

Salamanca Gto. a 13 de Septiembre del 2018

M. en I. HERIBERTO GUTIÉRREZ MARTÍN  
JEFE DE LA UNIDAD DE ADMINISTRACIÓN ESCOLAR  
PRESENTE.-

Por medio de la presente, se otorga autorización para proceder a los trámites de impresión, empastado de tesis y titulación al alumno(a) JORGE ULISES MUÑOZ MINJARES del Programa de Doctorado en INGENIERÍA ELÉCTRICA y cuyo número de NUA es 144196 del cual soy director. El título de la tesis es: Improving Estimates of Genome CNAs by developing Probabilistic Masks for Microarray Data


Hago constar que he revisado dicho trabajo y he tenido comunicación con los síndicales asignados para la revisión de la tesis, por lo que no hay impedimento alguno para fijar la fecha de examen de titulación.

ATENTAMENTE


  
Dr. Yuriy Semenovich Shmaliy  
DIRECTOR DE TESIS  
SECRETARIO

  
Dr. Yuriy Semenovich Shmaliy  
DIRECTOR DE TESIS

  
Dr. Oscar Gerardo Ibarra Manzano  
PRESIDENTE

  
Dr. Israel Alejandro Arriaga Trejo  
VOCAL

  
Dr. Gustavo Cerda Villafaña  
VOCAL

  
Dr. Luis Javier Morales Mendoza  
VOCAL



**UNIVERSIDAD DE GUANAJUATO**

---

---

**CAMPUS IRAPUATO – SALAMANCA  
DIVISIÓN DE INGENIERÍAS**

**Improving Estimates of Genome CNAs by  
developing Probabilistic Masks for  
Microarray Data**

**TESIS DOCTORAL**

**QUE PARA OBTENER EL GRADO:  
DOCTOR EN INGENIERÍA ELÉCTRICA**

**PRESENTA:**

**M.I. JORGE ULISES MUÑOZ–MINJARES**

**DIRECTORES:**

**DR. YURIY S. SHMALIY**

**SALAMANCA, GUANAJUATO**

**JUNE, 2018**



# Abstract

Copy Number Alterations (CNA)s are hallmarks of cancer, which are gains or losses in copies of Deoxyribonucleic Acid (DNA) sections. Nowadays, CNAs are routinely measured by different techniques for diagnostic and prognostic purposes. The array-Comparative Genomic Hybridization (aCGH), Array-Single Nucleotide Polymorphism (aSNP) and Next Generation Sequencing (NGS) are examples of technologies that enable cost-efficient high resolution detection of CNAs.

Intensive noise as well as technical and biological biases inherent to modern technologies of CNAs probing often cause inconsistency between the estimates provided by different methods. Efficient and accurate detection of the breakpoint positions in heterogeneous cancer samples measured under such conditions is a challenging practical and methodological problem. Despite the necessity of accurate CNA estimates, there is no much information regarding the estimation errors..

Based on studies of the confidence limits for noisy stepwise signals, an efficient algorithm has been developed for computing the upper and lower confidence boundary masks with a specific probability, in order to guarantee an existence of genomic changes within certain regions. This tool combined with estimates can give more information to medical experts about true CNAs structures.

The probabilistic confidence masks are initially designed based on the Skew Laplace distribution to represent jitter in the CNA breakpoints. Using experimental measurements, it is concluded that Laplace distribution is accurate when the segmental Signal-to-Noise Ratio (SNR) exceeds unity. In this work the experimental jitter distribution is simulated to different ranges in order to find approximations to actual distributions with minimal errors.

Following this procedure, three techniques are described to approximate the experimental jitter distribution: Heuristic approximation, parametrization of skew Laplace distribution, and asymmetric exponential power distribution. The confidence masks algorithm is designed and modified for each approximation. It is also tested by arrays: High-Resolution Comparative Genomic Hybridization and Single Nucleotide Polymorphism data.

Additionally, the confidence masks based on the exponential power distribution are tuned to the medical expert annotations of the training set of the breakpoints obtained by the standard circular binary segmentation algorithm. A comparison of modified confidence masks and experts annotations related to CNA profiles of neuroblastoma demonstrates an efficiency of the designed masks to improve the CNA estimates.

# Acknowledges

Foremost, I would like to express my sincere gratitude to my advisor Prof. Yuriy S. Shmaliy for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Oscar G. Ibarra–Manzano, Prof. Israel A. Arriaga Trejo, Gustavo Cerda–Villafaa and Prof. Luis J. Morales–Mendoza, for their encouragement, insightful comments, and hard questions.

My sincere thanks also goes to Dr. Tatiana Popova for offering me the summer stay opportunities in her group and leading me working on diverse exciting tasks.

To my family, particularly my parents Jorge Muoz and Angelica Minjares, and brothers Freddy, Karen, Erika and Marcos, thank you for your love, support, and unwavering belief in me. Without you, I would not be the person I am today.

Above all I would like to thank my wife Janette Perez for her love and constant support, for all the late nights and early mornings, and for keeping me sane over the past few months. But most of all, thank you for being my best friend. I owe you everything. Also, I want to thank to my children Isaac and Darleth, for giving me courage, bravery and strength throughout my everyday life.

Finally, the work reported in this thesis would not have been possible without the financial support of an Conacyt studentship (No. 254890/CVU 388890), for which I am grateful.

# Contents

<b>1</b>	<b>Introduction</b>	<b>18</b>
1.1	Background . . . . .	18
1.2	Motivation . . . . .	21
1.3	Research Objectives . . . . .	22
1.4	Scope of this work . . . . .	22
<b>2</b>	<b>Foundations</b>	<b>2</b>
2.1	DNA and Cancer . . . . .	2
2.2	DNA Microarray . . . . .	3
2.2.1	SNP microarray . . . . .	4
2.2.2	CGH microarray . . . . .	4
2.2.3	Next Generation Sequencing . . . . .	4
2.3	Algorithms for estimate CNAs . . . . .	6
2.3.1	Median Breakpoints detector . . . . .	6
2.3.2	Circular Binary Segmentation . . . . .	8
2.3.3	Pruned Exact Linear Time . . . . .	8
2.3.4	Binary Segmentation . . . . .	9
2.3.5	Segment Neighborhood . . . . .	10
<b>3</b>	<b>Jitter Distribution</b>	<b>11</b>
3.1	Jitter Distribution in Breakpoints . . . . .	11
3.1.1	Probabilities of events <b>A</b> and <b>B</b> . . . . .	15

---

3.1.2	Normalization . . . . .	19
3.1.3	Distribution verification by simulation . . . . .	21
3.2	Confidence UB and LB Masks . . . . .	23
3.2.1	Testing real measurements by the probabilistic confidence masks. . . . .	28
3.3	Limitation of Laplace-based Approximation . . . . .	29
<b>4</b>	<b>Improving Jitter</b>	<b>31</b>
4.1	Experimental Jitter Histogram . . . . .	31
4.2	Approximations of jitter pdf . . . . .	35
4.2.1	Heuristic Approximation . . . . .	37
4.3	Parametrization of Laplace Density . . . . .	40
4.4	Asymmetric Exponential Power Distribution . . . . .	46
4.4.1	Parameters Estimation for AEP distribution . . . . .	47
4.5	Comparison of Proposed Approximations . . . . .	51
<b>5</b>	<b>Modified Confidence Masks</b>	<b>53</b>
5.1	Confidence Masks for Hybrid approximation . . . . .	53
5.1.1	Testing Estimates by $\mathcal{B}_{l H}^{\text{UB}}$ and $\mathcal{B}_{l H}^{\text{LB}}$ Masks . . . . .	54
5.1.2	Improving CNAs Estimates . . . . .	57
5.2	Confidence Masks based on Laplace-parametrization . . . . .	59
5.2.1	Hybrid confidence masks . . . . .	59
5.2.2	Applications to SNP Array Probing . . . . .	60
5.3	Confidence Masks based on AEP distribution . . . . .	63
<b>6</b>	<b>Matching Expert’s Annotations</b>	<b>66</b>
6.1	Breakpoints Annotations as Gold Standard . . . . .	66
6.2	Match Cases . . . . .	68
6.2.1	Case 1–Perfect Match . . . . .	68
6.2.2	Case 2–Good Match . . . . .	68
6.2.3	Case 3–Wrong Match . . . . .	69
6.2.4	Case 3.1–Transitional Match . . . . .	71



---

<b>7 Algorithms comparizon using Confidence Masks</b>	<b>75</b>
7.1 Comparison of breakpoints estimators . . . . .	75
7.1.1 CNAs Size Analysis . . . . .	83
<b>8 Conclusions</b>	<b>87</b>
8.1 About Heuristic Approximation . . . . .	87
8.2 About Laplace–Parametrization . . . . .	88
8.3 About AEP approximation . . . . .	88
8.4 About Matching Expert’s Annotations . . . . .	89
8.5 About Comparative of algorithms using Confidence Masks . . . . .	89
<b>Bibliography</b>	<b>91</b>
<b>Appendices</b>	<b>102</b>
<b>Appendix A</b>	<b>103</b>
A.1 Analysis of Gaussian Process . . . . .	103
<b>Appendix B</b>	<b>108</b>
B.1 Skew Laplace Distribution . . . . .	108
<b>Appendix C</b>	<b>112</b>
C.1 Computational Algorithm . . . . .	112
<b>Appendix D</b>	<b>115</b>
D.1 Comparison of Approximations . . . . .	115
<b>Appendix E</b>	<b>117</b>

# List of Figures

- 1.1 Jitter caused by intensive noise in a typical representation of CNA (Gain and Loss), plotted Log2Ratio respect to probes  $n$ . . . . . 20
- 2.1 Structure of deoxyribonucleic Acid (DNA). . . . . 3
- 2.2 Chromosomes 1 – 10 represented in *Log2Ratio* from pancreatic adenocarcinoma genome. . . . . 5
- 2.3 Median-based denoising of the microarray measurement ... . . . . 7
- 2.4 Example of breakpoint estimated using the Pruned Exact Linear Time (PELT) method. . . . . 9
- 2.5 Schematic of Binary Segmentation (BINSEG) algorithm. . . . . 10
- 3.1 Jitter distributions computed with Maximum Likelihood and Skew Laplace distribution to a) SNR=0.1 and b) SNR =0.5. . . . . 13
- 3.2 Simulated CNAs with one breakpoint located at  $n = 50$  and standard deviations  $\sigma_l$  and  $\sigma_{l+1}$  of each segment... . . . . 14
- 3.3 Cases respect to values of standart deviation to compute the events  $A_i$  and  $B_i$ . . . . . 16
- 3.4 The discrete skew Laplace pdf (dashed) for different segmental SNRs;  $k = 0$  corresponds to  $i_l$ . . . . . 22
- 3.5 Total jitter probability  $P_J(\gamma)$  and probabilities  $P_k(\gamma)$  of the right jitter at  $k = 1$ ,  $k = 2$ , and  $k = 3$  for equal SNRs in the CNAs segments. . . . . 24
- 3.6 An example of UB mask  $\mathcal{B}_n^U$  and LB mask  $\mathcal{B}_n^L$  around the simulated CNA . 28
- 3.7 Median-based denoising of the microarray measurement... . . . . 30

4.1	Procedure to approximate the jitter distribution in the CNA breakpoints by simulating a stepwise signal in the presence of AWGN with different segmental SNRs. The breakpoint $i_l$ change its position from $\hat{i}_0$ to $\hat{i}_{200}$ seeking the best CNA estimate. . . . .	32
4.2	Jitter distributions computed with Maximum Likelihood and Skew Laplace distribution to a) SNR=0.1 and b) SNR =0.5. . . . .	33
4.3	A flowchart to approximate the jitter distribution in the CNA breakpoints by simulating a stepwise signal in the presence of WGN with different segmental SNRs... . . . .	35
4.4	Experimentally defined one-sided jitter probability densities (dotted) of the breakpoint location for equal segmental SNRs $\gamma$ in the range of $M = 400$ points with a true breakpoint at $n = 200$ . . . . .	36
4.5	Difference between experimental distributions obtained using a format of 4 (solid) and 9 (circles) decimal values. . . . .	36
4.6	Modified Bessel functions of the second kind, $K_\alpha(x)$ , for $\alpha = 0$ (solid), 1(dashed), 2 (dotted), 3 (dash-dotted) and 4 (solid-pointed). . . . .	37
4.7	Experimentally defined one-sided jitter probability densities (dotted) of the breakpoint location for equal segmental SNRs in the range of $M = 400$ points with a breakpoint at $n = 200$ . . . . .	38
4.8	Coefficients for the approximation functions: (a) $\alpha(\gamma)$ , (b) $\beta(\gamma)$ , and (c) $\epsilon(\gamma)$	41
4.9	Representation of a standard deviation constant $\sigma_l$ and the proposed $k$ -varying standard deviation function $\sigma_l(k)$ . . . . .	42
4.10	The proposed $k$ -varying variance functions $\sigma_l^2(k)$ used to parameterize the SkL pdf (3.39) . . . . .	44
4.11	Measured jitter pdf functions (circles) and the approximations by the SkL law (3.39) and by the SkL law parameterized . . . . .	45
4.12	Asymmetric Exponential Power Distribution to several parameters of shape $\alpha$ , for the symmetric case skew is set as $\kappa = 1$ , scale $\sigma = 1$ and zero mean. . . . .	47

---

4.13	The Kolmogorov-Smirnov distance between the empirical distribution function of the sample $S_N(x)$ and the cumulative distribution function of the reference distribution $F_0(x)$ . . . . .	48
4.14	Approximations of the (a) shape factor $\alpha_l(\gamma_l^\pm)$ and (b) scale factor $\sigma_l(\gamma_l^\pm)$ for jitter distribution using Asymmetric Exponential Power distribution. Measured data are dotted to a range of SNR from $\gamma^\pm = 0.1$ to $\gamma^\pm = 2$ and the curves to fit them are represented with a dashed line for both variables. . . . .	49
4.15	The approximations of the experimentally measured jitter pdf (dotted) for different SNR values using the AEP distribution (solid) for $\gamma_l^- = 0.3$ : (a) $\gamma_l^+ = 0.3$ and (b) $\gamma_l^+ = 0.8$ . . . . .	50
4.16	Error of ML estimator and proposed approximations. a) Minimum errors obtained of each approximations, b) errors of all approximations respect to the experimental jitter obtained with the detailed algorithm. . . . .	52
5.1	Upper boundaries $\mathcal{B}_l^{UB}$ , $\mathcal{B}_{l H}^{UB}$ and lower boundaries $\mathcal{B}_l^{LB}$ and $\mathcal{B}_{l H}^{LB}$ for the breakpoint $i_2$ of Chromosome 1 from database BLC_B1_T45.txt given $\vartheta = 3$ . Confidence bounds $\mathcal{B}_{l H}^{UB}$ and $\mathcal{B}_{l H}^{LB}$ (dash-dot) are based on the heuristic approximations and $\mathcal{B}_{l H}^{UB}$ and $\mathcal{B}_l^{UB}$ and $\mathcal{B}_l^{LB}$ and $\mathcal{B}_{l H}^{LB}$ (dotted) use the SkL distribution. . . . .	55
5.2	The $\mathcal{B}_{l H}^{UB}$ and $\mathcal{B}_{l H}^{LB}$ masks placed a) around the segmental level $a_{18}$ for several confidence probabilities. . . . .	56
5.3	Improving estimates of the CNAs obtained in Project GAP by removing some unlikely existing breakpoints: (a) original estimates, (b) even changes, $P = 50\%$ , (c) probable changes, $P = 75\%$ , (d) almost certain changes, $P = 93\%$ , and (e) 3-sigma sense, $P = 99.73\%$ . . . . .	58
5.4	Testing the ML estimate of the breakpoint $i_5$ location of the sample BLC_B1_T31 in the 13th Chromosome by the confidence masks. . . . .	61
5.5	Probes (points), CNAs estimates (solid), and confidence regions (dashed) provided by the hybrid masks for the 13th chromosome taken from <i>BLC_B1_T37</i> of GAP. . . . .	62

5.6	a) Chromosome 10 and b) Chromosome 19 of neuroblastoma copy number profile 207 with masks $\mathcal{B}_{l \alpha E}^{UB}$ and $\mathcal{B}_{l \alpha E}^{LB}$ applied to the CNA estimates (bold)	65
6.1	Annotations made by medical experts to Profile 44–Chromosome 1 of sample of neuroblastoma. . . . .	67
6.2	Chromosome 10 of neuroblastoma copy number profile 207 with masks $\mathcal{B}_{l \alpha E}^{UB}$ and $\mathcal{B}_{l \alpha E}^{LB}$ applied to the CNA estimates (bold) and expert’s annotations (striped)... . . . .	69
6.3	Estimate (bold), masks $\mathcal{B}_{l \alpha E}^{UB}$ and $\mathcal{B}_{l \alpha E}^{LB}$ , and expert’s annotations (striped) for neuroblastoma copy number profile 207... . . . .	70
6.4	The $\mathcal{B}_{l \alpha E}^{UB}$ and $\mathcal{B}_{l \alpha E}^{LB}$ masks around the chromosome 2 and 17 for neuroblastoma copy number profile 207. . . . .	71
6.5	The $\mathcal{B}_{l \alpha E}^{UB}$ and $\mathcal{B}_{l \alpha E}^{LB}$ masks around the chromosome 2 and 11 for neuroblastoma copy number profile 22 and 522 respectively. . . . .	72
7.1	Procedure diagram to compare the breakpoints estimator methods respect to an ideal estimation obtained using the Next Generation Sequencing technology. The modified probabilistic masks remove the breakpoints that unsatisfied a given probability, which is specified with the $\vartheta$ –sense. . . . .	76
7.2	Estimated CNAs–annotated as Normal, Gains or Loss– to measurements of Chromosome 4 from the Sample_1, which is plotted Genomic Position versus Log2 Ratio. The total of breakpoints located by the NGS (triangle down) is limited to segments greater than 1 Mega bases (plus sign). The points estimated by the CBS method (circles) can be removed using the modified confidence masks (cross) in a range of $\vartheta$ from 0.6745 to 20. . . .	79
7.3	Estimated CNAs–annotated as Normal, Gains or Loss– to measurements of Chromosome 8 from the Sample_1, which is plotted Genomic Position versus Log2 Ratio. The total of breakpoints located by the NGS (triangle down) is limited to segments greater than 1 Mega bases (plus sign). The points estimated by the CBS method (circles) can be removed using the modified confidence masks (cross) in a range of $\vartheta$ from 0.6745 to 20. . . .	80

---

7.4	Comparative of breakpoints estimated with Circular Binary Segmentation and Next Generation Sequencing algorithms employing a Venn diagram. Based on the suggestion of confidence masks $\mathcal{B}_{l \alpha E}^{UB}$ and $\mathcal{B}_{l \alpha E}^{LB}$ the breakpoints estimated by CBS method can be refined at a given probability, parameter $\vartheta$ .	81
7.5	Comparative of breakpoints estimated with Circular Binary Segmentation and Next Generation Sequencing algorithms employing a Venn diagram. Based on the suggestion of confidence masks $\mathcal{B}_{l \alpha E}^{UB}$ and $\mathcal{B}_{l \alpha E}^{LB}$ the breakpoints estimated by CBS method can be refined at a given probability, parameter $\vartheta$ .	82
7.6	Estimated breakpoints of CNAs using the methods a) CBS (circle) and b) PELT (square) and deleted change points at specified probability. The first value of both curves is the initial number of estimated breakpoints. . . . .	83
7.7	True positive rate against the false positive rate based on the results of comparison between a) CBS and b) PELT with NGS estimates at three thresholds: 1 (triangle), 2 (square), and 3 (circle) Mega base pairs for a range $\vartheta$ from 0.6745 to 20. . . . .	84
7.8	Size analysis of CNAs estimated by a) CBS and b) PELT tested with the probabilistic confidence masks $\mathcal{B}_{l \alpha E}^{UB}$ and $\mathcal{B}_{l \alpha E}^{LB}$ . . . . .	85

# List of Tables

- 3.1 Probabilistic measures for genomic changes . . . . . 27
- 4.1 MSEs produced by Laplace-based (3.39) and Bessel-based (4.3) approximations. . . . . 40
- 5.1 SNR regions for MBA, Laplace pdf (3.39), and (3.39) parameterized with (4.9), (4.10), and (4.11) to detect the right jitter  $k^-$  and the left jitter  $k^+$  with the minimum MSE. . . . . 59
- 6.1 Comparison between several profiles annotations with CNAs estimated by CBS and tested by confidence masks: CASE-I Excellent match and CASE-II Poor match. . . . . 73
- 7.1 Possible cases of comparison between a particular breakpoints detector – CBS or PELT– and Next Generation Sequencing estimation, represented with the symbols  $\bigcirc$  and  $\nabla$ , respectively. The case of True Negatives is given when a False Positive is removed using the confidence masks, it is illustrated with the symbol  $\otimes$ . . . . . 77
- C.1 Algorithm for computing the UB mask  $\mathcal{B}_n^U$  and LB mask  $\mathcal{B}_n^L$  via SNP array CNVs measurements  $y_n$  and the breakpoint locations estimates  $\hat{n}_l$ . Given: bound wideness ( $\vartheta$ -sigma). . . . . 112
- C.2 Algorithm for computing the KR jitter  $k_l^R$  and KL jitter  $k_l^L$ . Given:  $\hat{a}_j$ ,  $\hat{\sigma}_j$  and number  $L$  of breakpoints. . . . . 114

---

D.1	Typical MSEs produced by all the approximations proposed for different values of Signal to Noise Ratio $\gamma = \gamma_l^- = \gamma_l^+$ . . . . .	115
D.2	Typical MSEs produced by all the approximations proposed for different values of Signal to Noise Ratio $\gamma = \gamma_l^- = \gamma_l^+$ . . . . .	116
E.1	Part I. Left jitter $k_l^-$ and right jitter $k_l^+$ detected by different masks in the CNA breakpoints of the 13th chromosomal sample “BLC_ B1_ T37.txt” in the $3\sigma$ sense with the confidence probability of $P = 99.73\%$ . The chromosome is associated with breast cancer and all values are given for $\text{Log}_2$ Ratio. Here symbol “-” means that the jitter cannot be calculated by the masks. . . . .	118
E.2	Part II. Left jitter $k_l^-$ and right jitter $k_l^+$ detected by different masks in the CNA breakpoints of the 13th chromosomal sample “BLC_ B1_ T37.txt” in the $3\sigma$ sense with the confidence probability of $P = 99.73\%$ . The chromosome is associated with breast cancer and all values are given for $\text{Log}_2$ Ratio. Here symbol “-” means that the jitter cannot be calculated by the masks. . . . .	119
E.3	Part III. Left jitter $k_l^-$ and right jitter $k_l^+$ detected by different masks in the CNA breakpoints of the 13th chromosomal sample “BLC_ B1_ T37.txt” in the $3\sigma$ sense with the confidence probability of $P = 99.73\%$ . The chromosome is associated with breast cancer and all values are given for $\text{Log}_2$ Ratio. Here symbol “-” means that the jitter cannot be calculated by the masks. . . . .	120



# Acronyms

**CNA** Copy Number Alterations

**DNA** Deoxyribonucleic Acid

**aCGH** array-Comparative Genomic Hybridization

**aSNP** Array-Single Nucleotide Polymorphism

**NGS** Next Generation Sequencing

**SNR** Signal-to-Noise Ratio

**HR-CGH** High Resolution CGH

**WGS** Whole Genome Sequencing

**SkL** skew discrete Laplace

**CNVs** Copy Number Variations

**UB** Upper Bound

**LB** Lower Bound

**LRR** Log R ratios

**PCR** Polymerase Chain Reaction

**bp** base pairs

---

<b>CBS</b>	Circular Binary Segmentation
<b>BS</b>	Binary Segmentation
<b>PELT</b>	Pruned Exact Linear Time
<b>OP</b>	Optimal Partition
<b>BINSEG</b>	Binary Segmentation
<b>SEGNEIGH</b>	Segment Neighborhood
<b>ML</b>	Maximum Likelihood
<b>WGN</b>	Additive White Gaussian Noise
<b>pdf</b>	probability density function
<b>PWC</b>	piecewise constant
<b>LJB</b>	Left Jitter Bound $J_l^L$
<b>RJB</b>	Right Jitter Bound $J_l^R$
<b>UB mask</b>	Upper Bound mask $\mathcal{B}_n^U$
<b>LB mask</b>	Lower Bound mask $\mathcal{B}_n^L$
<b>OLS</b>	Ordinary Least Squares
<b>MATLAB</b>	MATrix LABoratory
<b>GHz</b>	GigaHertz
<b>RAM</b>	Random Access Memory
<b>MSE</b>	Mean Square Error
<b>MBA</b>	Modified Bessel-based approximation

**AEP** Asymmetric Exponential Power

**KS** Kolmogorov–Smirnov

**GAP** Genome Alteration Print

**BLC** basal–like carcinomas

**cghcbs** Comparative Genomic Hybridization–Circular Binary Segmentation

**TP** True Positive

**FP** False Positive

**FN** False Negative

**TN** True Negative

**ROC** Receiver Operating Characteristic

**TPR** True Positive Rate

**FPR** False Positive Rate

# Chapter 1

## Introduction

### 1.1 Background

Copy Number Alterations (CNAs) result from Somatic aberrations in DNA which represent an important class of structural variation of the sequence of the genetic chain DNA across diverse cancer types [1, 2, 3]. Cancer is well known as a disease of the genome and genomic aberrations of interest are mostly somatic aberrations [4, 5].

During the last decades, several technologies have been developed to measure the genome chromosomal structure such as, aCGH [6], High Resolution CGH (HR-CGH) [7], and Whole Genome Sequencing (WGS) [8] are among the most common.

Recently, the modern NGS-based technologies have provided a high resolution in sequencing, this quality produces that an estimator discovers more subtle chromosomal effects than before causing other problems [9]. Additionally, the NGS-based technologies are generally considered to be more difficult in use and still more expensive with respect to the aCGH microarrays. The aCGH microarrays have been developed as genome-wide assays for measurements of CNAs, using the fact that microarray fluorescence intensity is proportional to DNA copy number [10].

Nevertheless, due to complexity of cancer, cause that the microarrays are mostly contaminated by normal cells and that causes noise in CNAs measurements [11]. In the presence of the technological biases (quality of material and hybridization/sequencing) and the in-

tensive random noise [12, 13], there are increasing difficulties to detect the breakpoints in the CNAs. Consequently, under the intense noise, no estimator, optimal or robust, is able to produce a sufficiently accurate estimate [14].

The problem is complicated by an incapacity of generating multiple probing in short time [15] and thus to enhance the estimates statistically. No estimator can guarantee that detected changes exist with high probability, because single probing does not provide sufficient information. Thus, accurate identification of CNAs remains a challenging problem. Accordingly, as shown in [16], some small CNAs tested by the confidence masks may be identified with low probability and some others not detected.

The aCGH is one of the most modern techniques employing chromosomal microarray analysis to detect the CNAs at a resolution level of 5–10 kbp (kilo base pairs) [17]. In practice, the interpretation made by an expert biologist looking for CNAs can be provided easier if the normalized aCGH measurements are plotted against genomic position [18]. The CNAs data are represented in genomic position with the  $n$ th probe,  $n_l \in [1, M]$ , where  $M$  is the number of probes. In the CNAs picture, the  $n_l$ th discrete point corresponds to the  $l$ th edge or *breakpoint*. In microarray technique, the CNAs are often normalized and plotted as  $\log_2 R/G = \log_2 \text{Ratio}$ , where R and G are the fluorescent Red and Green intensities, respectively [19]. The CNAs levels are estimated by simple averaging between the breakpoints which reduces the variance of the segmental noise.

Uncertainty in the breakpoint location caused by the intensive noise is called “jitter”. Such phenomenon denotes a deviation from the true breakpoint location [20]. Figure 1.1a shows the possible estimates of breakpoints (Jittered signals) caused by noise embedded in the measurements of CNAs (the gain or loss). There have been developed a number of methods to refine the breakpoints, such as [21, 22, 23] and new mathematical models proposed to provide noise removal (denoising) while preserving edges from these microarray assays [24, 25, 26, 27, 28, 29]. Even so, detection of the breakpoint locations often becomes unavailable due to low segmental Signal to Noise Ratio (SNR). The values of segmental SNRs  $\gamma_l^-$  and  $\gamma_l^+$  are computed with the left and right measurements respect to the breakpoint.

In [14, 16, 30] we have shown that jitter in the CNAs breakpoints is distributed with

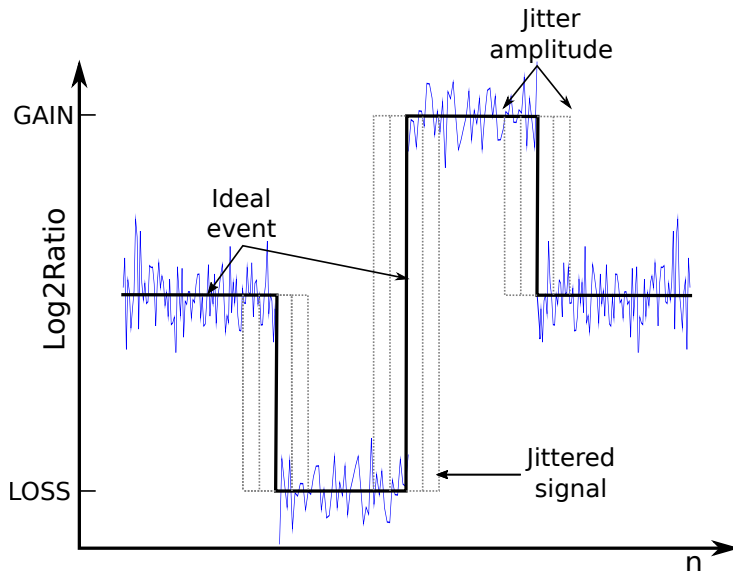


Figure 1.1: Jitter caused by intensive noise in a typical representation of Copy Number Alterations (Gain and Loss), plotted Log2Ratio respect to probes  $n$ . The ideal event (bold line), the measurements of CNA (noise signal) (line), and jitter (dashed line) are showed.

the discrete skew discrete Laplace (SkL) law and derived the confidence masks using the SkL. Later, we have shown that the SkL is adequate only when SNR values overcome unity [31]. Otherwise, the Laplace distribution becomes too rough and a more correct probabilistic model of jitter in the breakpoints is required. Such conditions include the low SNR,  $\gamma_i^-, \gamma_i^+ < 1$ , and extremely low,  $\gamma_i^-, \gamma_i^+ \ll 1$ .

Jan O. Korbel, Alexander Ekehart Urbanwe, *et.al.* have proposed in [32] an approach, called BreakPtr, for fine-mapping Copy Number Variations (CNVs). They statistically integrate both sequence characteristics and data from high-resolution comparative genome hybridization experiments in a discrete-valued, bivariate hidden Markov model. The incorporation of nucleotide-sequence information has allowed them to take into account the fact that recently duplicated sequences (e.g., segmental duplications) often coincide with the breakpoints.

Sergii Ivakhno *et.al.* [33] presented a novel approach, called CNAsseg, to identify CNAs from second-generation sequencing data. It uses depth of coverage to the estimate copy number states and flowcell-to-flowcell variability in cancer and normal samples to control

the false positive rate.

Popova T., Boeva, V., *et.al.* in [34] provided a data mining technique based on the GAP method which allows extraction of absolute copy numbers and allelic contents from the whole genome copy number variation and allelic imbalance profiles obtained by SNP arrays or NGS.

Szatkiewicz, J. P. *et.al.* in [35] presented a novel read–depth–based method, GENSENG, which uses a hidden Markov model and negative binomial regression framework to identify regions of discrete copy-number changes while simultaneously accounting for the effects of multiple confounders.

Van den Broek, Evert *et al.* in [36] developed GeneBreak method to systematically identify genes recurrently affected by the genomic location of chromosomal CNA-associated breaks by a genome-wide approach, which can be applied to DNA copy number data obtained by aCGH or by (low-pass) WGS.

## 1.2 Motivation

Modern technologies developed to produce the CNA profiles with high resolution are still very sensitive to additive white Gaussian noise. As a consequence, jitter is inherent to the breakpoints of measured genome somatic CNAs causing errors and ambiguities in the breakpoint detection with low signal-to-noise ratios (SNRs). When  $SNR > 1$ , it can statistically be described using the discrete skew Laplace distribution. Otherwise, if  $SNR < 1$ , better approximations are required to produce more accuracy.

Nowadays, no estimator–robust or optimal– is able to provide jitter-free estimation of segmental changes. Thus, in order to avoid wrong decisions, the estimates must be bounded by the confidence probability. Having the jitter distribution, it is easy to find a region within which the breakpoint exists for the required probability. Of practical importance are the confidence UB and LB masks, which can be created based on the segmental and jitter distributions for the given confidence probability.

A researcher or medical geneticist takes hours to detect aberrations in samples of DNA. The masks can serve as an auxiliary tool for medical experts to make decisions about the

---

CNA structures. Therefore, an analysis and improvement of jitter is required to reduce errors in the CNAs estimation.

### 1.3 Research Objectives

We propose and investigate several approximations for the jitter distribution in the CNA's breakpoints for low and extra low SNRs, and show that the approximations proposed fits the CNA probes much better than the Laplace distribution for any reasonable SNR value of practical interest. Then, a statistical theoretical model to compute the confidence masks is justified via the lower and upper confidence boundaries, using the suggested approximations for the given confidence probability. The estimated CNAs are tested by the masks for data obtained using microarray technology and show how to improve the CNAs estimates by removing some unlikely existing breakpoints.

### 1.4 Scope of this work

In Chapter 2, some biologic concepts about the genetic data are defined, including a description of the principal technologies of hybridization to obtain the measurements of CNAs and details about the algorithms to estimate changes in signals piecewise.

In Chapter 3 we provide a detailed description of initial algorithm for computing the probabilistic confidence masks for the confidence limits the stepwise signals measured in noise. Also, we derive the SkL distribution for the jitter in the breakpoints under the ideal conditions and show its limitations. This distribution is later used to compute the confidence upper and lower boundary masks in order to guarantee an existence of genomic changes with required probability.

In Chapter 4, we analyze errors caused by the fitting of SkL approximation based on data obtained by simulation for different segmental SNRs. Aimed at improving the approximation accuracy, we develop here three new approaches resulting in the following outputs: Heuristic approximation, parametrization of skew Laplace distribution, and asymmetric exponential power distribution.



---

In Chapter 5, the proposed approximations are adapted to the probabilistic confidence masks by modifying the initial equations and replacing the SkL. Then, the modified confidence masks are applied to microarrays data obtained with different technologies to test estimates of the Copy Number Alterations.

Based on the annotations made by experts, the probability of their observations is computed in Chapter 6, using the CNAs estimates of neuroblastoma and the probabilistic modified confidence masks.

In Chapter 7, the CNAs estimates are compared using the Circular Binary Segmentation and Pruned Exact Linear Time methods respect to the Next Generation Sequencing technology which are established as the ideal estimates. Also, it is given the global analysis of the deleted breakpoints and the length of CNAs at each level of probability

Finally, Chapter 8 is sketched the conclusions with respect to each approximation of jitter distribution proposed and modified confidence masks.

# Chapter 2

## Foundations

### 2.1 DNA and Cancer

The nucleus of each human cell contains 22 Chromosomes, plus the X chromosome (one in males, two in females) and, in males only, one Y chromosome, composed by a chemical substance called DNA stranded by Histones. A biomolecule of DNA is formed by four molecules, they are adenine (A), thymine (T), cytosine (C), and guanine (G). Simultaneously, each chromosome consists of two polynucleotide chains wound around each other in the form of a double helix formed by genes that contain complementary genetic information. In humans, genes vary in size from a few hundred DNA bases to more than 2 million bases. The Human Genome Project has estimated that humans have between 20,000 and 25,000 genes.

According to the World Health Organization *www.who.int*, cancer is one of the main causes of morbidity and mortality worldwide. This disease causes mutations in cell's ADN, modifying the structures and positions of genes. Such mutations are called Copy Number Aberrations (CNA).

CNAs are recurrently defined as gains and losses of large segments of the genome in size, ranging from a few kilobases to whole chromosomes. Somatic CNAs (SCNAs) that occur during the lifetime of an individual are a major contributor to cancer development, particularly for solid tumors [37].

