiply it by a constant we still have a martingale difference. Then, let us consider $\{D_i/\sqrt{(o_n-2)(v_n-1)}\}_{i=1}^{o_n}$. For the central limit theorem for martingale differences (Theorem 2.3 in [49]), if

$$\sup_{n\geq 1} \mathrm{E}\left[\max_{1\leq i\leq o_n}\left(\frac{D_i}{\sqrt{(o_n-2)(v_n-1)}}\right)^2\right] < \infty \tag{6.46}$$

$$\frac{1}{\sqrt{(o_n-2)(v_n-1)}} \max_{1\leq i\leq o_n} |D_i| \xrightarrow[n\to\infty]{P} 0, \tag{6.47}$$

and there exists a constant $\tau^2 \geq 0$ such that

$$\frac{1}{(o_n-2)(v_n-1)} \sum_{i=1}^{o_n} D_i^2 \xrightarrow[n\to\infty]{L^1} \tau^2, \tag{6.48}$$

then,

$$\frac{H(\mathscr{P}_n) - \mathrm{E}[H(\mathscr{P}_n)]}{\sqrt{(o_n-2)(v_n-1)}} = \frac{1}{\sqrt{(o_n-2)(v_n-1)}} \sum_{i=1}^{o_n} D_i \xrightarrow[n\to\infty]{\mathscr{D}} \mathscr{N}(0,\tau^2). \tag{6.49}$$

By using Jensen's inequality, a sufficient condition for (6.46) is

$$\mathrm{E}\left[\max_{1\le i\le o_n}\left(\frac{D_i}{\sqrt{(o_n-2)(v_n-1)}}\right)^2\right]\le\frac{1}{(o_n-2)(v_n-1)}\sum_{i=1}^{o_n}\mathrm{E}[D_i^2]<C_1<\infty,$$

for all $n\ge 1$. We can show the sufficient condition, by applying Proposition 6.2. Since $\delta_1=\delta_2=0$, $\delta_3=1$, $\delta_4=2$ and $\delta_5=2$ for a point process independent at distance L (see Proposition 6.3), the two constraints on the parameters required by the Proposition 6.2 hold:

$$\begin{aligned}(\beta+1)/2=\gamma&\ge\max\{(\delta_3\beta+1)/2,(1+\delta_1(1+\beta))/2,(\delta_5\beta+2)/4\}\\&=(\beta+1)/2\\(\beta+1)/2=\gamma&>\max\{(\delta_4\beta+1)/4,(1+\delta_2(1+\beta))/4\}=\beta/2+1/4.\end{aligned}$$

As a consequence, there exists a constant C such that

$$o_n^{-2\gamma}\sum_{i=1}^{o_n}\mathrm{E}[D_i^2]<C<\infty,\qquad\forall n\ge 1,$$

and then, by using also the monotonicity of the sequence $\{v_n/o_n^{\beta}\}$ and Property (6.4), the Inequality (6.46) is verified

$$\begin{aligned}\frac{1}{(o_n-2)(v_n-1)}\sum_{i=1}^{o_n}\mathrm{E}[D_i^2]&=\frac{o_n^{1+\beta}}{(o_n-2)(v_n-1)}o_n^{-2\gamma}\sum_{i=1}^{o_n}\mathrm{E}[D_i^2]\\&=\left(\frac{o_n-2}{o_n-2}+\frac{2}{o_n-2}\right)\frac{o_n^{\beta}}{v_n}\frac{v_n}{v_n-1}o_n^{-2\gamma}\sum_{i=1}^{o_n}\mathrm{E}[D_i^2]\\&\le 3C\max\left(\frac{o_1^{\beta}}{v_1},\frac{1}{c_2}\right)<\infty.\end{aligned}$$

Using Boole's and Markov's inequalities, we can see that a sufficient condition for (6.47) is that for any $\varepsilon>0$

$$\mathrm{P}\left(\max_{1\le i\le o_n}|D_i|\ge\sqrt{(o_n-2)(v_n-1)}\varepsilon\right)\le\frac{1}{\varepsilon^4(o_n-2)^2(v_n-1)^2}\sum_{i=1}^{o_n}\mathrm{E}[D_i^4]\underset{n\to\infty}{\longrightarrow}0.$$

We have already seen that, for $\gamma = (1/2)(\beta+1)$, Proposition 6.2 holds and thus,

$$o_n^{-4\gamma} \sum_{i=1}^{o_n} \mathrm{E}[D_i^4] \underset{n\to\infty}{\longrightarrow} 0.$$

By using the previous limit and Properties (6.3) and (6.4) of $\{v_n\}_{n\geq 1}$ and $\{o_n\}_{n\geq 1}$, we can verify the sufficient condition for Limit (6.47),

$$\begin{aligned}\frac{1}{\varepsilon^4 (o_n-2)^2 (v_n-1)^2} \sum_{i=1}^{o_n} \mathrm{E}[D_i^4] &= \frac{1}{\varepsilon^4} \frac{o_n^2}{(o_n-2)^2} \frac{o_n^{2\beta}}{(v_n-1)^2} o_n^{-4\gamma} \sum_{i=1}^{o_n} \mathrm{E}[D_i^4] \\ &\to \frac{1}{\varepsilon^4} \cdot 1 \cdot \frac{1}{c_2^2} \cdot 0 = 0, \qquad n\to\infty.\end{aligned}$$

Finally, Limit (6.48) follows from Proposition 6.6 with $\tau^2 = \mathrm{E}[M^2_{L\mathbf{u},L\mathbf{u}}] + 2\mathrm{E}[M_{L\mathbf{u},L\mathbf{u}} \sum_{j=2}^4 M_{(L,jL),L\mathbf{u}}]$ and thus Limit (6.49) holds.

It remains to show that the thesis derives from Limit (6.49). Thanks to Properties (6.3) and (6.4),

$$\begin{aligned}\frac{\sqrt{(o_n-2)(v_n-1)}}{n^{\alpha\gamma}} &= \frac{o_n^{\gamma}}{n^{\alpha\gamma}} \frac{\sqrt{o_n-2}}{o_n^{1/2}} \frac{\sqrt{v_n-1}}{o_n^{\beta/2}} \\ &\to c_1^{\gamma} \cdot 1 \cdot c_2^{1/2} = \sqrt{c_1^{\beta+1} c_2}, \qquad n\to\infty,\end{aligned}$$

and thus we obtain the thesis for $\sigma^2 = \tau^2 (c_1^{\beta+1} c_2)$. □

With a proof similar to the one of the previous theorem, but using Proposition 6.5, we can show the following result for a sequence of increasing rectangles with vertical side of fixed length.

Theorem 6.2. *Let $\mathscr{P}$ be a stationary point process such that $\mathscr{H}(\mathscr{P}\cap .)$ has finite fourth moment measure. Let us suppose that there exist $0 < l < \infty$ such that $\mathscr{P}\cap A$ and $\mathscr{P}\cap B$ are independent for each A, $B \in \mathfrak{B}^2$ with $d(A,B) > l$. Let $\{B_n\}_{n\geq 1}$ be a sequence of rectangles such that, for each n, B_n has vertical side V ($V \in \mathbb{R}_+$) and horizontal side $L \cdot o_n$, with $L = l$ ($o_n \in \mathbb{N}$ and $\{o_n\}_{n\geq 1}$ satisfies Limit (6.3)). Then,*

$$n^{-\frac{\alpha}{2}}(H(\mathscr{P}_n)-\mathrm{E}[H(\mathscr{P}_n)])\xrightarrow[n\to\infty]{\mathscr{D}}\mathscr{N}(0,\sigma^2),$$

where $\mathscr{P}_n=\mathscr{P}\cap B_n$, $\sigma^2=c_1[\mathrm{E}[(H(\mathscr{P}\cap\mathscr{R}_{L/2,V/2}(\mathbf{0}))-\mathrm{E}[H(\mathscr{P}\cap\mathscr{R}_{L/2,V/2}(\mathbf{0}))]$ $+\mathrm{E}[H(\mathscr{P}\cap\mathscr{R}_{L/2,V/2}(L\mathbf{e}_1))\,|\,\mathfrak{F}_{\mathbf{0}}]-\mathrm{E}[H(\mathscr{P}\cap\mathscr{R}_{L/2,V/2}(\mathbf{0}))\,|\,\mathfrak{F}_{-L\mathbf{e}_1}]]$, $\mathscr{R}_{l_1/2,l_2/2}(\mathbf{x})$ *is a rectangle with horizontal side* l_1*, vertical side* l_2 *and centered at* $\mathbf{x}$.

Proof. For simplicity, we prove the theorem in case $V/L\in\mathbb{N}$. The proof can be easily generalized to the case $V/L\notin\mathbb{N}$.

Analogously to the proof of Theorem 6.1, for each $n\geq 1$, we define a lattice of points $\{\mathbf{x}_{i,j}\}_{i,j}\in\mathbb{R}^2$ such that

$$B_n=\bigcup_{i=1}^{o_n}\bigcup_{j=1}^{v_n}Q_{L/2}(\mathbf{x}_{i,j}).$$

Moreover, due to the stationarity of the point process, we can assume that, without loss of generality, $\mathbf{x}_{i,j}\in\mathbb{Z}^2_L$, for all $i=1,\ldots,o_n$ and $j=1,\ldots,V/L$. For each $\mathbf{x}\in\mathbb{Z}^2_L$, we define the σ-algebra $\mathfrak{F}_{\mathbf{x}}=\sigma(\{\mathscr{P}\cap Q_{L/2}(\mathbf{y})\,|\,\mathbf{y}\in\mathbb{Z}^2_L,y_1\leq x_1\})$.

Similarly to the proof of Theorem 6.1, fixed n, we define a filtration $\{\mathfrak{G}_0,\ldots,\mathfrak{G}_{o_n}\}$, where $\mathfrak{G}_0$ is the trivial σ-algebra and $\mathfrak{G}_i=\mathfrak{F}_{\mathbf{x}_{i,1}}$, for $i=1,\ldots,o_n$. Therefore,

$$H(\mathscr{P}_n)-\mathrm{E}[H(\mathscr{P}_n)]=\sum_{i=1}^{o_n}E[H(\mathscr{P}_n)\,|\,\mathfrak{G}_i]-E[H(\mathscr{P}_n)\,|\,\mathfrak{G}_{i-1}]=:\sum_{i=1}^{o_n}D_i.$$

and, fixed n, $\{D_i\}_{i=1}^{o_n}$ is a martingale difference. Since if we multiply a martingale difference by a constant we still have a martingale difference, we consider $\{D_i/\sqrt{(o_n-2)}\}_{i=1}^{o_n}$. For the central limit theorem for martingale differences (Theorem 2.3 in [49]), we have that,

$$\frac{H(\mathscr{P}_n)-\mathrm{E}[H(\mathscr{P}_n)]}{\sqrt{(o_n-2)}}=\frac{1}{\sqrt{(o_n-2)}}\sum_{i=1}^{o_n}D_i\xrightarrow[n\to\infty]{\mathscr{D}}\mathscr{N}(0,\tau^2),\tag{6.50}$$

if

$$\sup_{n\geq 1} \mathrm{E}\left[\max_{1\leq i\leq o_n}\left(\frac{D_i}{\sqrt{(o_n-2)}}\right)^2\right] < \infty \tag{6.51}$$

$$\frac{1}{\sqrt{(o_n-2)}}\max_{1\leq i\leq o_n}|D_i| \xrightarrow[n\to\infty]{P} 0, \tag{6.52}$$

and there exists a constant $\tau^2 \geq 0$ such that

$$\frac{1}{(o_n-2)}\sum_{i=1}^{o_n} D_i^2 \xrightarrow[n\to\infty]{L^1} \tau^2. \tag{6.53}$$

Analogously to the proof of Theorem 6.1, two sufficient conditions of (6.51) and (6.52) can be derived by applying Jensen's inequality to (6.51) and Boole's and Markov's inequalities to (6.52),

$$\frac{1}{(o_n-2)}\sum_{i=1}^{o_n}\mathrm{E}[D_i^2] < C_1 < \infty, \qquad \text{for all } n \geq 1$$

$$\frac{1}{\varepsilon^4(o_n-2)^2}\sum_{i=1}^{o_n}\mathrm{E}[D_i^4] \longrightarrow_{n\to\infty} 0.$$

We can prove both conditions by using Jensen's inequality and Inequalities (6.12) and (6.13),

$$\begin{aligned}
\frac{1}{(o_n-2)}\sum_{i=1}^{o_n}\mathrm{E}[D_i^2] &= \frac{1}{(o_n-2)}\sum_{i=1}^{o_n}\mathrm{E}\left[\left(\sum_{j=1}^{V/L} M_{i,j}\right)^2\right] \\
&\leq \frac{V}{L(o_n-2)}\sum_{i=1}^{o_n}\sum_{j=1}^{V/L}\mathrm{E}\left[M_{i,j}^2\right] \\
&\leq \frac{o_n}{o_n-2}\left(\frac{V}{L}\right)^2 8E_{\mathbf{0},2} \\
&\leq 3\left(\frac{V}{L}\right)^2 8E_{\mathbf{0},2} < \infty, \qquad \text{for all } n \geq 1
\end{aligned}$$

$$\frac{1}{\varepsilon^4(o_n-2)^2}\sum_{i=1}^{o_n}\mathrm{E}[D_i^4] \le \frac{1}{\varepsilon^4(o_n-2)^2}\left(\frac{V}{L}\right)^3\sum_{j=1}^{V/L}\mathrm{E}\left[M_{i,j}^4\right]$$
$$\le \frac{o_n}{\varepsilon^4(o_n-2)^2}\left(\frac{V}{L}\right)^4 64E_{\mathbf{0},4}\xrightarrow[n\to\infty]{}0.$$

Finally, Limit (6.50) holds, since Proposition 6.5 guarantees Limit (6.53) with

$$\tau^2 = \mathrm{E}[(H(\mathscr{P}\cap\mathscr{R}_{L/2,V/2}(\mathbf{0}))-\mathrm{E}[H(\mathscr{P}\cap\mathscr{R}_{L/2,V/2}(\mathbf{0}))\,|\,\mathfrak{F}_{-L\mathbf{e}_1}]$$
$$+\,\mathrm{E}[H(\mathscr{P}\cap\mathscr{R}_{L/2,V/2}(L\mathbf{e}_1))\,|\,\mathfrak{F}_{\mathbf{0}}]-\mathrm{E}[H(\mathscr{P}\cap\mathscr{R}_{L/2,V/2}(\mathbf{0}))].$$

To obtain the thesis, we use Limit (6.50) and Property (6.3),

$$n^{-\frac{\alpha}{2}}(H(\mathscr{P}_n)-\mathrm{E}[H(\mathscr{P}_n)]) = \frac{\sqrt{(o_n-2)}}{n^{-\frac{\alpha}{2}}}\frac{1}{\sqrt{(o_n-2)}}\sum_{i=1}^{o_n}D_i\xrightarrow[n\to\infty]{\mathscr{D}}\mathscr{N}(0,c_1\tau^2).$$

□

6.5 Asymptotic normality of the estimators of the intensity

We now apply Theorem 6.1 to retrieve the asymptotic normality of the estimators described in Section 5.3, under some regularity conditions on the fibre processes involved.

Corollary 6.1. *Let Φ_1 and Φ_2 be two stationary and independent fibre processes with intensities $\mathrm{L}_{A,1}$ and $\mathrm{L}_{A,2}$, respectively. Moreover, let Φ_2 be also isotropic. Let us suppose that $\Phi_1\cap\Phi_2$ has finite fourth moment measure, there exists $0<l<\infty$ such that the point process $\Phi_1\cap\Phi_2$ is independent at distance l and Limit (6.38) in Lemma 6.2 is satisfied with $L=l$. Then, for any sequence $\{B_n\}_{n\ge1}$ of rectangular sets in $\mathbb{R}^2$, which satisfies Conditions 6.4,*

$$\sqrt{n}(\widehat{\mathrm{L}}_{A,1}(B_n) - L_{A,1}) \xrightarrow[n\to\infty]{\mathscr{D}} \mathscr{N}\left(0, \frac{\pi^2\tau^2}{4\mathrm{L}_{A,2}^2 l^4 c_1^{\beta+1} c_2}\right), \tag{6.54}$$

where $\tau^2 = \mathrm{E}[M^2_{L\mathbf{u},L\mathbf{u}}] + 2\mathrm{E}\left[M_{L\mathbf{u},L\mathbf{u}} \sum_{j=2}^{4} M_{L(1,j),L\mathbf{u}}\right]$.

Proof. Estimator $\widehat{\mathrm{L}}_{A,1}(B_n)$ is given by

$$\frac{\pi}{2\mathrm{L}_{A,2}} \frac{H(\mathscr{P}_n)}{\nu_2(B_n)},$$

where $H(\mathscr{P}_n)$= card$(\Phi_1 \cap \Phi_2 \cap B_n)$=$N_{\Phi_1\cap\Phi_2}(B_n)$, with $\mathscr{P}_n = \Phi_1 \cap \Phi_2 \cap B_n$, and thus H trivially satisfies Conditions 6.3.

Setting the parameter α of Equation (6.3) equal to $1/(\beta+1)$, we obtain that $\alpha\gamma = \alpha(\beta+1)/2 = 1/2$. Then,

$$\begin{aligned}
\frac{\sqrt{n}}{\nu_2(B_n)} &= \frac{\sqrt{n}}{l^2 o_n \nu_n} \\
&= \frac{n^{\alpha\gamma}}{l^2 o_n \nu_n} \\
&= \frac{1}{l^2 n^{\alpha\gamma}} \frac{n^{2\alpha\gamma}}{o_n \nu_n} \\
&= \frac{1}{l^2 n^{\alpha\gamma}} \frac{n^{\alpha}}{o_n} \frac{n^{\alpha\beta}}{\nu_n} \\
&= \frac{1}{l^2 n^{\alpha\gamma}} \frac{n^{\alpha}}{o_n} \frac{o_n^{\beta}}{\nu_n} \left(\frac{n^{\alpha}}{o_n}\right)^{\beta}
\end{aligned}$$

and we can write the left-hand side of (6.54) in the following way,

$$\begin{aligned}
\sqrt{n}(\widehat{\mathrm{L}}_{A,1}(B_n) - L_{A,1}) &= \frac{\sqrt{n}}{\nu_2(B_n)} \frac{\pi}{2\mathrm{L}_{A,2}} \left(H(\mathscr{P}_n) - \mathrm{E}[H(\mathscr{P}_n)]\right) \\
&= \frac{n^{\alpha}}{o_n} \frac{o_n^{\beta}}{\nu_n} \left(\frac{n^{\alpha}}{o_n}\right)^{\beta} \frac{\pi}{l^2 2\mathrm{L}_{A,2}} \frac{1}{n^{\alpha\gamma}} \left(H(\mathscr{P}_n) - \mathrm{E}[H(\mathscr{P}_n)]\right).
\end{aligned}$$

By using Theorem 6.1 and Limits (6.3) and (6.4), we obtain the thesis. □

Corollary 6.2. *Let Φ_1 and Φ_2 be two stationary and independent fibre processes with intensities $\mathrm{L}_{A,1}$ and $\mathrm{L}_{A,2}$, respectively. Let us define, for any $A \in \mathfrak{B}^2$,*

$$H(\Phi_1 \cap \Phi_2 \cap A) = \begin{cases} \sum_{y \in \Phi_1 \cap \Phi_2} \frac{\mathbb{I}_{A'}(y)}{\left|\sin(w(\mathrm{T}_y\Phi_2) - w(\mathrm{T}_y\Phi_1))\right|} & \text{if } A \cap \Phi_1 \cap \Phi_2 \neq \emptyset \\ 0 & \text{otherwise,} \end{cases} \tag{6.55}$$

and let $\Phi_1 \cap \Phi_2$ be such that $H(\Phi_1 \cap \Phi_2 \cap .)$ has finite fourth moment measure. Moreover, let us suppose that there exists $0 < l < \infty$ such that the point process $\Phi_1 \cap \Phi_2$ is independent at distance l and Limit (6.38) in Lemma 6.2 is satisfied with $L = l$. Then, for any sequence $\{B_n\}_{n \geq 1}$ of rectangular sets in $\mathbb{R}^2$, which satisfies Conditions 6.4,

$$\sqrt{n}(\widehat{\widehat{\mathrm{L}}}_{A,1}(B_n) - L_{A,1}) \xrightarrow[n \to \infty]{\mathscr{D}} \mathscr{N}\left(0, \frac{\tau^2}{\mathrm{L}_{A,2}^2 l^4 c_1^{\beta+1} c_2}\right), \tag{6.56}$$

where $\tau^2 = \mathrm{E}[M^2_{L\mathbf{u},L\mathbf{u}}] + 2\mathrm{E}\left[M_{L\mathbf{u},L\mathbf{u}} \sum_{j=2}^4 M_{L(1,j),L\mathbf{u}}\right]$.

Proof. Estimator $\widehat{\widehat{\mathrm{L}}}_{A,1}(B_n)$ can be rewritten as $H(\mathscr{P}_n)/(\mathrm{L}_{A,2}\nu_2(B_n))$, where $H(\Phi_1 \cap \Phi_2 \cap .)$ is defined in Equation (6.55) and $\mathscr{P}_n = \Phi_1 \cap \Phi_2 \cap B_n$. It can be easily proven that H satisfies Conditions 6.3.

Now, following the same steps of the proof of Corollary 6.1, we obtain that the left-hand side of (6.56) can be rewritten as

$$\begin{aligned} \sqrt{n}(\widehat{\widehat{\mathrm{L}}}_{A,1}(B_n) - L_{A,1}) &= \frac{\sqrt{n}}{\nu_2(B_n)\mathrm{L}_{A,2}} \left(H(\mathscr{P}_n) - \mathrm{E}[H(\mathscr{P}_n)]\right) \\ &= \frac{n^\alpha}{o_n} \frac{o_n^\beta}{\nu_n} \left(\frac{n^\alpha}{o_n}\right)^\beta \frac{1}{l^2 \mathrm{L}_{A,2}} \frac{1}{n^{\alpha\gamma}} \left(H(\mathscr{P}_n) - \mathrm{E}[H(\mathscr{P}_n)]\right). \end{aligned}$$

By using Theorem 6.1 and Limits (6.3) and (6.4), we obtain the thesis. □

With suitable conditions on the system of circles used, we can also define a sequence of estimators $\{\widehat{\mathrm{L}}_{A,n}^{circles}\}_{n\in\mathbb{N}}$ which is asymptotically normal (see Equation 5.6 for the definition of $\widehat{\mathrm{L}}_{A}^{circles}$).

Conditions 6.6.

Let $\{B_n\}_{n\geq 1}$ be a sequence of rectangular sets which satisfies Conditions 6.4. Then, we define $\{\mathscr{C}_n\}_{n\geq 1}$ as an increasing sequence of systems of circles which satisfies the following properties,

1. each circle has ray R, with $R \leq L/2$,
2. each circle must be all contained in one square $Q_{L/2}(\mathbf{x}_{i,j})$ of the grid that cover B_n, for each $n \geq 1$,
3. the number of circles contained in $Q_{L/2}(\mathbf{x}_{i,j})$ is N_L, for each $Q_{L/2}(\mathbf{x}_{i,j})$ of the grid of squares that cover B_n and for each $n \geq 1$.

In practice, we define a system $\mathscr{C}$ of N_L circles with ray R which are all contained in the square $Q_{L/2}(\mathbf{0})$. Then, we shift $\mathscr{C}$ by $\mathbf{x}_{i,j}$ (obtaining thus $\mathscr{C}_{i,j}$) in each square $Q_{L/2}(\mathbf{x}_{i,j})$ of the grid that covers B_n, for each $n \geq 1$. Thus, for any $n \geq 1$, $\mathscr{C}_n$ is the union of the systems $\{\mathscr{C}_{i,j}\}_{i,j}$.

Corollary 6.3. *Let Φ_1 be a stationary fibre processes with intensity $\mathrm{L}_{A,1}$. Let $\{B_n\}_{n\geq 1}$ be a sequence of rectangular sets which satisfies Conditions 6.4 and let $\{\mathscr{C}_n\}_{n\geq 1}$ be an increasing sequence of systems of circles which satisfies Conditions 6.6. Moreover, let us suppose that $\Phi_1 \cap \mathscr{C}_\infty$ has finite fourth moment measure. If there exists $0 < l < \infty$ such that the point process $\Phi_1 \cap \mathscr{C}_\infty$ is independent at distance l and, assuming $L = l$, Limit (6.38) in Lemma 6.2 is satisfied, then*

$$\sqrt{n}(\widehat{\mathrm{L}}_{A,n}^{circles} - L_{A,1}) \xrightarrow[n\to\infty]{\mathscr{D}} \mathscr{N}\left(0, \frac{\tau^2}{16R^2N_l^2c_1^{\beta+1}c_2}\right), \tag{6.57}$$

where $\widehat{\mathrm{L}}_{A,n}^{circles}$ is the estimator $\widehat{\mathrm{L}}_{A}^{circles}$ (Equation 5.6) corresponding to the system of circles $\mathscr{C}_n$ and $\tau^2 = \mathrm{E}[M_{L\mathbf{u},L\mathbf{u}}^2] + 2\mathrm{E}\left[M_{L\mathbf{u},L\mathbf{u}}\sum_{j=2}^{4} M_{L(1,j),L\mathbf{u}}\right]$.

Proof. First, let us observe that for each $n \geq 1$ the number of circles in the system $\mathscr{C}_n$ is $N_l o_n v_n$ (due to Conditions 6.6) and that $\mathscr{C}_\infty \cap B_n = \mathscr{C}_n$. Therefore, the estimator $\widehat{\mathrm{L}}_{A,n}^{circles}$ can be seen as

$$\frac{1}{4RN_l} \frac{H(\mathscr{P}_n)}{o_n v_n},$$

where $H(\mathscr{P}_n)$= card$(\Phi_1 \cap \mathscr{C}_\infty \cap B_n)$, with $\mathscr{P}_n = \Phi_1 \cap \mathscr{C}_\infty \cap B_n$, and thus H trivially satisfies Conditions 6.3.

Now, following the same steps of the proof of Corollary 6.1, we obtain that the left-hand side of (6.57) can be rewritten as

$$\begin{aligned}\sqrt{n}(\widehat{\mathrm{L}}_{A,n}^{circles} - L_{A,1}) &= \frac{\sqrt{n}}{4RN_l o_n v_n} \left(H(\mathscr{P}_n) - \mathrm{E}[H(\mathscr{P}_n)]\right) \\ &= \frac{n^\alpha}{o_n} \frac{o_n^\beta}{v_n} \left(\frac{n^\alpha}{o_n}\right)^\beta \frac{1}{4RN_l} \frac{1}{n^{\alpha\gamma}} \left(H(\mathscr{P}_n) - \mathrm{E}[H(\mathscr{P}_n)]\right).\end{aligned}$$

and, by using Theorem 6.1 and Limits (6.3) and (6.4), we obtain the thesis. □

Chapter 7
Applications

In this chapter, we apply estimators $\widehat{\mathrm{L}}_{A,1}$ (Equation (5.8)) and $\widehat{\mathrm{L}}_{A}^{circles}$ (Equation (5.6)) to both simulated and real data. Estimator $\widehat{\widehat{\mathrm{L}}}_{A,1}$ (Equation (5.10)) is not considered because it can be affected by computational problems, when the directions of the fibres at the intersection points are close and thus the denominator is close to zero.

In Section 7.1, we use simulated data to verify empirically the asymptotic properties (i.e. the speed of convergence) of the two estimators and we derive two methods to compute their variance when only one window of observation is available. The aim of this study is to verify the behavior of our estimators in presence of finite windows of observations, and thus finite samples. Moreover, since in real applications the window of observation is usually a digital image, we verify the asymptotic properties of the estimators also on simulated images of fibre processes and we compare them with their "theoretical" counterpart.

In Section 7.2, we apply the two estimators to images of processes of angiogenesis on mouse cornea. The mice were treated with different antibodies, which should inhibit the angiogenesis process. Therefore, we want to determine the best performing antibody in a quantitative way, by estimating the intensity of the corresponding fibre processes of angiogenesis and by comparing them.

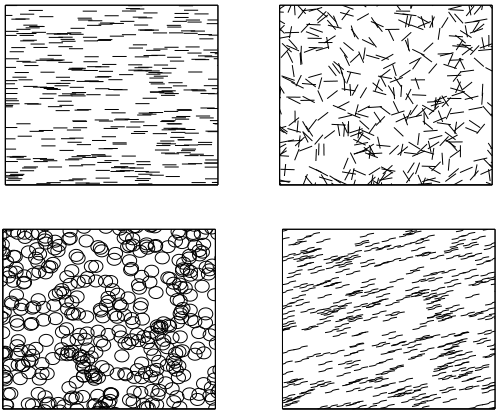

Fig. 7.1 Examples of realizations of the four types of process Φ_1 used in the simulations: on the top-left-hand side, the Poisson horizontal segment process; on the top-right-hand side, the Poisson segment process; on the bottom-left-hand side, the Poisson circle process; on the bottom-right-hand side, the Boolean process of arcs of parabola.

7.1 Simulations

In this section, we study the behavior of estimators $\widehat{\mathrm{L}}_{A,1}$ and $\widehat{\mathrm{L}}_A^{circles}$ on simulations, in order to verify empirically their asymptotic properties. In the simulations, we used Boolean fibre processes (see Section 5.4) to define both the fibre process under study Φ_1 and the test one Φ_2, and we use λ_1 and λ_2 to denote the intensity of the Poisson point process used to generate Φ_1 and Φ_2, respectively. In order to show that the asymptotic properties are independent from the shape of the fibre process Φ_1, we considered four types of process Φ_1 (see Figure 7.1):

- the Poisson horizontal segment process with length of the segments l_1,
- the Poisson segment process with length of the segments l_1 and uniform distribution of their orientation,
- the Poisson circle process with ray R_1,

- a Boolean process where the fibres are composed by two arcs of parabola.

Namely, in the last case, if the center of the fibre is $\mathbf{x}^0$, then the fibre can be represented in the following way:

$$x_2 = \begin{cases} x_2^0 + \sqrt{x_1 - x_1^0} \text{ if } x_1 \in [x_1^0, x_1^0 + l_x/2] \\ x_2^0 - \sqrt{x_1^0 - x_1} \text{ if } x_1 \in [x_1^0 - l_x/2, x_1^0), \end{cases}$$

where l_x is the width of the range of the x-values of the arcs of parabola.