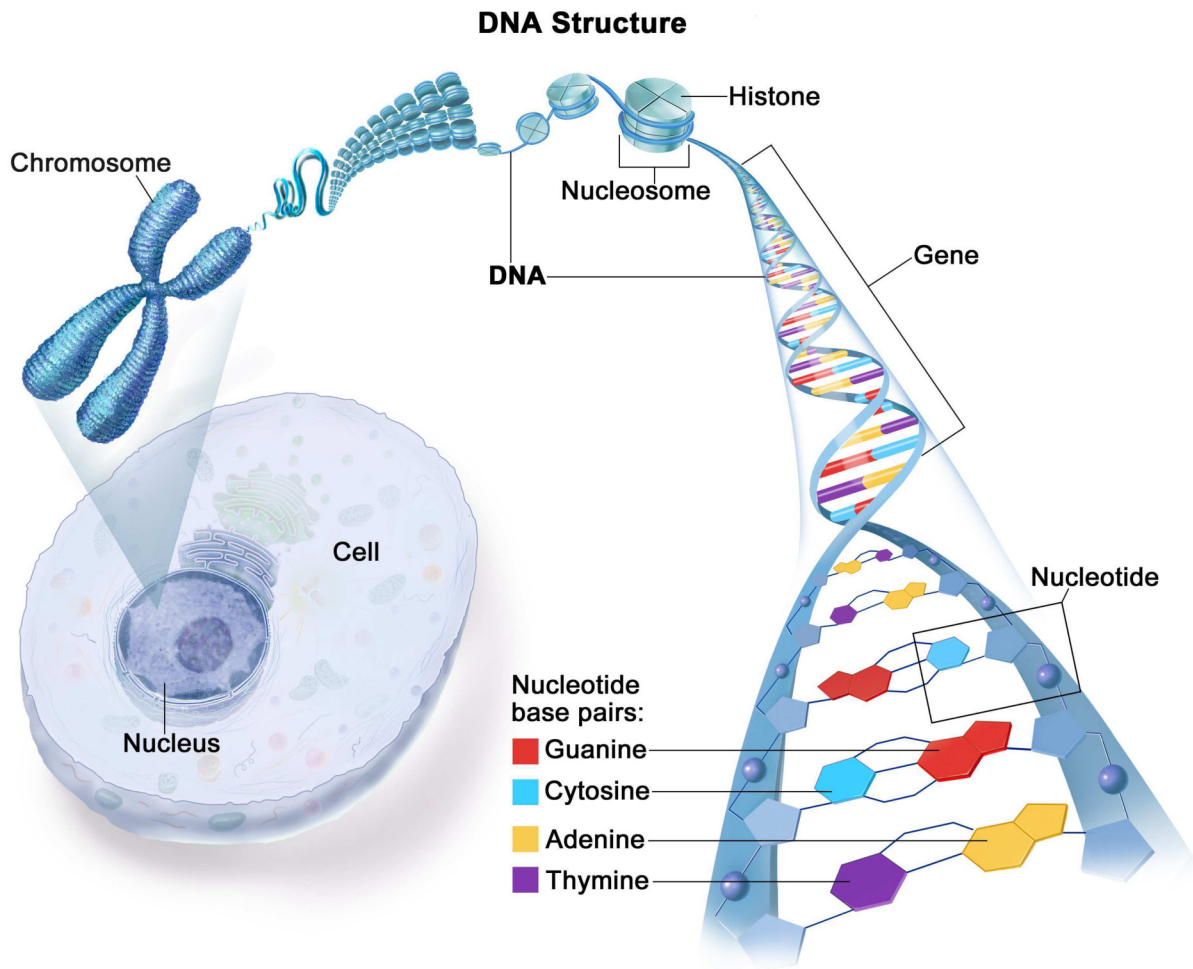


Figure 2.1: Structure of DNA. SITEMAN, Cancer Center.

## 2.2 DNA Microarray

DNA microarray is a technology to determine whether a sample of DNA from a living organism contains CNAs in genes. Basically, a microarray is an ordered arrangement of known or unknown DNA samples attached to a solid support [38]. Several techniques have been developed based on this methodology, being the most popular to analyze samples of genome.

### 2.2.1 SNP microarray

Single Nucleotide Polymorphism (SNP) array is a type of DNA microarray which is used to detect polymorphisms within a population. An SNP, a variation at a single site in DNA, is the most frequent type of variation in the genome. Currently, there are around 85 million SNPs that have been identified in the human genome [39]. The CNVs profiles are represented by the Log R ratios (LRR)s centered at zero for each sample.

The SNP arrays are presently one of the most efficient technologies for the identification of the CNAs [40]. The SNP method is considered as an NGS technology because gives maximal information about tumors. Although, SNP arrays have progressively replaced aCGH in samples of cancer is less common to analyze CNAs.

### 2.2.2 CGH microarray

The aCGH microarrays have been developed as genome-wide assays for measurements of CNAs, using the fact that the microarray fluorescence intensity is proportional to DNA copy number [10]. It is one of the most modern techniques providing a resolution of 5–10 kbp [17]. The CNAs data are represented in genomic position with the  $n$ th probe,  $n_l \in [1, M]$ , where  $M$  is the number of probes. In the CNAs picture, the  $n_l$ th discrete point corresponds to the  $i$ th edge or breakpoint. In microarray technique, the CNAs are often normalized and plotted as  $\log_2 R/G = \log_2 \text{Ratio}$ , where  $R$  and  $G$  are the fluorescent Red and Green intensities, respectively [19]. Figure 2.2 shows a structure based on aCGH data from pancreatic cancer study.

### 2.2.3 Next Generation Sequencing

The most modern technology of hybridization is the Next Generation Sequencing NGS (also known as massively parallel sequencing) that is revolutionizing the ability to characterize cancers. The NGS technique can recognize copy number aberrations and somatic rearrangements in an entire cancer genome at base pair resolution in a matter of weeks.

Comparative genomic hybridization and SNP array analysis have provided a wealth of

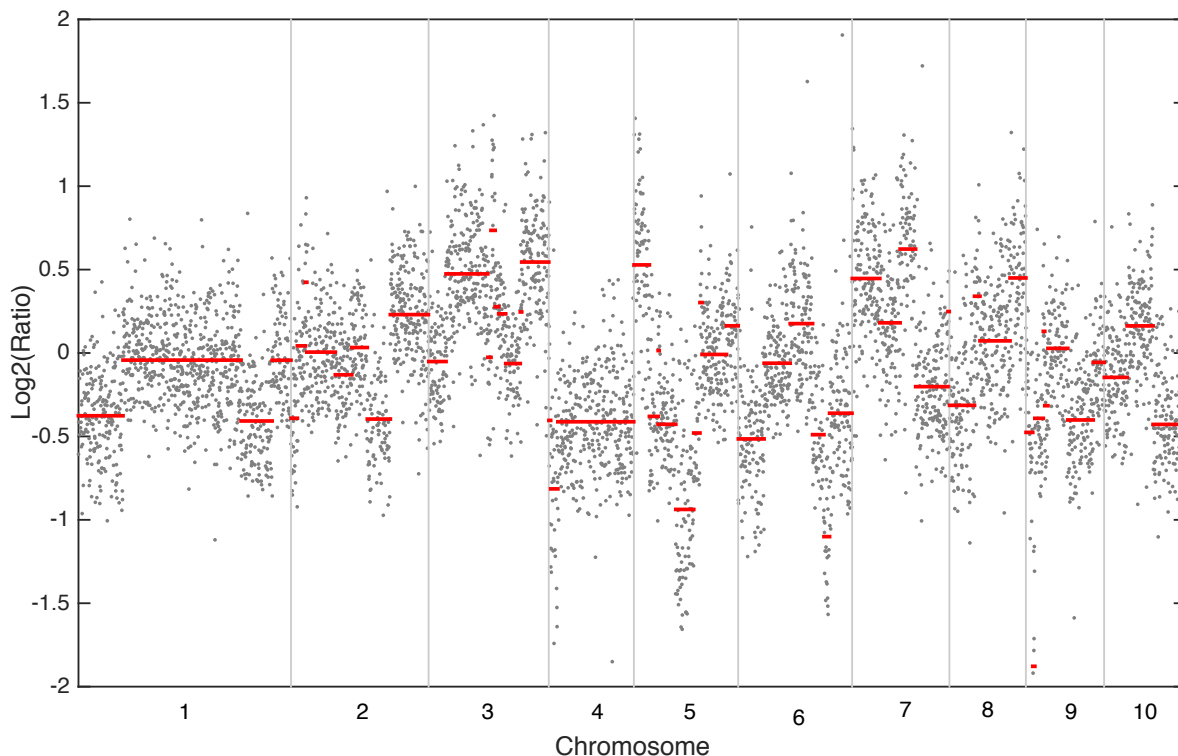


Figure 2.2: Chromosomes 1 – 10 represented in  $\text{Log}_2\text{Ratio}$  from pancreatic adenocarcinoma genome. Measurements and estimate CNAs are plotted in dotted and solid lines, respectively.

data on gene copy number aberrations in breast cancer and have helped identify potential therapeutic targets for subgroups of breast cancer patients; however, this technology does not provide any information about structural genomic aberrations and base pair mutations [41].

Perhaps more important than the sequencing throughput provided by this technology and its relative low cost compared with traditional sequencing methods is the type of data it generates[42]. Instead of long reads generated from a Polymerase Chain Reaction (PCR)–amplified sample, massively parallel sequencing methods provide much shorter reads ( 21 to 400 base pairs (bp)), but millions of them [43, 44, 45, 46, 47, 48].

## 2.3 Algorithms for estimate CNAs

Nowadays the estimation and evaluation of CNAs are fundamental tasks because physicians use this information to give a diagnostic and treatment to a particular disease. Many methods have been developed with this purpose following different statistical properties. Below, a technique proposed in [28] and the most popular algorithms are described.

### 2.3.1 Median Breakpoints detector

To apply the Median filter as a breakpoint detector, consider a measurement  $y_n$ ,  $n \in [1, M]$ . Because the median is efficient asymptotically in large measurement noise [49], apply the median

$$\hat{x} = \underset{y_j \in W}{\operatorname{argmin}} \sum_{i=1}^v |y_j - y_i| = \operatorname{med}\{y_i\}_{i=1}^{W_v} \quad (2.1)$$

sequentially  $v$  times with windows  $W_1 < W_2 < \dots < W_v$ , provided that each  $W_v$ ,  $v \in [1, V]$ , is odd,

$$\begin{aligned} \hat{x}_{n-\frac{W_1-1}{2}}^{(1)} &= \operatorname{med}(y_{n-i})_{i=0}^{W_1-1}, \\ \hat{x}_{n-\frac{W_2-1}{2}}^{(2)} &= \operatorname{med}(x_{n-i}^{(1)})_{i=0}^{W_2-1}, \\ &\vdots \\ \hat{x}_{n-\frac{W_v-1}{2}}^{(v)} &= \operatorname{med}(x_{n-i}^{(v)})_{i=0}^{W_v-1}, \end{aligned} \quad (2.2)$$

Until the CNAs structure becomes clear. If the computation time is not an issue, then the window length  $W_v$  can be increased as 3, 5, 7, ..., until the next window gives no change. Alternatively, follow the recommendations given in [50] regarding the window optimality. Figure 2.3 depicts the effect of median smoothing with  $v_1 = 11$ ,  $v_2 = 21$ ,  $v_3 = 41$ , and  $v_4 = 71$ . By (2.2), the final median estimate becomes  $\hat{x}_n^{\operatorname{med}} = \hat{x}_n^{(v)}$ . The locations of edges in the  $\hat{x}^{\operatorname{med}}$  structure determine the candidate breakpoints vector  $\mathcal{N}^* = [n_1^* n_2^* \dots n_L^*]^T$ . The estimate  $\hat{\mathcal{N}} = [\hat{n}_1 \hat{n}_2 \dots \hat{n}_L]^T$  of  $\mathcal{N}$  can then be found by adjusting  $\mathcal{N}^*$

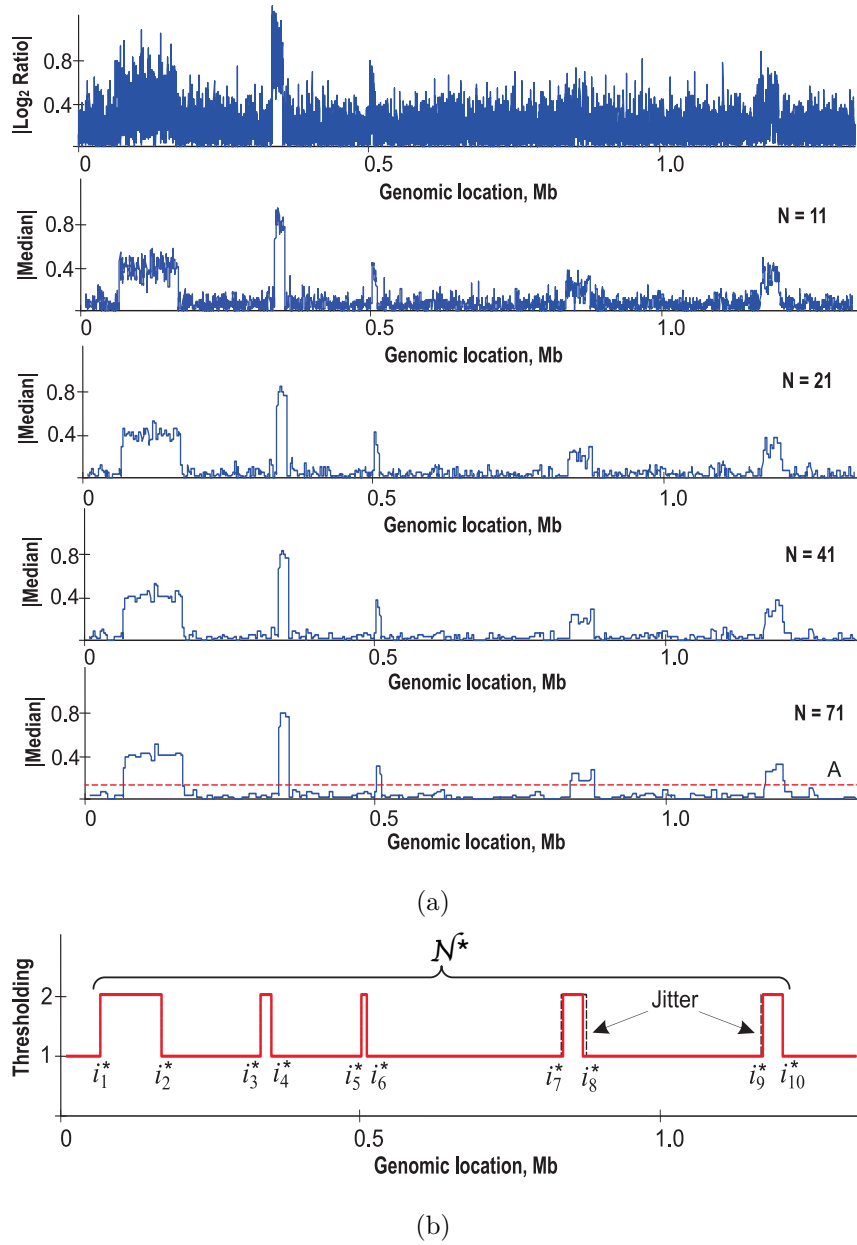


Figure 2.3: Median-based denoising of the microarray measurement: (a) subsequent smoothing of  $\log_2$  Ratio with  $v = 11, v = 21, v = 41,$  and  $v = 71$  and (b) threshold-based forming of a rectangular pulse train.

visually or using any of the optimization procedures such as the combined ML estimation algorithm [28]. Placing a threshold  $\mathbf{A}$  below the minimum copy number in  $|\hat{x}_n^{\text{med}}|$  will

form a pulse trains as shown in Figure 2.3b. Provided the candidate breakpoints  $\mathcal{N}^*$ , the component  $\hat{a}_l$  of the estimate  $\hat{\mathbf{a}}_l = [\hat{a}_1 \hat{a}_2 \dots \hat{a}_{L+1}]^T$  can be found by averaging.

This method is able to detect the breakpoints of CNAs with precision. However, two principal disadvantages of this procedure are a high time to select the adequate window  $W$  and that the breakpoints of little changes could be undetectable.

### 2.3.2 Circular Binary Segmentation

Circular Binary Segmentation (CBS) was one of the first algorithms used to estimate CNAs. In [51], it was developed a modification of Binary Segmentation (BS), which was called *circular binary segmentation*, to translate noisy intensity measurements into regions of equal copy number. This algorithm is based on the partitions of a genome into constant segments, detecting copy numbers alterations and the change-point (breakpoint).

Following the change-point method it is possible to estimate the CNAs. Let  $X_1, X_2, \dots$  be a sequence of random probes. An index  $i$  is called a breakpoint if  $X_1, \dots, X_i$  have a common distribution function  $F_0$  and  $X_{i+1}, \dots$  have a different common distribution function  $F_1$  until the next change-point (if one exists).

Subsequently, the CBS algorithm is modified in [52] to faster. The algorithm tests for breakpoints using a maximal  $t$ -statistic with a permutation reference distribution to obtain the corresponding  $P$ -value.

### 2.3.3 Pruned Exact Linear Time

The Pruned Exact Linear Time (PELT) method was introduced to find breakpoints and is computationally efficient in several applications, such as CNAs estimate. The PELT finds the minimum of the cost functions, such as the negative log likelihood, quadratic loss, cumulative sums or those based on both the segment log-likelihood and the length of the segment. Next, the Optimal Partition (OP) and location of breakpoints are obtained having a linear computational cost respect to the number of observations  $n$ , under mild conditions, so the computational efficiency of PELT is  $\mathcal{O}(n)$  [27]. Also, this procedure requires a penalty for inserted changepoints. In Fig. 2.4, we give an example of the



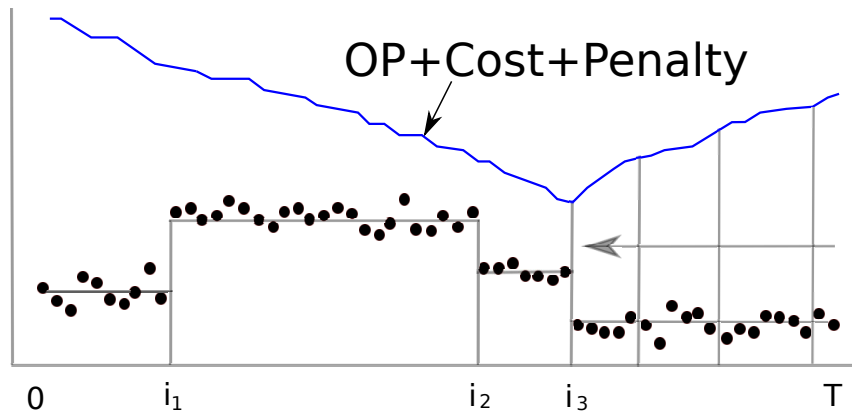


Figure 2.4: Example of breakpoint estimated using the Pruned Exact Linear Time (PELT) method.

breakpoint estimates using the PELT algorithm.

This method is more accurate than Binary Segmentation (BS) algorithm, but computationally it is not efficient. Even so, the authors argued that the statistical benefits of an exact segmentation outweigh is in the relatively small computational costs.

### 2.3.4 Binary Segmentation

The Binary Segmentation (BINSEG) uses techniques of cluster analysis to split measurements into reasonably homogeneous groups [53]. The BINSEG method is based on a technique of cluster analysis to separate the sample treatment means in an equitable design. Then, a likelihood ratio test set the significance of the difference among groups, *i.e.* this parameter determines the grade of segmentation. The BINSEG includes a high efficiency of the order of  $\mathcal{O}(n \log n)$ . A graphic example of the BINSEG algorithm is showed in Figure 2.5.

This algorithm was modified in [51] and called Circular Binary Segmentation (CBS) to translate noisy intensity measurements into regions of equal copy number. The method was evaluated by simulation and was demonstrated on cell line data with known copy number alterations and on a breast cancer cell line data set.

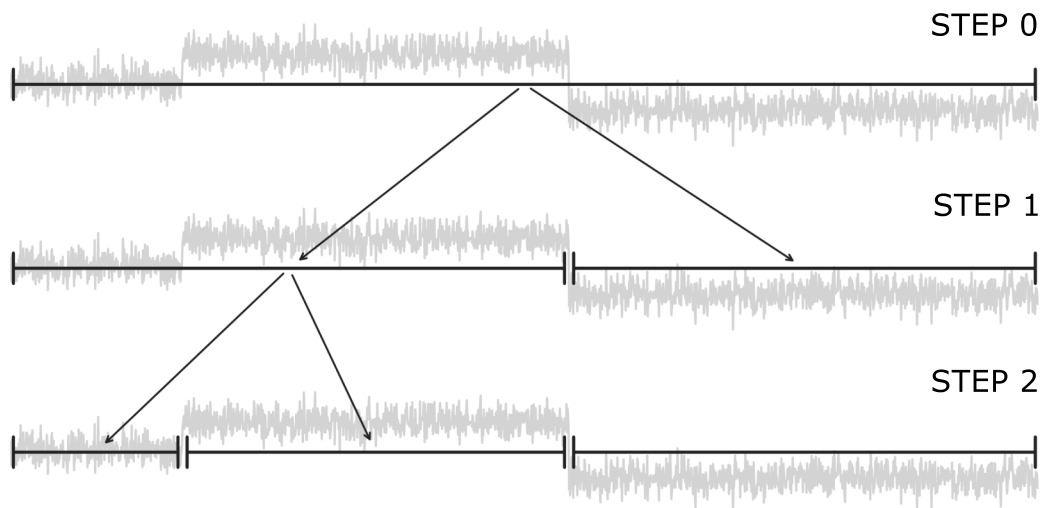


Figure 2.5: Schematic of Binary Segmentation (BINSEG) algorithm.

### 2.3.5 Segment Neighborhood

The Segment Neighborhood (SEGNEIGH) method is based on the concept of segment neighborhood which is defined as a set of contiguous residues that share common features [54]. So, the model of these features and the residuals that define the boundaries of each segment neighborhood need to be estimated using a defined algorithm. The least squares and maximum likelihood are used to compute the most critical features. The computational efficiency of this algorithm is  $\mathcal{O}(Kn^2)$  where  $n$  is the number of observations and  $K$  a repetition parameter.

# Chapter 3

## Jitter Distribution

### 3.1 Jitter Distribution in Breakpoints

Most generally, the CNAs estimation problem implies predicting the breakpoints locations  $I = [i_1 i_2 \dots i_L]^T \in \mathcal{R}^L$ , where  $i_l, l \in [1, L]$ , and the segmental changes  $\hat{\mathbf{a}}_l$  with a maximum possible accuracy and precision acceptable for medical applications. Several cases to estimate  $I$  and  $\hat{\mathbf{a}}_l$  were detected in the research made in [30]:

- **Case I:**  $I$  and  $\hat{\mathbf{a}}_l$  can easily be estimated if the number of probes is large in each neighboring segment and the edges are sharp.
- **Case II:** the component of  $I$  can be well detectable, but the estimate of the relevant component of  $a_l$  may be imprecise owing to a small number of probes.
- **Case III:** it is hard to estimate the components of  $I$  if the segmental differences between  $a_l$  and its CNAs neighborhoods are small.
- **Case IV:** the segmental estimates  $a_l$  may have enough precision, whereas the estimates of the edges not.

So, an analysis of the estimation errors caused by the segmental noise and jitter in the breakpoints is required.

Let us consider the microarray-based measurement of the CNAs in more detail. Fig. 3.1 gives several simulated examples around the  $l$ th breakpoint with different realizations of the measurement affected with white Gaussian noise having for simplicity equal segmental variances. The threshold (dashed) is placed equidistantly between the segmental changes. The breakpoint location is found by the Maximum Likelihood (ML) based on Gauss's ordinary least squares (OLS). The case (a) is ideal to mean that with such locations of the measured points the ML estimate will be jitter-free. If it happens that some left-neighboring to  $i_l$  points lie below the threshold (dashed), then the ML estimate will be found to the left of  $i_l$ ; two points to the left in the case (b). We call it the left jitter. If some right-neighboring points lie as in 3.1c, then the ML estimate will be found to the right of  $i_l$ ; two points to the right in the case (c). We call it the right jitter. Also, there may be observed some ambiguities as in the case (d) when the estimator gives two or more possible locations for the same breakpoint.

In order to derive the approximate jitter distribution for CNAs, a simulated measurement with one breakpoint and two constant segments is considered, as sketched in Fig. 3.2a. Here, the  $a_l$  and  $a_{l+1}$  segmental levels are contaminated with zero mean Additive White Gaussian Noise (WGN) [55, 56] having the variances  $\sigma_l^2$  and  $\sigma_{l+1}^2$  as shown in Fig. 3.2b. The segmental signal-to-noise ratios (SNRs) in the  $l$ th and  $(l+1)$ th segments are computed as in [57], respectively,

$$\begin{aligned}\gamma_l^- &= \frac{2(a_{l+1} - a_l)^2}{2\sigma_l} = \frac{(a_{l+1} - a_l)^2}{\sigma_l} = \frac{\Delta_l^2}{\sigma_l^2} \\ \gamma_l^+ &= \frac{2(a_{l+1} - a_l)^2}{2\sigma_{l+1}} = \frac{(a_{l+1} - a_l)^2}{\sigma_{l+1}} = \frac{\Delta_l^2}{\sigma_{l+1}^2},\end{aligned}\tag{3.1}$$

where  $\Delta_l = a_{l+1} - a_l$  is the segmental difference, which corresponds to the breakpoint  $i_l$  at  $n = 50$ .

Aiming to find the probability between intersection points  $\alpha$  and  $\beta$  of probability function densities, it is needed to represent the Gaussian pdfs of each segment  $l$  and  $l+1$  with  $p(x)$  and  $p(y)$ , respectively, and defined as :

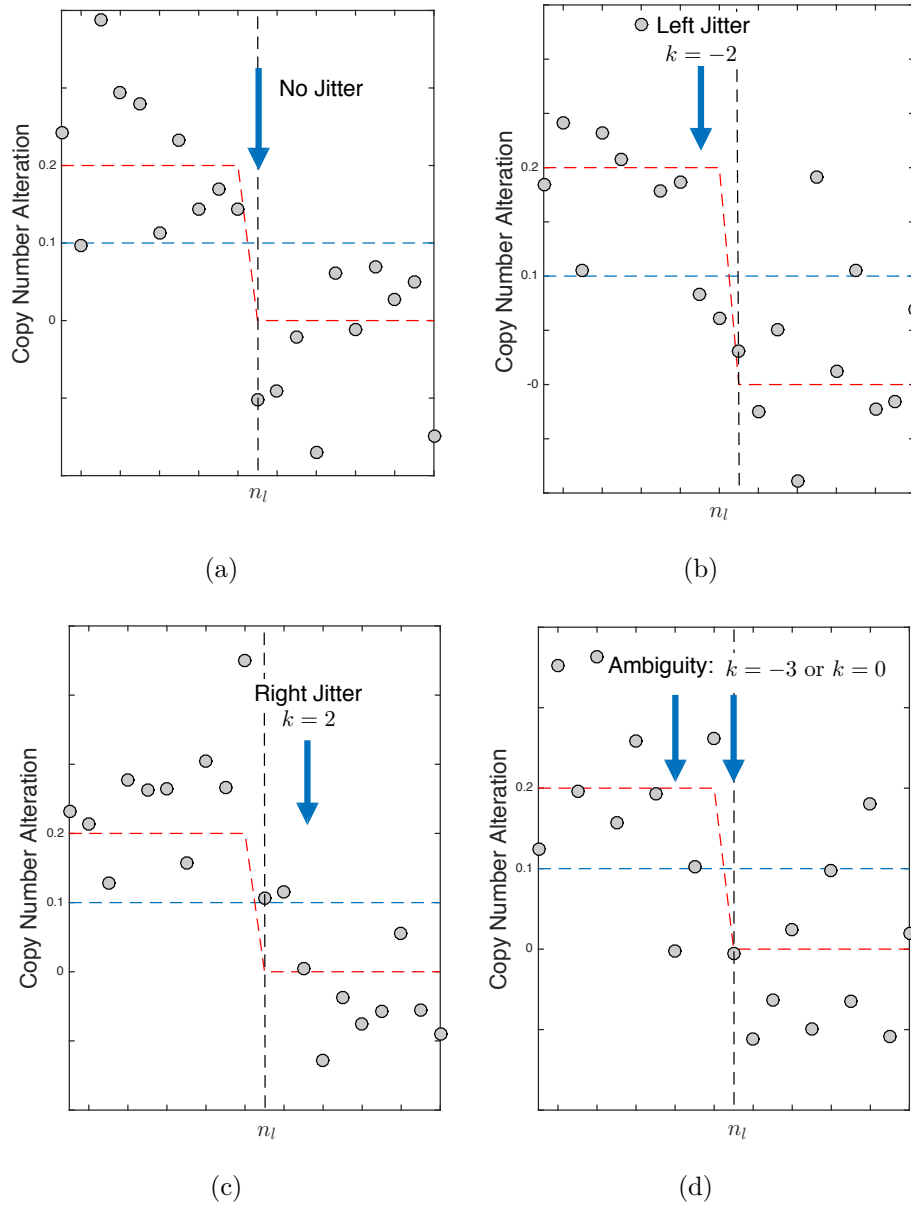


Figure 3.1: Jitter distributions computed with Maximum Likelihood and Skew Laplace distribution to a) SNR=0.1 and b) SNR =0.5. The ML (circled) is the jitter pdf obtained experimentally using a ML estimator via a histogram over  $50 \times 10^3$  runs and SkL (solid) is the skew Laplace distribution.

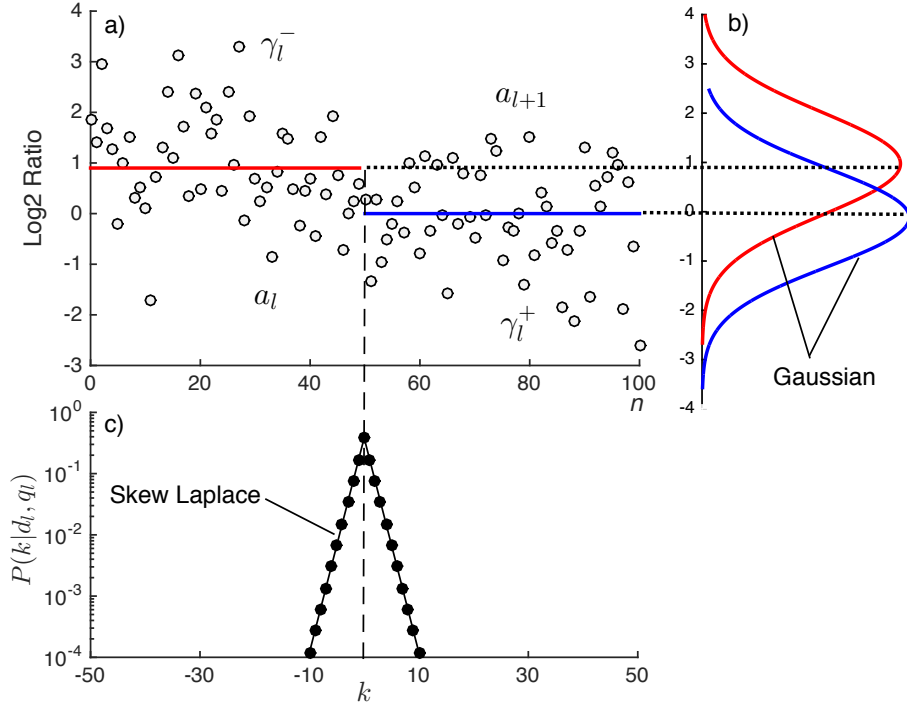


Figure 3.2: Simulated CNAs with one breakpoint located at  $n = 50$  and standard deviations  $\sigma_l$  and  $\sigma_{l+1}$  of each segment corresponding to constant values of SNRs  $\gamma_l^- \approx \gamma_l^+ = 1$ : (a) noise measurements, (b) segmental Gaussian distributions and (c) skew Laplace jitter distributions.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{(x-a_l)^2}{2\sigma_x^2}} \quad (3.2)$$

$$p(y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(y-a_{l+1})^2}{2\sigma_y^2}}. \quad (3.3)$$

Seeking the cross points between  $p(x)$  and  $p(y)$ , we arrive at the following equalities

$$\frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{(x-a_l)^2}{2\sigma_x^2}} = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(y-a_{l+1})^2}{2\sigma_y^2}}$$

$$\ln \frac{\sigma_y}{\sigma_x} = \frac{(x-a_l)^2}{2\sigma_x^2} - \frac{(y-a_{l+1})^2}{2\sigma_y^2}. \quad (3.4)$$

After simplifying equation (3.4) and grouping variables, we have,

$$x^2 \frac{\sigma_y^2 - \sigma_x^2}{2\sigma_x^2\sigma_y^2} + x \frac{a_{l+1}\sigma_x^2 - a_l\sigma_y^2}{\sigma_x^2\sigma_y^2} + \frac{a_l\sigma_y^2 - a_{l+1}\sigma_x^2}{2\sigma_x^2\sigma_y^2} - \ln \frac{\sigma_y}{\sigma_x} = 0. \quad (3.5)$$

Now, (3.5) can be represented as a squared binomial replacing the following variables:

$$\mathbf{a} = \frac{\sigma_y^2 - \sigma_x^2}{2\sigma_x^2\sigma_y^2} = \frac{\gamma_l^-}{2\Delta^2} - \frac{\gamma_l^+}{2\Delta^2} = \frac{\gamma_l^- - \gamma_l^+}{2\Delta} \quad (3.6)$$

$$\mathbf{b} = \frac{a_{l+1}\sigma_x^2 - a_l\sigma_y^2}{\sigma_x^2\sigma_y^2} = \frac{a_{l+1}\gamma_l^+}{\Delta^2} - \frac{a_l\gamma_l^-}{\Delta^2} = \frac{a_{l+1}\gamma_l^+ - a_l\gamma_l^-}{\Delta^2} \quad (3.7)$$

$$\mathbf{c} = \frac{a_l\sigma_y^2 - a_{l+1}\sigma_x^2}{2\sigma_x^2\sigma_y^2} - \ln \frac{\sigma_y}{\sigma_x} = \frac{a_l^2\gamma_l^- - a_{l+1}^2\gamma_l^+}{2\Delta^2} - \ln \sqrt{\frac{\gamma_l^-}{\gamma_l^+}} \quad (3.8)$$

which can be solved as a standard quadratic function

$$\begin{aligned} \alpha_l, \beta_l &= -\frac{\mathbf{b}}{2\mathbf{a}} \pm \frac{1}{2\mathbf{a}} \sqrt{\mathbf{b}^2 - 4\mathbf{a}\mathbf{c}} \\ &= \frac{a_l\gamma_l^- - a_{l+1}\gamma_l^+}{\gamma_l^- - \gamma_l^+} \mp \frac{1}{\gamma_l^- - \gamma_l^+} \\ &\quad \times \sqrt{(a_l - a_{l+1})^2\gamma_l^-\gamma_l^+ + 2\Delta_l^2(\gamma_l^- - \gamma_l^+) \ln \sqrt{\frac{\gamma_l^-}{\gamma_l^+}}}. \end{aligned} \quad (3.9)$$

If  $a_{l+1} = 0$  and  $a_l = \Delta$ , it can be represented with

$$\alpha, \beta = \frac{\Delta\gamma_l^-}{\gamma_l^-\gamma_l^+} \left[ 1 + \sqrt{1 - \frac{\gamma_l^- - \gamma_l^+}{\gamma_l^-} \left( 1 - \frac{2}{\gamma_l^-} \ln \sqrt{\frac{\gamma_l^-}{\gamma_l^+}} \right)} \right], \quad (3.10)$$

if  $\gamma_l^- \neq \gamma_l^+$ . For  $\gamma_l^- = \gamma_l^+$ , set  $\alpha_l = \Delta_l/2$  and  $\beta_l = \pm\infty$ .

### 3.1.1 Probabilities of events A and B

In order to compute the probability of jitter, three cases can be considered:  $\sigma_x > \sigma_y$ ,  $\sigma_x \geq \sigma_y$  and  $\sigma_x < \sigma_y$ , which are sketched in Fig. 3.3a, Fig. 3.3b and Fig. 3.3c, respectively.