The unbiasedness, the strong consistency and the asymptotic normality of $\widehat{\mathrm{L}}_{A,1}$ require that the test process Φ_2 is stationary and isotropic (see Section 5.3, Proposition 5.2 and Corollary 6.1, respectively). Therefore, we considered two types of test processes Φ_2: the Poisson segment process with length of the segments l_2 and uniform distribution of their orientation and the Poisson circle process with ray R_2. In the following, we will call $\widehat{\mathrm{L}}_{A,1,seg}$ the estimator $\widehat{\mathrm{L}}_{A,1}$ where Φ_2 is a Poisson segment process and $\widehat{\mathrm{L}}_{A,1,circ}$ the estimator $\widehat{\mathrm{L}}_{A,1}$ where Φ_2 is a Poisson circle process. Our tests over different choices for the geometry of Φ_1 and Φ_2 have the aim to reveal whether the speed of convergence of the estimators (both for what concerns consistency and asymptotic normality) is influenced by the geometric characteristics of the fibre processes under study, or instead depends only on their intensities. In fact, estimator $\widehat{\mathrm{L}}_{A,1}$ denotes a class of estimators which depend on the particular chosen test process Φ_2. The choice of Φ_2 can be crucial in case of a small window of observation, in order to reduce the variance of the estimator and thus obtain an accurate estimate of $\mathrm{L}_{A,1}$. The study of the dependence of the asymptotic properties of $\widehat{\mathrm{L}}_{A,1}$ on the parameters that characterize Φ_2 can give us an indication on how to choose Φ_2.

In the simulations, we fixed the parameters of the process Φ_1 and we varied the parameters of the process Φ_2, in order to observe the asymptotic behavior in the various cases. We set $\lambda_1 = 0.004$, $l_1 = 20$, $R_1 = 10$ and $l_x = 20$ and thus the true intensities of the corresponding processes are $\lambda_1 l_\varphi = 0.08, 0.2513, 21.52$, respectively, where l_φ represents the length of

a fibre of the process. Regarding the process Φ_2, we allowed λ_2 to vary in $[0.002, 0.008]$, $l_2 = 20$ or 60 and $R_2 = 10$ or 30. For the estimator $\widehat{\mathrm{L}}_A^{circles}$, we considered a grid of non-intersecting circles with the same ray R, $R \in [5, 30]$. Moreover, in all the simulations we considered squared windows of observation W with side: 100, 200, 300, 400 and 500. In the following, we will call dimension (or dim) of the window the length of its side.

In Subsection 7.1.1, we verify the asymptotic properties of the estimators on simulated data. Moreover, since in the applications often we have only one or few images of the process Φ_1, in Subsection 7.1.2 we study how to estimate the variance of the estimators $\widehat{\mathrm{L}}_{A,1}$ and $\widehat{\mathrm{L}}_A^{circles}$ in this situation. Finally, in Subsection 7.1.3 we simulate "digital images" of fibre processes (that is we use the 2D-box approximation given by the pixels to represent the fibres) and we observe the asymptotic behavior of the estimators in a situation closer to the real applications.

7.1.1 Behavior of the variance and asymptotic normality of the estimator

In order to verify empirically the speed of convergence in the consistency of $\widehat{\mathrm{L}}_{A,1}$ and $\widehat{\mathrm{L}}_A^{circles}$ and thus to observe the asymptotic behavior of their variance (also with respect to the parameters of the process Φ_2), we generated 100 observations of the process Φ_1, for each type of process Φ_1 and dimension of W, and we computed: $\widehat{\mathrm{L}}_{A,1,seg}$ (with both $l_2 = 20$ and $l_2 = 60$) and $\widehat{\mathrm{L}}_{A,1,circ}$ (with both $R_2 = 10$ and $R_2 = 30$) for 10 values of λ_2 in $[0.002, 0.008]$, and $\widehat{\mathrm{L}}_A^{circles}$ with 11 values of R in $[5, 30]$. For fixed Φ_1, *dim* and either (λ_2,l_2) or (λ_2,R_2) or R (depending on the estimator used), the variance was computed by using the sample variance of the values of the estimator computed over the 100 observations.

Regarding estimator $\widehat{\mathrm{L}}_{A,1}$, in Figures 7.2, 7.3, 7.4 and 7.5, we can see that, fixed any combination of shape of process Φ_2, type of process Φ_1 and dimension of W, the variance decreased when λ_2 increased. Moreover, in general, fixed λ_2, the variance of $\widehat{\mathrm{L}}_{A,1,seg}$ with $l_2 = 60$ was lower than the

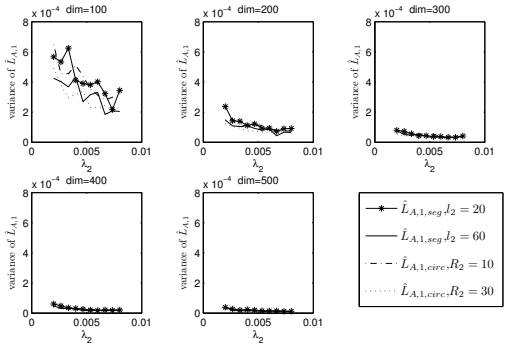

Fig. 7.2 Variance of the estimator $\widehat{\mathrm{L}}_{A,1}$ computed on 100 observations of the Poisson horizontal segment process Φ_1 ($\lambda_1 = 0.004$ and $l_1 = 20$). As test fibre process Φ_2, we used the Poisson segment process, with both $l_2 = 20$ and $l_2 = 60$, and the Poisson circle process, with both $R_2 = 10$ and $R_2 = 30$, each considering 10 values of λ_2 in the interval $[0.002, 0.008]$.

corresponding one with $l_2 = 20$, and the variance of $\widehat{\mathrm{L}}_{A,1,circ}$ with $R_2 = 30$ was lower than the corresponding one with $R_2 = 10$. In fact, the variance in Limit (6.54) is inversely proportional to $\mathrm{L}^2_{A,2}$, but also proportional to τ^2 which depends on Φ_2. Finally, fixed any combination of type of process Φ_2 and Φ_1, the variance decreased as the dimension increased (as should happen due to Limit (6.54)).

Regarding estimator $\widehat{\mathrm{L}}_A^{circles}$, first we observe that, in order to have a system of non-intersecting circles and have a high number of circles in the system, we used a grid of $N = M \times M$ circles with $M = dim/(2R)$. Therefore, the denominator in $\widehat{\mathrm{L}}_A^{circles}$ is $4RN = dim^2/2R$, i.e. enlarging the ray of the circles we do not decrease the variance. In fact, with a large R, we obtain a smaller sample of the point process of intersection and thus we augment the variance of $\#(\Phi_1 \cap \psi)$ (and then of $\widehat{\mathrm{L}}_A^{circles}$). We verified this behavior on the simulations. Figure 7.6 confirms our assumptions: fixed any combination of Φ_1 and dimension of W, the variance increased

as R increased. Moreover, fixed any combination of Φ_1 and R, the variance decreased as the dimension of the window increased, as was expected.

In conclusion, for any type of process Φ_1, at small dimension of the window of observation (such as 100 or 200), a considerable reduction of the variance of estimators $\widehat{\mathrm{L}}_{A,1,seg}$ and $\widehat{\mathrm{L}}_{A,1,circ}$ can be achieved by augmenting both λ_2 and either l_2 (for $\widehat{\mathrm{L}}_{A,1,seg}$) or R_2 (for $\widehat{\mathrm{L}}_{A,1,circ}$). Instead, in order to reduce the variance of $\widehat{\mathrm{L}}_A^{circles}$, we need to decrease R. For higher values of the dimension of the window, we cannot notice any significant difference among the variances of the estimators with respect to the parameters used.

Finally, by comparing the variances of estimators $\widehat{\mathrm{L}}_A^{circles}$ and $\widehat{\mathrm{L}}_{A,1,circ}$ for a fixed dimension of the image and a fixed value of the radius, we can notice that not always the variance of $\widehat{\mathrm{L}}_A^{circles}$ (computed on a deterministic grid of circles) is lower than the variance of $\widehat{\mathrm{L}}_{A,1,circ}$ (computed on a random grid of circles), as we could expected. For example, often in simulations $\mathrm{Var}(\widehat{\mathrm{L}}_A^{circles})$ and $\mathrm{Var}(\widehat{\mathrm{L}}_{A,1,circ})$ are close for R=R_2=30 and

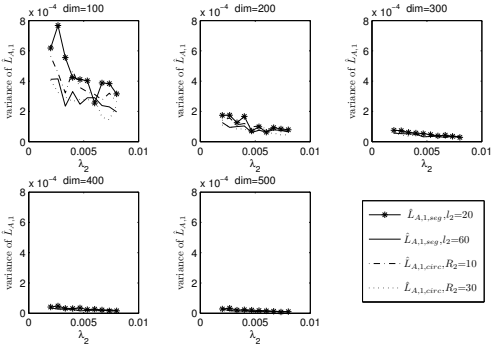

Fig. 7.3 Variance of the estimator $\widehat{\mathrm{L}}_{A,1}$ computed on 100 observations of the Poisson segment process Φ_1 ($\lambda_1 = 0.004$ and $l_1 = 20$). As test fibre process Φ_2, we used the Poisson segment process, with both $l_2 = 20$ and $l_2 = 60$, and the Poisson circle process, with both $R_2 = 10$ and $R_2 = 30$, each considering 10 values of λ_2 in the interval $[0.002, 0.008]$.

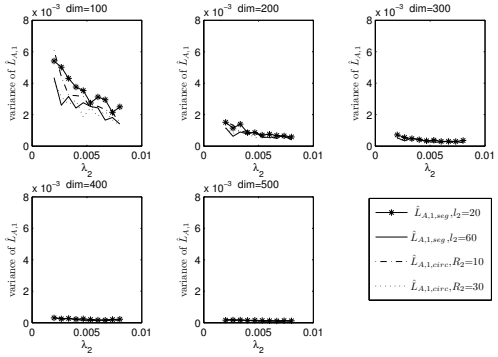

Fig. 7.4 Variance of the estimator $\widehat{\mathrm{L}}_{A,1}$ computed on 100 observations of the Poisson circle process Φ_1 ($\lambda_1 = 0.004$ and $R_1 = 10$). As test fibre process Φ_2, we used the Poisson segment process, with both $l_2 = 20$ and $l_2 = 60$, and the Poisson circle process, with both $R_2 = 10$ and $R_2 = 30$, each considering 10 values of λ_2 in the interval $[0.002, 0.008]$.

$\lambda_2 = 0.01$, therefore it is sufficient to increase λ_2 to obtain $\mathrm{Var}(\widehat{\mathrm{L}}_{A,1,circ}) < \mathrm{Var}(\widehat{\mathrm{L}}_A^{circles})$ (since $\mathrm{Var}(\widehat{\mathrm{L}}_A^{circles})$ decreases when λ_2 increases). Thus, $\widehat{\mathrm{L}}_A^{circles}$ (and $\widehat{\mathrm{L}}_{A,1,seg}$) can have a lower variance than $\widehat{\mathrm{L}}_A^{circles}$ by tuning suitably the parameters of Φ_2.

Using the same data, we also verified empirically the asymptotic normality of the estimators by a χ^2-test of goodness of fit with null hypothesis that the estimator is normally distributed (with mean and variance estimated from the sample). We found that, for the estimator $\widehat{\mathrm{L}}_{A,1}$, we did not reject the null hypothesis in 90% of the cases at level 0.05 and in 97.75% of the cases at level 0.01. Moreover, the cases in which we rejected the hypothesis did not show any particular dependence on the value of the parameters (such as λ_2 and the dimension of the window), but were due to the particular sample process. Considering the estimator $\widehat{\mathrm{L}}_A^{circles}$, we did not rejected the null hypothesis in 90.45% of the cases at level 0.05 and 97.27% of the cases at level 0.01. Therefore, for both estimators, we can approximate their distribution with a normal distribution, already at

dimension 100 of the window of observation. In Figure 7.7, we can see an example of asymptotic convergence of the distribution of the estimator $\widehat{\mathrm{L}}_{A,1}$.

7.1.2 Estimation of the variance on a single window of observation

As we mentioned previously, in many applications we have only one or few images of the process Φ_1 from which we can estimate its intensity. If we suppose that the point process of intersections is independent at distance l and we have a sufficiently large window of observation W, we can subdivide the window into independent subwindows that we can use as a sample of windows of observation for the estimation of τ^2 (see Equations (6.54) and (6.57)). Therefore, the size of W must guarantee a suffi-

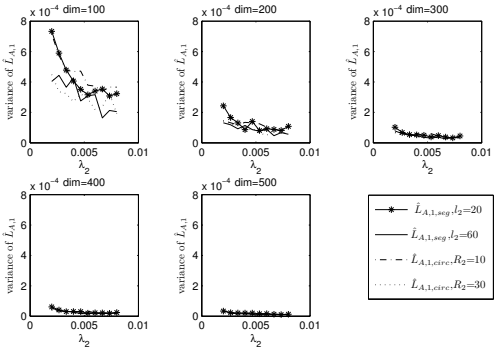

Fig. 7.5 Variance of the estimator $\widehat{\mathrm{L}}_{A,1}$ computed on 100 observations of the Poisson process of arcs of parabola Φ_1 ($\lambda_1 = 0.004$ and $l_x = 20$). As test fibre process Φ_2, we used the Poisson segment process, with both $l_2 = 20$ and $l_2 = 60$, and the Poisson circle process, with both $R_2 = 10$ and $R_2 = 30$, each considering 10 values of λ_2 in the interval $[0.002, 0.008]$.

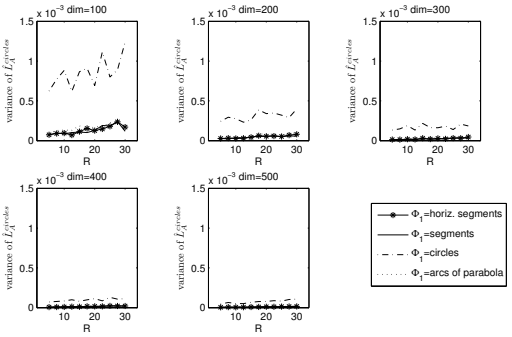

Fig. 7.6 Variance of the estimator $\widehat{\mathrm{L}}_A^{circles}$ computed on 100 observations of the fibre process Φ_1. As system of circles, we used systems of non-intersecting circles considering 11 values of the ray of the circles R in the interval $[5,30]$. As fibre process Φ_1, we used: the Poisson segment process with horizontal or uniform directions ($\lambda_1 = 0.004$ and $l_1 = 20$), the Poisson circle process ($\lambda_1 = 0.004$ and $R_1 = 10$) and the Poisson process of arcs of parabola ($\lambda_1 = 0.004$ and $l_x = 20$).

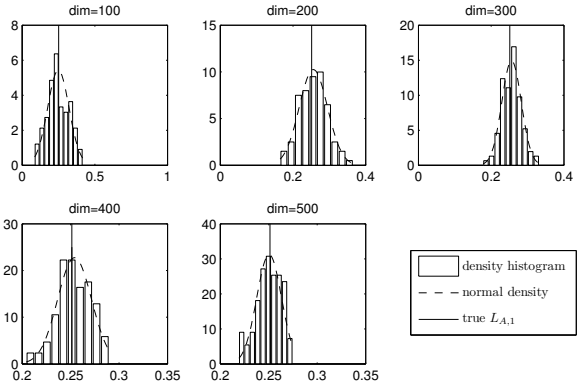

Fig. 7.7 Example of convergence of the distribution of estimator $\widehat{\mathrm{L}}_{A,1}$. As process Φ_1 we used the Poisson circle process ($\lambda_1 = 0.004$ and $R_1 = 10$) and, as fibre process Φ_2, we used the Poisson segment process, with both $l_2 = 20$ and $\lambda_2 = 0.002$.

ciently large number of subwindows. This assumption allows to estimate the variance of both $\widehat{L}_{A,1}$ and $\widehat{L}_A^{circles}$.

As an example, we simulated a Poisson horizontal segment process (Φ_1) in a window of observation of dimension 1000. To estimate the intensity, we applied $\widehat{L}_{A,1,seg}$ with $l_2 = 20$ and $\lambda_2 = 0.003$. Since the point process of the intersections is independent at distance $l = 20$, we divided the image in subimages (squares) of dimension 20 and, for the estimation of τ^2 (Equation (6.54)), we considered only the subimages with centers at distance greater or equal to $4l = 80$ (i.e. a sample of 144 subimages). In this way, the quantities $M_{i,j}\sum_{d=1}^{3} M_{i,j+d}$ and $M_{i',j'}\sum_{d=1}^{3} M_{i',j'+d}$ were independent for any i, j, i' and j'. Using the formula of the variance in Corollary 6.1, we obtained the following 95% confidence interval: $[0.0742, 0.0854]$, which is close to the one computed on 100 observations of Φ_1 in 100 windows W of side 1000 ($[0.0746, 0.0850]$).

In case W is not sufficiently large for the previous type of estimation, we are still able to have a rough estimate of the variance. In fact, for a fixed W, we can compute $\widehat{L}_{A,1}$ on n independent realizations of the process Φ_2 or we can compute $\widehat{L}_A^{circles}$ using n translated systems of circles (all the systems must be all contained inside W). Surely, the n values of the estimators will be correlated, but if we correct the estimate for the correlation we can still estimate the variance.

Given a window W of observation of the process Φ_1 and denoting by $\widehat{L}$ an estimator of the the intensity ($\widehat{L}_{A,1}$ or $\widehat{L}_A^{circles}$, in our case), we indicate with $\widehat{L}_i$ the i^{th} estimator of type $\widehat{L}$ computed on W, from the i^{th} realization of Φ_2, in case of $\widehat{L}_{A,1}$, and the i^{th} system of circles, in case of $\widehat{L}_A^{circles}$. $\overline{\widehat{L}}$ denotes the average of these estimators. Obviously, $\{\widehat{L}_i\}_{i=1}^n$ have all the same distribution and we call μ and σ^2, respectively, their mean and variance. Moreover, since they are calculated on the same window W and the same realization of the process Φ_1, they are not independent and we call c the covariance of any pair of estimators ($c = \mathrm{Cov}(\widehat{L}_i, \widehat{L}_j)$, for any $i \neq j$ and $i, j = 1, \ldots, n$). Due to this dependence, the sample variance of the estimators is an unbiased estimator for $\sigma^2 - c$. In fact,

$$
\begin{aligned}
&\frac{1}{n-1}\mathrm{E}\left[\sum_{i=1}^{n}\left(\widehat{L}_i-\overline{\widehat{L}}\right)^2\right]\\
&=\frac{1}{n-1}\mathrm{E}\left[\sum_{i=1}^{n}\widehat{L}_i^2-n\overline{\widehat{L}}^2\right]\\
&=\frac{1}{n-1}\left[n(\sigma^2+\mu^2)-\frac{1}{n}\mathrm{E}\left[\left(\sum_{i=1}^{n}\widehat{L}_i\right)^2\right]\right]\\
&=\frac{1}{n-1}\left[n(\sigma^2+\mu^2)-\frac{1}{n}\left(n\mathrm{E}[\widehat{L}_1^2]+n(n-1)\mathrm{E}[\widehat{L}_1\widehat{L}_2]\right)\right]\\
&=\frac{1}{n-1}\left[n(\sigma^2+\mu^2)-\frac{1}{n}(n(\sigma^2+\mu^2)+n(n-1)(c+\mu^2))\right]\\
&=\frac{1}{n-1}\left[n(\sigma^2+\mu^2)-\frac{1}{n}(n\sigma^2+n^2\mu^2+n(n-1)c)\right]\\
&=\frac{1}{n-1}[n(\sigma^2+\mu^2)-\sigma^2-n\mu^2-(n-1)c]\\
&=\frac{1}{n-1}[(n-1)\sigma^2-(n-1)c]\\
&=\sigma^2-c=\mathrm{Var}(\widehat{L}_i)-\mathrm{Cov}(\widehat{L}_i,\widehat{L}_j).
\end{aligned}
$$

Therefore, if $c\le c_1\sigma^2$ with $c_1<1$, then

$$
\mathrm{Var}(\widehat{L}_i)\le\frac{1}{1-c_1}\frac{1}{n-1}\mathrm{E}\left[\sum_{i=j}^{n}\left(\widehat{L}_j-\overline{\widehat{L}}\right)^2\right], \tag{7.1}
$$

for all $i=1,\dots,n$. Thus, if c_1 is small, we can obtain a good approximation of $\mathrm{Var}(\widehat{L}_i)$ with the sample variance of estimators computed on one single image. Instead, if $c\sim\sigma^2$ (that is c_1 is close to 1), we cannot retrieve σ^2 from the sample variance.

In order to have an idea of the order of magnitude of c_1, we performed some simulations to estimate c and σ^2 for our estimators, via the sample covariance and the sample variance of estimates over independent realizations of Φ_1. Similarly to the previous subsection, we generated 100 obser-

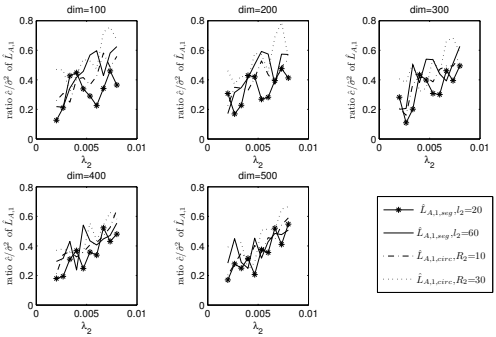

Fig. 7.8 Ratio $\hat{c}/\hat{\sigma}^2$ of the estimator $\widehat{\mathrm{L}}_{A,1}$ computed on 100 realizations of the Poisson horizontal segment process Φ_1 ($\lambda_1 = 0.004$ and $l_1 = 20$). As fibre process Φ_2, we used the Poisson segment process, with both $l_2 = 20$ and $l_2 = 60$, and the Poisson circle process, with both $R_2 = 10$ and $R_2 = 30$, each considering 10 values of λ_2 in the interval $[0.002, 0.008]$.

vations of the process Φ_1, for each type of process Φ_1 and dimension of W, and we computed (twice for each realization of Φ_1): $\widehat{\mathrm{L}}_{A,1,seg}$ (with both $l_2 = 20$ and $l_2 = 60$) and $\widehat{\mathrm{L}}_{A,1,circ}$ (with both $R_2 = 10$ and $R_2 = 30$) for 10 values of λ_2 in $[0.002, 0.008]$, and $\widehat{\mathrm{L}}_A^{circles}$ for 11 values of R in $[5, 30]$. For any estimator $\widehat{L}$, we call $\widehat{L}_{i,j}$ the i^{th} estimator $\widehat{L}$ computed on the j^{th} realization of Φ_1. As explained previously, in case of $\widehat{\mathrm{L}}_{A,1,seg}$ or $\widehat{\mathrm{L}}_{A,1,circ}$, the i^{th} estimator $\widehat{L}$ is calculated from the i^{th} realization of the corresponding test process Φ_2. In case of $\widehat{\mathrm{L}}_A^{circles}$, it is calculated by using the i^{th} system of circles (the systems must contain the same number of circles, with the same ray and the circles must be all contained in the window W). Therefore, for any $\widehat{L}$, the covariance c and the variance σ^2 are estimated as follows,

$$\hat{c} = \frac{1}{99} \sum_{j=1}^{100} \left(\widehat{L}_{1,j} - \overline{\widehat{L}}_1 \right) \left(\widehat{L}_{2,j} - \overline{\widehat{L}}_2 \right)$$

$$\hat{\sigma}^2 = \frac{1}{99}\sum_{j=1}^{100}\left(\widehat{L}_{1,j} - \overline{\widehat{L}}_1\right)^2,$$

where $\overline{\widehat{L}}_i = \sum_{j=1}^{100}\widehat{L}_{i,j}/100$, $i = 1, 2$.

In Figures 7.8, 7.9, 7.10 and 7.11, we can see the results for the estimator $\widehat{\mathrm{L}}_{A,1}$. The ratio $\hat{c}/\hat{\sigma}^2$ seemed to be independent of the dimension of the window. Instead, it increased as λ_2 increased with a rate that slightly depended on the shape of the two processes. Note that, for all $\lambda_2 \in [0.002, 0.008]$, usually $\hat{c} \leq 0.6\hat{\sigma}^2$. Therefore, if we are analyzing a fibre process that can be modeled as a Boolean fibre process with geometric characteristics similar to the ones of the four types of processes Φ_1 considered by us, then using a test process Φ_2 with intensity $\lambda_2 \in [0.002, 0.008]$, we can roughly estimate the variance by using the previous upper bound in Equation (7.1) with $c_1 = 0.6$, obtaining for any i,

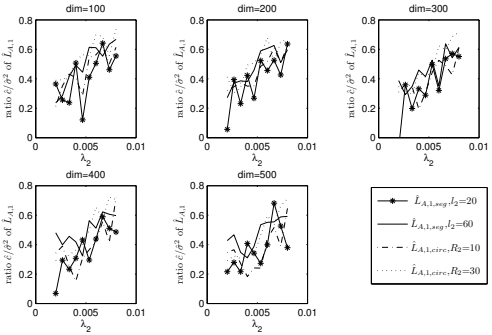

Fig. 7.9 Ratio $\hat{c}/\hat{\sigma}^2$ of the estimator $\widehat{\mathrm{L}}_{A,1}$ computed on 100 realizations of the Poisson segment process Φ_1 ($\lambda_1 = 0.004$ and $l_1 = 20$). As fibre process Φ_2, we used the Poisson segment process, with both $l_2 = 20$ and $l_2 = 60$, and the Poisson circle process, with both $R_2 = 10$ and $R_2 = 30$, each considering 10 values of λ_2 in the interval $[0.002, 0.008]$.

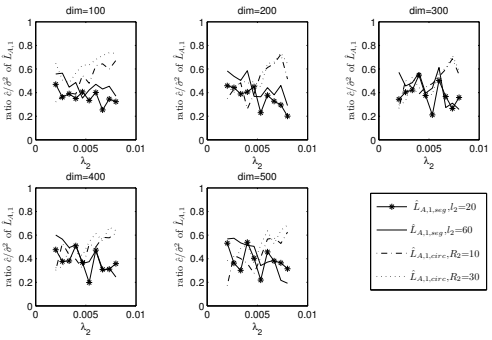

Fig. 7.10 Ratio $\hat{c}/\hat{\sigma}^2$ of the estimator $\widehat{\mathrm{L}}_{A,1}$ computed on 100 realizations of the Poisson circle process Φ_1 ($\lambda_1 = 0.004$ and $R_1 = 10$). As fibre process Φ_2, we used the Poisson segment process, with both $l_2 = 20$ and $l_2 = 60$, and the Poisson circle process, with both $R_2 = 10$ and $R_2 = 30$, each considering 10 values of λ_2 in the interval $[0.002, 0.008]$.

$$\widehat{\mathrm{Var}}(\widehat{L}_i) \approx 2.5 \frac{1}{n-1} \sum_{j=1}^{n} \left(\widehat{L}_j - \overline{\widehat{L}} \right)^2 . \tag{7.2}$$

In Figure 7.13, we can see an example of confidence intervals of $\mathrm{L}_{A,1}$ estimated in this way.

Instead, for the estimator $\widehat{\mathrm{L}}_A^{circles}$, we observed that the ratio $\hat{c}/\hat{\sigma}^2$ was always close to 1, for all values of the ray of the circles and all kind of process Φ_1 (Figure 7.12). Therefore, if we are using estimator $\widehat{\mathrm{L}}_A^{circles}$, we suggest not to use the sample variance of the estimates computed on the same realization of Φ_1, for estimating a confidence interval for the intensity.

Remark 7.1. It is evident, from the discussion by which we obtained the approximation (7.2), the importance of having a suitable model for a given real fibre process. The model can be simulated and the value of c_1 can be estimated form the simulations, allowing thus a better approximation of

the variance of the estimators of the intensity, even in presence of one single image of the real process.

7.1.3 Behavior of the estimator on simulated images of fibre processes

Since in many applications the window of observation is a digital image, we did the same study as in Subsections 7.1.1 and 7.1.2 on simulated images, in order to see if the properties of the estimators are maintained even with the computational issues due to the pixels, that is when each fibre is represented by its 2D-box pixel approximation in a digital image.

Let us suppose to have a digital black and white image of Φ_1, where, for example, the fibres are depicted in black and the background in white. In this case, instead of knowing the exact coordinates of the points belong-

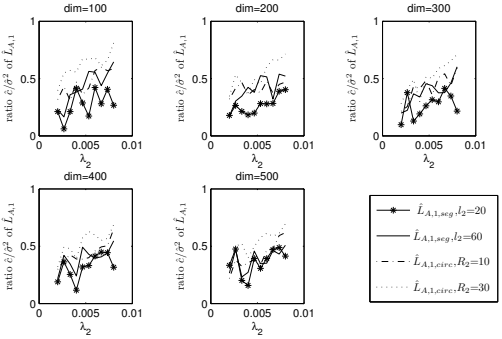

Fig. 7.11 Ratio $\hat{c}/\hat{\sigma}^2$ of the estimator $\widehat{\mathrm{L}}_{A,1}$ computed on 100 realizations of the Poisson process of arcs of parabola Φ_1 ($\lambda_1 = 0.004$ and $l_x = 20$). As fibre process Φ_2, we used the Poisson segment process, with both $l_2 = 20$ and $l_2 = 60$, and the Poisson circle process, with both $R_2 = 10$ and $R_2 = 30$, each considering 10 values of λ_2 in the interval $[0.002, 0.008]$.

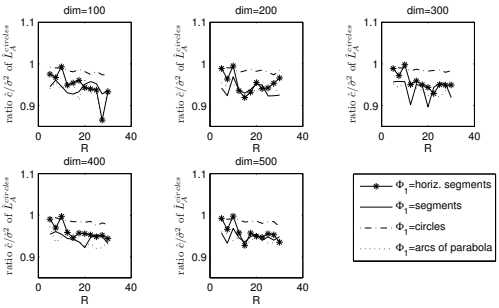

Fig. 7.12 Ratio $\hat{c}/\hat{\sigma}^2$ of the estimator $\widehat{L}_A^{circles}$ computed on 100 realizations of the fibre process Φ_1. As system of circles, we used systems of non-intersecting circles considering 11 values of the ray of the circles R in the interval $[5,30]$. As fibre process Φ_1, we used: the Poisson segment process with horizontal or uniform directions ($\lambda_1 = 0.004$ and $l_1 = 20$), the Poisson circle process ($\lambda_1 = 0.004$ and $R_1 = 10$) and the Poisson process of arcs of parabola ($\lambda_1 = 0.004$ and $l_x = 20$).

ing to the fibres, we only know that such points are located inside the black pixels. We necessarily need an algorithm that is able to identify the intersection points in this situation. Obviously, due to the pixel approximation of the fibre, we expect that the algorithm will not be able to identify all the intersections (that is the number of "false negative intersections" is not zero) and that sometimes it will detect an intersection where it does not exist. Therefore, we could have an under/over estimation of the intensity, respectively. We can verify if the estimation that we obtain is reliable and if the properties we found in the previous subsections hold also in this case, by studying the asymptotic properties of the estimators (computed with our algorithm) on simulated digital images and by comparing the results with the true theoretical values.