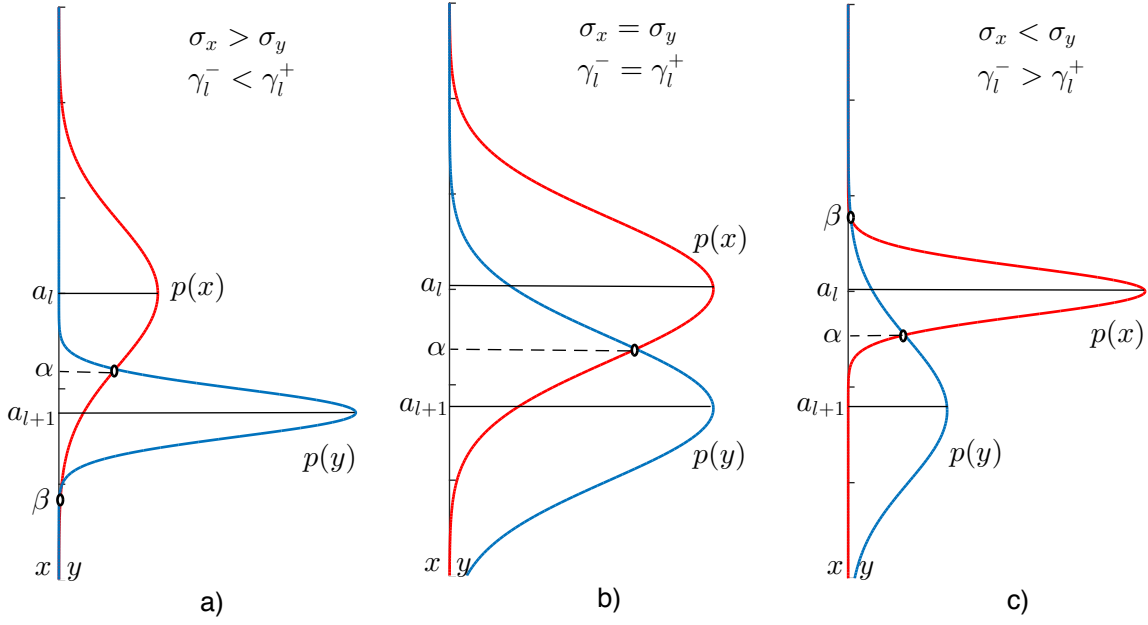


Figure 3.3: Cases for different values of standard deviation to compute the events  $A_i$  and  $B_i$ . a)  $\sigma_x > \sigma_y$ , b)  $\sigma_x = \sigma_y$  and c)  $\sigma_x < \sigma_y$

Following Fig. 3.3 and assuming different noise variances for  $\sigma_x^2$  and  $\sigma_y^2$ , the events  $A_i$  and  $B_i$  can be specified as

$$A_{i|n-N+1 \leq i \leq n} = \begin{cases} \alpha < y_i \wedge y_i < \beta & , \sigma_x > \sigma_y , \\ y_i < \beta & , \sigma_x = \sigma_y , \\ \alpha < y_i < \beta & , \sigma_x < \sigma_y , \end{cases} \quad (3.11)$$

$$B_{j|n+1 \leq j \leq n+N} = \begin{cases} \beta < y_i < \alpha & , \sigma_x > \sigma_y , \\ y_i < \alpha & , \sigma_x = \sigma_y , \\ y_i < \alpha \wedge y_i > \beta & , \sigma_x < \sigma_y , \end{cases} \quad (3.12)$$

where  $N$  is the segment length. Following the conjectures made in [58], we develop the next procedure. The event  $A_i$  means that measurements at the points  $n - N + 1 \leq i \leq n$  belong to the first segment, where  $n$  is its last point. Also, the event  $B_j$  means that the measurements at  $n + 1 \leq j \leq n + N$  totally belong to the second segment. The inverse



events are  $\bar{A}_i = 1 - A_i$  and  $\bar{B}_j = 1 - B_j$ . The events  $A_i$  and  $B_j$  can be united in two blocks

$$\mathbf{A} = \{A_{n-N+1}A_{n-N+2}\cdots A_n\} \quad (3.13)$$

$$\mathbf{B} = \{B_{n+1}B_{n+2}\cdots B_{n+N}\} \quad (3.14)$$

If the events  $\mathbf{A}$  and  $\mathbf{B}$  occur simultaneously, then the jitter at  $i_l$  will never occur. However, there may be found some events which do not obligatorily lead to jitter. For example, the second point in Fig. 3.1a lies below the threshold (dashed) but does not lead to jitter. Such events are ignored and the lower bound of the jitter free can be obtained with equation (3.15) and the jitter upper bound with (3.16) as

$$\check{P}(\mathbf{AB}) = \check{P}(A_{n-N+1}\cdots A_n B_{n+1}\cdots B_{n+N}) \quad (3.15)$$

$$\hat{P} = 1 - \check{P}(\mathbf{AB}). \quad (3.16)$$

Assuming that the noise in the measurements is additive white noise and that all of the events are thus independent, it is allowed to write (3.15) and (3.16) as

$$\hat{P} = 1 - P^N(\mathbf{A})P^N(\mathbf{B}) \quad (3.17)$$

$$\check{P} = P^N(\mathbf{A})P^N(\mathbf{B}). \quad (3.18)$$

Thinking that jitter occurs at some  $n + 1 \pm k$  point, we assign two additional blocks of events

$$\mathbf{A}_{-k} = \{A_{n-N+1}\cdots A_{n-k}\}, \quad (3.19)$$

$$\mathbf{B}_k = \{B_{n+1+k}\cdots B_{n+N}\}. \quad (3.20)$$

The probability  $\check{P}_{-k}$  that jitter occurs at the  $k$ th point to the left from  $n + 1$  (left jitter) and the probability  $\check{P}_k$  that jitter occurs at the  $k$ th point to the right from  $n + 1$  (right jitter) can be specified as, respectively,

$$\check{P}_{-k}(\mathbf{A}_{-k}\bar{A}_{n-k+1}\cdots\bar{A}_n\mathbf{B}) \quad (3.21)$$

$$= \check{P}_{-k}(A_{n-N+1}\cdots A_n - k\bar{A}_{n-k+1}\cdots\bar{A}_n B_{n+1}\cdots B_{n+N}), \quad (3.22)$$

$$\check{P}_k(\mathbf{A}_k\bar{B}_{n+1}\cdots\bar{B}_{n+k}\mathbf{B}_k) \quad (3.23)$$

$$= \check{P}_k(A_{n-N+1}\cdots A_n\bar{B}_{n+1}\cdots\bar{B}_{n+k}B_{n+1+k}\cdots B_{n+N}). \quad (3.24)$$

Assuming that the events are independent, the (3.21) and (3.23) can be simplified to

$$\check{P}_{-k} = P^{N-k}(A) [1 - P(A)]^k P^N(B), \quad (3.25)$$

$$\check{P}_k = P^N(A) [1 - P(B)]^k P^{N-k}(B). \quad (3.26)$$

Now,  $P(A)$  and  $P(B)$  can be specified for segmental Gaussian noise as, respectively,

$$P(A) = \begin{cases} 1 - \int_{\beta}^{\alpha} p_1(y) dy & , \sigma_x > \sigma_y, \\ \int_{\beta}^{\alpha} p_1(y) dy & , \sigma_x = \sigma_y, \\ \int_{\alpha}^{\beta} p_1(y) dy & , \sigma_x < \sigma_y, \end{cases} \quad (3.27)$$

$$P(B) = \begin{cases} \int_{\beta}^{\alpha} p_2(y) dy & , \sigma_x > \sigma_y, \\ \int_{\alpha}^{\beta} p_2(y) dy & , \sigma_x = \sigma_y, \\ 1 - \int_{\alpha}^{\beta} p_2(y) dy & , \sigma_x < \sigma_y. \end{cases} \quad (3.28)$$

Analyzing the Gaussian processes (see Appendix A), the probabilities  $P(A_l)$  and  $P(B_l)$  can be expressed in terms of the erf and erfc functions

$$P(A_l) = \begin{cases} 1 + \frac{1}{2}[\text{erf}(g_l^{\beta}) - \text{erf}(g_l^{\alpha})] & , \gamma_l^- < \gamma_l^+, \\ \frac{1}{2}\text{erfc}(g_l^{\alpha}) & , \gamma_l^- = \gamma_l^+, \\ \frac{1}{2}[\text{erf}(g_l^{\beta}) - \text{erf}(g_l^{\alpha})] & , \gamma_l^- > \gamma_l^+, \end{cases} \quad (3.29)$$

$$P(B_l) = \begin{cases} \frac{1}{2}[\operatorname{erf}(h_l^\alpha) - \operatorname{erf}(h_l^\beta)] & , \gamma_l^- < \gamma_l^+ , \\ 1 - \frac{1}{2}\operatorname{erfc}(h_l^\alpha) & , \gamma_l^- = \gamma_l^+ , \\ 1 + \frac{1}{2}[\operatorname{erf}(h_l^\alpha) - \operatorname{erf}(h_l^\beta)] & , \gamma_l^- > \gamma_l^+ , \end{cases} \quad (3.30)$$

where  $g_l^\beta = \frac{\beta_l - \Delta_l}{|\Delta_l|} \sqrt{\frac{\gamma_l^-}{2}}$ ,  $g_l^\alpha = \frac{\alpha_l - \Delta_l}{|\Delta_l|} \sqrt{\frac{\gamma_l^-}{2}}$ ,  $h_l^\beta = \frac{\beta_l}{|\Delta_l|} \sqrt{\frac{\gamma_l^+}{2}}$ ,  $h_l^\alpha = \frac{\alpha_l}{|\Delta_l|} \sqrt{\frac{\gamma_l^+}{2}}$ ,  $\operatorname{erf}(x)$  is the error function,  $\operatorname{erfc}(x)$  define the complementary error function.

### 3.1.2 Normalization

Referring to (3.25) and (3.26), we can now write a k-variant function as

$$\check{f}_k = \begin{cases} [P^{-1}(A) - 1]^{|k|} & , k < 0 , \\ 1 & , k = 0 , \\ [P^{-1}(B) - 1]^k & , k > 0 , \end{cases} \quad (3.31)$$

which turns out to be independent on  $N$ . Further normalization of  $\check{f}_k$  to have a unit area leads to the pdf  $p(k) = \frac{1}{\phi} \check{f}_k$ , where  $\phi$  is the sum of the values of  $\check{f}_k$  for all k,

$$\phi = 1 + \sum_{k=-1}^{-\infty} [P^{-1}(A) - 1]^{|k|} + \sum_{k=1}^{\infty} [P^{-1}(B) - 1]^k \quad (3.32)$$

$$= 1 + \sum_{k=-1}^{-\infty} \phi_{1k} + \sum_{k=1}^{\infty} \phi_{2k}, \quad (3.33)$$

where  $\phi_{1k} = [P^{-1}(A) - 1]^k$  and  $\phi_{2k} = [P^{-1}(B) - 1]^k$ . Because  $0.5 < P = P(A), P(B) < 1$ , we obtain  $\ln P < 0$ ,  $\ln P(1 - P) < 0$  and  $\ln(1 - P) < 1 - P$ . This allows us to find

$$\phi_{1k} = e^{-k|\ln(1-P(A))-\ln(P(A))|}, \quad (3.34)$$

$$\phi_{2k} = e^{-k|\ln(1-P(B))-\ln(P(B))|}. \quad (3.35)$$

By substituting equations (3.34) and (3.35) into (3.33) and expressing the summations a in short form as

$$\phi = 1 + \frac{1}{e^{|\ln \frac{1-P(A)}{P(A)}|} - 1} + \frac{1}{e^{|\ln \frac{1-P(B)}{P(B)}|} - 1}. \quad (3.36)$$

Given the conditions  $0.5 \leq P = P(A), P(B) < 1$ ,  $\frac{1-P(A,B)}{P(A,B)}$ , and  $\ln \left( \frac{1-P(A,B)}{P(A,B)} \right) \leq 0$ , can be represented equation (3.36) in terms of  $P(A)$  and  $P(B)$  as

$$\begin{aligned} \phi &= 1 + \frac{1}{\left(\frac{1-P(A)}{P(A)}\right)^{-1} - 1} + \frac{1}{\left(\frac{1-P(B)}{P(B)}\right)^{-1} - 1} \\ &= 1 - \frac{1-P(A)}{1-2P(A)} - \frac{1-P(B)}{1-2P(B)} \\ &= \frac{-P(A)}{1-2P(A)} - \frac{1-P(B)}{1-2P(B)} \\ &= \boxed{\frac{P(A) + P(B) - 1}{[1-2P(A)][1-2P(B)]}}. \end{aligned} \quad (3.37)$$

The jitter pdf  $p(k)$  becomes now

$$p(k) = \frac{1}{\phi} \begin{cases} [P^{-1}(B) - 1]^k & , k > 0, \\ 1 & , k = 0, \\ [P^{-1}(A) - 1]^{|k|} & , k < 0, \end{cases} \quad (3.38)$$

where  $\phi$  is specified by (3.37). Because (3.38) depends only on  $k$  points around the breakpoint and is uninfluenced by any other point, the probability is complete and we called it the *jitter probability*.

Introducing new variables  $d_l = P^{-1}(A) - 1$  and  $q_l = P^{-1}(B) - 1$ , we can write  $P(A) = 1(1 + d_l)$  and  $P(B) = 1(1 + q_l)$ , provide the transformations, and arrive at the conclusion that (3.38) is the discrete skew Laplace pdf recently shown in [59],

$$p(k|d_l, q_l) = \frac{(1-d_l)(1-q_l)}{1-d_l q_l} \begin{cases} d_l^k, & k \geq 0, \\ q_l^{|k|}, & k \leq 0, \end{cases} \quad (3.39)$$

where  $0 < d_l = e^{-\frac{\kappa_l}{\nu_l}} = P(B_l)^{-1} - 1 < 1$ ,  $0 < q_l = e^{-\frac{1}{\kappa_l \nu_l}} = P(A_l)^{-1} - 1 < 1$ ,  $\kappa_l = \sqrt{\frac{\ln x_l}{\ln(x_l/\mu_l)}}$ ,  $\nu_l = -\frac{\kappa_l}{\ln x_l}$ , and

$$x_l = \frac{\phi_l(1 + \mu_l)}{2(1 + \phi_l)} \left( 1 - \sqrt{1 + \frac{4\mu_l(1 - \phi_l^2)}{\phi_l^2(1 + \mu_l)^2}} \right), \quad (3.40)$$

$$\mu_l = \frac{P(A_l)[1 - P(B_l)]}{P(B_l)[1 - P(A_l)]}, \quad (3.41)$$

$$\phi_l = \frac{P(A_l) + P(B_l) - 1}{[1 - 2P(A_l)][1 - 2P(B_l)]}. \quad (3.42)$$

A complete analysis of  $\kappa$ ,  $\nu$  and equations (3.40) to (3.42) is provided in Appendix B.

### 3.1.3 Distribution verification by simulation

The discrete skew Laplace pdf (3.39) computed using (3.29) and (3.30) is shown in Fig. 3.4 for  $\gamma_l = 2.78$  and  $\gamma_l^+ = 6.25$ . It is seen that jitter may occur here at 7 points (five to the left and two to the right from  $k = 0$ ) allowing the jitter probability of 1%. With smaller SNR values, the jitter increases.

To find out how the skew Laplace distribution fits real measurements, a piecewise constant (PWC) signal with 200 points has been generated. The breakpoint is located at the sample 101 with known  $a_l = 1$  and  $a_{l+1} = 0$ . The SNRs were set the same values as in Fig. 3.4. Then, using a ML estimator [28], the breakpoint locations were found over  $10 \times 10^4$  realizations, plotted with a histogram, normalized to have unit area, and represented as the jitter pdf. To avoid ripples, we further repeated this procedure 9 times and averaged the results. The final jitter pdf circled in Fig. 3.4 has revealed some discrepancy with the Laplace distribution. This is explained by some probabilities ignored in the derivation of (3.38). The maximum approximation error of about 4% is seen here only at  $k = 0$ . The error diminishes with  $k$  that however cannot be seen in logarithmic measures in (3.4). Note that (3.39) also fits the experimental histograms reported for some CNAs breakpoints in [60]. Thus, we come up with a conclusion that the discrete skew Laplace distribution is a good approximation for jitter in the breakpoints of CNAs, although further investigations are necessary in order to find a more correct law.

Because a step is unity with integer  $k$  the jitter probability  $P_k(\gamma_i^-, \gamma_i^+)$  can be specified at the  $k$ th point utilizing (3.39) as

$$P_k(\gamma_i^-, \gamma_i^+) = p[k - |d(\gamma_i^-, \gamma_i^+), q(\gamma_i^-, \gamma_i^+)|]. \quad (3.43)$$

The jitter probability (3.43) is determined for  $i$ th breakpoint at zero assuming  $k \leq 0$ . To move (3.43) to  $i_l > 0$ , one can substitute  $k$  with  $ki_l$  and change  $k$  as  $k \leq i_l$ . Figure 3.5 sketches  $P_k(\gamma)$  for small and large equal SNRs  $\gamma = \gamma_i^- = \gamma_i^+$  affected by somatic alterations. Here, we show the probability of a jitter-free breakpoint (dashed) along with some values obtained by simulation in the SNR region most typical for the microarray

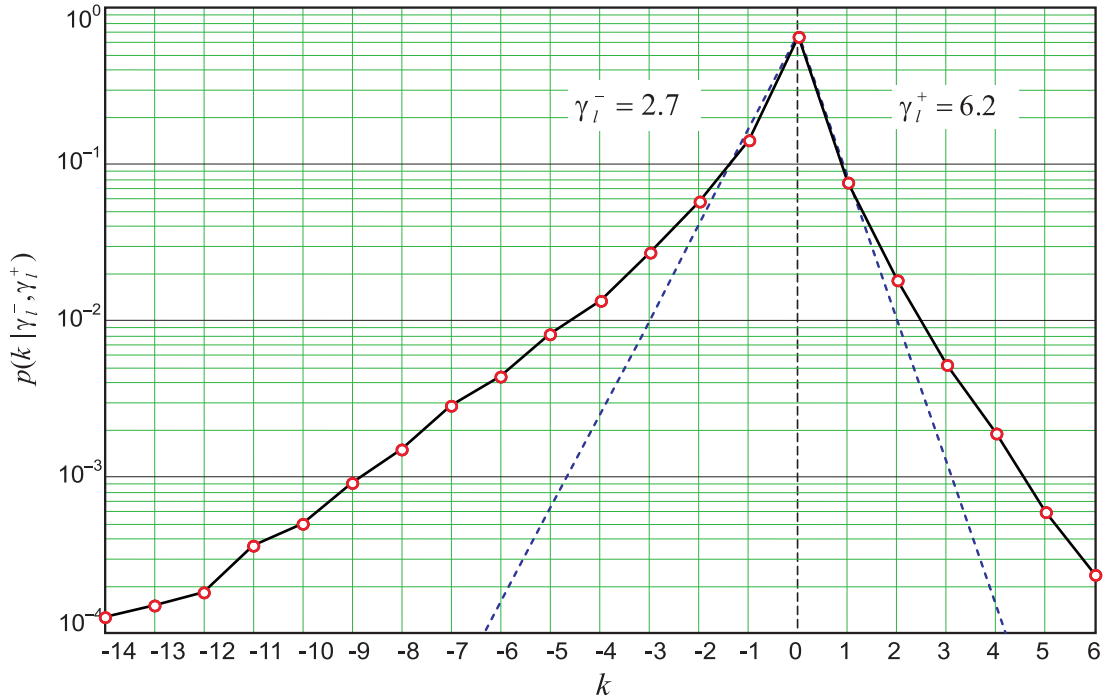


Figure 3.4: The discrete skew Laplace pdf (dashed) for different segmental SNRs;  $k = 0$  corresponds to  $i_l$ . By simulation, the breakpoint locations were found using a ML estimator over  $10 \times 10^5$  runs. Measurements and estimations were repeated 9 times. An average jitter pdf is circled. A discrepancy between the curves is due to some probabilities ignored while deriving the jitter distribution. The maximum approximation error is about 4% at  $k = 0$ . The difference between the functions is almost indistinguishable in linear scales.

measurements. As can be seen, (3.39) fits well the simulated data and we may continue on with some analysis.

The SNRs are extremely small if structural changes are present in small fraction of the tested cells and the value of  $\Delta$  is minimal, then the total jitter probability  $P_J(\gamma_i^-, \gamma_i^+) = P(\mathbf{A}_i \mathbf{B}_i)$  becomes almost unity, whereas the probability of the jitter-free breakpoint becomes almost zero. More precisely, the probability of jitter-free breakpoint and the jitter probability at  $k$ th point tend toward  $1/2N$  (see that the curves merge at  $\gamma_i = \gamma_i^+ = 10^{-2}$  in Fig. 3.43. All these tendencies are illustrated in Fig. 3.43.

With an increase in structural changes, the segmental SNRs also increase, and the probability of the jitter-free breakpoint (dashed in Fig. 3.5) naturally becomes larger and finally reaches unity when  $\gamma_i^- \rightarrow \infty$  and  $\gamma_i^+ \rightarrow \infty$ . It is seen in Fig. 3.5 that  $\gamma_i^- = \gamma_i^+ = 40$  makes the breakpoint jitter-free in the 3-sigma sense. With an increase in the SNRs from zero, the jitter probability initially increases. It then reaches a maximum and decreases, when the SNRs become relatively large. It also follows from Fig. 3.5 that the maximum jitter probability at  $k = 1$  corresponds to about unit SNRs. This case can be met most frequently in the microarrays of CNAs measurements within a typical SNR range of  $0.1, \dots, 100$  which follows from [61, 62, 63, 64, 65, 17, 66, 19, 60, 51, 67, 68, 69].

## 3.2 Confidence UB and LB Masks

The confidence masks is an algorithm of local segmentation profiles with confidence probabilities if noise in segments and jitter in the breakpoints are statistically specified [70]. The research made in [16, 30] has shown that the confidence masks become a unique tool to verify validity of the estimated CNAs with a given probability. Thus, the confidence boundaries should be provided as additional information for medical experts to give an accurate diagnostic. The confidence masks are based on the fundamental principle of *jitter distribution*, which is used to compute the Jitter and Segmental bounds.

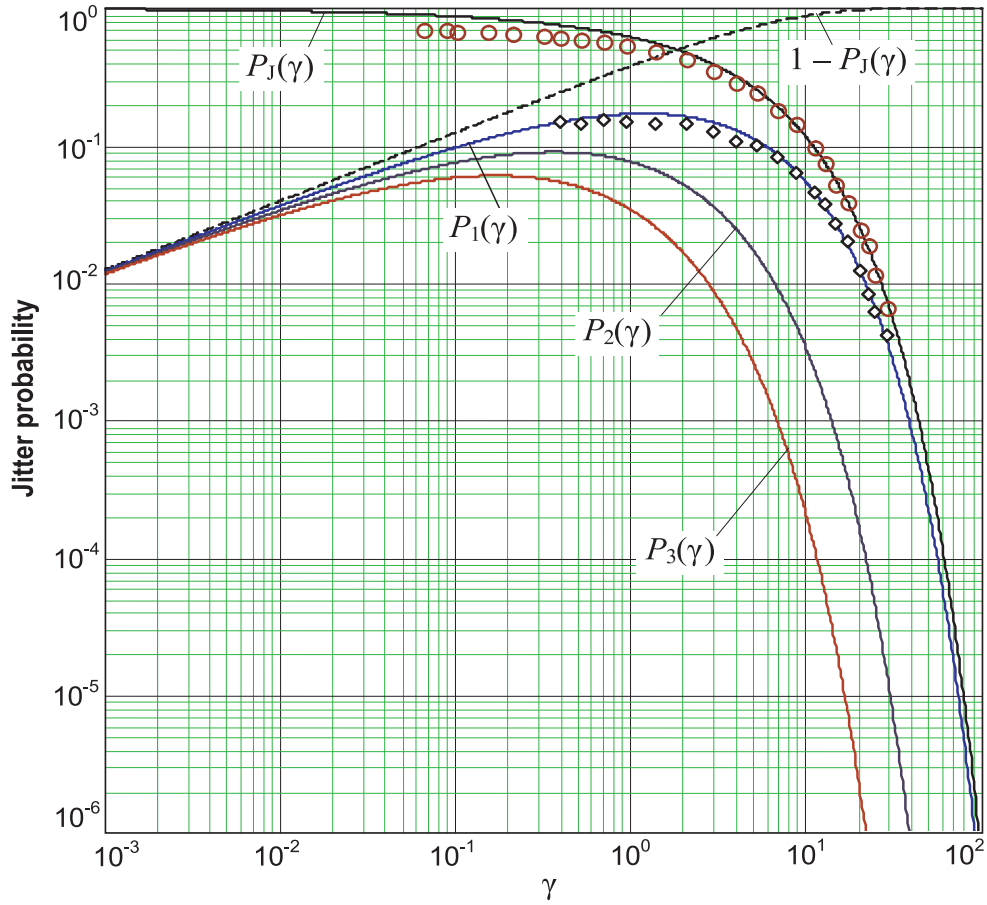


Figure 3.5: Total jitter probability  $P_J(\gamma)$  and probabilities  $P_k(\gamma)$  of the right jitter at  $k = 1$ ,  $k = 2$ , and  $k = 3$  for equal SNRs in the CNAs segments. The probability of a jitter-free breakpoint is dashed and some values obtained by simulation are depicted with diamond  $\diamond$  and circles  $\circ$ . A discrepancy between the theory and measurement is due to some probabilities ignored while deriving the jitter distribution.

### Jitter Bounds

The Left Jitter Bound  $J_l^L$  (LJB) and the right jitter bound Right Jitter Bound  $J_l^R$  (RJB) can be determined with respect to the  $l$ th breakpoint  $\hat{i}_l$  as follows. Consider the jitter distribution (3.39) for known  $\gamma_l^-$  and  $\gamma_l^+$ . Increase  $k$  in (3.39) from zero until  $p_k < \xi, \%$ . Accept the relevant value of  $k$  as the right jitter  $k_l^R$ . Next, reduce  $k$  from zero until  $p_k < \xi, \%$  and accept the relevant value of  $k$  as the left jitter  $k_l^L$ . Form the LJB and RJB



as

$$J_l^L \cong \hat{n}_l - k_l^R, \quad (3.44)$$

$$J_l^R \cong \hat{n}_l + k_l^L. \quad (3.45)$$

The equations to compute  $k_l^R$  and  $k_l^L$  are obtained by equating (3.39) to  $\xi$  and solving for  $k_l$ . Then, to find the right jitter  $k_l^R$  with a constant value of  $\xi$ , we set the next steps to  $k \geq 0$  as

$$\xi \leq \frac{(1-d_l)(1-q_l)}{1-d_lq_l} p^k = \frac{(1-d_l)(1-q_l)}{1-d_lq_l} e^{-\frac{\kappa}{\nu}k}, \quad (3.46)$$

$$e^{\frac{\kappa}{\nu}k} \leq \frac{(1-d_l)(1-q_l)}{\xi(1-d_lq_l)}, \quad (3.47)$$

$$\frac{\kappa}{\nu}k \leq \ln \frac{(1-d_l)(1-q_l)}{\xi(1-d_lq_l)}, \quad (3.48)$$

$$k \leq \frac{\nu}{\kappa} \ln \frac{(1-d_l)(1-q_l)}{\xi(1-d_lq_l)}. \quad (3.49)$$

Applying the same procedure to  $k_l^L$  and  $k \leq 0$ , we obtain

$$\xi \leq \frac{(1-d_l)(1-q_l)}{1-d_lq_l} q^{|k|} = \frac{(1-d_l)(1-q_l)}{1-d_lq_l} e^{-\frac{|k|}{\kappa\nu}}, \quad (3.50)$$

$$e^{\frac{|k|}{\kappa\nu}} \leq \frac{(1-d_l)(1-q_l)}{\xi(1-d_lq_l)}, \quad (3.51)$$

$$\frac{|k|}{\kappa\nu} \leq \ln \frac{(1-d_l)(1-q_l)}{\xi(1-d_lq_l)}, \quad (3.52)$$

$$k - \kappa\nu \leq \frac{\nu}{\kappa} \ln \frac{(1-d_l)(1-q_l)}{\xi(1-d_lq_l)}. \quad (3.53)$$

Thus,  $k_l^R$  and  $k_l^L$  can be defined as, respectively,

$$k_l^R = \left\lfloor \frac{\nu_l}{\kappa_l} \frac{1-d_l}{\xi(1-d_lq_l)} \right\rfloor, \quad (3.54)$$

$$k_l^L = \left\lfloor \frac{1}{\nu_l \kappa_l} \frac{1-d_l}{\xi(1-d_lq_l)} \right\rfloor, \quad (3.55)$$

to correspond to the right (superscript R) and the left (superscript L) jitter. Here,  $[x]$  means a maximum integer lower than or equal to  $x$ . Here, equal confidence intervals are allowed for thesegments and breakpoints.

### Segmental Bounds

Let us suppose that the estimate  $\hat{n}_l$  of the  $l$ th breakpoint location is available (see figure 3.2, at least it can be assigned visually. In view of white Gaussian noise in measurement  $y_v$ , simple averaging applied on an interval of  $N_l = \hat{n}_l - \hat{n}_{l-1}$  points, from  $\hat{a}_{l-1}$  to  $\hat{n}_l - 1$ , gives the best estimate for the  $l$ th segmental level:

$$\hat{a}_l = \frac{1}{N} \sum_{v=\hat{n}_{l-1}}^{\hat{n}_l-1} y_v, \quad (3.56)$$

which mean  $E\hat{a}_l = a_l$  and the variance  $\hat{\sigma}_l^2 = (\sigma_l^2/N_l)$ . Because  $\hat{\sigma}_l^2$  is commonly not negligible, segmental errors occur. The confidence UB and LB for segmental estimates can thus be specified in the  $\vartheta$ -sigma sense as:

$$\hat{a}_j^{\text{UB}} \cong \hat{a}_j + \epsilon = \hat{a}_j + \vartheta \sqrt{\frac{\sigma_j^2}{N_j}} = \hat{a}_j + \vartheta \hat{\sigma}_j, \quad (3.57)$$

$$\hat{a}_j^{\text{LB}} \cong \hat{a}_j - \epsilon = \hat{a}_j - \vartheta \sqrt{\frac{\sigma_j^2}{N_j}} = \hat{a}_j - \vartheta \hat{\sigma}_j, \quad (3.58)$$

where  $\vartheta$  indicates the bound wideness in terms of  $\hat{\sigma}_j$ . The probability  $\xi$  for the segmental estimate to exceed a threshold  $\epsilon$  strongly depends on the segmental length  $N_j$  and can be determined using equation (3.39) and  $\hat{a}_l$ , as

$$\xi(N_l) = 2 \int_{a_l + \epsilon}^{\infty} p_l(x) dx = \text{erfc} \left( \mu_l \sqrt{\frac{N_l}{2}} \right), \quad (3.59)$$

where  $\text{erfc}(x)$  is the complementary error function and  $\mu_l = (\epsilon/\sqrt{\sigma_l^2})$  is the normalized threshold. A distinctive feature of  $\xi$  is that it does not depend on the unknown  $a_l$ . By combining  $\epsilon$  and  $\mu_l$  in  $\xi(N_l)$  we obtain  $\xi(N_l) = \text{erfc}(\vartheta/\sqrt{2})$  and the confidence interval for segmental estimate becomes

$$P(N_l) = 1 - \xi(N_l) = 1 - \operatorname{erfc}\left(\frac{\vartheta}{\sqrt{2}}\right). \quad (3.60)$$

Table 3.1 gives several values of  $\vartheta$ ,  $P$ , and  $\xi$  for likely existing genomic changes (50%). As can be seen, the 1-sigma sense ( $\vartheta = 1$ ) occupies an intermediate position between the 50% probability (even chances) and 75% probability (probably existing changes). Here-with, the 2-sigma sense ( $\vartheta = 2$ ) can be treated as typical or almost certainly existing changes and 3-sigma ( $\vartheta = 3$ ) as certainly existing changes [16].

Table 3.1: Probabilistic measures for genomic changes

	$\vartheta$	$P(\%)$	$\xi(\%)$
Even chances	0.6745	50	50
1-Sigma	1	68.27	31.73
Probable	1.15035	75	25
Almost certain	1.81191	93	7
Typical confidence	1.96	95	5
2-Sigma	2	95.45	4.55
3-Sigma	3	99.73	0.27
Certain	$\infty$	100	0

By combining (3.44), (3.45), (3.57), and (3.58), the UB and LB confidence masks outline the region of existence for true CNAs. The algorithm for computing the Upper Bound mask  $\mathcal{B}_n^U$  (UB mask) and Lower Bound mask  $\mathcal{B}_n^L$  (LB mask) is developed in Appendix B Table C.1 and Table C.2 [71, 72].

Figure 3.6 shows an example of the probabilistic confidence masks applied to a simulated CNA from microarray data for different probabilities  $P$  taken from Table 3.1. As can be seen, the breakpoints are localized in the 1-sigma sense ( $P = 68.27\%$ ) with no errors, but the segmental level is detected with an error of about  $\pm 20\%$ . Admitting that the confidence probability of  $P = 68.27\%$  may not be sufficient for medical decisions, we

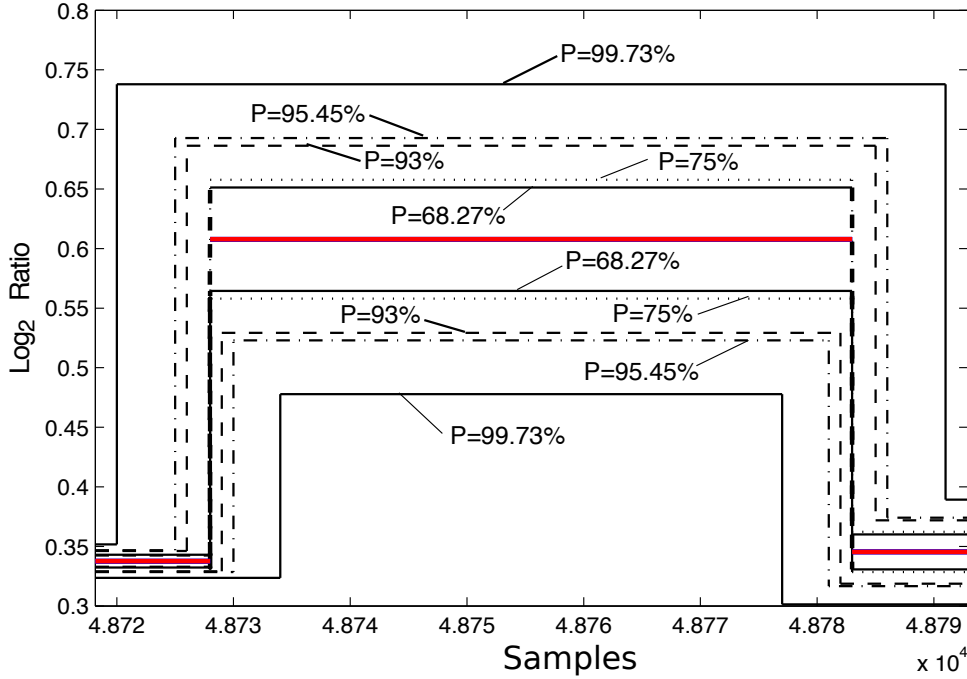


Figure 3.6: An example of UB mask  $\mathcal{B}_n^U$  and LB mask  $\mathcal{B}_n^L$  around the simulated CNA for confidence probabilities taken from Table 3.1.

apply the masks in the 3-sigma sense ( $P = 99.73\%$ ) and observe that the breakpoints can no longer be detected exactly and the segmental errors increase to about  $\pm 50\%$ . The CNA evidently exists in this case and there is a necessity of defining an exact value of  $P$  which is sufficient for medical needs.

### 3.2.1 Testing real measurements by the probabilistic confidence masks.

The first database processed is part of the 7th chromosome in archive "159A-vs-159D-cut of ROMA. It is shown to have 14 segments and 13 breakpoints (Figs. 3.7a and 3.7b). Observing Fig. 3.7a, the only breakpoint which location can be estimated with a high accuracy is  $i_1$ . Jitter in  $\hat{i}_6$  and  $\hat{i}_7$  is moderate. All other breakpoints have large jitter. It is seen that the UB mask covering 2nd-to-6th segments is almost uniform. Thus, there

is a probability that the 2nd–to–5th breakpoints do not exist. If to follow the LB mask, then locations of the 2nd–to–4th breakpoints can be predicted even with large errors. At least they can be supposed to exist. However, nothing definitive can be said about the 5th breakpoint and one may suppose that it does not exist. It is also hard to distinguish a true location of the 8th breakpoint. In Fig. 3.7b,  $i_{10}$ ,  $i_{12}$ , and  $i_{13}$  are well detectable owing to large segmental SNRs. The breakpoint  $i_9$  has a moderate jitter. In turn, the location of  $i_{11}$  is unclear. Moreover, there is a probability that  $i_{11}$  does not exist.

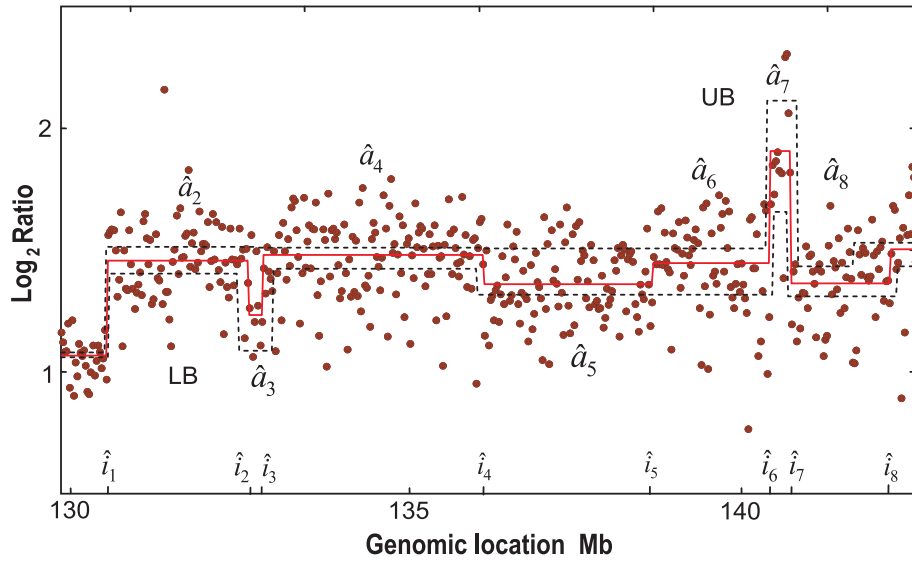
### 3.3 Limitation of Laplace-based Approximation

As it was described previously in [30], Jitter is inherent to measurements affected by intensive noise in all CNA’s breakpoints. When  $(\gamma_i^-$  or  $\gamma_i^+$  are lesser than 1, the jitter distribution is approximated with the discrete skew Laplace distribution [30]. If  $(\gamma_i^-, \gamma_i^+) < 1$ , an actual breakpoint may occur several points to the left or to the right of the candidate one detected by an estimator. Subtle chromosomal changes are often observed with  $(\gamma_i^-, \gamma_i^+) \ll 1$  and, for the required high confidence probability, the actual breakpoint can be found tens of points apart from the candidate one.