The algorithm for the identification of the intersection points is the following. After having simulated the centers of the fibres of Φ_2 (and their orientations, if the fibres are segments), we overlap once at a time a fibre of Φ_2 (here called γ) to the image of a realization of Φ_1 (here called φ_1). Then, we follow the pixels of γ in a consecutive way, identifying which of

them belong also to any of the fibres of φ_1 (see Figure 7.14). The number of intersections between γ and φ_1 is given by the number of disjoint sets of consecutive pixels along γ, that belong also to φ_1.

Remark 7.2. In the real application in Section 7.2, the number of intersections will be derived by counting the number of extended maxima to avoid the binarization of the image (see Section 7.2 for the explanation). Other approaches that could be used are discussed in that section with respect to the particular real images of fibre processes of our application. In fact, the aim of this section is to test on simulated images the same method that we will use in the real application.

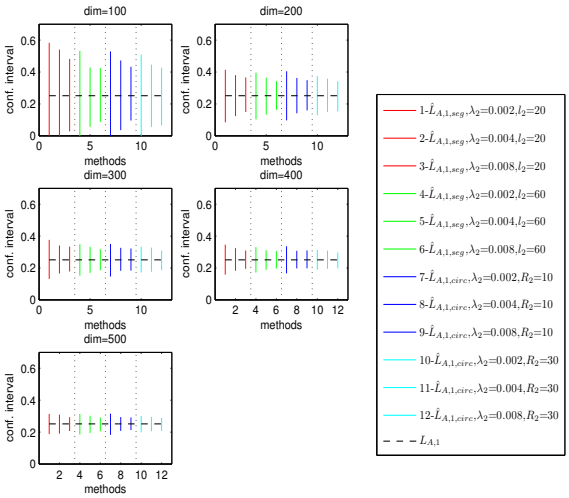

Fig. 7.13 Confidence intervals of the intensity of the Poisson circle process Φ_1 ($\lambda_1 = 0.004$ and $R_1 = 10$), computed by using the estimator $\widehat{\mathrm{L}}_{A,1}$. As fibre process Φ_2, we used the Poisson segment process, with both $l_2 = 20$ and $l_2 = 60$, and the Poisson circle process, with both $R_2 = 10$ and $R_2 = 30$, each considering three values of λ_2 (0.002, 0.004, 0.008).

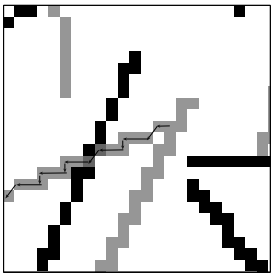

Fig. 7.14 Example that show how the algorithm works on digital images to detect the intersection points. The intersection point is detected as a set of contiguous pixels along a fibre of Φ_2, that belongs also to Φ_1 (but they may also be not contiguous in a fibre of Φ_1). In the image, Φ_1 is represented in black and Φ_2 in grey. See also Figure 7.19 for problems which may emerge in the identification of the number of intersections.

Similarly to Subsection 7.1.1, we generated 100 images of the process Φ_1, for each type of process Φ_1 and dimension of W, and we computed: $\widehat{\mathrm{L}}_{A,1,seg}$ (with both $l_2 = 20$ and $l_2 = 60$) and $\widehat{\mathrm{L}}_{A,1,circ}$ (with both $R_2 = 10$ and $R_2 = 30$), for $\lambda_2 \in \{0.002, 0.004, 0.008\}$, and $\widehat{\mathrm{L}}_A^{circles}$, for $R \in \{5, 10, 30\}$. We considered only three values of λ_2 and R, because in Subsection 7.1.1 we showed that the variance of the corresponding estimators changes in a continuous way with respect to these parameters. On each image we also computed $\widehat{\mathrm{L}}_A^{measure}$ (Equation (5.3)) as the product between the number of pixels in the fibres and a factor which represents the mean length of the fibres in a pixel. This factor has been computed via a linear regression, by comparing the estimated length of circumferences with different radii with their true length.

In Figures 7.15, 7.16, 7.17 and 7.18, we can see the confidence interval of the intensity at level 99%, computed for each type of process Φ_1 and each type of estimator used. As before, the variance of $\widehat{\mathrm{L}}_{A,1}$ decreased when both λ_2 and dim increased. Moreover, fixing the values of λ_2 and dim, the variance of $\widehat{\mathrm{L}}_{A,1,seg}$ decreased when l_2 increased and the variance

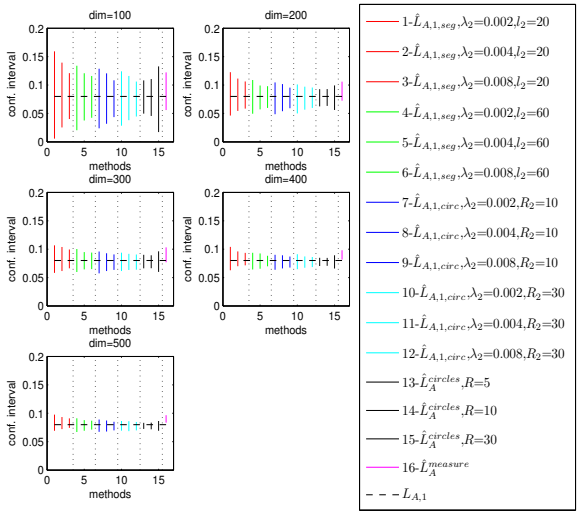

Fig. 7.15 Confidence intervals of the intensity of the Poisson horizontal segment process Φ_1 ($\lambda_1 = 0.004$ and $l_1 = 20$) at level 99%, computed by applying estimator $\widehat{\mathrm{L}}_{A,1}$ to digitized images of Φ_1. As fibre process Φ_2, we used the Poisson segment process, with both $l_2 = 20$ and $l_2 = 60$, and the Poisson circle process, with both $R_2 = 10$ and $R_2 = 30$, each considering three values of λ_2 (0.002, 0.004, 0.008). The figure reports also the confidence interval of $\widehat{\mathrm{L}}_A^{circles}$ (with $R = 5$, 10, 30) and of $\widehat{\mathrm{L}}_A^{measure}$.

of $\widehat{\mathrm{L}}_{A,1,circ}$ decreased when R_2 increased. On the contrary the variance of $\widehat{\mathrm{L}}_A^{circles}$ increased when R increased, but it decreased when *dim* increased.

Unfortunately, the estimators sometimes systematically overestimated (in particular when Φ_1 is the Poisson segment process or the Poisson process of arcs of parabola) or underestimated the intensity $\mathrm{L}_{A,1}$ (when Φ_1 is the Poisson circle process). These phenomena are due to the discretization of the fibres in pixels and to the algorithm, by which the intersections are identified. In fact, the loss of the intersection points happens, for example, when two "fibres" are close and, in the pixel resolution, they become a

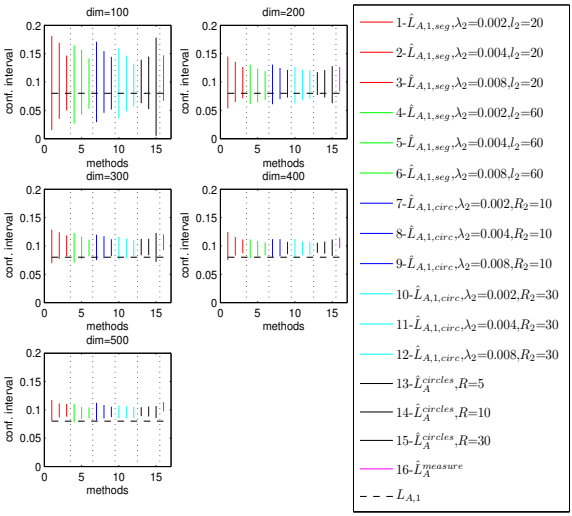

Fig. 7.16 Confidence intervals of the intensity of the Poisson segment process Φ_1 ($\lambda_1 = 0.004$ and $l_1 = 20$) at level 99%, computed by applying estimator $\widehat{L}_{A,1}$ to digitized images of Φ_1. As fibre process Φ_2, we used the Poisson segment process, with both $l_2 = 20$ and $l_2 = 60$, and the Poisson circle process, with both $R_2 = 10$ and $R_2 = 30$, each considering three values of λ_2 (0.002, 0.004, 0.008). The figure reports also the confidence interval of $\widehat{L}_A^{circles}$ (with $R = 5$, 10, 30) and of $\widehat{L}_A^{measure}$.

unique region (see for example Figure 7.19). Viceversa, a higher number of intersections can be counted for certain slopes of the fibres, that lead to a discretization of the fibre in pixels such that the pixels of the intersection are not contiguous on the fibre of Φ_2 (see again Figure 7.19). In fact, if the intersection point is represented by disjoint sets of contiguous pixels, then the program counts as many intersections as the number of these disjoint sets of pixels. On the other hand, when the fibres are not frequently tangent and their slopes are not "problematic" (like for the Poisson horizontal segment process), all estimators performed well (see Figure 7.15). By using other approaches of image analysis (like skeletons [76] and curve

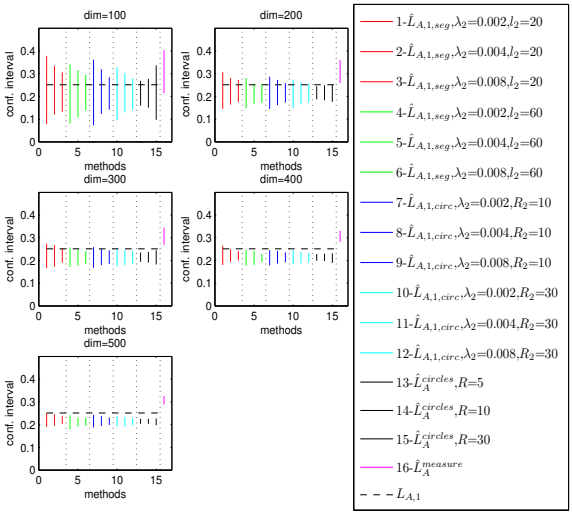

Fig. 7.17 Confidence intervals of the intensity of the Poisson circle process Φ_1 ($\lambda_1 = 0.004$ and $R_1 = 10$) at level 99%, computed by applying estimator $\widehat{\mathrm{L}}_{A,1}$ to digitized images of Φ_1. As fibre process Φ_2, we used the Poisson segment process, with both $l_2 = 20$ and $l_2 = 60$, and the Poisson circle process, with both $R_2 = 10$ and $R_2 = 30$, each considering three values of λ_2 (0.002, 0.004, 0.008). The figure reports also the confidence interval of $\widehat{\mathrm{L}}_A^{circles}$ (with $R = 5$, 10, 30) and of $\widehat{\mathrm{L}}_A^{measure}$.

fitting [16]), we could reduced the bias of the estimation, but these methods are not suitable for the analysis of our real images in Section 7.2, as it will be explained later.

In the applications (like in Section 7.2), usually, the fibres of Φ_1 have a width larger than a single pixel. This fact may help in the counting of the intersection points, since the intersections with thick fibres only seldom produces non-contiguous sets of pixels. In any case, since the over/underestimation depends on the shape of Φ_1, all estimators can still be used to compare the intensities of processes with the same shape (because the bias is the same).

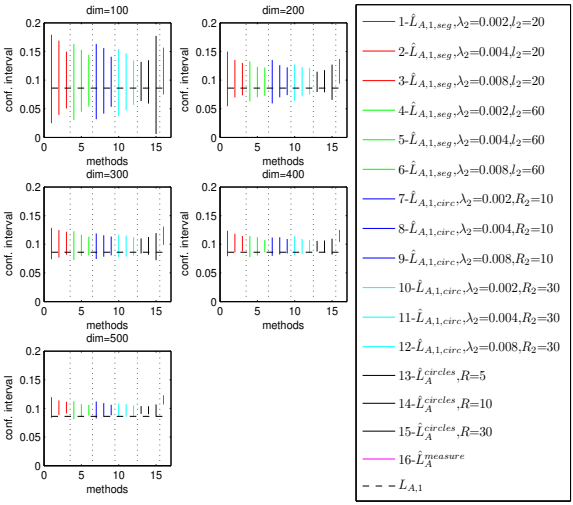

Fig. 7.18 Confidence intervals of the intensity of the Poisson process of arcs of parabola Φ_1 ($\lambda_1 = 0.004$ and $l_x = 20$) at level 99%, computed by applying estimator $\widehat{L}_{A,1}$ to digitized images of Φ_1. As fibre process Φ_2, we used the Poisson segment process, with both $l_2 = 20$ and $l_2 = 60$, and the Poisson circle process, with both $R_2 = 10$ and $R_2 = 30$, each considering three values of λ_2 (0.002, 0.004, 0.008). The figure reports also the confidence interval of $\widehat{L}_A^{circles}$ (with $R = 5$, 10, 30) and of $\widehat{L}_A^{measure}$.

Note that, in all experiments, the estimated variance of $\widehat{L}_A^{measure}$ was always very low, but the estimator showed always a positive bias. Thus, the use of estimators $\widehat{L}_{A,1}$ and $\widehat{L}_A^{circles}$ may be preferable in all applications where the bias must be small.

Analogously to Subsection 7.1.1, we used the same data also to verify empirically the asymptotic normality of the estimators computed on digital images. With this purpose, we applied a χ^2-test of goodness of fit with null hypothesis that the estimator is normally distributed (with mean and variance estimated from the sample). Concerning the estimator $\widehat{L}_{A,1}$, similarly to the theoretical simulations, we did not reject the null hypothesis

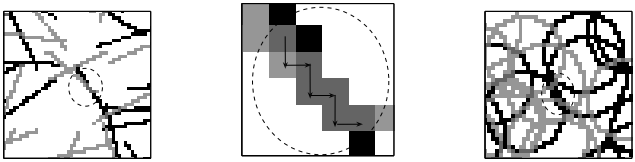

Fig. 7.19 On the left-hand side, an example of overestimation of the intensity: in the dotted circle we can see that the pixels of the intersection are not consecutive along the fibre of Φ_2 (that is intersecting Φ_1 at that point). A zoomed view of the intersection point is given in the image in the center. On the right-hand side, an example of underestimation of the intensity: in the dotted circle we observed examples where the fibres of Φ_1 are tangent to each other, forming a unique region, so that a lower number of intersections is counted. In the three images, Φ_1 is represented in black and Φ_2 in grey.

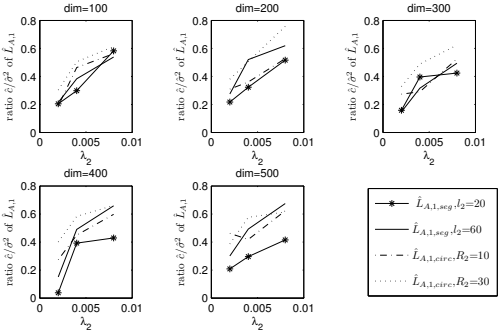

Fig. 7.20 Ratio $\hat{c}/\hat{\sigma}^2$ of the estimator $\widehat{\mathrm{L}}_{A,1}$ computed on 100 images of the Poisson horizontal segment process Φ_1 ($\lambda_1 = 0.004$ and $l_1 = 20$). As fibre process Φ_2, we used the Poisson segment process, with both $l_2 = 20$ and $l_2 = 60$, and the Poisson circle process, with both $R_2 = 10$ and $R_2 = 30$, each considering the values of λ_2 in $\{0.002, 0.004, 0.008\}$.

in 91% of the cases at level 0.05 and in 99% of the cases at level 0.01. Moreover, considering the estimator $\widehat{\mathrm{L}}_A^{circles}$, we did not rejected the null hypothesis in 93% of the case at level 0.05 and 100% of the cases at level

0.01. Therefore, the assumption of asymptotic normality of the estimators holds also in case of "digital images" and even for small dimension of the window of observation.