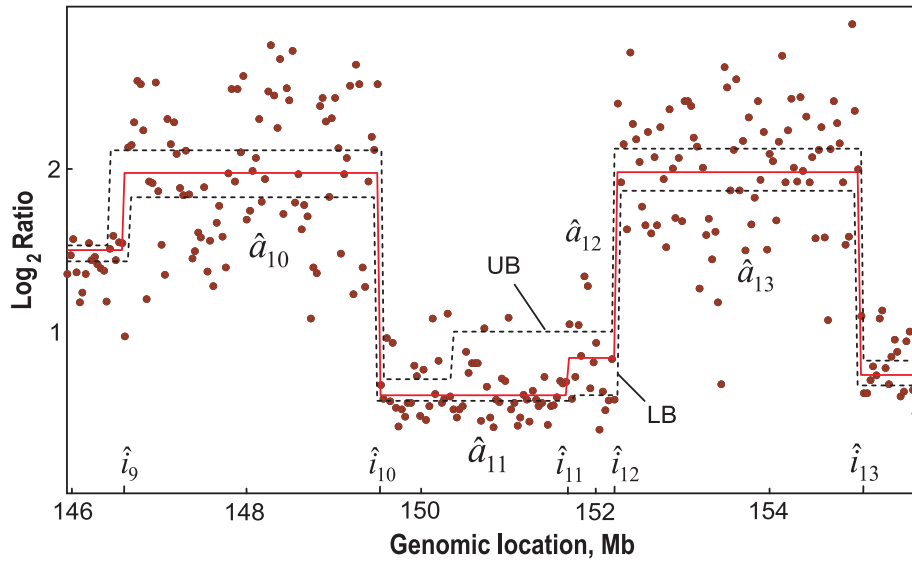
An extensive analysis has confirmed that the Skew Laplace distribution becomes highly inefficient when subtle CNAs reveal a SNR level much less than 1  $(\gamma_i^-, \gamma_i^+) \ll 1$  [58]. The research of SkL pdf (3.39) in applications to jitter in the CNA-like signals measured in WGN allowed making the following statements. The SkL-based approximation is

- Acceptable when  $\gamma_i^-, \gamma_i^+ > 1$  and very accurate if  $\gamma_i^-, \gamma_i^+ \gg 1$ ;
- Also acceptable if at least one of the SNRs exceeds unity,  $\gamma_i^- > 1$  or  $\gamma_i^+ > 1$ , and very accurate if  $\gamma_i^- \gg 1$  or  $\gamma_i^+ \gg 1$ ;
- Inaccurate when  $\gamma_i^-, \gamma_i^+ < 1$  and unacceptable if  $\gamma_i^-, \gamma_i^+ \ll 1$ .

An overall conclusion that can be made following [57, 73, 58] is that the SkL-based approximation (3.39) fits only easily seen breakpoints. The chromosomal changes are not brightly pronounced, the SkL should not be used to make decisions about the CNAs



(a)



(b)

Figure 3.7: Median-based denoising of the microarray measurement: (a) subsequent smoothing of  $\log_2$  Ratio with  $N = 11, N = 21, N = 41,$  and  $N = 71$  and (b) threshold-based forming of a rectangular pulse train.

structures [16, 30]. Therefore a more accurate approximation is required to avoid a wrong behavior of this tool.

# Chapter 4

## Improving Jitter

Previously in Chapter 3, we have derived the discrete skew Laplace function to represent the jitter distribution in the CNAs breakpoints. However, the SkL approximation has appeared to be inaccurate when  $\gamma_l^-, \gamma_l^+ < 1$  and unacceptable if  $\gamma_l^-, \gamma_l^+ \ll 1$ . In this chapter, the errors caused by SkL approximation are analyzed based on the experimental distribution following the procedure described below. Also, several approximations are proposed to fit the jitter distribution with the minimum error following three strategies. First, a heuristic approximation is developed, second a theory to parametrize the SkL distribution is proposed, and finally, a mathematical approximation is applied to fit the experimental jitter distribution.

### 4.1 Experimental Jitter Histogram

To find the experimental jitter histogram, we set an ideal CNA  $x_n$  of length  $n = 400$  with two constant levels  $a_l = 1$  and  $a_{l+1} = 0$  and one breakpoint  $i_l$  at  $n = 200$ . The CNA measurement is defined as  $y_n = x_n + v_n$ , where  $v_n$  is a vector of noise WGN with the variance  $\sigma^2$  corresponding to the given  $\gamma$ .

We change the breakpoint position  $i_l$  at  $n = 200$  in a range of  $n \pm 100$  or  $-100 \leq k \leq 100$  producing one estimate in each point. The variable  $\hat{y}_i$  represent the CNA estimated, which is defined as

$$\hat{y}_{n|\hat{i}} = [\bar{y}_{1:n-|2k|}, \bar{y}_{n-|2k|+1:n}] \quad (4.1)$$

where  $\bar{y}$  is the mean of measurements in a given lapse conditioned by the breakpoint  $\hat{i}$ . Next, we use the Ordinary Least Squares (OLS) method to find the CNA estimation  $\hat{y}_{n|\hat{i}}$  with minimal error of the measurement  $y_n$  obtaining the error variable  $S_k$ . The breakpoint location is detected when the Mean Square Error (MSE)  $S_k$  in the ML estimate reaches a minimum. Then, the vector that define the jitter histogram  $H_k$  is increased one unit at a specific position  $k$  based on the best CNA estimation. The calculated histogram  $H_k$  is normalized to produce the discrete jitter pdf. This procedure is summarized in Figure 4.1.

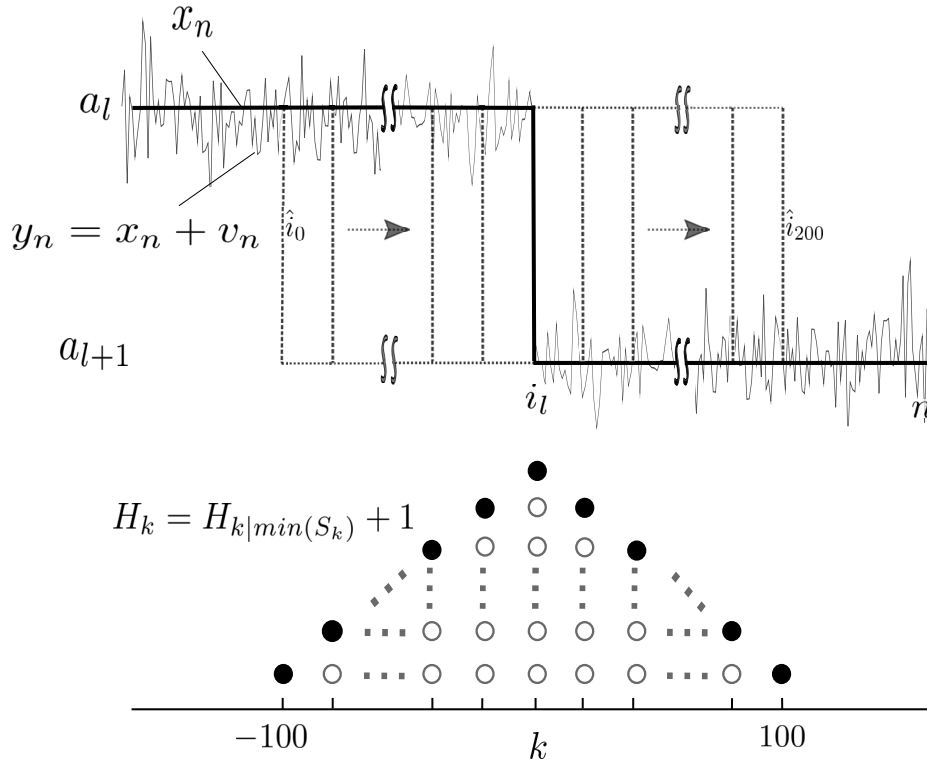
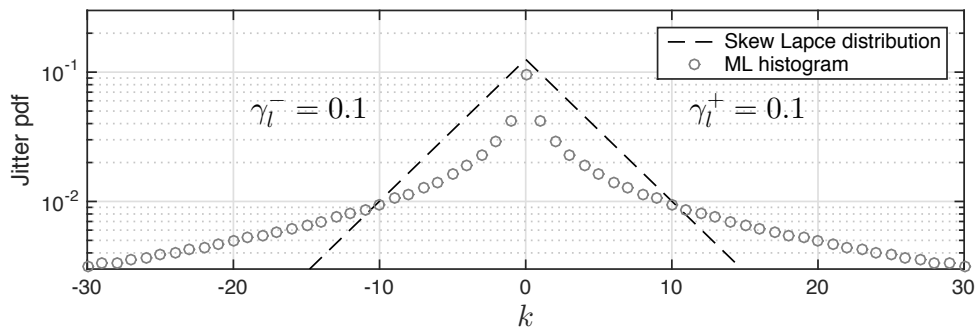


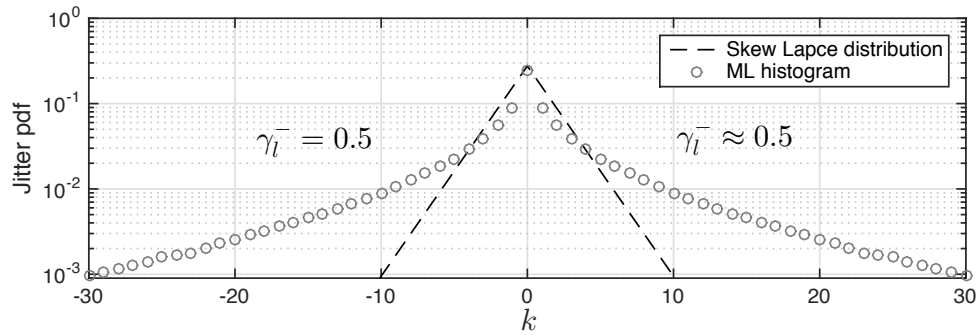
Figure 4.1: Procedure to approximate the jitter distribution in the CNA breakpoints by simulating a stepwise signal in the presence of AWGN with different segmentsl SNRs. The breakpointn  $i_l$  change its position from  $\hat{i}_0$  to  $\hat{i}_{200}$  seeking the best CNA estimate.



In the first simulation, the ML estimate was provided by  $50 \times 10^3$  times averaging of each vector of noise generated with a constant SNR. For each value of SNR, the histogram was plotted as a number of the events in the  $k$  scale. To smooth ripples, such a step was done in 9 runs and the estimates were averaged. To be accepted as an experimentally defined *jitter pdf*, the histogram obtained was normalized for a unit area as shown in Fig. 4.2a and Fig. 4.2b with circles to SNR  $\gamma_l = 0.1$  and  $\gamma_l = 0.5$ , respectively.



(a)



(b)

Figure 4.2: Jitter distributions computed with Maximum Likelihood and Skew Laplace distribution to a) SNR=0.1 and b) SNR =0.5. The ML (circled) is the jitter pdf obtained experimentally using a ML estimator via a histogram over  $50 \times 10^3$  runs and SkL (solid) is the skew Laplace distribution.

At the second stage, the MATrix LABoratory (MATLAB)-based algorithm [71] was run for the simulation using a computer based on Intel Core i5, 2.5 GHz. The computation

time required to produce a histogram was about 12.7 hours. To make it possible to operate faster, in [74] the algorithm was modified removing “for” cycles and did not save variables in Random Access Memory (RAM) memory, increasing the times averaging to  $10 \times 10^5$  for each SNR. Thereby, the computation time was significantly reduced and the jitter histogram computed with a improved accuracy in a wide range of  $k$ .

The left part of (Fig. 4.3, A1) is a flowchart of the procedure described above and illustrated in Figure 4.1 that allows getting faster the jitter histogram. Comparing the algorithm A0 designed in [71] and the modified algorithm A1, it was found that A0 consumes more time and that A1 is computationally more efficient. In average, the algorithm A1 operates about 28 times faster than A0 and requires about 27 min to produce one histogram. The simulated one-sided jitter distributions provided by the sub-algorithm A1 for equal segmental values of SNR are shown in Fig. 4.4.

Referring to the necessity of estimating the CNAs with low segmental SNRs [12] and taking into account that the Laplace distribution (3.39) is sufficiently accurate when the SNR values exceed unity [13, 75], so the jitter is investigated in the region of  $0.1 \leq \gamma_i^- = \gamma_i^+ \leq 1.37$ . As can be seen in Fig. 4.4, a decrease in the SNRs makes the actual jitter distribution less straight in the logarithmic scale and the SkL has thus limited applications for low segmental SNRs.

Subsequently, the resolution of each iteration was enhanced to generate the noise step from 4 to 9 decimal values, short and long formats, respectively, to diminish the accumulated error. Figure 4.5 shows the difference between the generated histograms using different formats in simulation. The detailed algorithm generate a complete jitter histogram of symmetric values of SNR  $\gamma_i^- = \gamma_i^+$  from 0.1 to 1.37 illustrated in Figure 4.4. It should be noted that the jitter pdf expands significantly its hedge with respect to  $k$  when  $\gamma_i < 1$ .

In summary, let us notice that the three stages are implemented to simulate jitter histogram according to averaging simulations:  $50 \times 10^3$ ,  $10 \times 10^4$  and  $10 \times 10^4$  with long

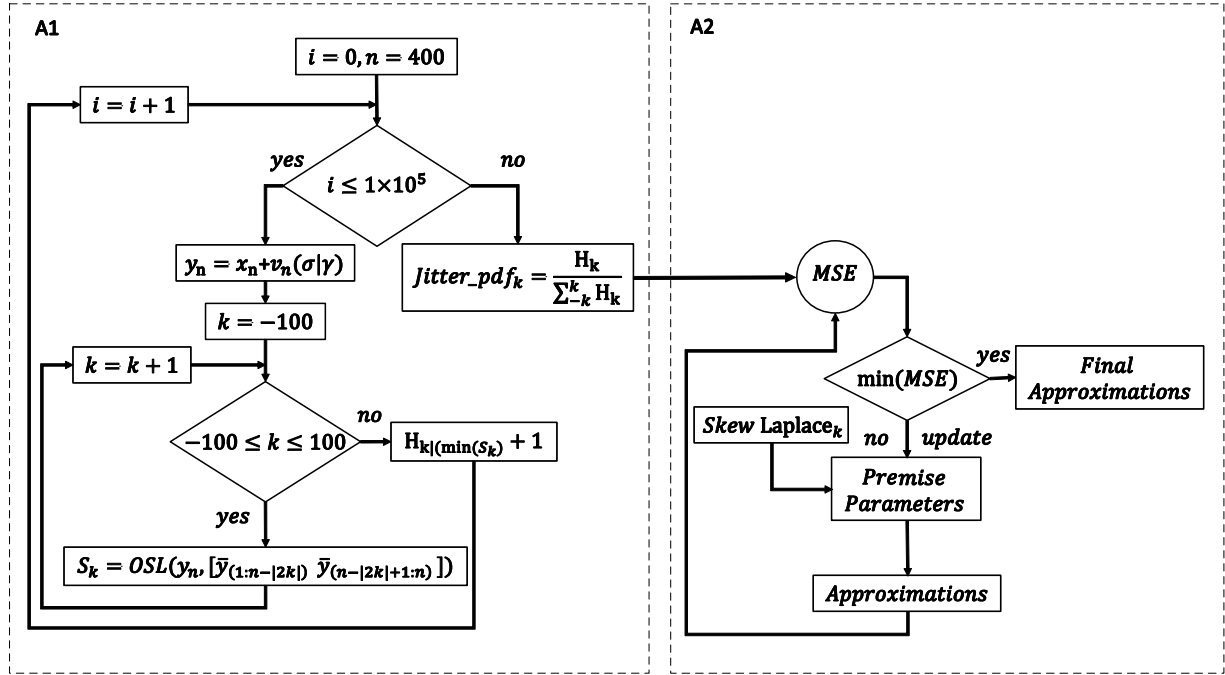


Figure 4.3: A flowchart to approximate the jitter distribution in the CNA breakpoints by simulating a stepwise signal in the presence of AWGN with different segmentsl SNRs: A1 provides the jitter histogram and A2 provides the jitter distribution approximation by minimizing the MSE. An example of signal  $y_n$  is given in Fig. 4.1.

format –which are labeled as *slow*, *fast* and *detailed* algorithms.

## 4.2 Approximations of jitter pdf

In this section, several methodologies are proposed to obtain a more accurate approximations of the jitter distribution than the Skew Laplace distribution. First, an heuristic approximation is developed based on the Bessel functions of the second kind of zeroth order. This approximation requires the estimation of constants and functions using the MSE with respect to the histograms generated with a *slow* algorithm used as a reference. Next, the parametrization of the SkL distribution is implemented to obtain a set of approximations using different functions, which modify the constant behavior of  $\sigma$ . The approximations obtained in such a way fit the measurements of a *fast* algorithm. Finally,

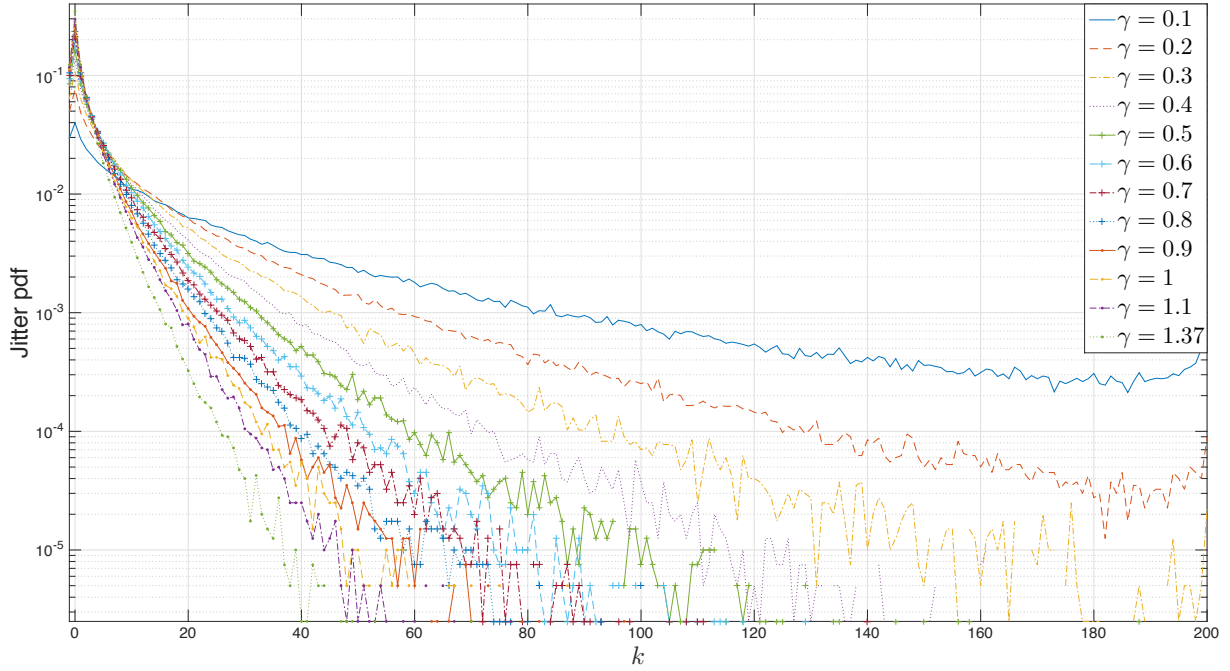


Figure 4.4: Experimentally defined one-sided jitter probability densities (dotted) of the breakpoint location for equal segmental SNRs  $\gamma$  in the range of  $M = 400$  points with a true breakpoint at  $n = 200$ . The experimental density functions were found using the ML estimator. The histogram was built over over  $10 \times 10^5$  runs repeated 10 times and averaged.

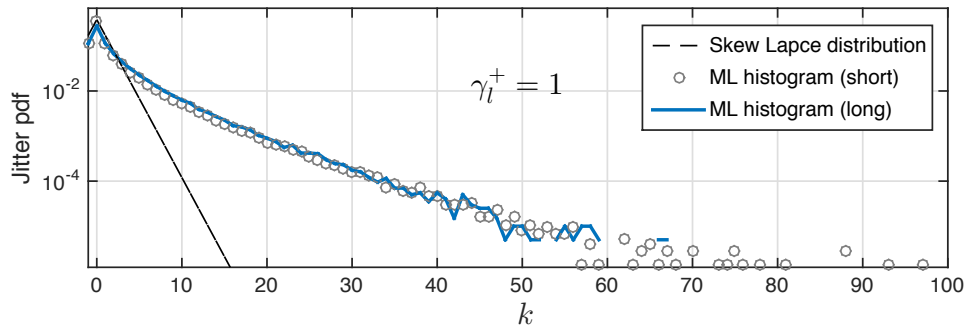


Figure 4.5: Difference between experimental distributions obtained using a format of 4 (solid) and 9 (circles) decimal values.

an Asymmetric Exponential Power Distribution is applied to fit the experimental jitter distribution obtained with a more accurate algorithm *detailed* based on the estimation of few constants.

### 4.2.1 Heuristic Approximation

A preliminary analysis has shown that, among the available special functions, the modified Bessel function of the second kind  $K_0(x)$  and zeroth order is a good candidate to fit the experimentally jitter measured. Figure 4.6 show some examples of the modified Bessel functions of the second kind  $K_\alpha(x)$  for  $\alpha = 0, 1, 2, 3$  and 4.

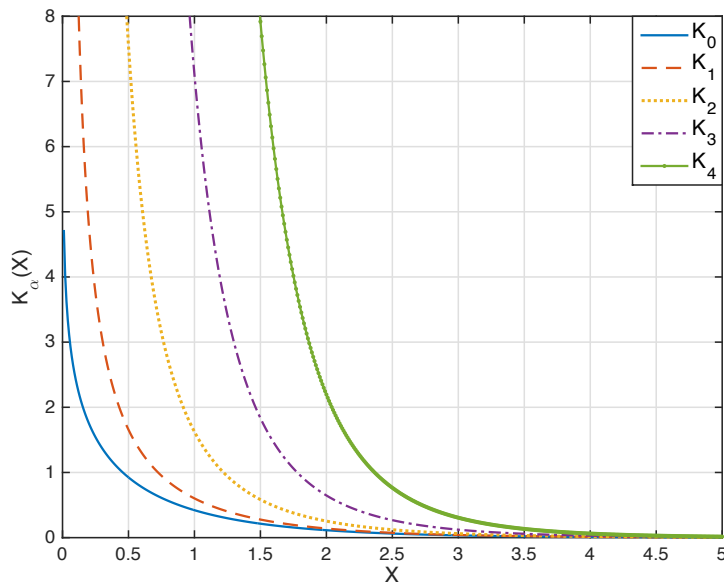


Figure 4.6: Modified Bessel functions of the second kind,  $K_\alpha(x)$ , for  $\alpha = 0$  (solid), 1(dashed), 2 (dotted), 3 (dash-dotted) and 4 (solid-pointed).

For the function proposed, the measured densities were obtained with a *slow* method and shown in Fig. 4.7. In our approximation, we use the following form of  $K_0(x)$ ,

$$\begin{aligned}
K_0[x(k)] &= \int_0^{\infty} \cos[x(k) \sinh t] dt \\
&= \int_0^{\infty} \frac{\cos[x(k)t]}{\sqrt{t^2 + 1}} dt > 0, x(k) > 0,
\end{aligned} \tag{4.2}$$

in which variable  $x(k)$  depends on index  $k$  which represents a discrete departure from the assumed breakpoint location (see Fig. 4.1). The equation (4.2) describes decaying functions and diverges at  $x = 0$  with the singularity being of logarithmic type [76]. Because  $K_0[x(k)]$  is a positive-valued for  $x(k) > 0$  smooth function decreasing with  $x$  to zero, this is used to approximate the measured probability densities shown in Fig. 4.7.

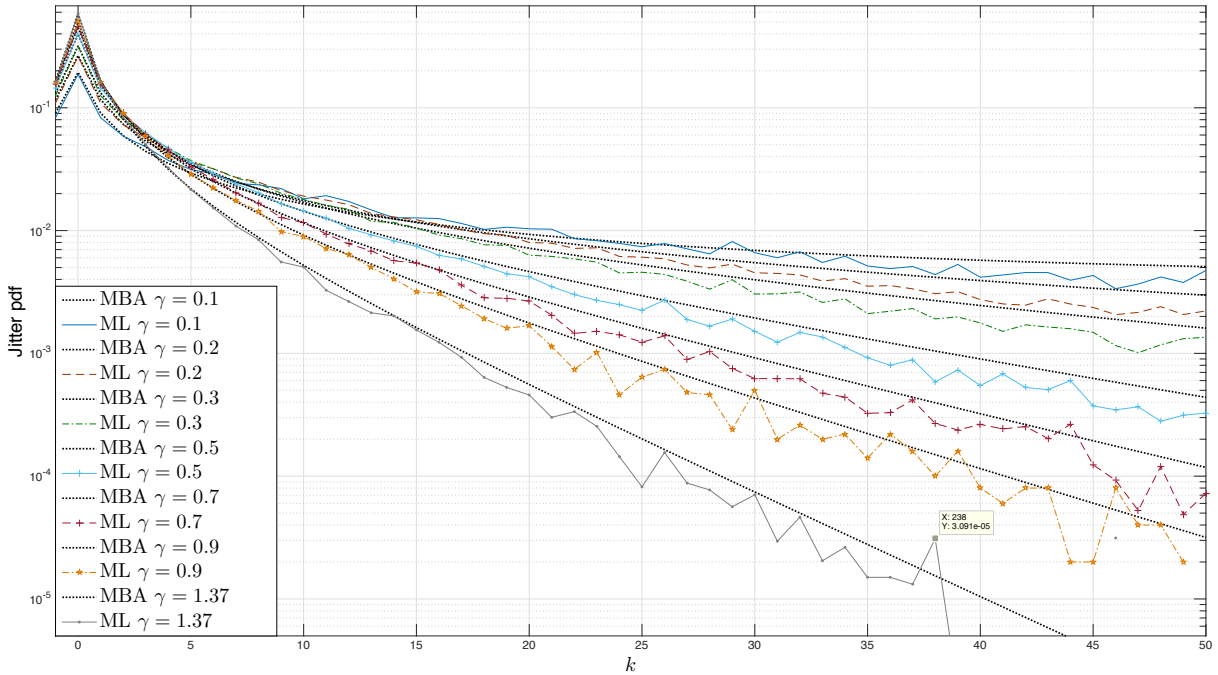


Figure 4.7: Experimentally defined one-sided jitter probability densities (dotted) of the breakpoint location for equal segmental SNRs in the range of  $M = 400$  points with a breakpoint at  $n = 200$ . The histogram was obtained using the ML estimator based in the slog algorithm and plotted over  $50 \times 10^3$  runs repeated 9 times and averaged. Approximations (dotted) are provided using the proposed MBA.

### Approximation

In order to use (4.2) as an approximating function

$$B(k|\gamma) = K_0[x(k)] \quad (4.3)$$

conditioned on  $\gamma$  for the one-sided jitter probability densities shown in Fig. 4.7 using the *slow* algorithm, a variable  $x$  is represented via  $k$  as  $x(k, \gamma) = \ln(\Phi(k, \gamma))$  in a way such that small  $k \geq 0$  correspond to large values of  $x$  and visa versa. Among several candidates, it has been found empirically that the function  $\Phi(k, \gamma)$  fits the histograms with highest accuracy,

$$\Phi(k, \gamma) = (|k| + 1)^{\beta + \alpha|k|} \left[ \frac{1 + \sqrt{\gamma}}{\gamma} - \epsilon \right], \quad (4.4)$$

if to set  $\gamma = \gamma_l^-$  for  $k < 0$ ,  $\gamma = \frac{\gamma_l^- + \gamma_l^+}{2}$  for  $k = 0$ , and  $\gamma = \gamma_l^+$  for  $k > 0$ , and represent the coefficients  $\alpha(\gamma)$ ,  $\beta(\gamma)$ , and  $\epsilon(\gamma)$  as

$$\alpha(\gamma) = a_0\gamma + a_1, \quad (4.5)$$

$$\beta(\gamma) = \gamma(b_0\gamma^{b_1} + a_0) + b_2, \quad (4.6)$$

$$\epsilon(\gamma) = c_0\gamma^{c_1} + c_2, \quad (4.7)$$

where  $a_0 = 0.02737$ ,  $a_1 = -4.5 \times 10^{-3}$ ,  $b_0 = 0.3425$ ,  $b_1 = -0.3413$ ,  $b_2 = 0.808$ ,  $c_0 = 0.8865$ ,  $c_1 = -1.033$  and  $c_2 = -1.233$  were found in the mean square error sense. These values were found in several iterations until the MSE reached a minimum.

The most appropriate values of  $\alpha(\gamma)$ ,  $\beta(\gamma)$ , and  $\epsilon(\gamma)$  computed for various symmetric SNRs  $\gamma_l^- = \gamma_l^+$  are sketched in Table 4.1 and given in Figure 4.8. The points in Fig. 4.8a, Fig. 4.8b, and Fig. 4.8c are the best values to each SNRs, and the solid lines are the fitted functions (4.5), (4.6), and (4.7) using the curve fitting tool from MATLAB. The MSE of approximations for functions  $\alpha(\gamma)$ ,  $\beta(\gamma)$ , and  $\epsilon(\gamma)$  generated are  $3.852 \times 10^{-5}$ ,  $2.35 \times 10^{-4}$  and  $4.226 \times 10^{-4}$  respectively. Table 4.1 shows this information in a range of SNRs from

Table 4.1: MSEs produced by Laplace-based (3.39) and Bessel-based (4.3) approximations.

$\gamma$	$\alpha(\gamma)$	$\beta(\gamma)$	$\epsilon(\gamma)$	$\gamma$	$\alpha(\gamma)$	$\beta(\gamma)$	$\epsilon(\gamma)$
0.1	-0.0018	0.886	8.32 5	0.7	0.0146	1.098	0.048
0.2	0.0001	0.932	3.438	0.8	0.0173	1.125	-0.116
0.3	0.0037	0.971	1.840	0.9	0.0201	1.152	-0.244
0.4	0.0064	1.0062	1.050	1.0	0.0228	1.177	-0.346
0.5	0.0092	1.0386	0.581	1.1	0.0255	1.2028	-0.498
0.6	0.0119	1.0691	0.269	1.37	0.0310	1.257	-0.556

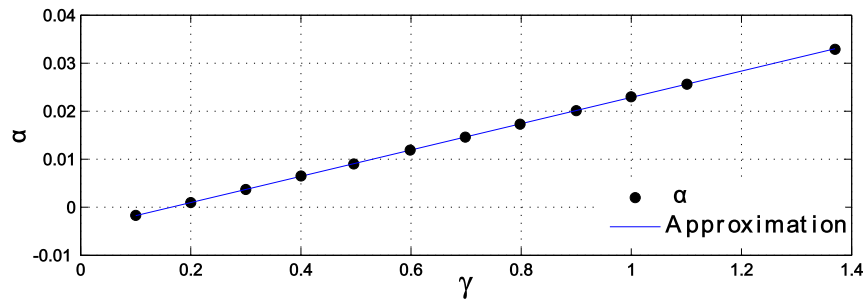
$\gamma = 0.1$  to  $\gamma = 1.37$ .

### 4.3 Parametrization of Laplace Density

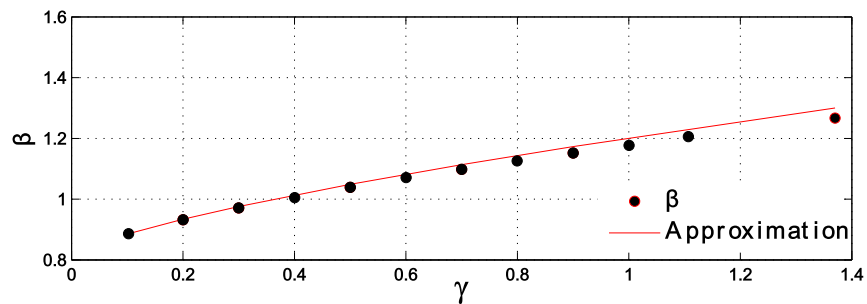
The SkL pdf (3.39) still can be applied in a parameterized form as follows. An increase in the discrete-step index  $k$  diminishes the effect of the segmental noise on jitter in the breakpoint. For example, noise at  $l - 10$  has a smaller effect on  $i_l$  than noise at  $l - 1$ . In Figure 4.9 is illustrated this theory using a simulated CNA with a breakpoint  $i_l$  at  $n = 10$ . The constant standard deviation  $\sigma_l$  and the  $k$ -varying standard deviation function  $\sigma_l(k)$  are plotted (dashed) to see the graphical difference.

To provide the same effect of noise at any point  $l \pm k$  on  $i_l$  as required by the derivation of the SkL-based approximation [58], the noise variances must be increased with  $k$ . That makes the variances,  $\sigma_l^2(k)$  and  $\sigma_{l+1}^2(k)$ ,  $k$ -variant and the SkL pdf (3.39) parameterized with  $k$ . Consequently, the probabilities computed of events A and B using the equations (3.29) and (3.30) to compute the SkL (3.39) distribution are inconstant and it can be established that  $P(A)_{|k=n} > P(A)_{|k=n-1}$ ,  $P(B)_{|k=n+1} > P(B)_{|k=n+2}$  and  $P(A)_{|k=n} \gg P(A)_{|k=n-10}$ ,  $P(B)_{|k=n+1} \gg P(B)_{|k=n+10}$ , see Fig. 4.9.

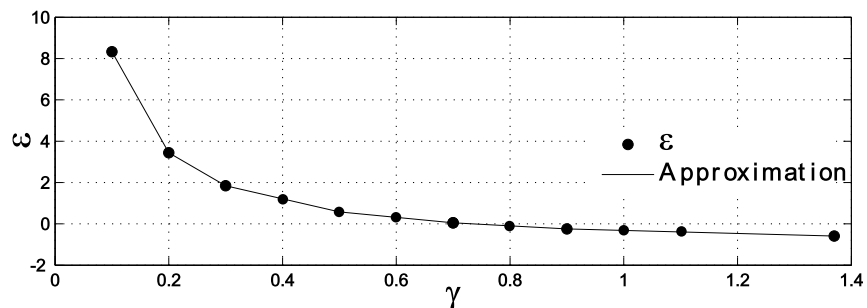




(a)



(b)



(c)

Figure 4.8: Coefficients for the approximation functions: (a)  $\alpha(\gamma)$ , (b)  $\beta(\gamma)$ , and (c)  $\epsilon(\gamma)$ . Actual values are dotted and the mean square error of approximations is a)  $3.852 \times 10^{-5}$ , b)  $2.35 \times 10^{-4}$  and c)  $4.226 \times 10^{-4}$ .

Because exact analytic functions are unavailable for  $\sigma_t^2(k)$  and  $\sigma_{t+1}^2(k)$ , in this section these functions are investigated numerically and find reasonable approximations in the minimum MSE sense based on simulations.

To this end, we redefine the  $k$ -varying segmental SNRs as

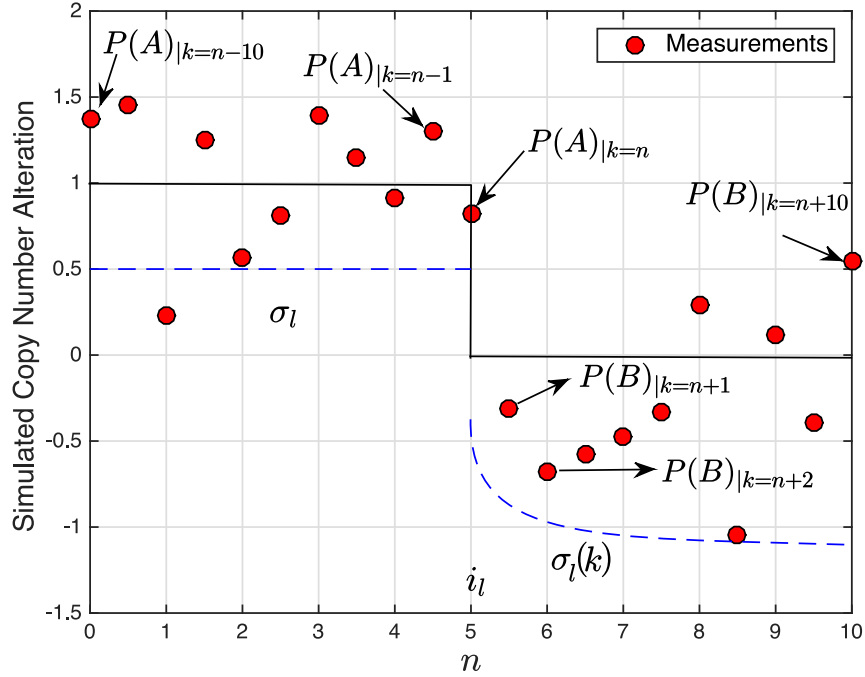


Figure 4.9: Representation of a standard deviation constant  $\sigma_l$  and the proposed  $k$ -varying standard deviation function  $\sigma_l(k)$  used to parametrize the SkL pdf (3.39). To this modification, it must keep the relationships  $P(A)|_{k=n} > P(A)|_{k=n-1}$ ,  $P(B)|_{k=n+1} > P(B)|_{k=n+2}$  and  $P(A)|_{k=n} \gg P(A)|_{k=n-10}$ ,  $P(B)|_{k=n+1} \gg P(B)|_{k=n+10}$ .

$$\gamma_l^-(k) = \frac{\Delta^2}{\sigma_l^2(k)}, \quad \gamma_l^+(k) = \frac{\Delta^2}{\sigma_{l+1}^2(k)}, \quad (4.8)$$

where  $\sigma_l^2 \triangleq \sigma_l^2(0)$ ,  $\sigma_{l+1}^2 \triangleq \sigma_{l+1}^2(0)$ ,  $\gamma_l^- \triangleq \gamma_l^-(0)$ , and  $\gamma_l^+ \triangleq \gamma_l^+(0)$ . Otherwise, when  $k \neq 0$ , it is assigned

$$\sigma_l^2(k) = \sigma_l^2 [1 + f_l(k)],$$

where  $f_l(k)$  is a function to be specified later.

### First Bessel-based Approximation

Testing several non-conventional functions has revealed that the modified Bessel equation  $K_\nu(x)$  of the second kind and fractional order  $\nu = 0.5$  is a good candidate to approximate

the measured jitter histogram, because it is positive-valued for  $x(k) > 0$ , smooth, and decreases with  $x$  to zero. We use the representation of  $K_\nu(x)$  defined in equation (4.2).

Based on simulations, it has been found that the following parameterizing function makes the SkL pdf (3.39) accurate in fitting the jitter histogram for any  $k$ ,

$$\sigma_l^2(k) = \sigma_l^2 \left[ 1 + K_{1/2}^{-1} \left( \log_{k+1}^{\alpha(\gamma_l^-)^b} \right) \right], \quad (4.9)$$

if to assign  $a = 0.6951$  and  $b = -0.1296$ . In fact,  $k = 0$  turns the parameterized SkL to (3.39) and, by  $\gamma_l^-, \gamma_l^+ \gg 1$ , it also converges to (3.39). An important property of the SkL parameterized with (4.9) is that it shows that when  $\gamma_l^+ \rightarrow 0$  and  $\gamma_l^- \rightarrow 0$  then  $\sigma_l^2(k) \rightarrow \infty$  and  $\sigma_{l+1}^2(k) \rightarrow \infty$  and  $i_l$  thus cannot be localized or, most likely, does not exist.