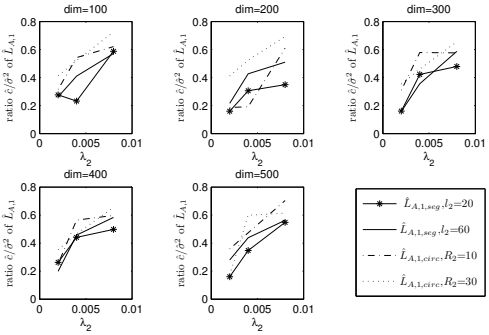

Fig. 7.21 Ratio $\hat{c}/\hat{\sigma}^2$ of the estimator $\widehat{\mathrm{L}}_{A,1}$ computed on 100 images of the Poisson segment process Φ_1 ($\lambda_1 = 0.004$ and $l_1 = 20$). As fibre process Φ_2, we used the Poisson segment process, with both $l_2 = 20$ and $l_2 = 60$, and the Poisson circle process, with both $R_2 = 10$ and $R_2 = 30$, each considering the values of λ_2 in $\{0.002, 0.004, 0.008\}$.

We wanted also to verify that, even in the case of simulated images, c/σ^2 (the ratio between the covariance of two estimators $\widehat{\mathrm{L}}_{A,1}$, computed on the same window of observation, and the variance of $\widehat{\mathrm{L}}_{A,1}$) was lower than 0.6 when $\lambda_2 \in [0.002, 0.008]$. Thus, we generated 100 realizations of the process Φ_1, for each type of process Φ_1 and dimension of W, and we computed (twice for each realization of Φ_1): $\widehat{\mathrm{L}}_{A,1,seg}$ (with both $l_2 = 20$ and $l_2 = 60$) and $\widehat{\mathrm{L}}_{A,1,circ}$ (with both $R_2 = 10$ and $R_2 = 30$) for $\lambda_2 \in \{0.002, 0.004, 0.008\}$. Figures 7.20, 7.21, 7.22 and 7.23 show that almost always $\hat{c}/\hat{\sigma}^2$ was lower than 0.6 (or very close to it) thus, in the real applications in Section 7.2, we will use Equation (7.2) to estimate the variance of the estimator, assuming that the model which can represent our real fibre process has geometric characteristics not much different from the ones that we tested. Obviously, in presence of a model for fibre

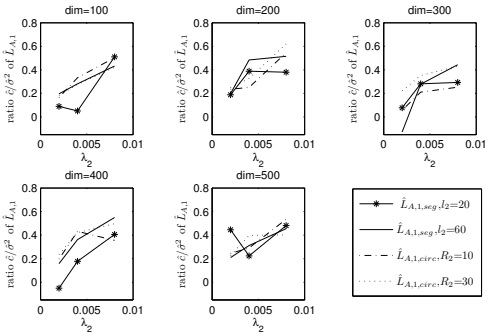

Fig. 7.22 Ratio $\hat{c}/\hat{\sigma}^2$ of the estimator $\widehat{\mathrm{L}}_{A,1}$ computed on 100 images of the Poisson circle process Φ_1 ($\lambda_1 = 0.004$ and $R_1 = 10$). As fibre process Φ_2, we used the Poisson segment process, with both $l_2 = 20$ and $l_2 = 60$, and the Poisson circle process, with both $R_2 = 10$ and $R_2 = 30$, each considering the values of λ_2 in $\{0.002, 0.004, 0.008\}$.

processes of angiogenesis, we could obtain a more reliable approximation of the upper bound of the variance of our estimators. An example of confidence interval computed with the variance estimated from Equation (7.2) is reported in Figure 7.24. We can notice that the intervals estimated in Figure 7.24 are not smaller than the ones estimated on 100 images of Φ_1 in Figure 7.17.

7.2 Applications to angiogenesis

In solid tumors, cell proliferation is helped by the formation of a vascular network (angiogenesis) around the tumor. In fact, the vessels supply nutrient to the cells, allowing the growth of the tumor. Therefore, a challenge in cancer research is to find an antibody which is able to inhibit the angiogenesis.

The protein VE-Cadherin plays a fundamental role in the creation of new vessels, thus the inactivation of this protein can inhibit the angio-

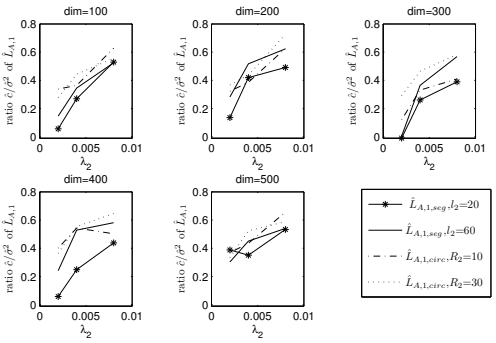

Fig. 7.23 Ratio $\hat{c}/\hat{\sigma}^2$ of the estimator $\widehat{\mathrm{L}}_{A,1}$ computed on 100 images of the Poisson process of arcs of parabola Φ_1 ($\lambda_1 = 0.004$ and $l_x = 20$). As fibre process Φ_2, we used the Poisson segment process, with both $l_2 = 20$ and $l_2 = 60$, and the Poisson circle process, with both $R_2 = 10$ and $R_2 = 30$,each considering the values of λ_2 in $\{0.002, 0.004, 0.008\}$.

genesis. In collaboration with an USA company, at IFOM (FIRC institute of molecular oncology foundation, Milan) several anti VE-Cadherin antibodies have been developed [14]. In order to determine which of these antibodies was more able to inhibit the formation of new vessels, they performed two types of *in vivo* experiments on mouse cornea. In the first type, they implanted on a mouse cornea a pellet, containing an angiogenic factor (called hrFGF-2) together with an antibody (*non-systemic treatment*). In the experiments of the second type, the antibody has been injected intraperitoneally to the mouse starting from the day after the pellet implantation, and thus the pellet contained only the angiogenic factor (*systemic treatment*). In both cases, photos of the eyes of the mice have been taken after 6 days from the pellet implantation (see, e.g. Figure 7.25).

The mice were treated with two types of antibodies: either the antibody nonimmune rat IgG (Rt-IgG), which has no inhibitory effect (i.e. is a placebo), or one of the developed anti VE-Cadherin antibodies. The anti VE-Cadherin antibodies used in the non-systemic treatment were 10G4, 8D6, 13E6 and 6D10, while in the systemic treatment they were 19E6

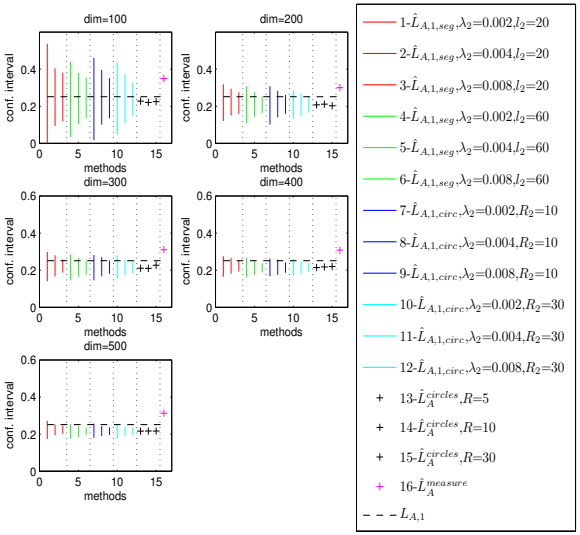

Fig. 7.24 Confidence intervals of the intensity of the Poisson circle process Φ_1 ($\lambda_1 = 0.004$ and $R_1 = 10$), computed on an image of Φ_1 by applying estimator $\widehat{L}_{A,1}$. As fibre process Φ_2, we used the Poisson segment process, with both $l_2 = 20$ and $l_2 = 60$, and the Poisson circle process, with both $R_2 = 10$ and $R_2 = 30$, each considering three values of λ_2 (0.002, 0.004, 0.008).

and E4B9. For each antibody, a sample of two images (of two eyes) was available. The small sample size is due to both the cost and the ethical issues related to the experiments on animals. We were not involved in the planning of the experiment.

In order to quantify the effect of a specific antibody in the inhibition of the angiogenic process (induced by hrFGF-2) from the images of the vessels, we can estimate one or more parameters that characterize their geometry. We modeled the capillaries as a stationary planar fibre process and, to compare the behavior of the antibodies, we estimated the intensity of the corresponding processes.

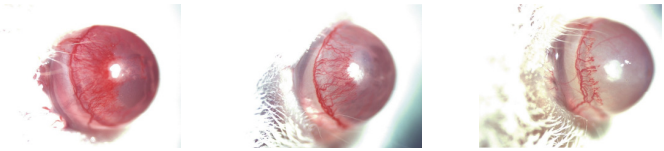

Fig. 7.25 On the left-hand side, the image of a mouse eye non-systemically treated with the antibody rat-IgG. In the center, the image of a mouse eye non-systemically treated with the antibody 6D10. On the right-hand side, the image of a mouse eye systemically treated with the antibody 19E6. Credits to E. Dejana et al. [14].

For the estimation of the intensity of the process of the vessels, we had to identify the capillaries in the images. Since the capillaries were of a vivid red color, while the background of the eye was pale red, we used the saturation values to select the capillaries. The saturation represents the degree of purity/intensity of a color and its range of values is in $[0,1]$ (1 means a pure color). Therefore, the capillaries have a higher saturation (close to 1) than the background (see Figure 7.26). Instead of using a threshold on the saturation values, which does not allow a reliable detection of thin capillaries, we computed the number of intersections by counting the number of *extended maxima* [76] in the saturation values along the fibre of the test process ($\widehat{\mathrm{L}}_{A,1,seg}$ and $\widehat{\mathrm{L}}_{A,1,circ}$) or the system of circles (for $\widehat{\mathrm{L}}_{A}^{circles}$). The extended maxima are defined as the regional maxima of the h-maxima transform. In practice, we first suppress all maxima whose depth is smaller than h, then we consider all connected components of pixels with a constant value and whose pixels at the boundaries have a lower value. The the choice of the h parameter is less critical than the choice of the threshold for the binarization of the image and, in our case, an expert of image analysis set $h = 0.03$ for all images. By counting the number of extended maxima, we automatically avoid false multiple intersections for thick fibres.

Other algorithms for image analysis that could be used to identify the fibres are, for example, skeletons [76], curve fitting [16] and filament segmentation [23, 24]. Our images of angiogenesis represent a very challeng-

ing problem in image analysis and therefore some tools for image analysis are not applicable. Skeletons would require a binarization of the image, which we are currently avoiding. Moreover, this morphological technique has a critical dependency on the threshold parameter for the binarization. The capillaries in our images show a complex and cluttered structure, thus they are hard to detect with sufficient reliability by curve fitting and other high-level approaches. Instead, filament segmentation is a recent and more sophisticated technique that could improve the calculation of the number of intersections. Nevertheless, we decided to not apply this algorithm since it is recent and not standardized and its code was not available. We will try to apply this technique as future work.

Fig. 7.26 Image analysis of a mouse eye non-systemically treated with the antibody rat-IgG. On the left-hand side, the original image, in the center, the saturation values of the image, on the right-hand side, the selected window of observation W.

Moreover, in order to consider the vessels as a stationary fibre process, for each eye, we selected, as window of observation W, only the part of the eye over the limbic vessel (i.e. the round vein) occupied by the capillaries (see Figure 7.26), where the process looked stationary.

To achieve a low variance, we used $\lambda_2 = 0.008$ for both $\widehat{\mathrm{L}}_{A,1,seg}$ and $\widehat{\mathrm{L}}_{A,1,circ}$ and we decided to increase the values of l_2 and R_2 (see Subsection 7.1.1 for notations and definitions). Since the portion of eye which contains the capillaries can be always included in a rectangle of sides 250×450 pixels (and sometimes is quite smaller than the area of that rectangle), we set l_2=100 and R_2=50. Concerning the estimator $\widehat{\mathrm{L}}_A^{circles}$, in order to obtain a low variance we chose a small value of R (see Subsection 7.1.1) but not too small so that the circles are not all contained in the

fibres (since the capillaries have a width larger than a pixel). Therefore, we set $R = 10$. We did not compute $\widehat{\mathrm{L}}_A^{measure}$ since it did not perform well in the experiments on simulated images in Subsection 7.1.3 and the capillaries have a variable width, usually larger than a pixel (and thus it is not possible to measure the length of the fibres by counting the pixels).

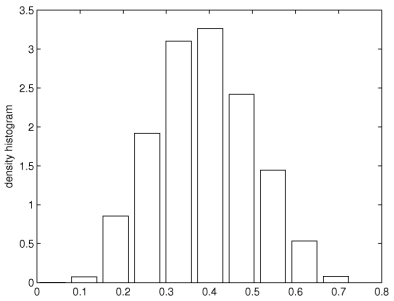

Fig. 7.27 Density histogram of the saturation values of the pixels inside the windows of observation inside all images of mouse eyes.

In order to better compare the geometry of the angiogenic processes, we also calculated or estimated other geometric parameters:

- the area fraction (ratio between the area of W and the area of the whole eye),
- h_{max} fraction (ratio between the height of the bounding-box of W and the height of the bounding-box of the whole eye, where the bounding-box is defined as the circumscribed rectangle),
- mean width fraction (defined as the ratio between the area fraction and h_{max} fraction),
- mean capillary width (estimated as the ratio of the area occupied by the capillaries and their estimated length).

To estimate the area occupied by the capillaries, we used a threshold computed on the basis of the saturation values of the pixels inside the windows of observations. As we said previously, the capillaries can be identified by a high value of the saturation of their pixels. But, since the capillaries are often thin, most of the pixels in W (the ones around the

borders of the capillaries) have a saturation value that lays between the one of the capillaries and the one of the background. Therefore, we estimated the density of the saturation values of the pixels (see Figure 7.27) and we computed a confidence interval around the mean value of the saturation (which represents the saturation value of the pixels around the borders of the capillaries). Then, the upper bound of this confidence interval was used as a threshold to determine the area occupied by the capillaries: all pixels with saturation value above this threshold were considered as belonging to the capillaries (see Figure 7.28).

Fig. 7.28 Example of computation of the area occupied by the capillaries. On the left-hand side, the saturation values of the pixels in the window of observation, on the right-hand side, the estimated area occupied by the capillaries.

7.2.1 Results

Since we had only a sample of two images per treatment, we decided to compute a confidence interval of the intensity for each eye. Moreover, due to the dimension of the window of observation, we could not estimate the variance of the estimators by using a sample of subimages (see Subsection 7.1.2). Thus, we could estimate only the intervals based on the estimators $\widehat{\mathrm{L}}_{A,1,seg}$ and $\widehat{\mathrm{L}}_{A,1,circ}$, by simulating, and overlapping to Φ_1, 100 i.i.d. copies of the process Φ_2 (see Subsection 7.1.2). In Figures 7.29

Table 7.1 Geometric parameters computed on the images of eyes of mice non-systemically treated. The mean capillary width was computed, by using as intensity the value estimated by $\widehat{\mathrm{L}}_{A,1,circ}$.

antibody	eye	area fraction	h_{max} fraction	mean width fraction	mean capil. width	$\widehat{\mathrm{L}}_{A,1,seg}$	$\widehat{\mathrm{L}}_{A,1,circ}$	$\widehat{\mathrm{L}}_{A}^{circles}$
Rt-IgG	1	0.278	0.760	0.366	6.999	0.093	0.088	0.101
Rt-IgG	2	0.343	0.769	0.446	10.535	0.095	0.089	0.098
10G4	1	0.272	0.669	0.407	12.182	0.078	0.073	0.084
10G4	2	0.311	0.737	0.422	6.468	0.099	0.096	0.108
8D6	1	0.146	0.526	0.278	1.689	0.091	0.086	0.105
8D6	2	0.165	0.603	0.274	1.847	0.096	0.092	0.103
13E6	1	0.288	0.775	0.372	2.949	0.091	0.088	0.099
13E6	2	0.195	0.550	0.355	0.952	0.092	0.087	0.098
6D10	1	0.110	0.395	0.279	0.609	0.094	0.093	0.105
6D10	2	0.136	0.496	0.274	1.243	0.093	0.088	0.101

and 7.30, we can observe that generally $\widehat{\mathrm{L}}_{A,1,seg}$ had a larger variance than $\widehat{\mathrm{L}}_{A,1,circ}$. Usually, the interval based on $\widehat{\mathrm{L}}_{A,1,circ}$ was almost included in the one based on $\widehat{\mathrm{L}}_{A,1,seg}$. Instead, $\widehat{\mathrm{L}}_{A}^{circles}$ seemed to overestimate the intensity, maybe because its variance is high due to the small number of circles in the test system.

Table 7.1 and Figure 7.29 report the results obtained on the images of non-systemic treatments. In Figure 7.29, we can observe that the antibody 10G4 had contrasting effects: in eye 1, the intensity decreased, but in eye 2, it increased. Also for the antibody 8D6 we had similar results, but in this case the variance is so large that we cannot say that the intensities corresponding to the mice treated with 8D6 are significantly different from the control ones, of the mice treated with Rt-IgG. The estimated intensities corresponding to the treatments with 13E6 and 6D10 were similar to the intensity of the capillaries in the control eyes.

In Table 7.1, we can see the effects of the treatments on the other geometric parameters. We observed a reduction of the area of observation for all treatments, apart from 10G4, and 8D6 and 6D10 achieved the small-

est area. But in the eyes of the mice treated with 6D10, the width of the capillaries was smaller than in the eyes of the mice treated with 8D6. The thinner capillaries supply less nutrient to the "tumor" (i.e. the pellet) and thus are less able to favor its growth. Therefore, the best non-systemic treatment was the one using the antibody 6D10.

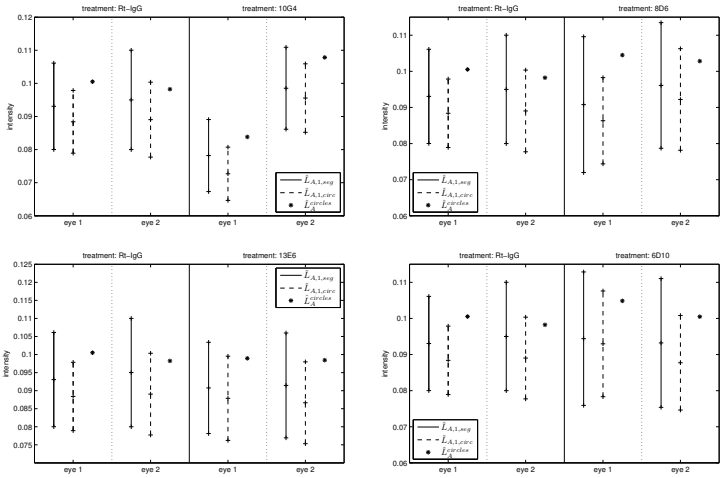

Fig. 7.29 Estimated confidence interval for the intensity of the capillaries at level 95%, computed on eyes non-systemically treated. Since we had to compute the interval on one image and it was small, we could compute it only for the estimators $\widehat{\mathrm{L}}_{A,1,seg}$ and $\widehat{\mathrm{L}}_{A,1,circ}$. We report also the value of $\widehat{\mathrm{L}}_{A}^{circles}$.

Regarding the eyes treated in a systemic way, we discovered that the antibody 19E6 had the effect to slightly reduce the intensity of the capillaries (Figure 7.30). Instead, the antibody E4B9 had contrasting results, since in one eye the intensity decreased and in the other one it increased (Figure 7.30). Moreover the other geometric characteristics of the process showed some differences, by the effects of these two treatments (see Table 7.2). In both cases we had a reduction of the fraction of area occupied

by the window of observation. But, for the antibody 19E6, the main reason of the reduction of the area of W was the decreased length of the capillaries (that can be seen through the reduction of the mean width fraction). Therefore, even if the capillaries in eye 1 were wider, they were not able to reach the "tumor" and thus to potentially favor its growth. For all these reasons, the antibody 19E6 performed better among the antibodies used in the systemic treatments.

For both types of treatments, the results we found were coherent with the qualitative results of the biomedical experts. In any case, due to the small sample size, the best selected antibodies should be tested on other mice in order to confirm their properties in inhibiting angiogenic processes.

Table 7.2 Geometric parameters computed on the images of eyes of mice systemically treated. The mean capillary width was computed, by using as intensity the value estimated by $\widehat{\mathrm{L}}_{A,1,circ}$.

antibody	eye	area fraction	h_{max} fraction	mean width fraction	mean capil. width	$\widehat{\mathrm{L}}_{A,1,seg}$	$\widehat{\mathrm{L}}_{A,1,circ}$	$\widehat{\mathrm{L}}_{A}^{circles}$
Rt-IgG	1	0.365	0.777	0.469	6.761	0.097	0.093	0.104
Rt-IgG	2	0.299	0.797	0.375	4.548	0.092	0.087	0.098
19E6	1	0.124	0.761	0.163	9.499	0.084	0.079	0.092
19E6	2	0.110	0.696	0.157	2.183	0.090	0.085	0.107
E4B9	1	0.179	0.591	0.302	3.878	0.078	0.072	0.083
E4B9	2	0.162	0.586	0.276	4.445	0.101	0.097	0.108

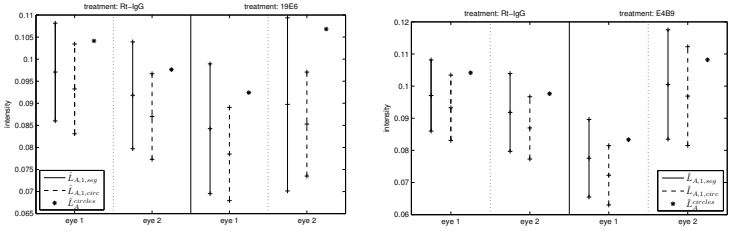

Fig. 7.30 Estimated confidence interval for the intensity of the capillaries at level 95%, computed on eyes systemically treated. Since we had to compute the interval on one image and it was small, we could compute it only for the estimators $\widehat{\mathrm{L}}_{A,1,seg}$ and $\widehat{\mathrm{L}}_{A,1,circ}$. We report also the value of $\widehat{\mathrm{L}}_{A}^{circles}$.

Conclusion

In the second part of the Thesis, I first proved a central limit theorem (CLT) for positive functionals of a point process independent at distance l and then I derived the asymptotic normality of the estimators $\widehat{\mathrm{L}}_A^{circles}$, $\widehat{\mathrm{L}}_{A,1}$ and $\widehat{\widehat{\mathrm{L}}}_{A,1}$, under suitable conditions. The Theorem ensures that, in presence of large samples, the distribution of the estimators can be approximated with a Gaussian, and thus (asymptotic) confidence intervals for the intensity can be computed. The structure of the proof of my CLT recalls the one of Theorem 3.1 in [61], where the authors showed a CLT for a functional of a Poisson or Binomial point process. Their proof was based on the fact that the process has independent increments, but this property does not hold for point processes derived from the intersection between two fibre processes (for $\widehat{\mathrm{L}}_{A,1}$ and $\widehat{\widehat{\mathrm{L}}}_{A,1}$) or between a fibre process and a fibre system (for $\widehat{\mathrm{L}}_A^{circles}$), since the fibres have length different from zero and points obtained by intersection with the same fibre are obviously correlated. Therefore, I generalized the proof to relax the hypothesis over the point process, requiring only the independence in Borel sets with distance greater than l. This hypothesis can be verified, for example, if the fibres are generated independently and have finite maximum length.

For practical applications (where the window of observation is finite and thus the number of fibres and of intersection points is finite), I verified on simulations the speed of convergence of the distribution of the estimators to the normal distribution in dependence of the parameters of the test process ($\widehat{L}_{A,1}$) or the system of circles ($\widehat{L}_A^{circles}$). The aim was to show if even with samples of "small size" the approximation with the asymptotic distribution was good. I observed that already with windows of relatively small size, if compared with the fibre process under study (side of window 100, length of a fibre 20 and mean number of fibres in the window 40), the distribution of the estimators is well approximated by a Gaussian, and their variance can be reduced by a suitable choice of the parameters of the test process. I also tested these properties on simulated images of fibre processes, since the behavior of the estimators could change because of the pixel approximation of the fibres. Moreover, since in practice only one or few images of the fibre process under study are available, I derived a method for the approximation of the variance of $\widehat{L}_{A,1}$ in presence of one single image, so that it is always possible to compute a confidence interval for the intensity $L_{A,1}$. Finally, I applied both estimators $\widehat{L}_A^{circles}$ and $\widehat{L}_{A,1}$ on images of angiogenic processes on the cornea of mice. On these images I estimated the intensity and other geometric characteristics of the fibre process and, by comparing these quantities for the different processes, I found results coherent with the qualitative conclusions of experts in biology.

The application to real images has two critical aspects: 1) the partition of the image in parts where the process is stationary, 2) the usage of a tool of image analysis which allows to count the intersections between the fibres. In the real images we used, we solved the first issue with the selection of the part of the image of interest by an expert, but, as future work, we could define some statistical techniques to look for a suitable and maybe better partition of the window of observation. We managed the second problem by using the saturation values of the pixels to identify the capillaries and by employing the notion of extended maxima for the identification of the intersections between fibres. This technique leads to bias in the estimation of the number of intersections, therefore in the future we will try to apply other techniques, like, e.g. a recent and more sophisticated method which is called filament segmentation [23, 24].

References

1. Affymetrix. BRLMM: an Improved Genotyping Calling Method for the GeneChip Human Mapping 500K Array Set. 2006.
2. Affymetrix. CNAT 4.0: Copy Number and Loss of Heterozygosity Estimation Algorithms for the GeneChip Human Mapping 10/50/100/250/500K Array Set. 2007.
3. D.G. Albertson, C. Collins, F. McCormick, and J.W. Gray. Chromosome aberrations in solid tumors. *Nature genetics*, 34(4):369–376, 2003.
4. M.D. Bacolod, G.S. Schemmann, S. Wang, R. Shattock, S.F. Giardina, Z. Zeng, J. Shia, R.F. Stengel, N. Gerry, J. Hoh, T. Kirchhoff, B. Gold, M.F. Christman, K. Offit, W.L. Gerald, D.A. Notterman, J. Ott, P.B. Paty, and F. Barany. The Signatures of Autozygosity among Patients with Colorectal Cancer. *Cancer Research*, 68(8):2610–2621, 2008.
5. S. Beà, I. Salaverria, L. Armengol, M. Pinyol, V. Fernández, E.M. Hartmann, P. Jares, V. Amador, L. Hernández, A. Navarro, G. Ott, A. Rosenwald, X. Estivill, and E. Campo. Uniparental disomies, homozygous deletions, amplifications and target genes in mantle cell lymphoma revealed by integrative high-resolution whole genome profiling. *Blood*, 113(13):3059–3069, 2009.
6. V. Beněs and J. Rataj. *Stochastic Geometry: selected topics*. Boston Kluwer Academic Publishers, Dordrecht, 2004.
7. A.T. Benjamin and J.J. Quinn. *Proofs that really count: the art of combinatorial proof*. Mathematical Association of America, Washington, D.C., 2003.
8. R. Beroukhim, M. Lin, Y. Park, K. Hao, X. Zhao, L.A. Garraway, E.A. Fox, E.P. Hochberg, I.K. Mellinghoff, M.D. Hofer, A. Descazeaud, M.A. Rubin, M. Meyerson, W.H. Wong, W.R. Sellers, and C. Li. Inferring Loss-of-Heterozygosity from

Unpaired Tumors Using High-Density Oligonucleotide SNP Arrays. *PLOS Computational Biology*, 2(5):323–332, 2006.

9. K.W. Broman and J.L. Weber. Long homozygous chromosomal segments in reference families from the Centre d'Etude du Polymorphisme Humain. *American journal of human genetics*, 65:1493–1500, 1999.
10. D. Capello, M. Scandurra, G. Poretti, P.M.V. Rancoita, M. Mian, A. Gloghini, C. Deambrogi, M. Martini, D. Rossi, T.C. Greiner, W.C. Chan, M. Ponzoni, S. Montes Moreno, M.A. Piris, V. Canzonieri, M. Spina, U. Tirelli, G. Inghirami, A. Rinaldi, E. Zucca, R. Dalla Favera, F. Cavalli, L.M. Larocca, I. Kwee, A. Carbone, G. Gaidano, and F. Bertoni. Genome wide DNA-profiling of HIV-related B-cell lymphomas. *British Journal of Haematology*, 148(2):245–255, 2010.
11. M.P. Do Carmo. *Differential geometry of curves and surfaces*. Prentice-Hall, Englewood Cliffs, 1976.
12. B. Carvalho, H. Bengtsson, T. Speed, and R.A. Irizarry. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics*, 8:485–499, 2007.
13. S. Colella, C. Yau, J.M. Taylor, G. Mirza, H. Butler, P. Clouston, A.S. Bassett, A. Seller, C.C. Holmes, and J. Ragoussis. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research*, 35(6):2013–2025, 2007.
14. M. Corada, L. Zanetta, F. Orsenigo, F. Breviario, M.G. Lampugnani, S. Bernasconi, F. Liao, D.J. Hicklin, P. Bohlen, and E. Dejana. A monoclonal antibody to vascular endothelial-cadherin inhibits tumor angiogenesis without side effects on endothelial permeability. *Blood*, 100:905–911, 2002.
15. D.J. Daley and D. Vere-Jones. *An introduction to the theory of point processes*. Springer-Verlag, New York, 1988.
16. P. Dierckx. *Curve and Surface Fitting with Splines*. Oxford University Press Inc., New York, 1996.
17. A. Dutt and R. Beroukhim. Single nucleotides polymorphism array analysis of cancer. *Current Opinion in Oncology*, 19:43–49, 2007.
18. P.H.C. Eilers and R.X. de Menezes. Quantile smoothing of array CGH data. *Bioinformatics*, 21(7):1146–1153, 2005.
19. F. Forconi, G. Poretti, I. Kwee, E. Sozzi, D. Rossi, P.M.V. Rancoita, D. Capello, A. Rinaldi, E. Zucca, D. Donatella Raspadori, V. Spina, F. Lauria, G. Gaidano, and F. Bertoni. High density genome-wide DNA profiling reveals a remarkably stable profile in hairy cell leukaemia. *British Journal of Haematology*, 141:622–630, 2008.
20. F. Forconi, A. Rinaldi, I. Kwee, E. Sozzi, D. Raspadori, P.M.V. Rancoita, M. Scandurra, D. Rossi, C. Deambrogi, D. Capello, E. Zucca, D. Marconi, R. Bomben, V. Gattei, F. Lauria, G. Gaidano, and F. Bertoni. Genome-wide DNA analysis identifies recurrent imbalances predicting outcome in chronic lymphocytic leukaemia with 17p deletion. *British Journal of Haematology*, 143(4):532–536, 2008.