### Second Bessel-based Approximation

The second approximation was obtained employing the same Bessel function (4.2), but with another variable,

$$\sigma_l^2(k) = \sigma_l^2 \left[ 1 + K_{\frac{1}{2}} \left( \frac{1}{(k+1)^{\beta(\gamma_l^-)} - 1} \right) \right], \quad (4.10)$$

where  $\beta(\gamma_l^-) = \sqrt{2}/\gamma_l^{0.1734}$ . Testing (4.10) by simulations has shown that this function can produce more accuracy for certain values of SNR and that (4.9) can be more accurate otherwise, although both (4.9) and (4.10) can be applied to any  $k$ .

### Functional approximation

A simple approximation has appeared by using a power function of

$$\sigma_l^2(k) = \sigma_l^2 \frac{1}{2} \left[ 1 + k^{\alpha(\gamma_l^-)^b} \right]^2, \quad (4.11)$$

where  $a = 0.436$  and  $b = -0.1575$ . An analysis has shown that (4.11) is about 10 times more accurate than (4.9) and (4.10) when  $k > 1$ , but cannot be applied to  $k = 0$  or  $k = 1$ .

Functions (4.9) (dashed), (4.10) (solid), and (4.11) (circled and dotted) are sketched in Fig. 4.10 for  $\gamma = 0.1, 1.0, 5.0$  in the range of  $0 \leq k \leq 50$ . As can be seen, the proposed  $k$ -varying variances are consistent, but produce different errors in the  $k$ -domain. Note that an exact function  $\sigma_l^2(k)$  is still unavailable.

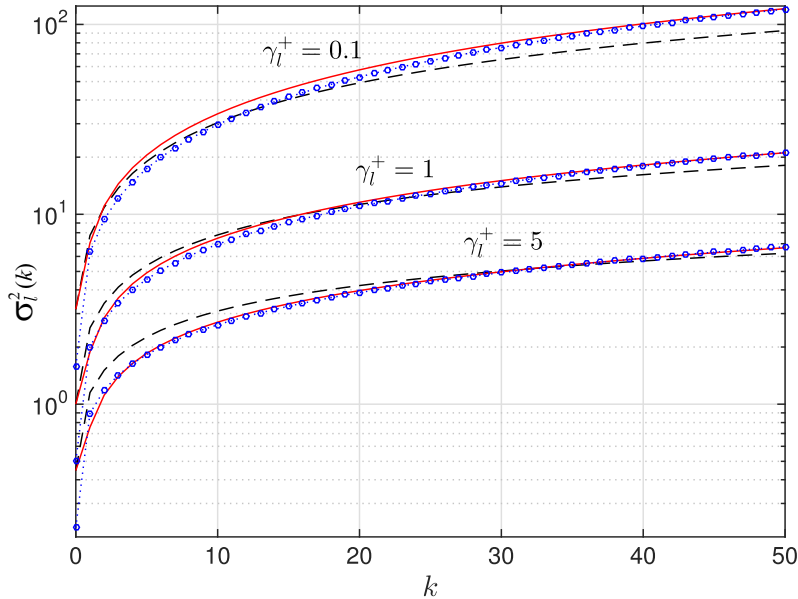
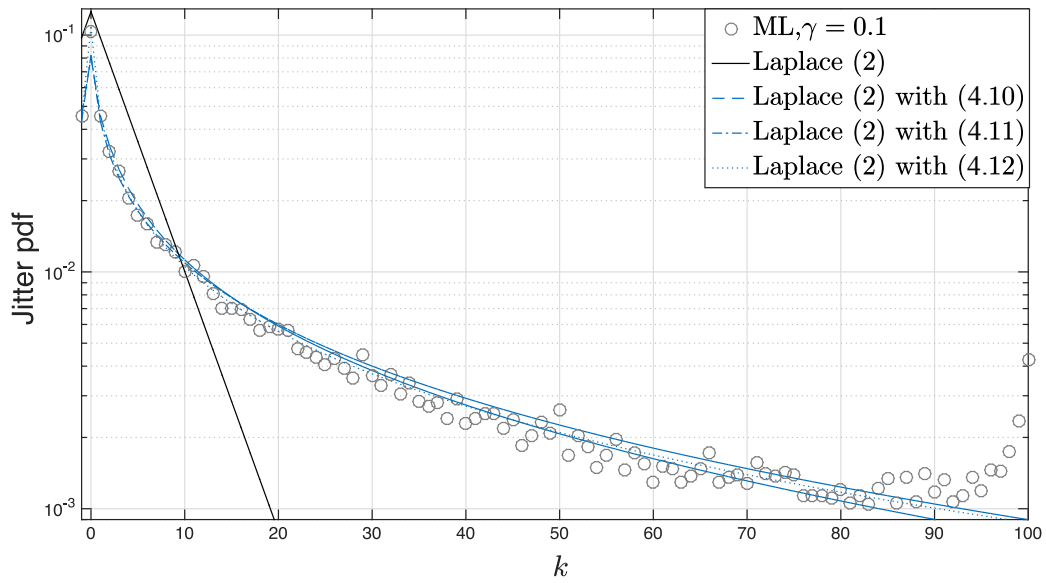
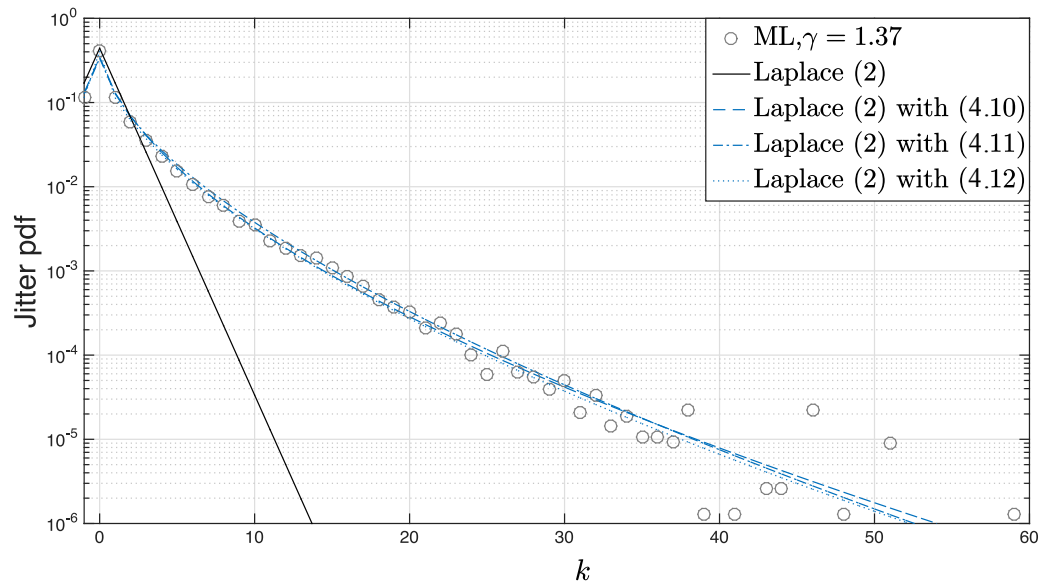


Figure 4.10: The proposed  $k$ -varying variance functions  $\sigma_l^2(k)$  used to parametrize the SkL pdf (3.39) for equal low ( $\gamma = 0.1$ ), normal ( $\gamma = 1$ ), and large ( $\gamma = 5$ ) SNRs: (4.9) is dashed, (4.10) is solid, and (4.11) is circled and dotted.

Figures 4.11a and 4.11b show the measurements obtained to symmetric SNRs  $\gamma_l^\pm = 0.1$  and  $\gamma_l = 1.37$  using the fast algorithm and the proposed approximations based on equations (4.9), (4.10) and (4.11). Figure 4.11a illustrates the simulation of the lowest value of SNR  $\gamma = 0.1$  (circles). It can also be noticed a notorious difference between measurements (circles) with the SkL (solid) distribution and a minimal error with respect to the proposed approximations using the parametrization with (4.9) (dashed), (4.10) (dash-dot) and (4.11) (dotted). Setting a SNR  $\gamma = 1.37$  the difference between measurements (circles) with SkL distribution (solid) and is lower than  $\gamma = 0.1$ , but the error of estimation is still significant, see Fig. 4.11b. Also, this picture shows a good fit of the parameterize approximations with (4.9) (dashed), (4.10) (dash-dot) and (4.11) (dotted).



(a)



(b)

Figure 4.11: Measured jitter pdf functions (circles) and the approximations by the SkL law (3.39) and by the SkL law parameterized with (4.9), (4.10), and (4.11) for equal segmental SNRs  $\gamma_l = \gamma_l^- = \gamma_l^+$ , a) low SNR  $\gamma_l = 0.1$  and b) normal SNR  $\gamma_l = 1.37$ .

## 4.4 Asymmetric Exponential Power Distribution

Analyzing the measured jitter pdf obtained with the detailed algorithm, we conclude that it could be approximated with a sub-Laplacian distribution such as the Asymmetric Exponential Power (AEP) distribution [77] –which is a generalization of the Gaussian and Laplace laws. A random variable  $y_l$  associated with the  $l$ th breakpoint is said to have the AEP function, if for the shape parameter  $\alpha_l > 0$ , scale factor  $\sigma_l > 0$ , location  $\theta_l = 0$ , and skew factor  $\kappa_l > 0$ , a variable  $y_l$  is distributed with

$$p(k|\bar{p}_l, \bar{q}_l) = \frac{\alpha_l}{\sigma_l \Gamma\left(\frac{1}{\alpha_l}\right)} \frac{\kappa_l}{1 + \kappa_l^2} \begin{cases} \bar{p}_l^{k\alpha_l}, & k \geq 0, \\ \bar{q}_l^{|k|\alpha_l}, & k \leq 0, \end{cases} \quad (4.12)$$

where  $\bar{p}_l = e^{-\frac{\kappa_l \alpha_l}{\sigma_l}}$ ,  $\bar{q}_l = e^{-\frac{1}{\kappa_l \sigma_l \alpha_l}}$ , and  $\Gamma(x)$  is the Gamma function.

In a special case of  $\gamma_l^- = \gamma_l^+$ , the skew factor becomes  $\kappa_l = 1$  and the AEP distribution symmetric. In the other special case of  $\kappa_l \neq 1$ , letting  $\alpha_l = 1$  transforms (4.12) to the discrete skew Laplace distribution (3.39) [77], which alternative form is

$$p(k|\bar{p}_l, \bar{q}_l)_{\alpha_l=1} = \frac{1}{\sigma_l} \frac{\kappa_l}{1 + \kappa_l^2} \begin{cases} \bar{p}_l^k, & k \geq 0, \\ \bar{q}_l^{|k|}, & k \leq 0. \end{cases} \quad (4.13)$$

Using values of  $\alpha = 0.5, 1, 1.5, 2, 2.5, 10$ ,  $\kappa = 1$ ,  $\sigma = 1$  and zero mean, some distributions are plotted in the Figure 4.12. The AEP function is a good candidate to approximate the jitter distribution because have a mathematical support and exhibit a great flexibility modifying only three parameters, given that location  $\theta_l$  always is zero.

An example of applications of (4.12) as an approximation is given in [78], where factors  $\alpha_l$ ,  $\kappa_l$  and  $\sigma_l$  were estimated in the maximum likelihood sense for the simulated growth distribution.

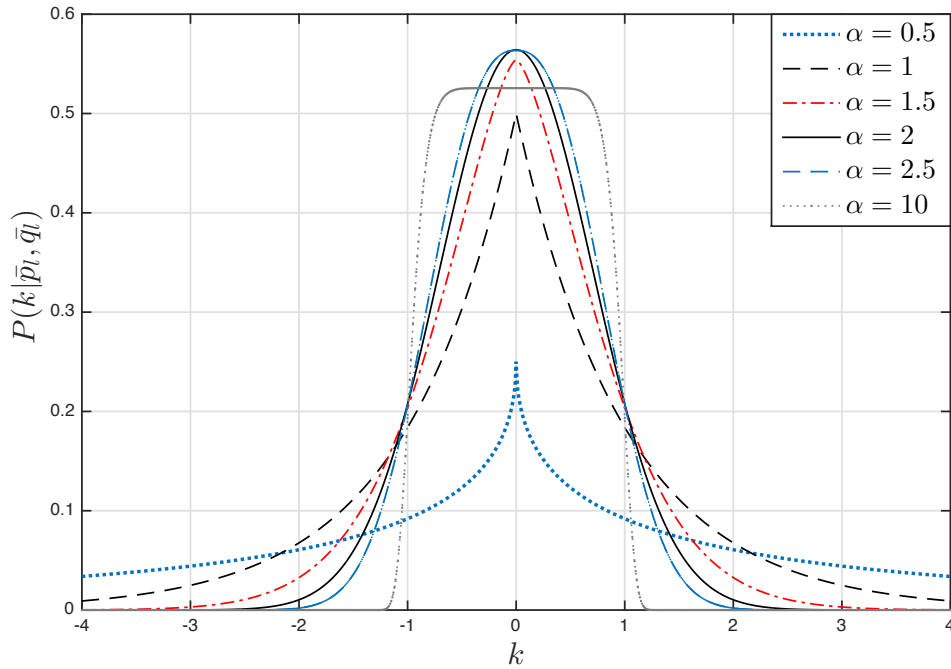


Figure 4.12: Asymmetric Exponential Power Distribution to several parameters of shape  $\alpha$ , for the symmetric case skew is set as  $\kappa = 1$ , scale  $\sigma = 1$  and zero mean. The function AEP function has two special cases: when  $\alpha = 1$  represents the Laplace distribution and if  $\alpha = 2$  the normal distribution can be computed.

#### 4.4.1 Parameters Estimation for AEP distribution

In order to approximate the jitter distribution with (4.12) in an optimal way, one needs to find  $\alpha_l$ ,  $\kappa_l$  and  $\sigma_l$  as functions of  $\gamma_l^-$  and  $\gamma_l^+$  to provide the best fit. These constants can be found by fitting the histograms with the highest accuracy by minimizing the Kolmogorov–Smirnov (KS) distance [78] defined as

$$d_{KS} = \max |F_0(x) - S_N(x)|, \quad (4.14)$$

where  $F_0(x)$  is the population cumulative distribution of (4.12) and  $S_N(x)$  is the observed cumulative step function. The KS distance  $d_{KS}$  between  $S_N(x)$  and  $F_0(x)$  functions is illustrated in Figure 4.13. The Kolmogorov–Smirnov distance is a common statistical

measure called the “test of goodness of fit” or the Kolmogorov-Smirnov test (KS-test). The parameter  $d_{KS}$  is computed by (4.15) and selects the minimum one to set the most appropriate values of  $\alpha_l$ ,  $\kappa_l$ , and  $\sigma_l$  for various symmetric SNRs  $\gamma_l^- = \gamma_l^+$ .

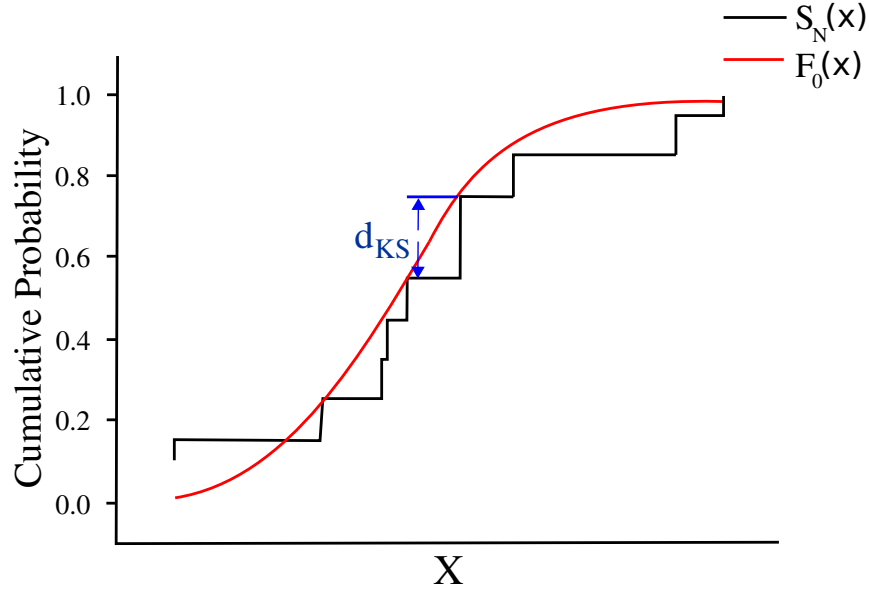


Figure 4.13: The Kolmogorov-Smirnov distance between the empirical distribution function of the sample  $S_N(x)$  and the cumulative distribution function of the reference distribution  $F_0(x)$ .

Also, constants  $\alpha_l$  and  $\sigma_l$  are approximated in the MSE sense as

$$\alpha_l(\gamma_l) = 1 - \frac{a_1}{\gamma_l^{b_1}}, \quad (4.15)$$

$$\sigma_l(\gamma_l) = a_2 \gamma_l^{b_2}, \quad (4.16)$$

where  $a_1 = 0.389$ ,  $b_1 = 0.1394$ ,  $a_2 = 1.142$  and  $b_2 = -0.6289$ . Note that, for the asymmetric case  $\gamma_l^- \neq \gamma_l^+$ , the shape and scale factors depends on the parameters individually estimated, so  $\alpha_l(\gamma_l^\pm)$  and  $\sigma_l(\gamma_l^\pm)$  are provided by equations (4.17) and (4.18).

$$\alpha_l(\gamma_l^\pm) = \frac{\alpha_l(\gamma_l^+) + \alpha_l(\gamma_l^-)}{2}, \quad (4.17)$$

$$\sigma_l(\gamma_l^\pm) = \frac{\sigma_l(\gamma_l^+) + \sigma_l(\gamma_l^-)}{2}. \quad (4.18)$$



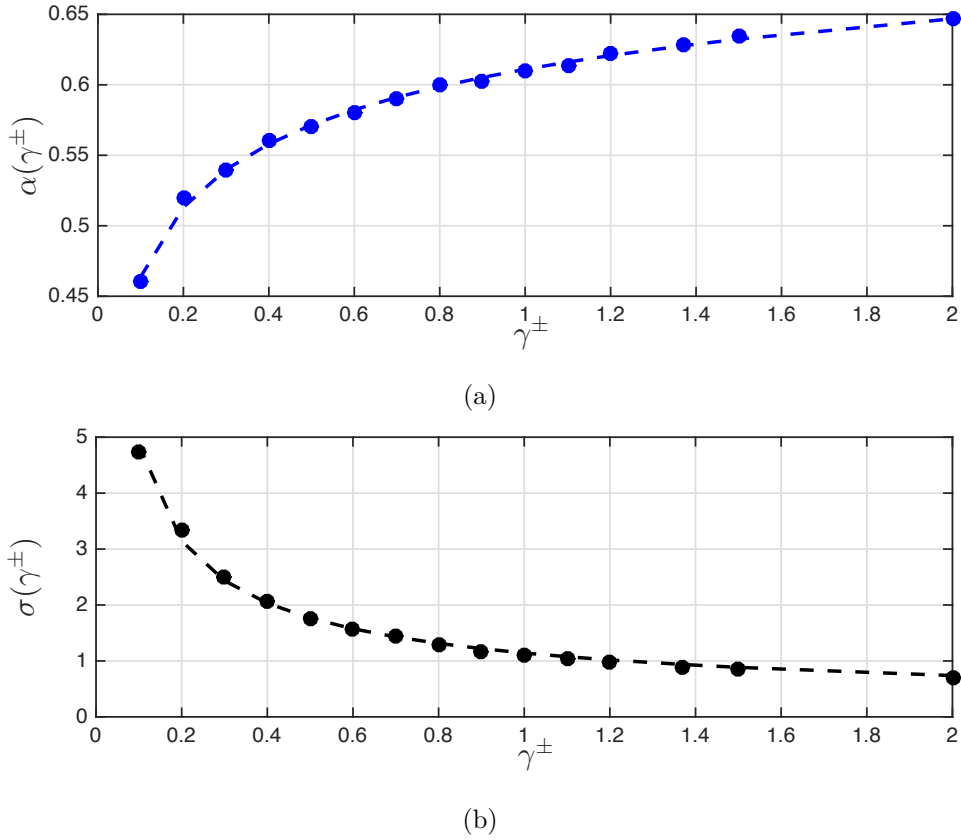


Figure 4.14: Approximations of the (a) shape factor  $\alpha_l(\gamma_l^\pm)$  and (b) scale factor  $\sigma_l(\gamma_l^\pm)$  for jitter distribution using Asymmetric Exponential Power distribution. Measured data are dotted to a range of SNR from  $\gamma^\pm = 0.1$  to  $\gamma^\pm = 2$  and the curves to fit them are represented with a dashed line for both variables.

Figure 4.14a and Fig. 4.14b sketch the measured data (dotted) and the approximations of  $\alpha_l(\gamma_l^\pm)$  and  $\sigma_l(\sigma_l^\pm)$  (dashed) obtained using SkL (3.39) and AEP distribution (4.12). To find  $\kappa_l$ , the equation proposed in [77] is modified.

$$\kappa_l = \left[ \frac{\bar{X}_{\alpha_l}^-}{\bar{X}_{\alpha_l}^+} \right]^{\frac{1}{2(\alpha_l+1)}} \quad (4.19)$$

by substituting  $\bar{X}_{\alpha_l}^- = \gamma_l^+$  and  $\bar{X}_{\alpha_l}^+ = \gamma_l^-$  for the asymmetric case and  $\kappa_l = 1$  otherwise.

The AEP distribution applied to symmetric  $\gamma_l^- = 0.3$ ,  $\gamma_l^+ = 0.3$  and asymmetric case  $\gamma_l^- = 0.3$ ,  $\gamma_l^+ = 0.8$  are plotted in Figures 4.15a and 4.15b. For the symmetric case, Figure

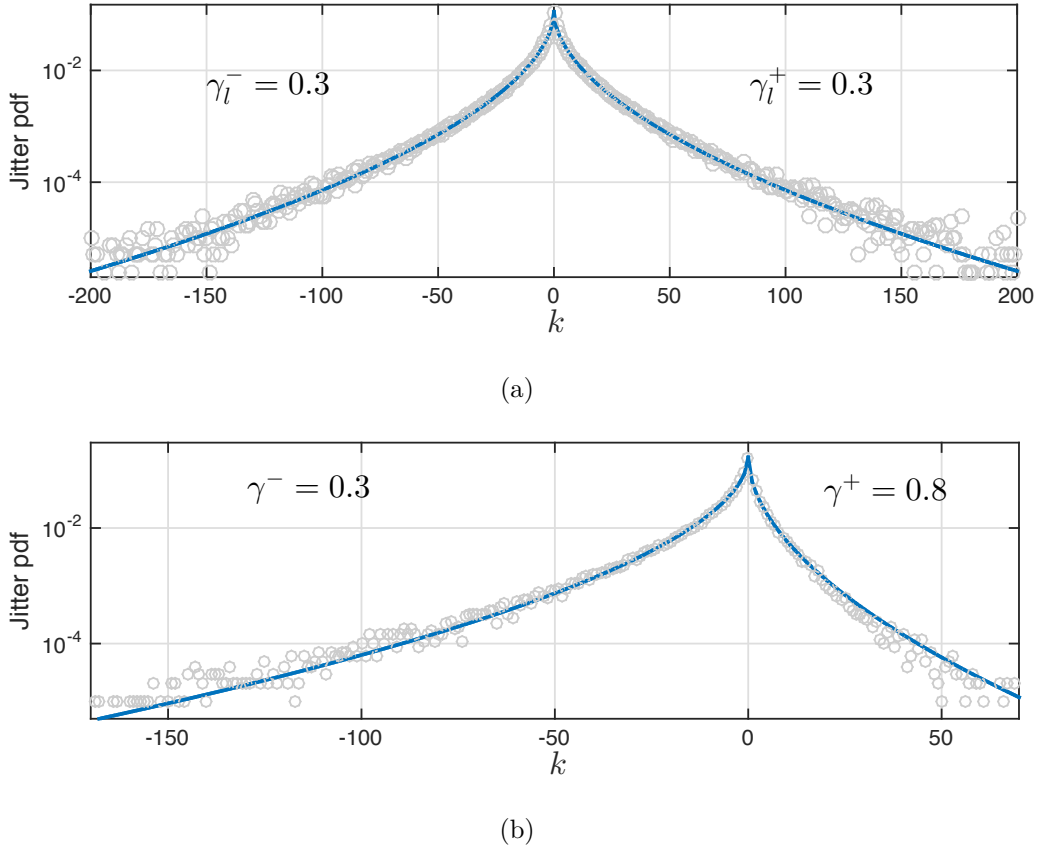


Figure 4.15: The approximations of the experimentally measured jitter pdf (dotted) for different SNR values using the AEP distribution (solid) for  $\gamma_l^- = 0.3$ : (a)  $\gamma_l^+ = 0.3$  and (b)  $\gamma_l^+ = 0.8$ . Measurement data are provided by averaging  $10^4$  runs using the detailed algorithm for ML estimator.

4.15a, the parameters estimated were shape  $\alpha = 0.54$ , skew  $\kappa = 1$  and scale  $\sigma = 2.49$ . For the asymmetric case given in Figure 4.15b, the parameter of shape, scale and skew were computed as  $\alpha_{|\gamma^-=0.3, \gamma^+=0.8} = 0.57$ ,  $\sigma_{|\gamma^-=0.3, \gamma^+=0.8} = 1.89$  and  $\kappa_{|\gamma^-=0.3, \gamma^+=0.8} = 1.36$ , respectively.

## 4.5 Comparison of Proposed Approximations

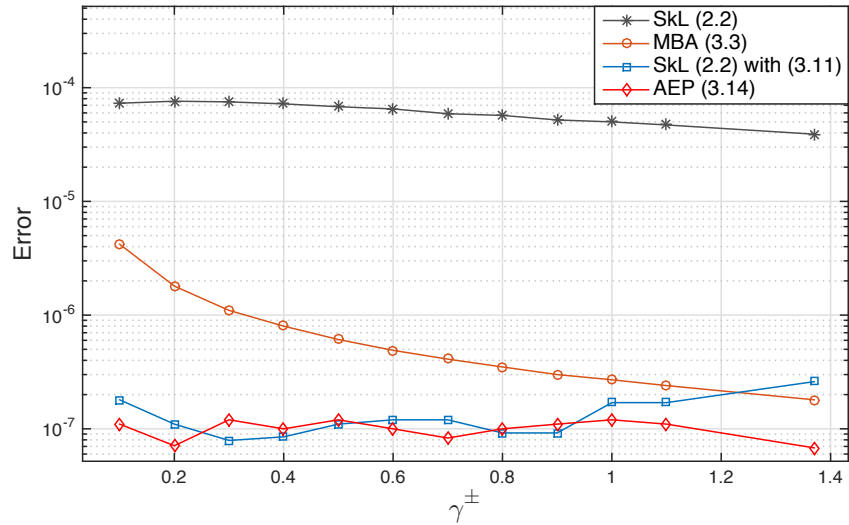
The errors caused by the propose approximations depend on the algorithm used to generate the experimental jitter distribution. It is worth mentioning that the parameters estimated to each approximation were obtained using a particular algorithm: slow, fast or detailed. The Tables D.1 and D.2, Appendix D, shows errors of each distributions computed with the MSE defined as

$$MSE(\hat{x}) = E[(\hat{x} - x)^2] \quad (4.20)$$

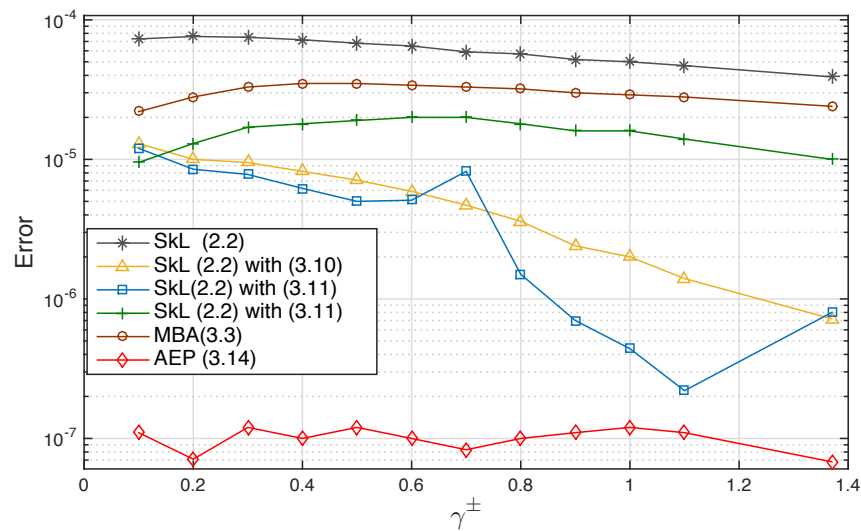
where  $\hat{x}$  is the approximation proposed and  $x$  is the experimental jitter pdf obtained with the ML estimator. Complete tables D.1 and D.2, given in Appendix D, compile the mse of each approximation and a comparison with the measurements of long format.

The MBA distribution was designed to estimate the jitter histogram obtained with the slow algorithm. Here, the mean square error is notoriously lesser than Skew Laplace to the same measurements, column 2 to 3 section I of Table D.1. By the fast algorithm, we found the parametrized approximations, which produce less mean square error compared with the Skew Laplace distribution. The values of MSE of a range of SNR are recorded in the column 4 to 7 section I Table D.1. In the same table, Section II we collected the MSE of all proposed distributions compared with the measurements of the detailed algorithm.

Figures 4.16a and 4.16b illustrate a graphic summary of tables D.1 and D.2, in Appendix D, which were presented in [79, 80]. The minimum error of each approximation is plotted in 4.16a, it can be noticed that the Skew Laplace distribution have the higher error. A general error comparison of proposed approximations and the experimental histogram based on the detailed algorithm is given in 4.16b. The AEP and the SkL distributions show the lower and higher errors, respectively.



(a)



(b)

Figure 4.16: Error of ML estimator and proposed approximations. a) Minimum errors obtained of each approximations, b) errors of all approximations respect to the experimental jitter obtained with the detailed algorithm.

# Chapter 5

## Modified Confidence Masks

In this Chapter, the proposed approximations, to fit the experimental jitter distributions showed in Chapter 4, are adapted to the confidence masks. Consequently, the initial algorithm computing the Upper Bound and Lower Bound is modified using these functions. The modified algorithm is applied to microarrays data obtained with Single Nucleotide Polymorphism and Comparative Genomic Hybridization technologies to test estimates of Copy Number Alterations.

### 5.1 Confidence Masks for Hybrid approximation

The UB mask  $\mathcal{B}_{l|H}^{\text{UB}}$  and LB mask  $\mathcal{B}_{l|H}^{\text{LB}}$  for the heuristic (Bessel-based) approximation can be formed using the same equations as for the Laplace distribution described in Section 3.2. In doing so, we suppose that the Laplace pdf (3.39) is equal to the approximating function  $B_l(k)$  equation 4.3 at  $k = 0$ ,

$$p(k = 0|d_l, q_l) = B_l(k = 0), \quad (5.1)$$

that gives us  $B_l(k = 0) = \frac{1}{\phi_l}$  where  $\phi_l$  is the parameter of normalization defined in 3.42 . Next, the probabilities  $P(A_l)_H$  at  $k = -1$  and  $P(B_l)_H$  at  $k = 1$  are defined based on the heuristic approximation as

$$P(A_l)_H = \frac{B_l(k=0)}{B_l(k=-1) + B_l(k=0)}, \quad (5.2)$$

$$P(B_l)_H = \frac{B_l(k=0)}{B_l(k=1) + B_l(k=0)}. \quad (5.3)$$

Then equations (5.2) and (5.3) are substituted into (3.40), (3.41), and (3.42), and  $\kappa_{l|H}$  and  $\nu_{l|H}$  can be calculated. That allows us to specify the right-hand jitter  $k_{l|H}^R$  and left-hand jitter  $k_{l|H}^L$  by, respectively,

$$k_{l|H}^R = \left\lfloor \frac{\nu_{l|H}}{\kappa_{l|H}} \ln \frac{1}{\xi B_l(k=0)} \right\rfloor, \quad (5.4)$$

$$k_{l|H}^L = \left\lfloor \nu_{l|H} \kappa_{l|H} \ln \frac{1}{\xi B_l(k=0)} \right\rfloor. \quad (5.5)$$

Finally, we define the jitter left boundary  $J_{l|H}^L$  and right boundary  $J_{l|H}^R$  as, respectively,

$$J_{l|H}^L \cong \hat{n}_l - k_{l|H}^R, \quad (5.6)$$

$$J_{l|H}^R \cong \hat{n}_l + k_{l|H}^L, \quad (5.7)$$

and use in the algorithm [16] previously designed for the confidence masks based on the Laplace distribution.

### 5.1.1 Testing Estimates by $\mathcal{B}_{l|H}^{UB}$ and $\mathcal{B}_{l|H}^{LB}$ Masks

Our purpose now is to test the complete CNA estimates by the probabilistic masks. Specifically, the probes employed are from 1st chromosome available from “BLC\_B1\_T45.txt” obtained using the SNP array technology.

Inherently, the more accurate Bessel-based approximation extends the jitter probabilistic boundaries with respect to the Laplace-based ones, especially for low SNRs. This is illustrated in Fig. 5.1, where the estimates of the 1st chromosome were tested by  $\mathcal{B}_l^{UB}$ ,  $\mathcal{B}_l^{LB}$ ,  $\mathcal{B}_{l|H}^{UB}$  and  $\mathcal{B}_{l|H}^{LB}$  for  $\vartheta = 3$  (confidence probability  $P = 99.73\%$ ).

In Fig. 5.2a, the masks  $\mathcal{B}_{l|H}^{UB}$  and  $\mathcal{B}_{l|H}^{LB}$  are showed and placed in the vicinity of segment  $\hat{a}_{18}$  for several confidence probabilities:  $\vartheta = 0.6745$  ( $P = 50\%$ ),  $\vartheta = 1$  ( $P = 68.27\%$ ),

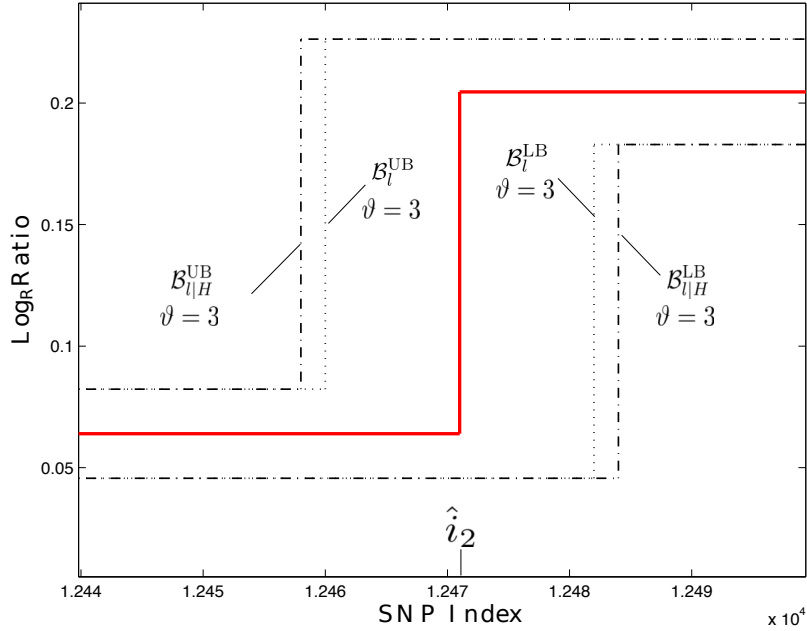
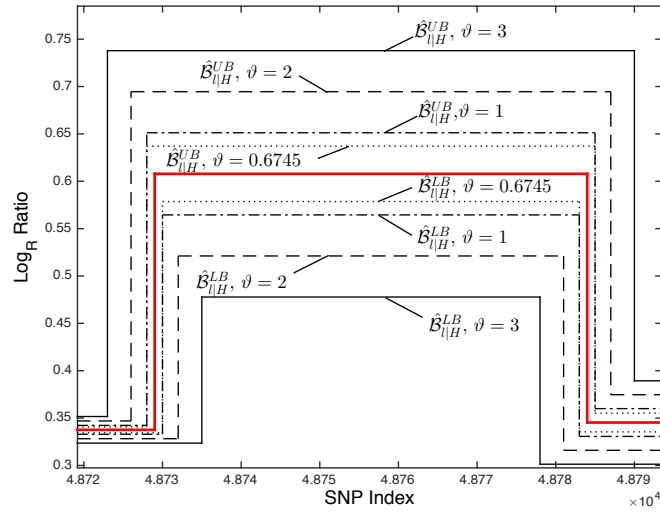


Figure 5.1: Upper boundaries  $\mathcal{B}_l^{UB}$ ,  $\mathcal{B}_{l|H}^{UB}$  and lower boundaries  $\mathcal{B}_l^{LB}$  and  $\mathcal{B}_{l|H}^{LB}$  for the breakpoint  $i_2$  of Chromosome 1 from database BLC\_B1\_T45.txt given  $\vartheta = 3$ . Confidence bounds  $\mathcal{B}_{l|H}^{UB}$  and  $\mathcal{B}_{l|H}^{LB}$  (dash-dot) are based on the heuristic approximations and  $\mathcal{B}_l^{UB}$  and  $\mathcal{B}_l^{LB}$  (dotted) use the SkL distribution.