21. J. Fridlyand, A.M. Snijders, D. Pinkel, D.G. Albertson, and A.N. Jain. Hidden Markov Models approach to the analysis of array CGH data. *Journal of Multivariate Analysis*, 90:132–153, 2004.
22. L.P. Gondek, R. Tiu, C.L. O'Keefe, M.A. Sekeres, K.S. Theil, and J.P. Maciejewski. Chromosomal lesions and uniparental disomy detected by SNP arrays in MDS, MDS/MPD, and MDS-derived AML. *Blood*, 111(3):1534–1542, 2008.
23. G. Gonzalez, F. Fleuret, and P. Fua. Automated Delineation of Dendritic Networks in Noisy Image Stacks. In *Computer Vision ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV*. Springer-Verlag, LNCS 5305, 2008.
24. G. Gonzalez, F. Fleuret, and P. Fua. Learning rotational features for filament detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1582–1589, 2009.
25. D.L. Hartl and E.W. Jones. *Genetics: analysis of genes and genomes*. Jones and Bartlett Publishers, Sadbury, 2009.
26. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data mining, Inference and Prediction*. Springer-Verlag, New York, 2001.
27. G. Hodgson, J.H. Hager, S. Volik, S. Hariono, M. Wernick, D. Moore, N. Nowak, D.G. Albertson, D. Pinkel, C. Collins, D. Hanahan, and J.W. Gray. Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nature Genetics*, 29:459–464, 2001.
28. L. Hsu, S.G. Self, D. Grove, T. Randolph, K. Wang, J.J. Delrow, L. Loo, and P. Porter. Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, 6(2):211–226, 2005.
29. J. Huang, A. Gusnanto, K. O'Sullivan, J. Staaf, Å. Borg, and Y. Pawitan. Robust smooth segmentation approach for array CGH data analysis. *Bioinformatics*, 23(18):2463–2469, 2007.
30. J. Huang, W. Wei, J. Zhang, G. Liu, G.R. Bignell, M.R. Stratton, P.A. Futreal, R. Wooster, K.W. Jones, and M.H. Shapero. Whole Genome DNA Copy Number Changes Identified by High Density Oligonucleotide Arrays. *Human Genomics*, 1(4):287–299, 2004.
31. P. Hupé, N. Stransky, J.P. Thiery, F. Radvanyi, and E. Barillot. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, 20(18):3413–3422, 2004.
32. M. Hutter. Bayesian Regression of Piecewise Constant Functions. In J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. David, D. Heckerman, A.F.M. Smith, and M. West, editors, *Bayesian Statistics: Proceedings of the Eighth Valencia International Meeting*. Universitat de València and International Society for Bayesian Analysis, 2007.
33. M. Hutter. Exact Bayesian Regression of Piecewise Constant Functions. *Bayesian Analysis*, 2(4):635–664, 2007.

34. G. Ivanoff. Central limit theorems for point processes. *Stochastic Processes and their Applications*, 12:171–186, 1982.
35. L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990.
36. N. Kawamata, S. Ogawa, M. Zimmermann, B. Niebuhr, C. Stocking, M. Sanada, K. Hemminki, G. Yamatomo, Y. Nannya, R. Koeler, T. Flohr, C.W. Miller, J. Harbott, W.D. Ludwing, M. Stanulla, M. Shrappe, C.R. Bartram, and H.P. Koeffler. Cloning of genes involving in chromosomal translocations by high-resolution single nucleotide polymorphism genomic microarray. *Proceedings of the National Academy of Sciences*, 105(33):11921–11926, 2008.
37. G.C. Kennedy, H. Matsuzaki, S. Dong, W.M. Liu, J. Huang, G. Liu, X. Su, M. Cao, W. Chen, J. Zhang, W. Liu, G. Yang, X. Di, T. Ryder, Z. He, U. Surti, M.S. Phillips, M.T. Boyce-Jacino, S.P. Fodor, and K.W. Jones. Large-scale genotyping of complex DNA. *Nature biotechnology*, 21(10):1233–1237, 2003.
38. R.W. Koenker and G.W. Basset. Four (pathological) examples in asymptotic statistics. *The American statistician*, 38:209–212, 1984.
39. I.S. Kohane, A.T. Kho, and A.J. Butte. *Microarrays for an Integrative Genomics*. MIT Press, Cambridge, MA, 2002.
40. D. Kotzot. Complex and segmental uniparental disomy (UPD): review and lessons from rare chromosomal complements. *Journal of Medical Genetics*, 38:497–507, 2001.
41. W.R. Lai, M. Johnson, R. Kucherlapati, and P.J. Park. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, 21(19):3763–3770, 2005.
42. Y. Lai and H. Zhao. A statistical method to detect chromosomal regions with DNA copy number alterations using SNP-array-based CGH data. *Computational Biology and Chemistry*, 29:47–54, 2005.
43. L.H. Li, S.F. Ho, C.H. Chen, W.C. Wei, C.Y. Wong, L.Y. Li, S.I. Hung, W.H. Chung, W.H. Pan, M.T.M. Lee, F.J. Tsai, C.F. Chang, J.Y. Wu, and Y.T. Chen. Long Contiguous Stretches of Homozygosity in the Human Genome. *Human Mutation*, 27(11):1115–1121, 2006.
44. J.R. Lupski. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in Genetics*, 14(10):417–422, 1998.
45. J. Marioni, N. Thorne, and S. Tavare. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, 22:1144–1146, 2006.
46. A.G. de Nooij-van Dalen, V.H.A. van Buuren-van Seggelen, P.H.M. Lohman, and M. Giphart-Gassler. Chromosome Loss With Concomitant Duplication and Recombination Both Contribute Most to Loss of Heterozygosity In Vitro. *Genes, chromosomes and cancer*, 21:30–38, 1998.

47. The international HapMap Consortium. The International HapMap Project. *Nature*, 426:789–796, 2003.
48. The international HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:851–862, 2007.
49. D.L. McLeish. Dependent central limit theorems and invariance principles. *The Annals of Probability*, 2:620–628, 1974.
50. J. Mecke. Formulas for Stationary Planar Fibre Processes III - Intersections with Fibre Systems. *Mathematische Operationsforschung und Statistik, Ser. Statistics*, 12:201–210, 1981.
51. J. Mecke and D. Stoyan. Formulas for Stationary Planar Fibre Processes I - General Theory. *Mathematische Operationsforschung und Statistik, Ser. Statistics*, 11:267–279, 1980.
52. A. Micheletti and P.M.V. Rancoita. A central limit theorem for functionals of stationary and independent at distance *l* point processes. 2009. Submitted.
53. A. Micheletti and P.M.V. Rancoita. Estimators of the intensity of stationary planar fibre processes. In V. Capasso, G. Aletti, and A. Micheletti, editors, *Stereology and Image Analysis. Ecs10: Proceeding of the 10th European Conference of ISS., The MIRIAM Project Series, Vol. 4*, pages 131–136, Bologna, 2009. ESCULAPIO Pub. Co.
54. S.K. Murthy, L.M. DiFrancesco, R. Travis Ogilvie, and D.J. Demetrick. Loss of Heterozygosity Associated with Uniparental Disomy in Breast Carcinoma. *Modern Pathology*, 15(12):1241–1250, 2002.
55. K. Nakao, K.R. Mehta, J. Fridlyand, D.H. Moore, A.N. Jain, A. Lafuente, J.W. Wiencke, J.P. Terdiman, and F.M. Waldman. High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carginogenesis*, 25(8):1345–1357, 2004.
56. M.A. Newton and Y. Lee. Inferring the Location and Effect of Tumor Suppressor Genes by Instability-Selection Modelling of Allelic-Loss Data. *Biometrics*, 56:1088–1097, 2000.
57. B. Nilsson, M. Johansson, A. Heyden, S. Nelander, and T. Fioretos. An improved method for detecting and delineating genomic regions with altered gene expression in cancer. *Genome Biology*, 9(1), 2008.
58. J. Ohser. A Remark on the Estimation of the Rose Directions of Fibre Processes. *Mathematische Operationsforschung und Statistik, Ser. Statistics*, 12:581–585, 1981.
59. A.B. Olshen, E.S. Venkatraman, R. Lucito, and M. Wigler. Circular Binary Segmentation for the Analysis of Array-based DNA Copy Number Data. *Biostatistics*, 5(4):557–572, 2004.
60. S. Pelengaris and M. Khan. *The molecular biology of cancer*. Wiley-Blackwell, 2006.
61. M.D. Penrose and J.E. Yukich. Central limit theorems for some graphs in computational geometry. *The Annals of Applied Probability*, 11:1005–1041, 2001.

62. K. Petersen. *Ergodic theory*. Cambridge University Press, Cambridge, 1983.
63. F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J.J. Daudin. A Statistical Approach for Array CGH Data Analysis. *BMC Bioinformatics*, 6(27), 2005.
64. P.M.V. Rancoita. Statistics of fibre processes. Applications to angiogenesis. Master's thesis, Università degli Studi di Milano, 2006. In italian.
65. P.M.V. Rancoita, M. Hutter, F. Bertoni, and I. Kwee. Bayesian DNA copy number analysis. *BMC Bioinformatics*, 10(10), 2009.
66. P.M.V. Rancoita, M. Hutter, F. Bertoni, and I. Kwee. Bayesian joint estimation of CN and LOH aberrations. In S. Omatu et al., editor, *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living, 10th International Work-Conference on Artificial Neural Networks, IWANN 2009 Workshops, Salamanca, Spain, June 10-12, 2009. Proceedings, Part II*, pages 1109–1117. Springer-Verlag, LNCS 5518, 2009.
67. P.M.V. Rancoita, M. Hutter, F. Bertoni, and I. Kwee. An integrated bayesian analysis of LOH and copy number data. 2009. Submitted.
68. A. Rinaldi, D. Capello, M. Scandurra, T.C. Greiner, W.C. Chan, D. Rossi, M. Bhagat, G. Paulli, P.M.V. Rancoita, G. Inghirami, M. Ponzoni, S. Montes Moreno, M.A. Piris, M. Mian, E. Chigrinova, E. Zucca, R. Dalla Favera, G. Gaidano, I. Kwee, and F. Bertoni. Single nucleotide polymorphism-arrays provide new insights in the pathogenesis of post-transplant diffuse large B-cell lymphoma (PT-DLBCL). *British Journal of Haematology*, 2010. Articles online in advance of print.
69. A. Rinaldi, I. Kwee, M. Taborelli, C. Largo, S. Uccella, V. Martin, G. Poretti, G. Gaidano, G. Calabrese, G. Martinelli, L. Baldini, G. Pruneri, C. Capella, E. Zucca, F.E. Cotter, J.C. Cigudosa, C.V. Catapano, M.G. Tibiletti, and F. Bertoni. Genomic and expression profiling identifies the B-cell associated tyrosine kinase Syk as a possible therapeutic target in mantle cell lymphoma. *British Journal of Haematology*, 132:303–316, 2006.
70. D. Rossi, M. Cerri, D. Capello, C. Deambrogi, F.M. Rossi, A. Zucchetto, L. De Paoli, S. Cresta, S. Rasi, V. Spina, S. Franceschetti, M. Lunghi, C. Vendramin, R. Bomben, A. Ramponi, G. Monga, A. Conconi, C. Magnani, V. Gattei, and G. Gaidano. Biological and clinical risk factors of chronic lymphocytic leukaemia transformation to Richter syndrome. *British Journal of Haematology*, 142:202–215, 2008.
71. M. Scandurra, D. Rossi, C. Deambrogi, P.M.V. Rancoita, E. Chigrinova, M. Mian, M. Cerri, S. Rasi, F. Sozzi, E. Forconi, M. Ponzoni, S. Montes-Moreno, M.A. Piris, G. Inghirami, E. Zucca, V. Gattei, A. Rinaldi, I. Kwee, Gaidano G., and F. Bertoni. Genomic profiling of Richter's syndrome: recurrent lesions and differences with *de novo* diffuse large B-cell lymphomas. *Hematological Oncology*, 2009. Articles online in advance of print.

72. R.B. Scharpf, G. Parmigiani, J. Pevsner, and I. Ruczinski. Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. *Annals of Applied Statistics*, 2(2):687–713, 2008.
73. A. Schwandtke. Second-Order Quantities for Stationary Weighted Fibre Processes. *Mathematische Nachrichten*, 139:321–334, 1988.
74. A. Sen and M.S. Srivastava. On test for detecting a change in mean. *Annals of Statistics*, 3:98–108, 1975.
75. R.M. Simon, E.L. Korn, L.M. McShane, M.D. Radmacher, G.W. Wright, and Y. Zhao. *Design and analysis of DNA microarray investigations*. Springer, New York, 2004.
76. P. Soille. *Morphological Image Analysis. Principles and Applications*. Springer-Verlag, Berlin, Heidelberg, New York, 1999.
77. D. Stoyan. On the Second-Order Analysis for Stationary planar fibre processes. *Mathematische Nachrichten*, 102:189–199, 1981.
78. D. Stoyan, W.S. Kendall, and J. Mecke. *Stochastic Geometry and its Applications. Second Edition*. John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore, 1995.
79. J.A. Veltman, J. Fridlyand, S. Pejavar, A.B. Olshen, J.E. Korkola, S. DeVries, P. Carroll, W.L. Kuo, D. Pinkel, D. Albertson, C. Cordon-Cardo, A.N. Jain, and F.M. Waldman. Array-based Comparative Genomic Hybridization for Genome-Wide Screening of DNA Copy Number in Bladder Tumors. *Cancer Research*, 63:2872–2880, 2003.
80. B.A. Walker and G.J. Morgan. Use of single nucleotide polymorphism-based mapping arrays to detect copy number changes and loss of heterozygosity in multiple myeloma. *Clinical Lymphoma & Myeloma*, 7:186–191, 2006.
81. K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S.F.A. Grant, H. Hakonarson, and M. Bucan. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, 17(11):1665–1674, 2007.
82. R.A. Weinberg. *The biology of cancer*. Garland science, New York, 2007.
83. E. Whittaker. On a New Method of Graduation. In *Proceedings of the Edinburgh Mathematical Society*, volume 41, pages 63–75. Cambridge University Press, 1923.
84. H. Willenbrock and J. Fridlyand. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, 21(7):4084–4091, 2005.
85. L.Y. Wu, X. Zhou, F. Li, X. Yang, and C.C. Chang. Conditional random pattern algorithm for LOH inference and segmentation. *Bioinformatics*, 25:61–67, 2009.
86. M. Zähle. Random processes of Hausdorff rectifiable closed sets. *Mathematische Nachrichten*, 108:49–72, 1982.
87. X. Zhao, C. Li, J.G. Paez, K. Chin, P.A. Jänne, T.H. Chen, L. Girard, J. Minna, D. Christiani, C. Leo, J.W. Gray, W.R. Sellers, and M. Meyerson. An Integrated View of Copy Number and Allelic Alterations in the Cancer Genome Using Single Nucleotide Polymorphism Arrays. *Cancer Research*, 64:3060–3071, 2004.