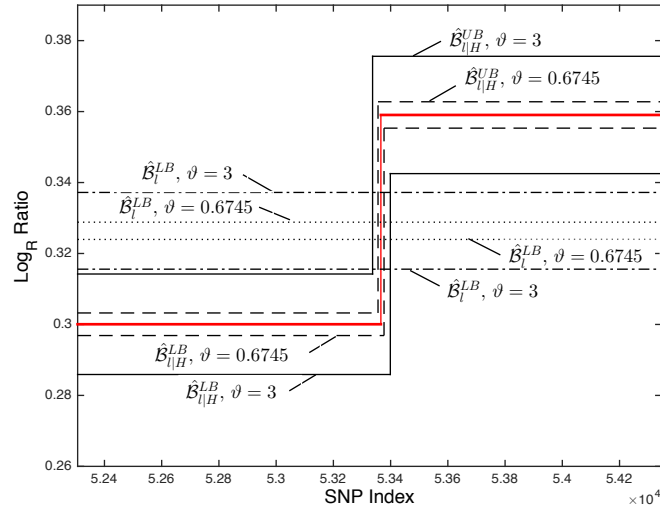
$\vartheta = 2$  ( $P = 95.45\%$ ), and  $\vartheta = 3$  ( $P = 99.73\%$ ). What the masks suggest here is that the CNA evidently exists with high probability, but the segmental levels and the breakpoint locations cannot be estimated with high accuracy, owing to low SNRs.

It also worth emphasizing on a special case when the masks  $\mathcal{B}_l^{UB}$  and  $\mathcal{B}_l^{LB}$  are not able to confirm or deny an existence of segmental changes with high probability, owing to an inability of computing the Laplace-based masks for extremely low SNRs. Figure 5.2b illustrates such situations. Just on the contrary, the masks  $\mathcal{B}_{l|H}^{UB}$  and  $\mathcal{B}_{l|H}^{LB}$  can be computed for any reasonable SNR.

A conclusion that can be made based on the results illustrated in Fig. 5.1, Fig. 5.2a and Fig. 5.2b is that the Bessel-based probabilistic masks can be used to improve estimates of the chromosomal changes at low and extra low values of SNR for the required probability that it is done in the next section.



(a)



(b)

Figure 5.2: The  $\mathcal{B}_{l|H}^{UB}$  and  $\mathcal{B}_{l|H}^{LB}$  masks placed a) around the segmental level  $a_{18}$  for several confidence probabilities and b) around the breakpoint  $i_{20}$  for  $\vartheta = 0.6745$  and  $\vartheta = 3$ . The CNA in a) exists with high probability, but the segmental levels and the breakpoint locations cannot be estimated with high accuracy.



### 5.1.2 Improving CNAs Estimates

As has been shown before, not all of the detected chromosomal changes have the same confidence, which means that there is a probability that some breakpoints do not exist. In order to improve the CNA estimates for the required confidence, the following methodology can be used:

1. Obtaining estimates of the CNA using the standard CBS algorithm [51, 52] or any other algorithm.
2. Computing masks  $\mathcal{B}_{i|H}^{UB}$  and  $\mathcal{B}_{i|H}^{LB}$  for the given confidence probability  $P, \%$  and bound the estimates.
3. If the masks reveal double *uniformities*, in UB and LB, in a gap of any three neighbouring breakpoints, then remove the intermediate breakpoint and estimate the segmental level between the survived breakpoints by simple averaging. The CNAs estimated in such a way will be valid for the given confidence  $P, \%$ .

Finally, Figures 5.3a–5.3e show an application of this methodology to the CNA structure detected in frames of the Project Genome Alteration Print (GAP) [34].

A number of hardly recognized small chromosomal changes are illustrated in (Fig. 5.3a) and the aim is tested them by the proposed masks  $\mathcal{B}_{i|H}^{UB}$  and  $\mathcal{B}_{i|H}^{LB}$ .

In doing so, first equal confidence probabilities start at  $P = 50\%$  for each estimate to exist or not and find out that three breakpoints demonstrate no detectability. These breakpoints are removed and depicted their locations with “×”. Reasoning similarly, four breakpoints are removed to retain only probable changes, by  $P = 75\%$ , nine breakpoints to show a picture combined with almost certain changes, by  $P = 93\%$ , and ten breakpoints in the 3-sigma sense,  $P = 99.73\%$ . Observing the results, it is inferred that the masks are able to correct only the estimates obtained under low SNRs. The relevant chromosomal sections S1–S7 are circled in Fig. 5.3. It is not surprising, because changes existing with high SNRs are seen visually. They thus can easily be detected with high confidence by an estimator.

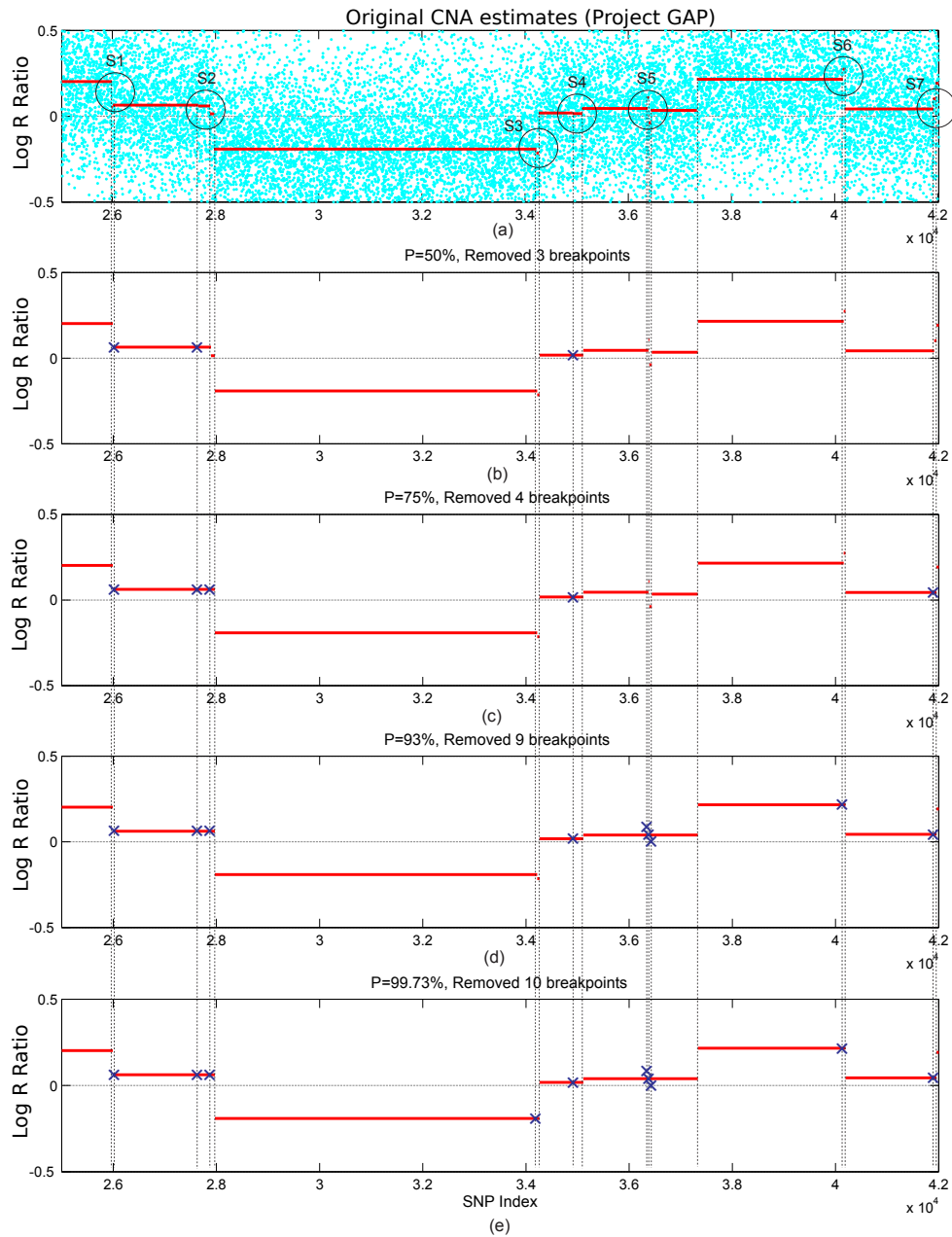


Figure 5.3: Improving estimates of the CNAs obtained in Project GAP by removing some unlikely existing breakpoints: (a) original estimates, (b) even changes,  $P = 50\%$ , (c) probable changes,  $P = 75\%$ , (d) almost certain changes,  $P = 93\%$ , and (e) 3-sigma sense,  $P = 99.73\%$ .

## 5.2 Confidence Masks based on Laplace–parametrization

It follows from an analysis of errors produced by the proposed approximations that the most reliable results can be achieved when developing the confidence masks worked out in 3.2 to be hybrid by using different approximations in diverse regions of the segmental SNRs.

### 5.2.1 Hybrid confidence masks

Based on the MSEs produced by the approximations (Table D.1), in Table 5.1 the segmental SNR regions are selected, in which the MBA developed in [16], Laplace pdf (3.39), and Laplace pdf (3.39) parameterized with (4.9), (4.10), and (4.11) are most successful in detecting the right jitter  $k^-$  and the left jitter  $k^+$  in the minimum MSE sense.

Table 5.1: SNR regions for MBA, Laplace pdf (3.39), and (3.39) parameterized with (4.9), (4.10), and (4.11) to detect the right jitter  $k^-$  and the left jitter  $k^+$  with the minimum MSE.

$\gamma_i^-, \gamma_i^+$	$k^-, k^+$	pdf	SNR region		
			0.1...0.9	0.9...1.37	> 1.37
=	Any	(3.39) with (4.9)	–	X	–
		(3.39) with (4.10)	X	–	–
	> 1	(3.39) with (4.11)	X	X	–
≠	Any	(MBA) [16]	X	X	–
	Any	(3.39)	–	–	X

Table 5.1 suggests that for  $\gamma_i^- = \gamma_i^+$  and  $|k| \geq 0$ , the Laplace pdf (3.39) parameterized with (4.9) is most accurate in the SNR region of  $0.9 \dots 1.37$ , while that with (4.10) produces better accuracy in  $0.1 \dots 0.9$ . The Laplace pdf parameterized with (4.11) is also accurate when  $0.1 < \text{SNR} < 1.37$ , but it loses accuracy at  $k = 0$  and  $k = 1$ . When  $\gamma_i^- \neq \gamma_i^+$ , the MBA is preferable in the SNR region of  $0.1 \dots 1.37$  and the skew Laplace pdf (3.39) can be used otherwise for any departure index  $k$ .

Following the above provided analysis of Table 5.1, better accuracy in the confidence UB mask and LB mask designed in [16] can be achieved if these masks are made hybrid. The difference between the hybrid masks and the basic ones [16] is in the parametrization of (3.39) and in the conditions introduced for the SNR values  $\gamma_i^-$  and  $\gamma_i^+$ . With such modifications, the basic masks can be used straightforwardly and readers can consult [16] for a detailed description of the basic algorithm.

## 5.2.2 Applications to SNP Array Probing

In this subsection, we tested experimentally the parameterized Laplace density (3.39) and several confidence masks by the SNP array-based CNAs probing data taken from database BLC\_BLT31 available from the project GAP [12].

### Confidence of the Breakpoint Location

To emphasize again on a practical importance of the hybrid confidence masks, in Fig. 5.4a and Fig. 5.4b it is showed a part of the 13rd chromosome [12] consisting of a single a single breakpoint  $i_5$  and two segments with the segmental SNR values of  $\gamma_5^- \cong 1.46$  and  $\gamma_5^+ \cong 1.5$  as investigated in [12].

The candidate breakpoint was detected using the ML estimator and then the ML estimates were tested by different masks based on the Laplace pdf (3.39), Laplace pdf parameterized with (4.10), and the one designed in [16] for the confidence probability  $P = 99.73\%$ . The MBA and (3.39) parameterized with (4.10) shown in Fig. 5.4b more accurately approximate the jitter distribution. Therefore, the regions of possible breakpoint locations produced by these approximations (Fig. 5.4a) must be accepted as more

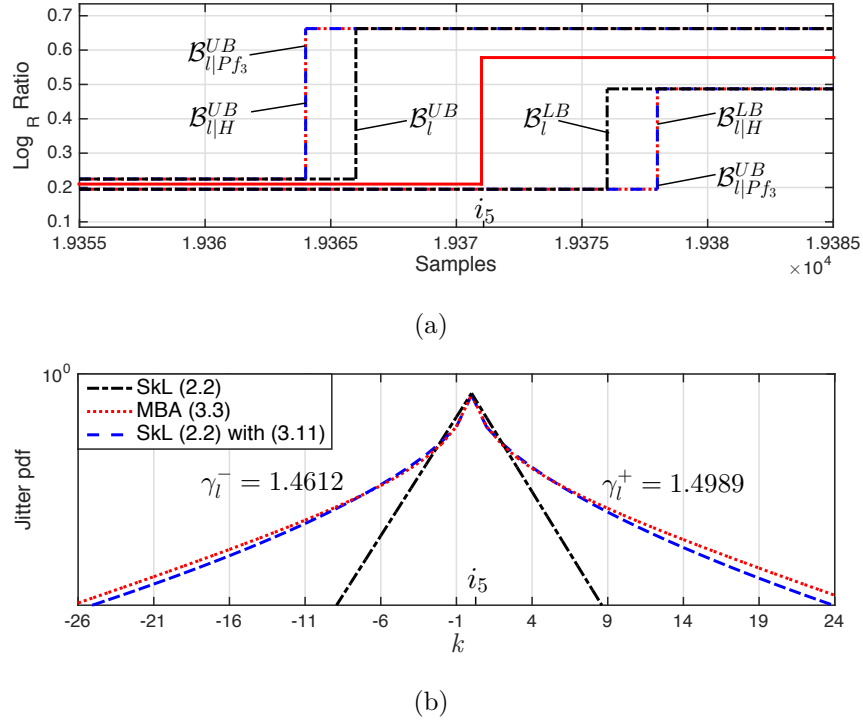


Figure 5.4: Testing the ML estimate of the breakpoint  $i_5$  location of the sample BLC\_BLT31 in the 13th Chromosome by the confidence masks: (a) ML estimate (solid) and different upper bound (UB) and lower bound (LB) masks and (b) jitter distribution in  $i_5$  approximated with the Laplace pdf (3.39), Laplace pdf parameterized with 3.45, and MBA.

realistic. As can be seen, these regions are wider than that produced by the Laplace pdf (3.39).

### Chromosome Probing by SNP Array

Now, the confidence masks are applied to test the estimates of the breakpoint locations in the complete chromosome 13th of the profile BLC\_BLT31 taken from the series of basal-like carcinomas (BLC) available from the project GAP [12]. This series are included in a study of primary breast carcinomas (40 cases) and two cell lines measured on a 300K Illumina SNP-arrays (Human Hap300-Duo). The Copy-Number Alteration profile

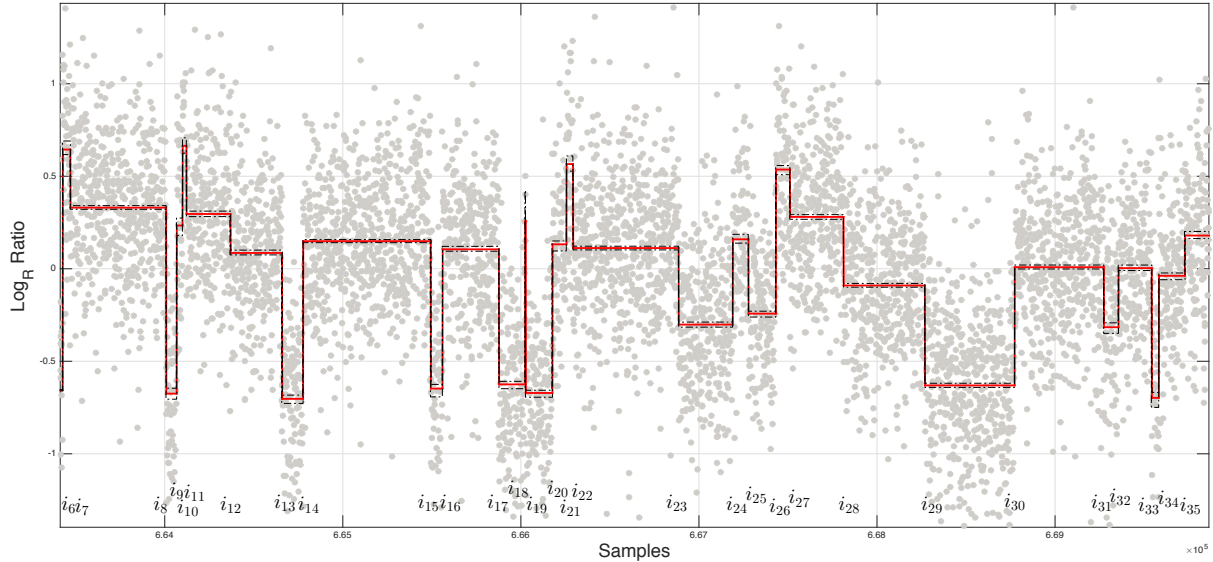


Figure 5.5: Probes (points), CNAs estimates (solid), and confidence regions (dashed) provided by the hybrid masks for the 13th chromosome taken from *BLC-BI-T37* of GAP. The breakpoint locations were detected using the algorithm *cghcbs*

is represented by the Log R ratios centered at zero for each sample. The estimates were obtained using the algorithm Comparative Genomic Hybridization–Circular Binary Segmentation (*cghcbs*) [51] available in MATLAB, which suggests that the chromosome has 59 segments and 58 breakpoints as shown in Fig. 5.5.

It follows from Fig. 5.5 that the confidence intervals are wider for the segmental levels than for the breakpoints. Therefore, this figure is supplied with Tables E.1 Part I and Table E.2 Part II given in Appendix E, in which the left jitter  $k_l^-$  and the right jitter  $k_l^+$  are estimated for the confidence probability  $P = 99.73\%$  in the  $(3\sigma)$  sense [16].

Tables E.1 and E.2 suggest that the masks often produce unequal jitter estimates and that the difference between the estimates can be in several points, as in the case of  $l = 7$  or  $l = 27$ . Large jitter in  $i_1$ ,  $i_{40}$ ,  $i_{43}$ , and  $i_{44}$  was detected only by the MBA. But the MBA was unsuccessful in detecting any jitter in a larger number of the breakpoints such as  $i_8$ ,

$i_9$ ,  $i_{14}$ ,  $i_{18-20}$ ,  $i_{26}$ ,  $i_{48-51}$ ,  $i_{57}$ , and  $i_{58}$ , while other masks provided it with near similar errors. One can also notice that extra low SNR values made the jitter unavailable for bounding by all of the masks, as in the cases of  $i_4$ ,  $i_5$ , and  $i_{41}$ .

Jitter computed by the hybrid masks is put to the two last columns of Tables E.1 and E.2. Because the hybrid masks combine the most accurate outputs of the particular masks, the left and right jitter computed by the hybrid masks can be considered as most reliable. What the hybrid masks suggest is that jitter in the breakpoints of this chromosome ranges from 1 point to tens of points and thus an actual breakpoint can be defined specifying tens points apart from the candidate one provided by an estimator.

### 5.3 Confidence Masks based on AEP distribution

For the AEP-based approximation (4.11), the confidence masks can easily be modified using the equations given in [16] for the SkL (3.39), in which case  $J_l^L$  (3.44) and  $J_l^R$  (3.45) can be defined specifying  $k_l^R(\vartheta)$  and  $k_l^L(\vartheta)$  as [58]

$$k_l^R = \left\lfloor \frac{\nu_l \ln \frac{(1-p_l)(1-q_l)}{\xi(1-p_l q_l)}}{\kappa_l} \right\rfloor, \quad (5.8)$$

$$k_l^L = \left\lfloor \nu_l \kappa_l \ln \frac{(1-p_l)(1-q_l)}{\xi(1-p_l q_l)} \right\rfloor, \quad (5.9)$$

where  $\lfloor x \rfloor$  means a maximum integer lower than or equal to  $x$ . Note that functions (5.8) and (5.9) were obtained in [58] by equating (3.39) to  $\xi(N_l) = \text{erfc}(\vartheta/\sqrt{2})$  and solving for  $k_l$ .

For the AEPD-based approximation (4.11), the UB and LB masks can be formed by replacing  $p_l$  and  $q_l$  with, respectively,

$$\bar{p}_l = e^{-\frac{\kappa_l^{\alpha_l}}{\sigma_l^{\alpha_l}}}, \quad (5.10)$$

$$\bar{q}_l = e^{-\frac{1}{\kappa_l^{\alpha_l} \sigma_l^{\alpha_l}}}. \quad (5.11)$$

That allows specifying the right-hand jitter  $k_{l|\alpha E}^R$  and the left-hand jitter  $k_{l|\alpha E}^L$  with, respectively,

$$k_{l|\alpha E}^R = \left\lceil \frac{\sigma_l}{\kappa_l} \ln \frac{(1 - \bar{p}_l)(1 - \bar{q}_l)}{\xi(1 - \bar{p}_l \bar{q}_l)} \right\rceil, \quad (5.12)$$

$$k_{l|\alpha E}^L = \left\lceil \sigma_l \kappa_l \ln \frac{(1 - \bar{p}_l)(1 - \bar{q}_l)}{\xi(1 - \bar{p}_l \bar{q}_l)} \right\rceil. \quad (5.13)$$

Provided (5.12) and (5.13), the jitter left boundary  $J_{l|\alpha E}^L$  and right boundary  $J_{l|\alpha E}^R$  can be finally defined as, respectively,

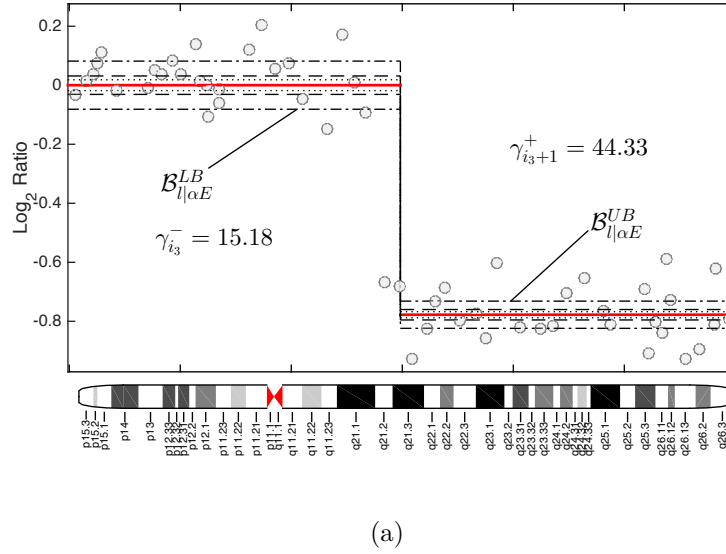
$$J_{l|\alpha E}^L \cong \hat{n}_l - k_{l|\alpha E}^R, \quad (5.14)$$

$$J_{l|\alpha E}^R \cong \hat{n}_l + k_{l|\alpha E}^L. \quad (5.15)$$

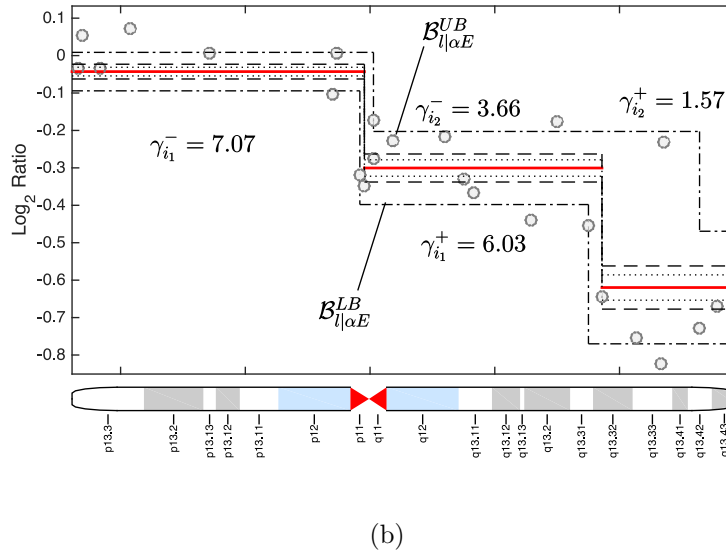
Replacing the equations (5.14) and (5.14) into the algorithm designed in [16], developed in Appendix C, for the SkL-based the confidence masks, it is obtained the  $\mathcal{B}_{l|\alpha E}^{UB}$  and  $\mathcal{B}_{l|\alpha E}^{LB}$  boundaries.

Figures, 5.6a and 5.6b show the confidence masks based on the asymmetric exponential power distribution applied to CNAs of Chromosomes 10 and 19 from neuroblastoma copy number profile 207. The  $\mathcal{B}_{l|\alpha E}^{UB}$  and  $\mathcal{B}_{l|\alpha E}^{LB}$  bounds computed around the breakpoint  $i_3$ , plotted in Figure 5.6a, suggest that this breakpoint can remain at high probability  $\approx 1$ . In the same figure, it can be noticed that the values of SNR are much greater than one  $\gamma_{i_3}^- = 15.18$  and  $\gamma_{i_3}^- = 44.33$ . Unlike, in the Figure 5.6b the breakpoint  $i_2$  possibly do not exist at a greater probability that the breakpoints  $i_1$ , in spite of that  $\gamma_{i_2}^- = 3.66$ .





(a)



(b)

Figure 5.6: a) Chromosome 10 and b) Chromosome 19 of neuroblastoma copy number profile 207 with masks  $\mathcal{B}_{l|\alpha E}^{UB}$  and  $\mathcal{B}_{l|\alpha E}^{LB}$  applied to the CNA estimates (bold). The confidence probabilities are:  $P = 0.5$  ( $\vartheta = 0.6745$ ) (dotted),  $P = 0.75$  ( $\vartheta = 1.15035$ ) (dashed), and  $P = 0.9973$  ( $\vartheta = 3$ ) (dash-dot).

# Chapter 6

## Matching Expert's Annotations

In this Chapter, the efficiency of the proposed AEP-based confidence masks is demonstrated. First, we use the standard circular binary segmentation algorithm [51, 52] to estimate the CNAs in some neuroblastoma copy number profiles. Then the masks are applied and the confidence probabilities founded, which match annotations made by experts. Finally, an analysis of similarities and discrepancies between the regions outlined by masks and experts annotations is done. Throughout this study, we exploit the database of 575 annotated neuroblastoma copy number profiles as a public benchmark available for testing new algorithms [81].

### 6.1 Breakpoints Annotations as Gold Standard

The confidence masks derived and developed in [16, 30] are intended to correct the CNA estimates for the given confidence probability  $P$ . The probability ranges as  $0.5 < P < 1.0$ , but its exact value acceptable for medical needs is still not specified. One way is to specify  $P$  using the breakpoint annotations provided by experts as the *gold standard* [81].

The annotations are counts of breakpoints in genomic regions made by visual inspection of the noisy signal. Observing each region, expert biologists determine whether or not it contains a breakpoint based on their expertise. Let us notice that visual annotations have been used successfully for object recognition in photos and cell phenotype recognition in

microscopy [82, 83]. An example of annotations for the breakpoints is shown in Fig. 6.1 as related to the Profile 44–Chromosome 1 taken from aCGH microarray experiments on neuroblastoma tumors.

Here, data are separated on several regions, which were annotated by the experts as having  $>0$ –Breakpoints, 1–Breakpoint, and 0–Breakpoints. Although the confidence probability was not specified, it can easily be deduced that the probability is high in each annotation.

To specify the expert's probability, the masks must be applied to the CNA estimates and the confidence probability increased until the masks reach the same decision as that by the experts. We present in [84, 85] an example of this procedure to match several algorithms with the annotations made by experts.

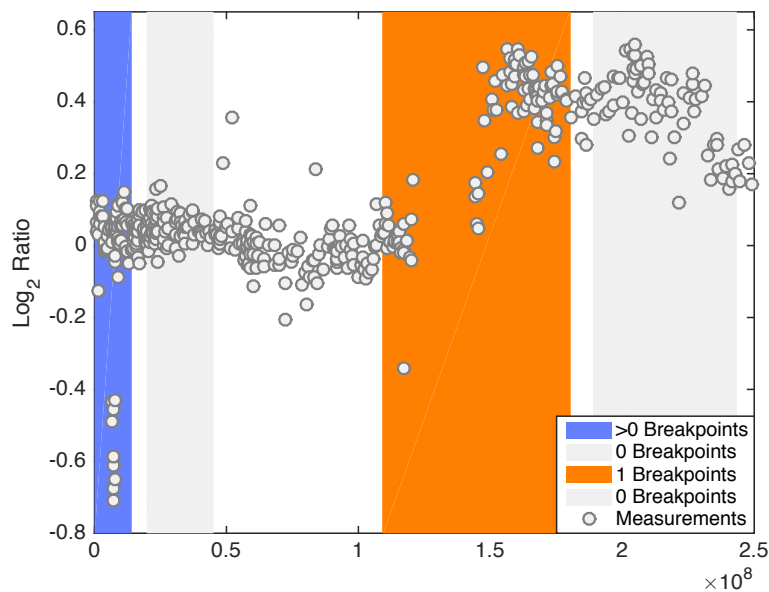


Figure 6.1: Annotations made by medical experts to Profile 44–Chromosome 1 of sample of neuroblastoma. Regions with different annotations are separated. Experts suggested that there exist  $>0$ –Breakpoints, 1–Breakpoint, and 0–Breakpoints. Data were obtained using the aCGH microarrays and plotted as Log<sub>2</sub>Ratio versus genomic position.

## 6.2 Match Cases

The Upper Bound mask based on Asymmetric Exponential Power will be denoted as  $\mathcal{B}_{l|\alpha E}^{\text{UB}}$  and the Lower Bound mask as  $\mathcal{B}_{l|\alpha E}^{\text{LB}}$ . The following expert's annotations will be taken into account for the particular chromosomal regions, which will be classified with colored stripes as follows:

- **0–Breakpoints** means that there are no breakpoints.
- **1–Breakpoint** means that there exists a single breakpoint.
- **$\geq 1$ –Breakpoints** means that there are one or more breakpoints.

### 6.2.1 Case 1–Perfect Match

The Chromosome 10 of profile 207 is analyzed in Fig. 6.2 using masks  $\mathcal{B}_{l|\alpha E}^{\text{UB}}$  and  $\mathcal{B}_{l|\alpha E}^{\text{LB}}$ . The confidence probabilities are:  $P = 0.5$  ( $\vartheta = 0.6745$ ) (dotted),  $P = 0.75$  ( $\vartheta = 1.15035$ ) (dashed), and  $P = 0.9973$  ( $\vartheta = 3$ ) (dash-dot). The expert's annotations (striped) are 0–Breakpoints and 1–Breakpoints. For convenience, the genomic position is represented here with the chromosome ideogram related to *Homo Sapiens*. The breakpoints of CNAs plotted in Figure 6.2 are rough and obvious to the naked eye. For this reason, the CNA estimates tested by the masks and the experts annotations match each other when  $\vartheta < 12.8$  with an extremely high probability of  $P \approx 1 - 1.11 \times 10^{-16}$ .

### 6.2.2 Case 2–Good Match

An analysis of chromosome 11 and chromosome 19 for neuroblastoma copy number profile 207 is represented in Fig. 6.3a and Fig. 6.3b, respectively. Two annotations to this measurements were made 0–Breakpoints and  $\geq 1$ –Breakpoints (striped). The masks  $\mathcal{B}_{l|\alpha E}^{\text{UB}}$  and  $\mathcal{B}_{l|\alpha E}^{\text{LB}}$  tested the CNAs using the confidence probabilities are  $P = 0.5$  (dotted),  $P = 0.75$  (dashed), and  $P = 0.9973$  (dash-dot). Here, it was demonstrated a good match with the annotations with  $\vartheta < 5.8$  (dash-dot) that corresponds to the very high probability of  $P = 1 - 6.63 \times 10^{-9}$ .

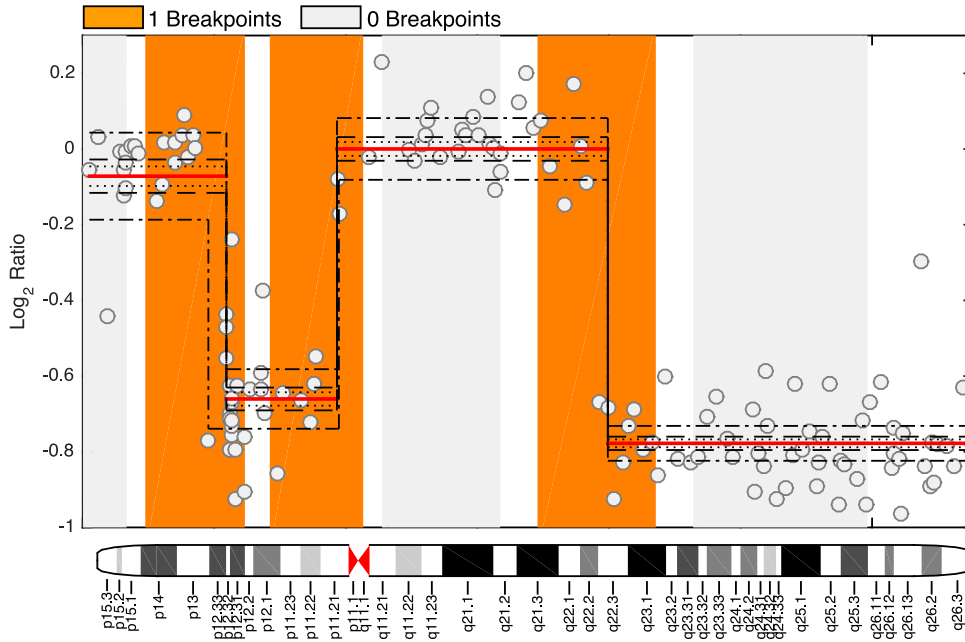
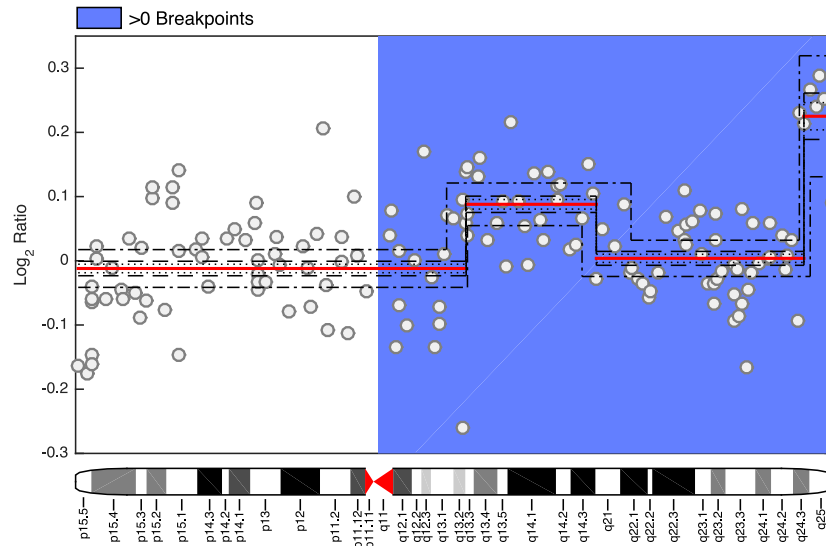


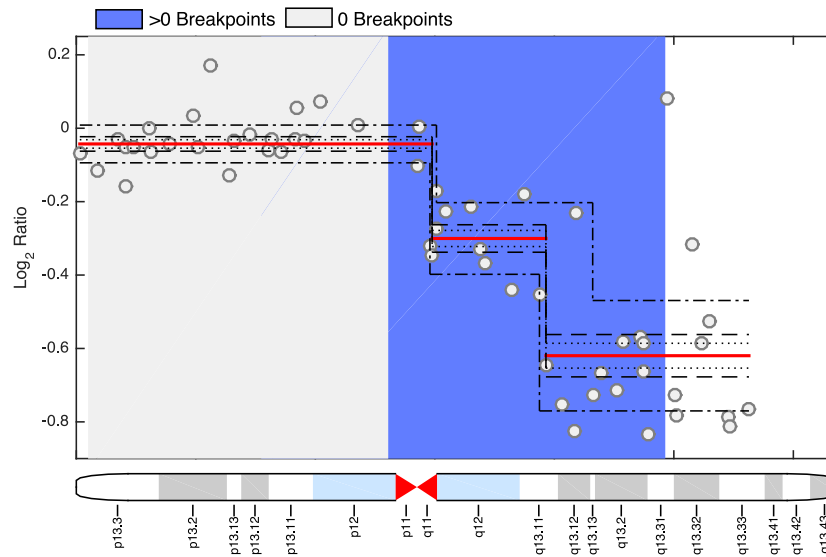
Figure 6.2: Chromosome 10 of neuroblastoma copy number profile 207 with masks  $\mathcal{B}_{l|\alpha E}^{UB}$  and  $\mathcal{B}_{l|\alpha E}^{LB}$  applied to the CNA estimates (bold) and expert’s annotations (striped): 0–Breakpoints and 1–Breakpoints. The confidence probabilities are:  $P = 0.5$  ( $\vartheta = 0.6745$ ) (dotted),  $P = 0.75$  ( $\vartheta = 1.15035$ ) (dashed), and  $P = 0.9973$  ( $\vartheta = 3$ ) (dash-dot).

### 6.2.3 Case 3–Wrong Match

Estimates obtained for chromosomes 2 and 17 of profile 207 suggest that some annotations made by experts are definitely wrong. In fact, within the region from 116305178 to 242918939 bp annotated by experts as “normal,” an estimator has discovered two breakpoints (Fig. 6.4a). For this case, the bounds  $\mathcal{B}_{l|\alpha E}^{UB}$  and  $\mathcal{B}_{l|\alpha E}^{LB}$  at a confidence probabilities of  $P = 0.9973$  or  $\vartheta = 3$ , Figures 6.4a (dotted) and Fig. 6.4b (dash-dot), suggest that all breakpoints still exist. However, the confidence masks discard these breakpoints with the probability of  $P = 1 - 1.44 \times 10^{-4}$  and  $P = 1 - 4.13 \times 10^{-5}$ . In the measurement of chromosome 17 (Fig. 6.4b), a single but not annotated breakpoint was found at the genomic position 60401416. The masks match this annotation with  $\vartheta = 9.7$ .



(a)



(b)

Figure 6.3: Estimate (bold), masks  $\mathcal{B}_{l|\alpha E}^{UB}$  and  $\mathcal{B}_{l|\alpha E}^{LB}$ , and expert's annotations (striped) for neuroblastoma copy number profile 207: (a) chromosome 11 and (b) chromosome 19. Experts have recognized two regions as having 0-Breakpoints and >0-Breakpoints. The confidence probabilities are  $P = 0.5$  (dotted),  $P = 0.75$  (dashed), and  $P = 0.9973$  (dash-dot).

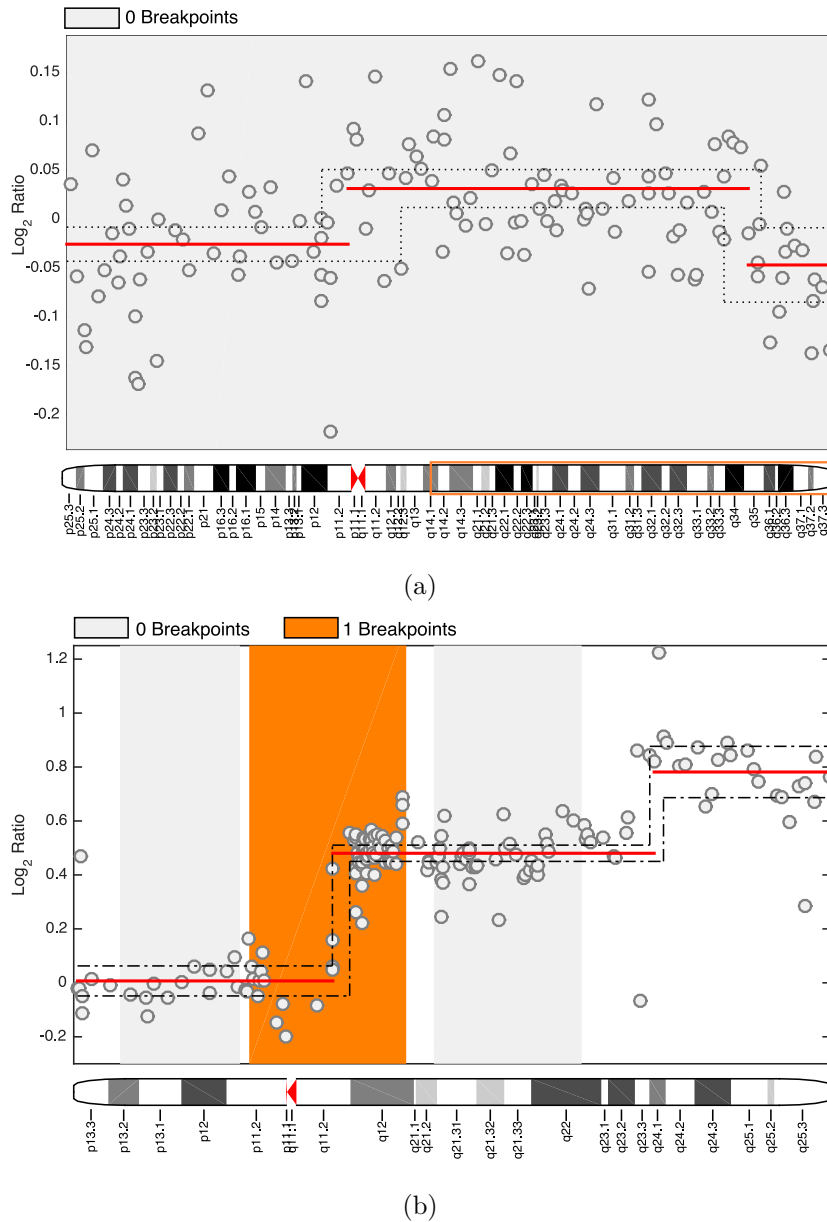
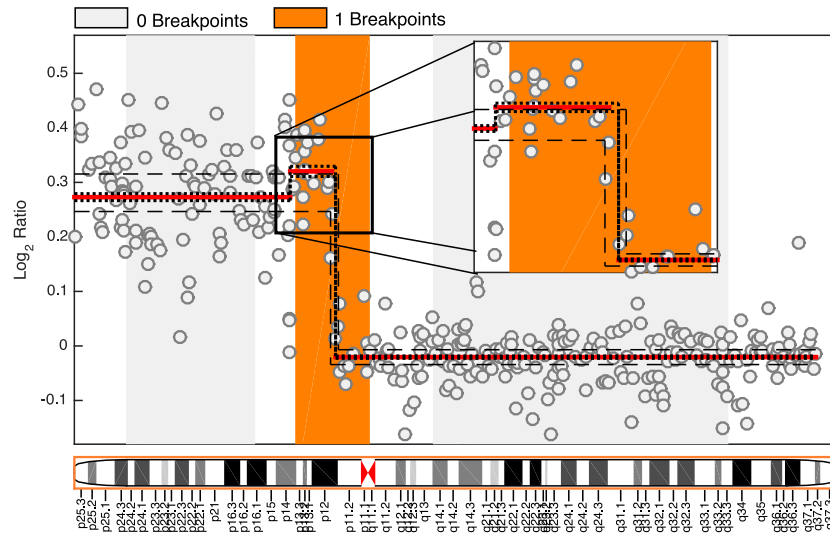


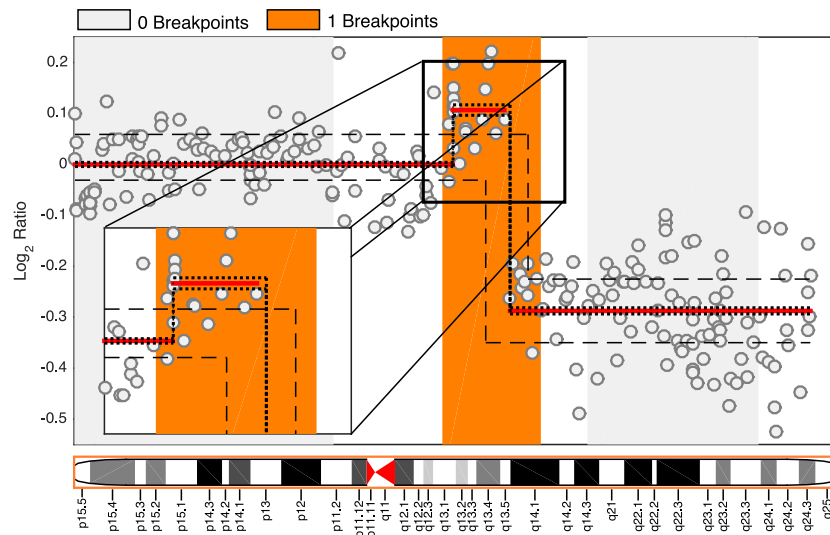
Figure 6.4: The  $\mathcal{B}_{l|\alpha E}^{UB}$  and  $\mathcal{B}_{l|\alpha E}^{LB}$  masks around the chromosome 2 and 17 for neuroblastoma copy number profile 207. The confidence probabilities are calculated with  $P = 0.9973$  a) (dotted) and b) (dash-dot).

### 6.2.4 Case 3.1–Transitional Match

Finally, it is considered a chromosome 2 of profile 12 (Fig. 6.5a) and chromosome 11 of profile 522 (Fig. 6.5b), which demonstrate that the expert probability can be determined



(a)



(b)

Figure 6.5: The  $\mathcal{B}_{l|\alpha E}^{UB}$  and  $\mathcal{B}_{l|\alpha E}^{LB}$  masks around the chromosome 2 and 11 for neuroblastoma copy number profile 22 and 522 respectively. a) The confidence masks (dashed) with a probability of  $P = 1 - 4.13 \times 10^{-3}$  suggest to remove the first breakpoint and b) the proposed algorithm with a probability of  $1 - 5.23 \times 10^{-10}$  also suggest to clear the first breakpoint(dashed).



in the transitional match. As can be seen, the experts failed to identify the breakpoints located at 70563419 and 69360977. The breakpoint discovered by an estimator at 70563419 in the chromosome 2 (Fig. 6.5a) was discarded by the masks with the probability of  $P = 1 - 4.13 \times 10^{-3}$ . In turn, the breakpoint discovered at 69360977 in the chromosome 11 (Fig. 6.5b) was discarded with the probability of  $P = 1 - 5.23 \times 10^{-10}$ .

By analyzing the results provided above, two important conclusions are obtained:

- All of the breakpoints identified by experts correspond to a high level of the segmental SNR,  $\gamma \geq 1$ , i.e. the CNAs exist at a high probability. Such breakpoints do not need the confidence masks to adjust the estimates.
- Some CNAs, which were estimated for low SNR values,  $\gamma < 1$ , but not annotated by experts, can be disagreed by the mask to match the experts annotations and, thereby, determine the experts confidence probability.

Table 6.1: Comparison between several profiles annotations with CNAs estimated by CBS and tested by confidence masks: CASE-I Excellent match and CASE-II Poor match.

Profile	Case	Chromosome	Profile	Case	Chromosome
1	I	2,3,6,7,11,15,17,19,X.	57	I	1,2,3,7,8,13,14,19.
	II	1,4,9.		II	4,5,12,16,17.
8	I	3,7,11,12,18.	162	I	1,2,3,5,7,11,17,19.
	II	1,2,4,17.		II	4,8,22.
12	I	1,3,6,11,14,22.	207	I	2,3,4,5,10,12,15,18,19.
	II	2,4,17,X.		II	1,17,21,22.
22	I	3, 9,11,12,13,17,19,21.	316	I	1,2,3,11,15,17,19,20.
	II	1,2,4.		II	4,7,12.
44	I	2,3,4,5,7,11,17,19.	522	I	2,11 ,21.
	II	1		II	1,3,4,17

An overall comparison between the CNA estimates provided by CBS and the experts annotations is given in Table 6.1. All cases are separated here into two classes. The Case I representing an excellent match, which means that the estimates perfectly match the annotations and no additional analysis is required using the masks. The Case II corresponds to a poor match, which means that the number of breakpoints identified by an estimator differs from that annotated by the experts. In this case, the masks must be applied to determine the experts confidence probability.

# Chapter 7

## Algorithms comparizon using Confidence Masks

In this Chapter, we compare the CNAs breakpoint estimates produced using the Circular Binary Segmentation and Pruned Exact Linear Time methods. To reach this goal, the breakpoints estimated with the Next Generation Sequencing technology are established as a reference, in other words, as ideal estimates. Then the modified confidence masks based on the AEP distribution is applied at several levels of probability to clear false positives estimated by CBS and PELT and to increase their accuracy. Also, we provide a complete analysis of the deleted breakpoints and the length of CNAs at each level of probability in the  $\vartheta$ -sense.

### 7.1 Comparison of breakpoints estimators

Due to the relevance of identification and classification of CNAs to identify diseases, many methods have been proposed to estimate the breakpoints and segmental constants in the CNAs with highest precision using the most powerful technologies of hybridization. However, locations and lengths of CNAs estimated using well-elaborated methods are often contradictory due to extensive variability of measurements and performance of the algorithms. Still much less attention is given to the estimation accuracy and it is difficult

to select the best estimator. We illustrate in Figure 7.1 a procedure to compare the breakpoints estimated with any method, for this example the CBS and PELT, with respect to an ideal estimation.

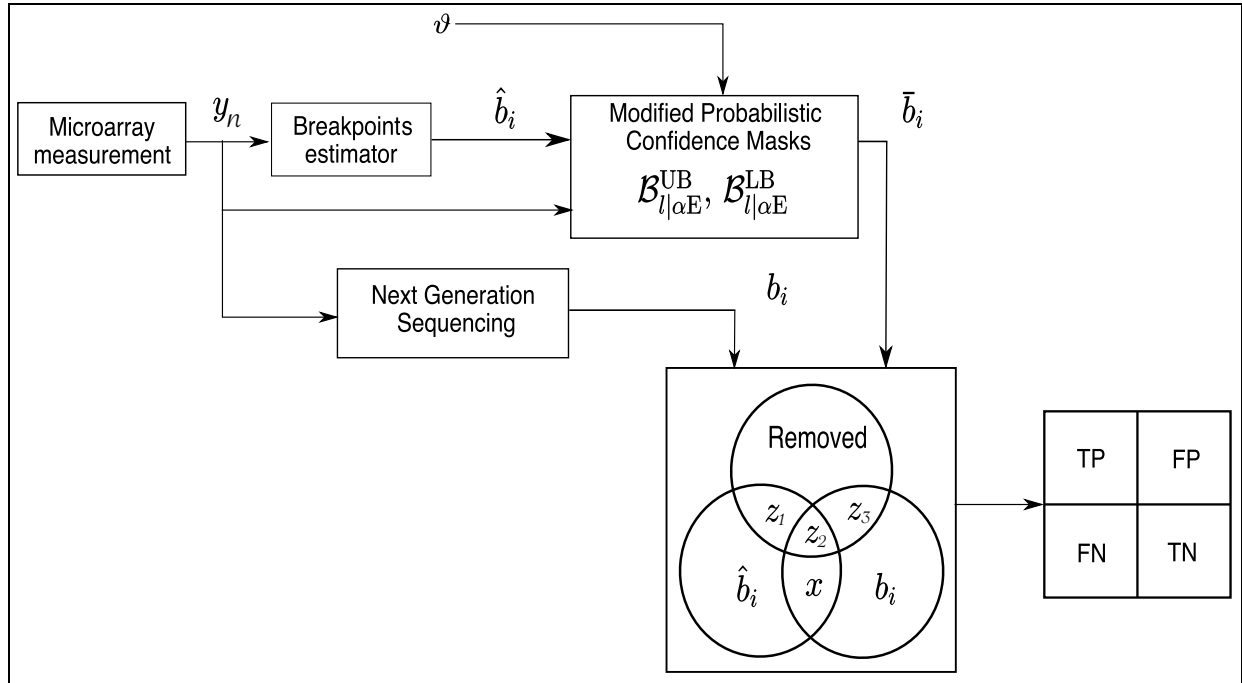


Figure 7.1: Procedure diagram to compare the breakpoints estimator methods respect to an ideal estimation obtained using the Next Generation Sequencing technology. The modified probabilistic masks remove the breakpoints that unsatisfied a given probability, which is specified with the  $\vartheta$ -sense.

The procedure showed in Fig. 7.1 implied several steps. First, the microarray measurement  $y_n$  is processed to estimate the breakpoints  $\hat{b}_i$  of CNAs with CBS and PELT methods. The copy number data considered in this chapter to illustrate the method were provided by Institute Curie, Paris, France Ovarian cancer samples were sequenced using shallow Whole genome sequencing technology; copy number estimations were obtained using ControlFree tool [86]; copy number profiles were randomly modified for anonymization. Sample\_1 and Sample\_2 were chosen to test our method because they represent highly altered variants of cancer genomic profiles. The package *changepoint* implements this methods in R, which

is a free software environment for statistical computing and graphics. The breakpoints  $\hat{b}_i$  are obtained with the command `cpt.mean` specifying the method BINSEG and PELT, respectively. To both algorithms, it is specified a manual penalty `value="log(n)` and a minimal length segment of 5 points to both algorithms. Next, the modified probabilistic confidence masks based on AEP distribution remove some breakpoints that do not exceed a threshold of probability.

The ideal estimation of breakpoints  $b_i$  is given by the method Next Generation Sequencing, which is the high resolution technology most reliable that exists to this day. However, the comparison is delimited to breakpoints  $b_i$  of segments with a length of 1, 2 and 3 Mbp seeking the best conditions for each algorithm. This restriction is based on the argument that many times medical experts made their analysis and diagnostics according to a specific length of estimated CNAs.

Finally, the matched breakpoints  $x$ , where the location of  $\hat{b}_i = b_i$ , the breakpoints removed of  $\hat{b}_i$  and  $x$  represented as  $z_1$  and  $z_2$  respectively, are showed with a Venn diagram. It can be noted that  $z_2 = z_3$  because the breakpoints  $b_i$  are unprocessed with the confidence masks, *i. e.* the unique points estimated with the NGS that can be deleted are the match points. So, it is obtained four possible results True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) points, which are represented in Table 7.1. The True Negative points are obtained erasing a False Positive using the confidence masks.

The Figures 7.2 and 7.3 show the measurement of Chromosomes 4 and 18 from Sam-

Table 7.1: Possible cases of comparison between a particular breakpoints detector –CBS or PELT– and Next Generation Sequencing estimation, represented with the symbols  $\circ$  and  $\nabla$ , respectively. The case of True Negatives is given when a False Positive is removed using the confidence masks, it is illustrated with the symbol  $\otimes$ .

Method	TP	FP	TN	FN
Breakpoint estimator	$\circ$	$\circ$	$\otimes$	
Reference NGS	$\nabla$			$\nabla$

ple\_1, respectively. The comparison between CBS  $\hat{b}_i$  (circles) and NGS  $b_i$  (plus sign) breakpoints is illustrated in the Figure 7.2. The estimated breakpoints of NGS are delimited at 1 Mega bases (Mb) in this example because the reason mentioned above. It can be noticed that the CNAs obtained with this restriction are quite obvious to the naked eye. For this reason, some estimated breakpoints are removed (cross) using the confidence masks at a very high probability, for example the breakpoint  $\hat{b}_9$  likely do no exist in the sigma sense of  $\vartheta = 14$ . However, the segments around the breakpoint  $\hat{b}_9$  were estimated as normal CNAs by the NGS algorithm.

A comparizon of the PELT  $\hat{b}_i$  (circles) and NGS  $b_i$  (plus sign) breakpoints given in the figure 7.3. It is worth mentioning that the PELT algorithm estimates more breakpoints than the CBS algorithm for the sames parameters. Here, several estimated breakpoints are removed in a range of sigma sense of  $\vartheta \in [0.6745, 20]$  or a probability of 50% to  $\approx 100\%$ . The measurements of CNAs have many oscillations close to the centromere of the DNA chain, which are detected by the PELT algorithm and completely removed with modified confidence masks.

Using a Venn diagram it is possible to visualize the global results of comparison of the 16 patients of 23 Chromosomes from microarray database. Figure 7.4 shows the results for the CBS method at several values of  $\vartheta$ , respectively. This Venn diagrams show a total of 1073 breakpoints estimated by the CBS method, 672 points estimated by the NGS algorithm, 167 match points, in other words the breakpoints estimated that shared the same position. An amount of 18 estimated points by CBS are deleted at the minimum probability of 50% and a total of 659 deleted breakpoints at a maximum probability  $\approx 100\%$ . Based on the results of figure 7.4, all match breakpoints of NGS delimited at 1Mb exist at a probability of  $< 1 - 2.78 \times 10^{-5}$  or at  $< \vartheta = 4.19$  where the first 5 points are removed. After this probability the match breakpoints begin to be erased by the probabilistic confidence masks  $\mathcal{B}_{l|\alpha E}^{UB}$  and  $\mathcal{B}_{l|\alpha E}^{LB}$  finishing with 53 points removed at  $\vartheta = 20$ .

Following the same procedure, the PELT algorithm obtained a total of 3002 CNAs breakpoints, 2.79 times more points than CBS. The figure 7.5 gives the comparison of breakpoints estimated with CBS and NGS algorithms employing a Venn diagram . In the comparison PELT versus NGS, the match points are increased to 353. However, at

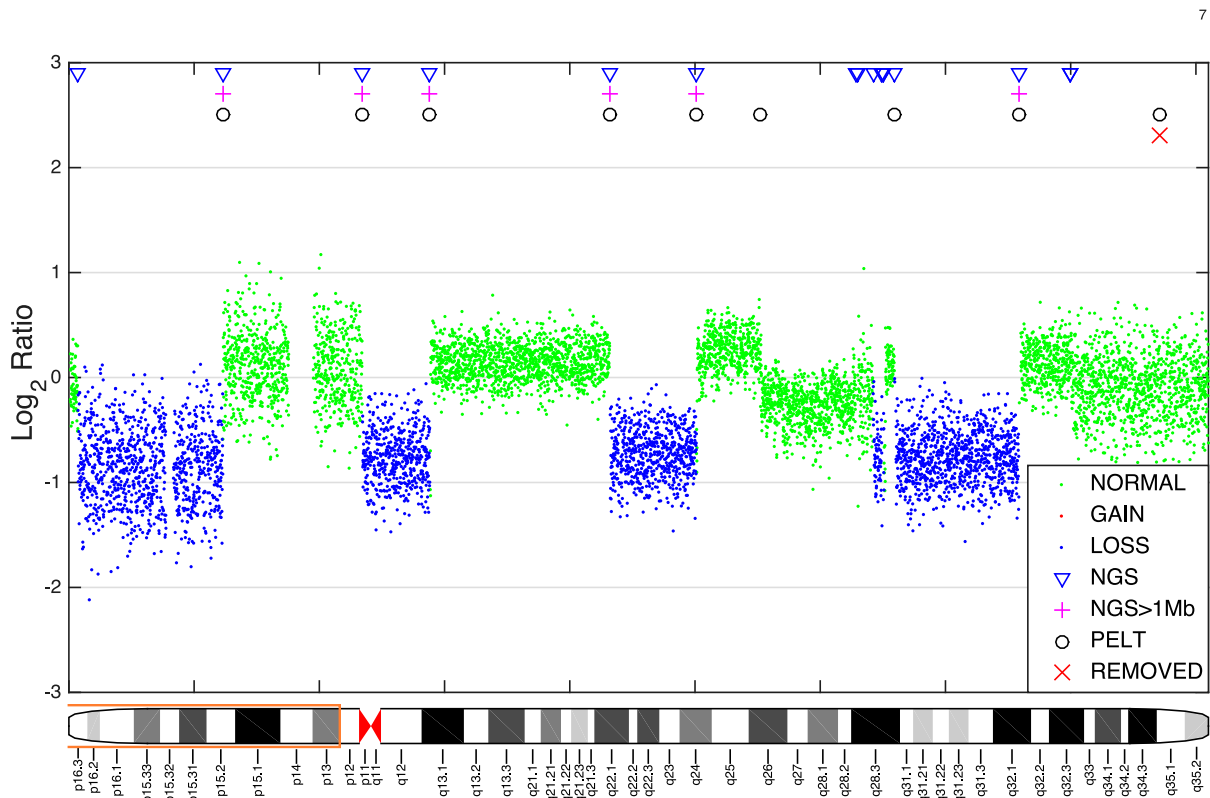


Figure 7.2: Estimated CNAs—annotated as Normal, Gains or Loss—to measurements of Chromosome 4 from the Sample\_1, which is plotted Genomic Position versus  $\text{Log}_2$  Ratio. The total of breakpoints located by the NGS (triangle down) is limited to segments greater than 1 Mega bases (plus sign). The points estimated by the CBS method (circles) can be removed using the modified confidence masks (cross) in a range of  $\vartheta$  from 0.6745 to 20.

the probability of  $1 - 2.78 \times 10^{-5}$  the confidence masks suggest that 13 match points not exist. Initially, only 4 points estimated by PELT are deleted at a probability of 50%, 2421 breakpoints at the maximum probability  $\approx 100\%$  and 207 match points to the same level.

The number of breakpoints removed with the modified confidence masks can be summarized in the figure 7.6. Here it is plotted the estimated breakpoints of CBS (circle) and PELT (square) respect to a level of specified probability in the  $\sigma$ -sense. The initial number of estimated change points in both curves decrease when the probability increased from

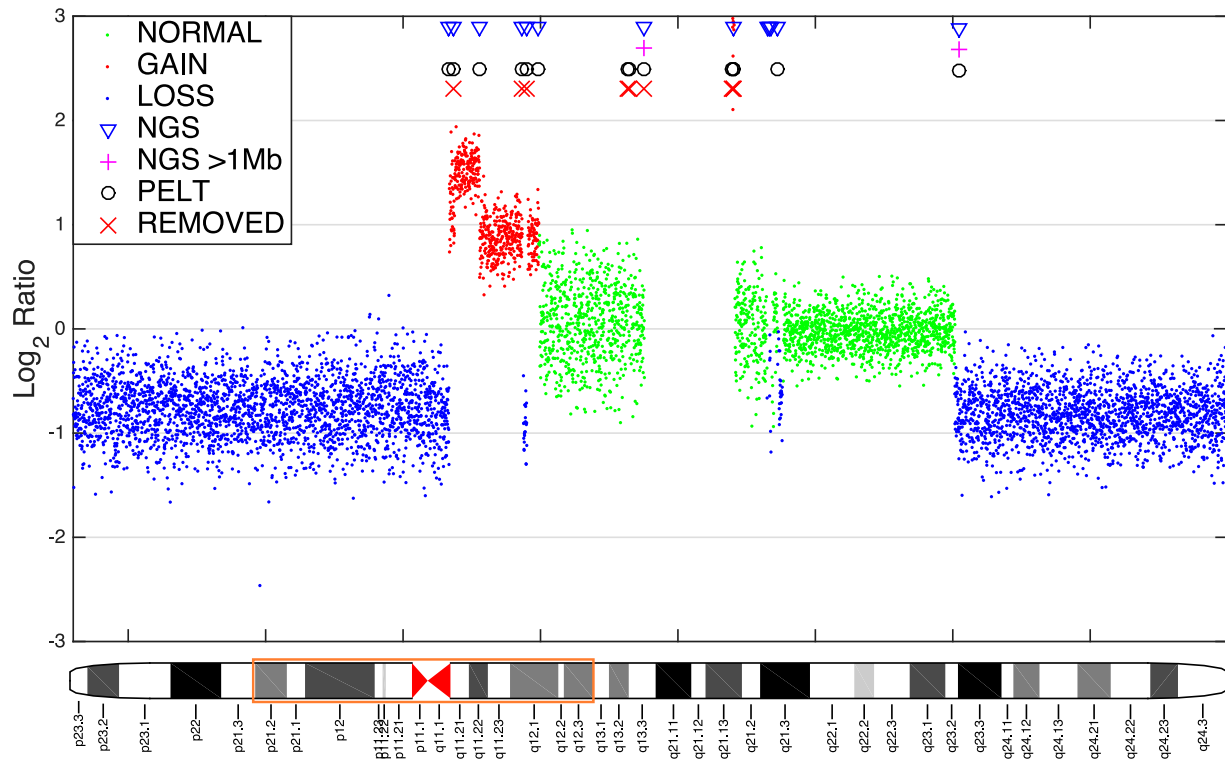


Figure 7.3: Estimated CNAs—annotated as Normal, Gains or Loss— to measurements of Chromosome 8 from the Sample\_1, which is plotted Genomic Position versus Log2 Ratio. The total of breakpoints located by the NGS (triangle down) is limited to segments greater than 1 Mega bases (plus sign). The points estimated by the CBS method (circles) can be removed using the modified confidence masks (cross) in a range of  $\vartheta$  from 0.6745 to 20.

50% to  $\approx 100\%$ . The analysis showed in figure 7.6 can be useful to set the breakpoints at a required probability.

Based on the computed points TP, FN, TN, and FP it is possible to show the results using a Receiver Operating Characteristic (ROC) curve. The ROC space is created plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) [87]. The TPR and FPR are also known as sensitivity and as the fall-out or probability of false alarm, respectively. These parameters are calculated using the below equations



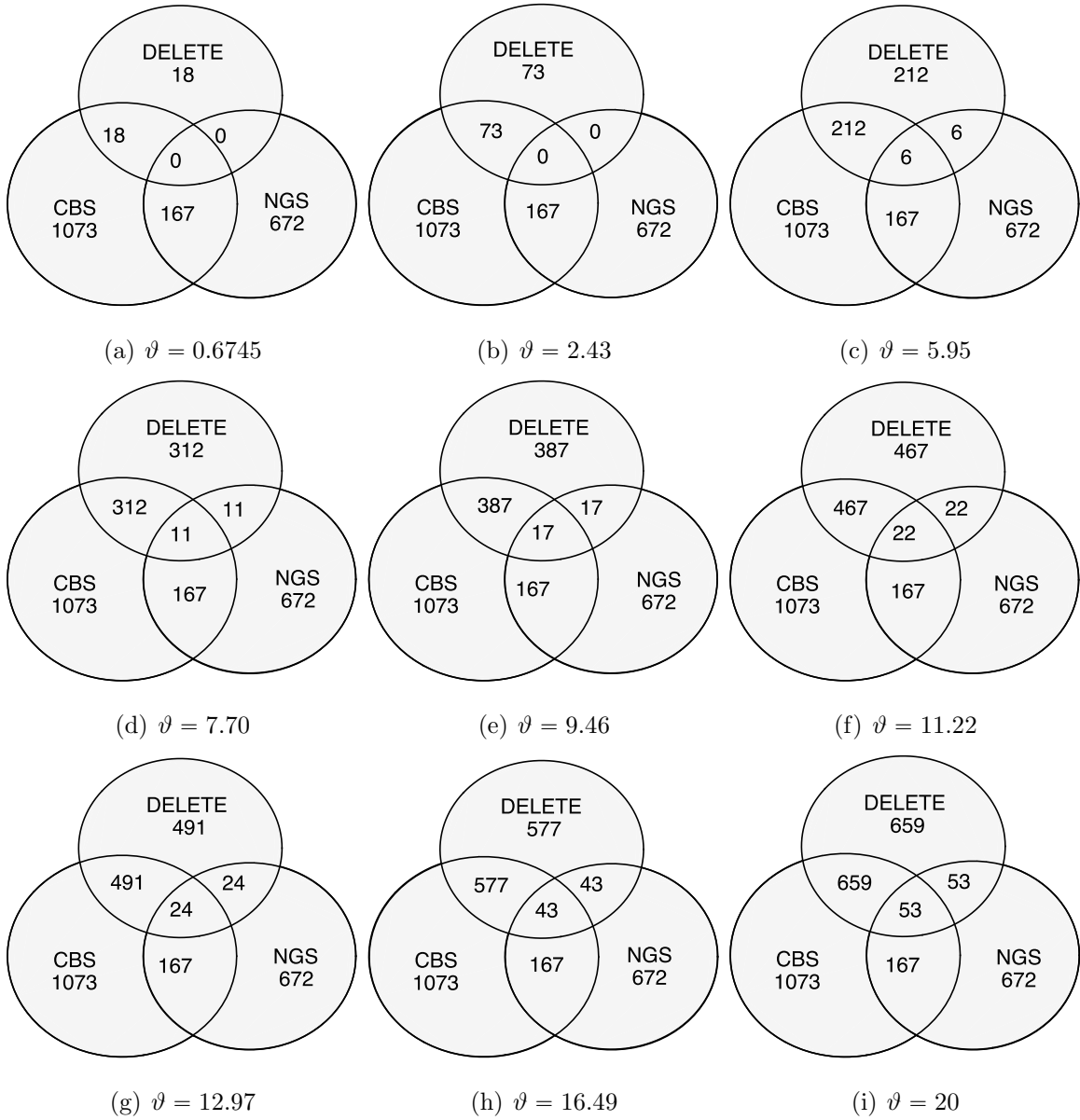


Figure 7.4: Comparative of breakpoints estimated with Circular Binary Segmentation and Next Generation Sequencing algorithms employing a Venn diagram. Based on the suggestion of confidence masks  $\mathcal{B}_{l|\alpha E}^{UB}$  and  $\mathcal{B}_{l|\alpha E}^{LB}$  the breakpoints estimated by CBS method can be refined at a given probability, parameter  $\vartheta$ .

$$TPR = \frac{TP}{P} = \frac{TP}{TP+FN} \quad (7.1)$$

$$FPR = \frac{FP}{N} = \frac{FP}{FP+TN} = 1 - TNR \quad (7.2)$$

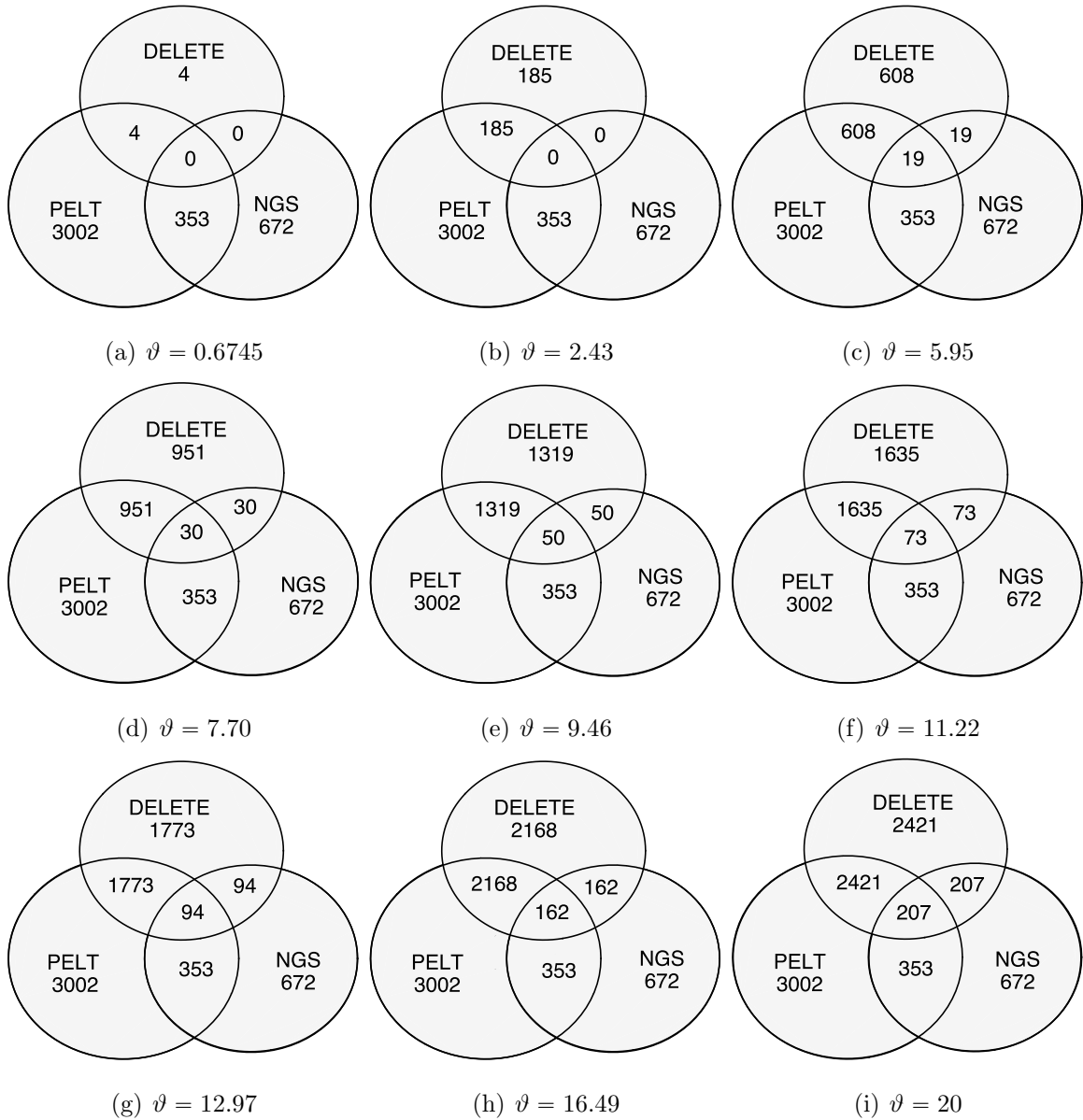


Figure 7.5: Comparative of breakpoints estimated with Circular Binary Segmentation and Next Generation Sequencing algorithms employing a Venn diagram. Based on the suggestion of confidence masks  $\mathcal{B}_{l|\alpha E}^{UB}$  and  $\mathcal{B}_{l|\alpha E}^{LB}$  the breakpoints estimated by CBS method can be refined at a given probability, parameter  $\vartheta$ .

where TNR is defined as the specificity and computed as  $TNR = \frac{TN}{TN + FP}$ , P the number of real positive cases and N the number of real cases in the data. So, the Figures

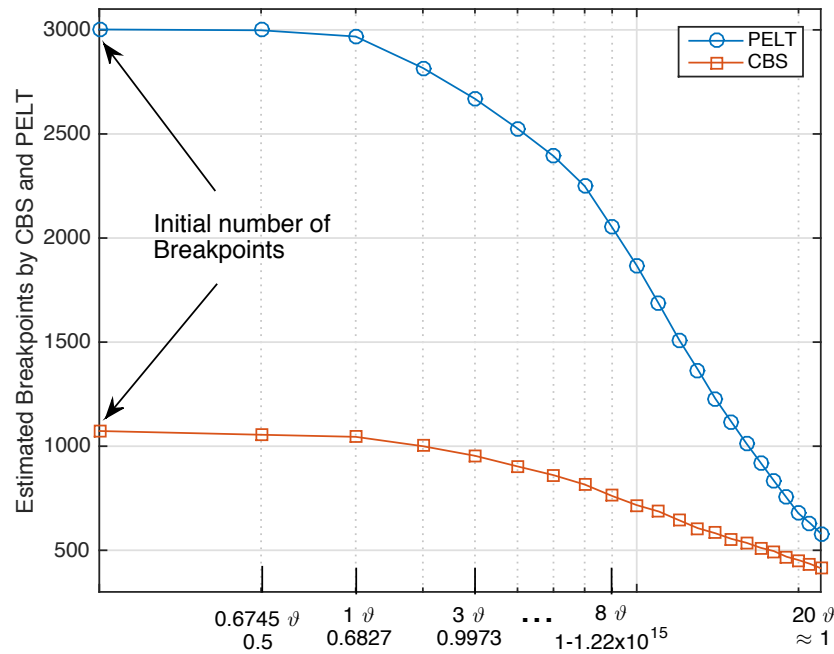
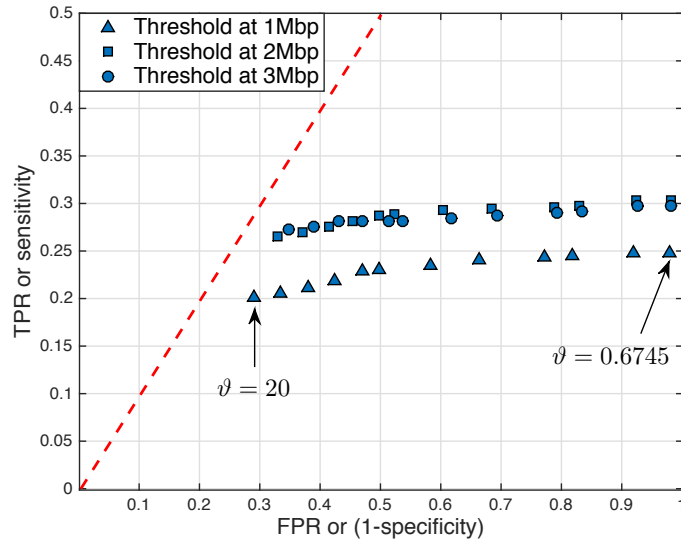


Figure 7.6: Estimated breakpoints of CNAs using the methods a) CBS (circle) and b) PELT (square) and deleted change points at specified probability. The first value of both curves is the initial number of estimated breakpoints.

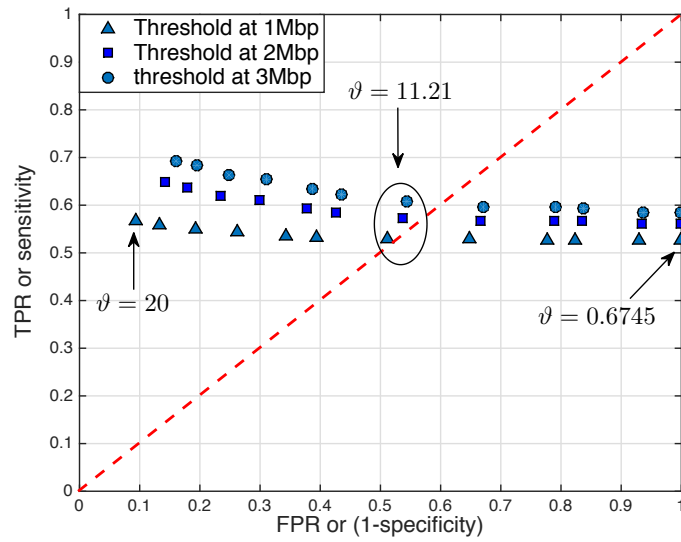
7.7a and 7.7b show the results of comparison in the range of  $\vartheta$  from 0.6745 to 20 in the ROC space of each method with three delimiters for CNAs of NGS: 1 (triangle), 2 (square), and 3 (circle) Mbp. The curve generated by CBS algorithm is far from being a good estimator to the CNA measurements from the Sample.1, because any results in the ROC space cross to the region of better classification (upper region). Nevertheless, the results exhibited by the PELT estimations suggest to be a good estimator at a value of  $\vartheta = 11.21$ , showed with a circle in the figure 7.7b.

### 7.1.1 CNAs Size Analysis

As mentioned above, sometimes the studies of medical experts are based on a definite size of CNAs to give a diagnostic of a particular disease. Because the probabilistic confidence masks are applied to test CNAs according to several intervals of confidence, there exist a

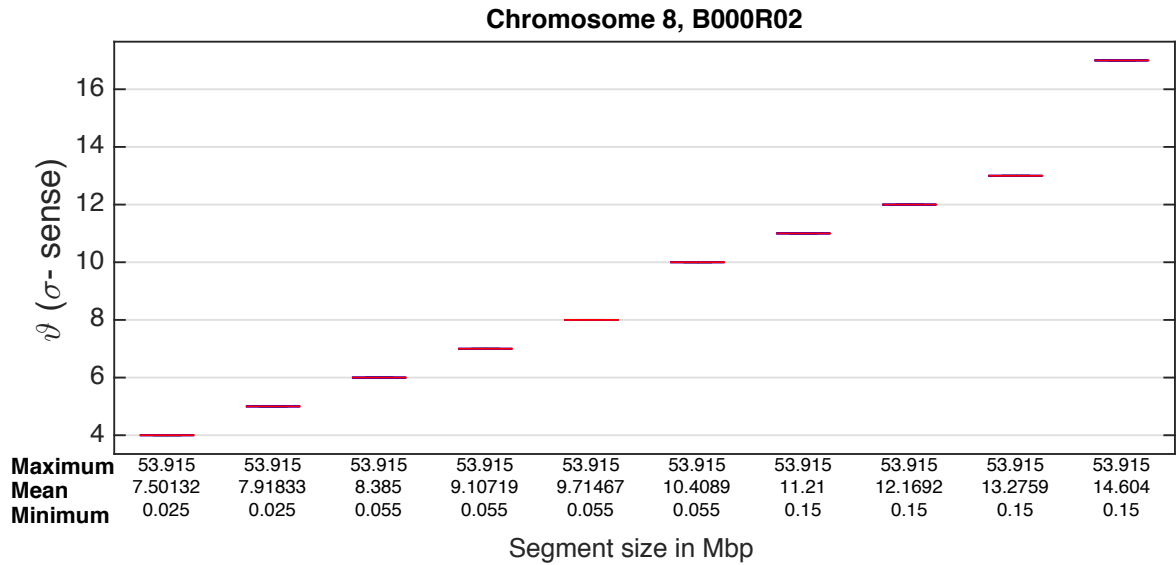


(a) CBS

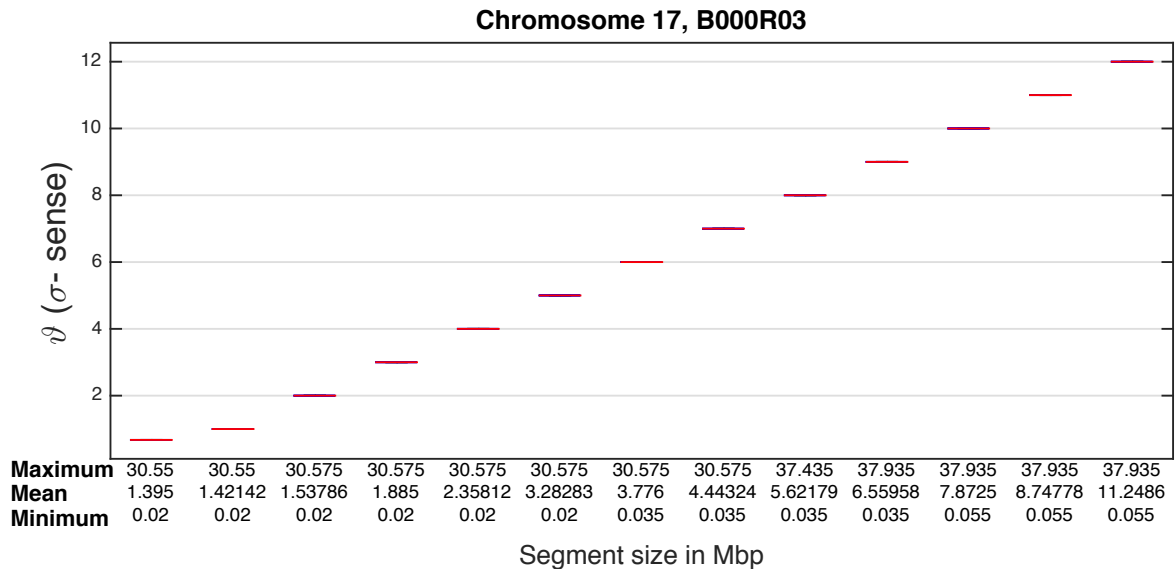


(b) PELT

Figure 7.7: True positive rate against the false positive rate based on the results of comparison between a) CBS and b) PELT with NGS estimates at three thresholds: 1 (triangle), 2 (square), and 3 (circle) Mega base pairs for a range  $\vartheta$  from 0.6745 to 20.



(a) CBS



(b) PELT

Figure 7.8: Size analysis of CNAs estimated by a) CBS and b) PELT tested with the probabilistic confidence masks  $\mathcal{B}_{l|\alpha E}^{UB}$  and  $\mathcal{B}_{l|\alpha E}^{LB}$ .

direct influence over the resulting CNAs at each level of probability.

To exemplify this affirmation, the breakpoints estimated by CBS of Chromosome 8 from

---

Sample\_1 and by PELT of Chromosome 17 from Sample\_2 were tested with the confidence masks  $\mathcal{B}_{l|\alpha E}^{UB}$  and  $\mathcal{B}_{l|\alpha E}^{LB}$  increasing the confidence intervals  $\vartheta$  from 0.6745 to 20. Next, the CNAs size is inspected at each level finding the maximum, the mean and the minimum length of tested segments. The figures 7.8a and 7.8b illustrate the analysis of CNAs size. The maximum segment showed in the figure 7.8a estimated by the CBS method do not change its length at any level of probability. The mean and minimum value of CNAs are increased 1.94 and 5.99 times, respectively. For the case of PELT estimates, Figure 7.8b, the maximum segment increase its length 7.385 Mbp and the minimum segment have a little difference of 0.035 Mbp. The mean of CNAs estimated with PELT method and tested by confidence masks is 9.8536 Mbp greater, growing 8.063 times.

# Chapter 8

## Conclusions

### 8.1 About Heuristic Approximation

The Bessel function-based heuristic approximation of the jitter distribution in the breakpoints is more accurate than the Laplace-based one. This is particularly true for low and extra low SNR values ( $\gamma_i^-, \gamma_i^+$ ) often observed in probes of small chromosomal changes. Note that, when  $\text{SNR} \ll 1$ , the Laplace distribution often is computed in complex numbers.

The confidence probabilistic masks formed with the Bessel-based approximation give a more correct picture for possible locations of chromosomal changes on a probabilistic field. These masks argue that the CNA estimates may be improved when the SNR reaches low values. Several estimates of chromosomal changes obtained in Project GAP using the SNP technology were tested by the masks and improved accordingly by removing some unlikely existing breakpoints.

Even though the heuristic approximation have less error than Laplace distribution, the Bessel-based distribution shows mathematical disadvantages and a simple analytic form for the jitter distribution should be sought in order to use it in the probabilistic masks.

## 8.2 About Laplace–Parametrization

The parametrization of the Laplace density provided by several approximations of the  $k$ -varying segmental noise variance has demonstrated higher accuracy in bounding jitter in the breakpoints with given confidence probability. The parametrization has appeared to be especially efficient for low and extra low segmental SNR values, when the break-point locations are unrecognized visually. The mathematical support of the skew Laplace distribution gives a great advantage of this technique in contrast to the Heuristic approximation.

The hybrid confidence masks combining best outputs of the particular masks have demonstrated an ability to bound the jitter with a high accuracy for practically all segmental SNR values observed in chromosomal probing. That was confirmed by testing the masks by a chromosome sample having 59 segments and 58 breakpoints and associated with breast cancer.

It has also been revealed that the left and right jitter in the breakpoints correlate each other. The parametrization of the Laplace density can be provided with more accuracy when accounting the correlation properties of the  $k$ -varying segmental noise variances. This problem is under investigation.

## 8.3 About AEP approximation

The AEPD (asymmetric exponential power distribution) has appeared to be more accurate than the SkL discrete skew Laplace and the above described distributions in the approximation of jitter distribution, obtained by the ML estimator, in the breakpoints of the CNAs.

This is particularly true for  $\text{SNR} < 1$  and  $\text{SNR} \ll 1$  values often observed in probes of small CNAs using CGH microarrays. Another advantage of AEPD is that if  $\text{SNR} \gg 1$  the parameter  $\alpha \approx 1$  and this distribution converges to the Skew Laplace distribution.

Concerning to the confidence masks, the computed upper and lower bounds decreased the error generated by other distributions in order to avoid uncertainties and false deci-



sions. To corroborate, the confidence masks were applied to real data obtained using the micro-array of HR-CGH coinciding correctly with the annotations established by medical experts. Referring to this approximation, It is necessary to seek a simple relationships between the AEPD parameters and the segmental SNRs to use them in the confidence probabilistic masks.

## 8.4 About Matching Expert's Annotations

The comparison of AEPD-based confidence probabilistic masks and experts annotations on the testing set of CNA profiles of neuroblastoma show improvement of the CNA estimates. This result implied that modified confidence masks can give additional information to biologists for diagnostic and prognostic purpose.

The CNAs estimated by the standard CBS algorithm were tested by the confidence masks and improved accordingly by removing some unlikely existing breakpoints and thus matching better with the annotations. Based on this procedure, It has been specified the probability  $P_\epsilon$  of the *gold standard* as  $P_\epsilon^{\min} = 0.9998 < P_\epsilon < P_\epsilon^{\max} \lesssim 1$  in the  $3.21 - \sigma$  interquartile with an average probability of  $\bar{P}_\epsilon = 1 - 1.41 \times 10^{-12}$ . Talking about this procedure, we propose using several algorithms to detect breakpoints and evaluate its efficiency.

## 8.5 About Comparative of algorithms using Confidence Masks

The CNAs estimated by Circular Binary Segmentation (CBS) and Pruned Exact Linear Time (PELT) algorithms were compared with the estimates generated by the most modern technology that exists, Next Generation Sequencing (NGS). The methods to find CNAs are based on dissimilar statistical foundations, for this reason the estimates from each technique are uncorrelated. Aiming to obtain better results, the CNAs of NGS were restricted using the thresholds of 1, 2, and 3 Megabase pairs.

---

Testing the CNAs estimates obtained by CBS and PELT using the probabilistic confidence masks  $\mathcal{B}_{l|\alpha E}^{UB}$  and  $\mathcal{B}_{l|\alpha E}^{LB}$ , it was enhanced the matching of these algorithms respect to the NGS estimates. Modulating the parameter  $\vartheta$  from 0.6745 to 20 it was showed that the algorithm PELT is better estimator than CBS to the database provided by the Institute of Curie in all cases. The best versions of algorithm PELT are found when the confidence masks cleared its estimations at a value of  $\vartheta = 11.21$ . The probability is represented in the  $\sigma$ -sense because it has very high values  $\approx 100\%$

Finally, we show the impact on the length of CNA estimated at several levels of probability using the confidence masks. This procedure give a general vision that small segments in size can remain at high probabilities because the confidence masks use more parameters that the CNAs length  $N$ .

# Bibliography

- [1] Nicholas A Graham, Aspram Minasyan, Anastasia Lomova, Ashley Cass, Nikolas G Balanis, Michael Friedman, Shawna Chan, Sophie Zhao, Adrian Delgado, James Go, et al. Recurrent patterns of dna copy number alterations in tumors reflect metabolic selection pressures. *Molecular systems biology*, 13(2):914, 2017.
- [2] Robert Weinberg. *The biology of cancer*. Garland science, 2013.
- [3] Darrin Stuart and William R Sellers. Linking somatic genetic alterations in cancer to therapeutics. *Current opinion in cell biology*, 21(2):304–310, 2009.
- [4] Barbara Weir, Xiaojun Zhao, and Matthew Meyerson. Somatic alterations in the human cancer genome. *Cancer cell*, 6(5):433–438, 2004.
- [5] Donna G Albertson, Colin Collins, Frank McCormick, and Joe W Gray. Chromosome aberrations in solid tumors. *Nature genetics*, 34(4):369, 2003.
- [6] Farahnaz Forozan, Ritva Karhu, Juha Kononen, Anne Kallioniemi, and Olli-P Kallioniemi. Genome screening by comparative genomic hybridization. *Trends in Genetics*, 13(10):405–409, 1997.
- [7] Michael R Speicher and Nigel P Carter. The new cytogenetics: blurring the boundaries with molecular biology. *Nature Reviews Genetics*, 6(10):782, 2005.
- [8] Pauline C Ng and Ewen F Kirkness. Whole genome sequencing. In *Genetic variation*, pages 215–226. Springer, 2010.

- 
- [9] J Munoz-Minjares and YS Shmaliy. The role of optimal detection of cnas and error analysis using next generation sequencing. *Next Generat Sequenc & Applic*, 3(141):2, 2016.
- [10] Daniel Pinkel, Richard Seagraves, Damir Sudar, Steven Clark, Ian Poole, David Kowbel, Colin Collins, Wen-Lin Kuo, Chira Chen, Ye Zhai, et al. High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nature genetics*, 20(2):207, 1998.
- [11] Fatima Zare, Michelle Dow, Nicholas Monteleone, Abdelrahman Hosny, and Sheida Nabavi. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC bioinformatics*, 18(1):286, 2017.
- [12] Tatiana Popova, Valentina Boeva, Elodie Manié, Yves Rozenholc, Emmanuel Barillot, and Marc-Henri Stern. Analysis of somatic alterations in cancer genome: from snp arrays to next generation sequencing, 2013.
- [13] Jorge Muñoz Minjares and Yuriy S Shmaliy. Improving estimates of the breakpoints in genome copy number alteration profiles with confidence masks. *Biomedical Signal Processing and Control*, 31:238–248, 2017.
- [14] Jorge Munoz-Minjares, Jesus Cabal-Aragon, and Yuriy S Shmaliy. Jitter probability in the breakpoints of discrete sparse piecewise-constant signals. In *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, pages 1–5. IEEE, 2013.
- [15] Andreas R Tobler, Sabine Short, Mark R Andersen, Teodoro M Paner, Jason C Briggs, Stephen M Lambert, Priscilla P Wu, Yiwen Wang, Alexander Y Spoonde, Ryan T Koehler, et al. The snplex genotyping system: a flexible and scalable platform for snp genotyping. *Journal of biomolecular techniques: JBT*, 16(4):398, 2005.
- [16] Jorge Munoz-Minjares, Jesús Cabal-Aragón, and Yuriy S Shmaliy. Confidence masks for genome dna copy number variations in applications to hr-cgh array measurements. *Biomedical Signal Processing and Control*, 13:337–344, 2014.

- 
- [17] Hua Ren, Wendy Francis, Amber Boys, Anderly C Chueh, Nick Wong, Phung La, Lee H Wong, Jacinta Ryan, Howard R Slater, and KH Andy Choo. Bac-based pcr fragment microarray: High-resolution detection of chromosomal deletion and duplication breakpoints. *Human mutation*, 25(5):476–482, 2005.
- [18] Philippe La Rosa, Eric Viara, Philippe Hupé, Gaëlle Pierron, Stéphane Liva, Pierre Neuvial, Isabel Brito, Séverine Lair, Nicolas Servant, Nicolas Robine, et al. Vamp: visualization and analysis of array-cgh, transcriptome and other molecular profiles. *Bioinformatics*, 22(17):2066–2073, 2006.
- [19] Yee Hwa Yang, Sandrine Dudoit, Percy Luu, David M Lin, Vivian Peng, John Ngai, and Terence P Speed. Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic acids research*, 30(4):e15–e15, 2002.
- [20] Aastha Joshi. Speech emotion recognition using combined features of hmm & svm algorithm. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(8), 2013.
- [21] Philippe Hupé, Nicolas Stransky, Jean-Paul Thiery, François Radvanyi, and Emmanuel Barillot. Analysis of array cgh data: from signal ratio to gain and loss of dna regions. *Bioinformatics*, 20(18):3413–3422, 2004.
- [22] Claire Lemaitre, Eric Tannier, Christian Gautier, and Marie-France Sagot. Precise detection of rearrangement breakpoints in mammalian chromosomes. *BMC bioinformatics*, 9(1):286, 2008.
- [23] Kim Wong, Thomas M Keane, James Stalker, and David J Adams. Enhanced structural variant and breakpoint detection using svmerge by integration of multiple detection methods and local assembly. *Genome biology*, 11(12):R128, 2010.
- [24] Robert Tibshirani and Pei Wang. Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics*, 9(1):18–29, 2007.

- 
- [25] Erez Ben-Yaacov and Yonina C Eldar. A fast and flexible method for the segmentation of aCGH data. *Bioinformatics*, 24(16):i139–i145, 2008.
- [26] Holger Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010.
- [27] Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of change-points with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.
- [28] J Munoz-Minjares, O Ibarra-Manzano, and YS Shmaliy. Maximum likelihood estimation of dna copy number variations in hr-cgh arrays data. *Proc. 12th WSEAS Int. Conf. on Signal Process., Comput. Geometry and Artif. Vision (ISCGAV'12), Proc. 12th WSEAS Int. Conf. on Systems Theory and Sc. Comput. (ISTASC'12), Istanbul, Turkey.*, pages 45–50, 2012.
- [29] Yuriy S Shmaliy and Luis J Morales-Mendoza. FIR smoothing of discrete-time polynomial signals in state space. *IEEE Transactions on Signal Processing*, 58(5):2544–2555, 2010.
- [30] Jorge Munoz-Minjares, Yuriy S Shmaliy, and Jesús Cabal-Aragón. Confidence limits for genome dna copy number variations in hr-cgh array measurements. *Biomedical Signal Processing and Control*, 10:166–173, 2014.
- [31] Jorge Munoz-Minjares and Yuriy S Shmaliy. Approximate jitter probability in the breakpoints of genome copy number variations. In *Electrical Engineering, Computing Science and Automatic Control (CCE), 2013 10th International Conference on*, pages 128–131. IEEE, 2013.
- [32] Jan O Korbel, Alexander Eckehart Urban, Fabian Grubert, Jiang Du, Thomas E Royce, Peter Starr, Guoneng Zhong, Beverly S Emanuel, Sherman M Weissman, Michael Snyder, et al. Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. *Proceedings of the National Academy of Sciences*, 104(24):10110–10115, 2007.

- 
- [33] Sergii Ivakhno, Tom Royce, Anthony J Cox, Dirk J Evers, R Keira Cheetham, and Simon Tavaré. Cnasega novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics*, 26(24):3051–3058, 2010.
- [34] Tatiana Popova, Elodie Manié, Dominique Stoppa-Lyonnet, Guillem Rigaiil, Emmanuel Barillot, and Marc Henri Stern. Genome alteration print (gap): a tool to visualize and mine complex cancer genomic profiles obtained by snp arrays. *Genome biology*, 10(11):R128, 2009.
- [35] Jin P Szatkiewicz, WeiBo Wang, Patrick F Sullivan, Wei Wang, and Wei Sun. Improving detection of copy-number variation by simultaneous bias correction and read-depth segmentation. *Nucleic acids research*, 41(3):1519–1532, 2012.
- [36] Evert van den Broek, Stef van Lieshout, Christian Rausch, Bauke Ylstra, Mark A van de Wiel, Gerrit A Meijer, Remond JA Fijneman, and Sanne Abeln. Genebreak: detection of recurrent dna copy number aberration-associated chromosomal break-points within genes. *F1000Research*, 5, 2016.
- [37] Cancer Genome Atlas Research Network et al. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456):43, 2013.
- [38] Colin S Cooper. Applications of microarray technology in breast cancer research. *Breast Cancer Research*, 3(3):158, 2001.
- [39] Michael R Barnes. Human genetic variation: databases and concepts. *FOR GENETICISTS*, page 39, 2003.
- [40] LJ Engle, CL Simpson, and JE Landers. Using high-throughput snp technologies to study cancer. *Oncogene*, 25(11):1594, 2006.
- [41] David SP Tan, Maryou BK Lambros, Rachael Natrajan, and Jorge S Reis-Filho. Getting it right: designing microarray (and not microawry) comparative genomic hybridization studies for cancer research. *Laboratory Investigation*, 87(8):737, 2007.

- 
- [42] Jorge S Reis-Filho. Next-generation sequencing. *Breast Cancer Research*, 11(3):S12, 2009.
- [43] Erik Pettersson, Joakim Lundeberg, and Afshin Ahmadian. Generations of sequencing technologies. *Genomics*, 93(2):105–111, 2009.
- [44] R John and Wayne W Grody. Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. *The Journal of Molecular Diagnostics*, 10(6):484–492, 2008.
- [45] Karl V Voelkerding, Shale A Dames, and Jacob D Durtschi. Next-generation sequencing: from basic research to diagnostics. *Clinical chemistry*, 55(4):641–658, 2009.
- [46] Tracy Tucker, Marco Marra, and Jan M Friedman. Massively parallel sequencing: the next big thing in genetic medicine. *The American Journal of Human Genetics*, 85(2):142–154, 2009.
- [47] Melissa J Fullwood, Chia-Lin Wei, Edison T Liu, and Yijun Ruan. Next-generation dna sequencing of paired-end tags (pet) for transcriptome and genome analyses. *Genome research*, 19(4):521–532, 2009.
- [48] Olena Morozova and Marco A Marra. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5):255–264, 2008.
- [49] Tukey J W. Exploratory data analysis. 1971.
- [50] E Arias-Castro and Donoho D L. Does median filtering truly preserve edges better than linear filtering? *Annals of Statistics*, 3(37):1172–1206, 2009.
- [51] Adam B Olshen, ES Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572, 2004.
- [52] ES Venkatraman and Adam B Olshen. A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, 23(6):657–663, 2007.



- 
- [53] Andrew Jhon Scott and M Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, pages 507–512, 1974.
- [54] Ivan E Auger and Charles E Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bulletin of mathematical biology*, 51(1):39–54, 1989.
- [55] Jorge Munoz-Minjares, Yuriy S Shmaliy, and A J Cabal. Noise studies in measurements and estimates of stepwise changes in genome dna chromosomal structures. *Adv. App Pure Math*, 2014.
- [56] Roger Pique-Regi, Antonio Ortega, Ahmed Tewfik, and Shahab Asgharzadeh. Detecting changes in dna copy number: Reviewing signal processing techniques. *IEEE Signal Processing Magazine*, 29(1):98–107, 2012.
- [57] Yuriy S Shmaliy. On the multivariate conditional probability density of a vector perturbed by gaussian noise. *IEEE Transactions on Information Theory*, 53(12):4792–4797, 2007.
- [58] J Munoz-Minjares, J Cabal-Aragon, and YS Shmaliy. Effect of noise on estimate bounds for genome dna structural changes. *WSEAS Trans. on Biology and Biomedicine*, 11:52–61, 2014.
- [59] Tomasz J Kozubowski and Seidu Inusah. A skew laplace distribution on integers. *Annals of the Institute of Statistical Mathematics*, 58(3):555–571, 2006.
- [60] Franck Picard, Stephane Robin, Marc Lavielle, Christian Vaisse, and Jean-Jacques Daudin. A statistical approach for array cgh data analysis. *BMC bioinformatics*, 6(1):27, 2005.
- [61] Paweł Stankiewicz and James R Lupski. Structural variation in the human genome and its role in disease. *Annual review of medicine*, 61:437–455, 2010.
- [62] Karen Buysse, Barbara Delle Chiaie, Rudy Van Coster, Bart Loeys, Anne De Paepe, Geert Mortier, Frank Speleman, and Björn Menten. Challenges for cnv interpretation

- in clinical molecular karyotyping: lessons learned from a 1001 sample experience. *European journal of medical genetics*, 52(6):398–403, 2009.
- [63] Geert Vandeweyer and R Frank Kooy. Detection and interpretation of genomic structural variation in health and disease. *Expert review of molecular diagnostics*, 13(1):61–82, 2013.
- [64] A John Iafrate, Lars Feuk, Miguel N Rivera, Marc L Listewnik, Patricia K Donahoe, Ying Qi, Stephen W Scherer, and Charles Lee. Detection of large-scale variation in the human genome. *Nature genetics*, 36(9):949, 2004.
- [65] Jonathan R Pollack, Therese Sørli, Charles M Perou, Christian A Rees, Stefanie S Jeffrey, Per E Lonning, Robert Tibshirani, David Botstein, Anne-Lise Børresen-Dale, and Patrick O Brown. Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences*, 99(20):12963–12968, 2002.
- [66] Alexander Eckehart Urban, Jan O Korb, Rebecca Selzer, Todd Richmond, April Hacker, George V Popescu, Joseph F Cubells, Roland Green, Beverly S Emanuel, Mark B Gerstein, et al. High-resolution mapping of dna copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, 103(12):4534–4539, 2006.
- [67] Peter J Campbell, Philip J Stephens, Erin D Pleasance, Sarah O’Meara, Heng Li, Thomas Santarius, Lucy A Stebbings, Catherine Leroy, Sarah Edkins, Claire Hardy, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature genetics*, 40(6):722, 2008.
- [68] Jared T Simpson, Rebecca E McIntyre, David J Adams, and Richard Durbin. Copy number variant detection in inbred strains from short read sequence data. *Bioinformatics*, 26(4):565–567, 2009.

- 
- [69] Robert Lucito, John Healy, Joan Alexander, Andrew Reiner, Diane Esposito, Maoyen Chi, Linda Rodgers, Amy Brady, Jonathan Sebat, Jennifer Troge, et al. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome research*, 13(10):2291–2305, 2003.
- [70] Jorge Munoz-Minjares, Jesus Cabal-Aragon, and Yuriy S Shmaliy. Probabilistic bounds for estimates of genome dna copy number variations using hr-cgh microarray. In *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, pages 1–5. IEEE, 2013.
- [71] Jorge Munoz Minjares and Yuriy S Shmaliy. Bounding errors in estimates of genome copy number variations using snp array. *International Journal of Biology and Biomedical Engineering*, 9:127–132.
- [72] Jorge Munoz-Minjares and Yuriy S Shmaliy. An algorithm for bounding errors in estimate of genome cnvs using snp array technology. In *Advances in Artificial Intelligence and Soft Computing, (MICAI), 2015 Proceedings of the 14th Mexican International Conference on*, pages 1–5. Springer, 2015.
- [73] Jorge Munoz-Minjares and Yuriy S Shmaliy. Approximate jitter probability in the breakpoints of genome copy number variations. In *Electrical Engineering, Computing Science and Automatic Control (CCE), 2013 10th International Conference on*, pages 128–131. IEEE, 2013.
- [74] Jorge Munoz-Minjares, Yuriy S Shmaliy, Luis J Morales-Mendoza, Miguel Vazquez-Olguin, and Carlos Lastre-Dominguez. Accurate jitter computation in cna breakpoints using hybrid confidence masks with applications to snp array probing. *IEEE Access*, 2017.
- [75] Jorge Munoz-Minjares, Yuriy S Shmaliy, Re Olivera-Reyna, and O Vite-Chavez. Improving approximation of jitter probability in the breakpoints of simulated copy number alterations. In *Electrical Engineering, Computing Science and Automatic Control (CCE), 2016 13th International Conference on*, pages 1–5. IEEE, 2016.

- 
- [76] Walter Greiner and Joachim Reinhardt. *Quantum electrodynamics*. Springer Science & Business Media, 2008.
- [77] Abraham Ayebo and Tomasz J Kozubowski. An asymmetric generalization of gaussian and laplace laws. *Journal of Probability and Statistical Science*, 1(2):187–210, 2003.
- [78] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.
- [79] Jorge Munoz-Minjares, Yuriy S Shmaliy, Ro Olivera-Reyna, Re Olivera-Reyna, and RJ Perez-Chimal. Jitter representation in scena breakpoints using asymmetric exponential power distribution. In *Electrical Engineering, Computing Science and Automatic Control (CCE), 2017 14th International Conference on*, pages 1–5. IEEE, 2017.
- [80] Jorge Munoz-Minjares, Yuriy S Shmaliy, Ro Olivera-Reyna, Re Olivera-Reyna, and O Vite-Chavez. Approximation of jitter probability in the breakpoints using aep distribution and confidence masks of scena. In *Power, Electronics and Computing (ROPEC), 2017 IEEE International Autumn Meeting on*, pages 1–6. IEEE, 2017.
- [81] Toby Dylan Hocking, Gudrun Schleiermacher, Isabelle Janoueix-Lerosey, Valentina Boeva, Julie Cappo, Olivier Delattre, Francis Bach, and Jean-Philippe Vert. Learning smoothing models of copy number profiles using breakpoint annotations. *BMC bioinformatics*, 14(1):164, 2013.
- [82] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008.
- [83] Thouis R Jones, Anne E Carpenter, Michael R Lamprecht, Jason Moffat, Serena J Silver, Jennifer K Grenier, Adam B Castoreno, Ulrike S Eggert, David E Root, Polina Golland, et al. Scoring diverse cellular morphologies in image-based screens with

- iterative feedback and machine learning. *Proceedings of the National Academy of Sciences*, 106(6):1826–1831, 2009.
- [84] JM Munoz-Minjares and YS Shmaliy. Matching confidence masks with experts annotations for estimates of chromosomal copy number alterations. *Journal of Genetic Disorders*, 1(1:9):1–3, 2017.
- [85] Jorge Muñoz-Minjares, Yuriy S Shmaliy, Tatiana Popova, and RJ Perez-Chimal. Matching confidence masks with experts annotations for estimates of chromosomal copy number alterations. In *International Conference on Bioinformatics and Biomedical Engineering*, pages 85–94. Springer, 2018.
- [86] Valentina Boeva, Tatiana Popova, Kevin Bleakley, Pierre Chiche, Julie Cappo, Gudrun Schleiermacher, Isabelle Janoueix-Lerosey, Olivier Delattre, and Emmanuel Barillot. Control-freec: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, 28(3):423–425, 2011.
- [87] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

# Appendices

# Appendix A

## A.1 Analysis of Gaussian Process

To find the probabilities of events  $\mathbf{A}_l$  and  $\mathbf{B}_l$  the following analysis of gaussian process is described. First, it is defined the equations in terms of  $\gamma$ :

$$p_1(y) = \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{(y-\Delta)^2}{2\sigma_x^2}} \quad (\text{A.1})$$

$$= \sqrt{\frac{\gamma_l^-}{2\pi\Delta^2}} e^{-\frac{\gamma_l^- (y-\Delta)^2}{2\Delta^2}} \quad (\text{A.2})$$

$$p_2(y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{y^2}{2\sigma_y^2}} \quad (\text{A.3})$$

$$= \sqrt{\frac{\gamma_l^+}{2\pi\Delta^2}} e^{-\frac{\gamma_l^+ y^2}{2\Delta^2}} \quad (\text{A.4})$$

**Case I** :  $\sigma_x > \sigma_y$ ,  $\gamma_l^- < \gamma_l^+$ . Defining the limits and substituting the gaussian function, it is established that

$$P(\mathbf{A}_l) = 1 - \int_{\beta}^{\alpha} p_1(y) dy \quad (\text{A.5})$$

$$= 1 - \sqrt{\frac{\gamma_l^-}{2\pi\Delta^2}} \int_{\beta}^{\alpha} e^{-\frac{\gamma_l^- (y-\Delta)^2}{2\Delta^2}} dy \quad (\text{A.6})$$

Now, using the definition of erf and erfc

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \quad (\text{A.7})$$

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt \quad (\text{A.8})$$

and setting a change of variable  $t = \sqrt{\frac{\gamma_l^-}{2\Delta^2}}(y - \Delta)$  and modifying the limits, the probability of  $\mathbf{A}_l$  can be represented as:

$$P(\mathbf{A}_l) = 1 - \frac{1}{\sqrt{\pi}} \int_{g_l^\beta}^{g_l^\alpha} e^{-t^2} dt \quad (\text{A.9})$$

where  $g_l^\beta = \frac{\beta_l - \Delta_l}{|\Delta_l|} \sqrt{\frac{\gamma_l^-}{2}}$  and  $g_l^\alpha = \frac{\alpha_l - \Delta_l}{|\Delta_l|} \sqrt{\frac{\gamma_l^-}{2}}$ . Finally, the equation (A.9) is represented in the form

$$\boxed{P(\mathbf{A}_l) = 1 + \frac{1}{2} [\operatorname{erf}(g_l^\beta) - \operatorname{erf}(g_l^\alpha)]}. \quad (\text{A.10})$$

In the same way, defining the limits and substituting the gaussian function, the probability of event  $\mathbf{B}_l$  is defined as:

$$P(\mathbf{B}_l) = \int_{\beta}^{\alpha} p_2(y) dy = \frac{1}{\Delta} \sqrt{\frac{\gamma_l^+}{2\pi}} \int_{\beta}^{\alpha} e^{-\frac{\gamma_l^+ y^2}{2\Delta^2}} dy \quad (\text{A.11})$$

setting a change of variable  $t = \sqrt{\frac{\gamma_l^+}{2\Delta^2}} y$  and modifying the limits, the probability of  $\mathbf{B}_l$  can be represented as:

$$P(\mathbf{B}_l) = \frac{1}{\sqrt{\pi}} \int_{h_l^\beta}^{h_l^\alpha} e^{-t^2} dt \quad (\text{A.12})$$

where  $h_l^\beta = \frac{\beta_l}{|\Delta_l|} \sqrt{\frac{\gamma_l^+}{2}}$ ,  $h_l^\alpha = \frac{\alpha_l}{|\Delta_l|}$ . So. the the equation (A.12) is represented in the form



$$\boxed{P(\mathbf{B}_l) = \frac{1}{2}[\text{erf}(h_l^\alpha) - \text{erf}(h_l^\beta)]}. \quad (\text{A.13})$$

**Case II:**  $\sigma_x = \sigma_y$ ,  $\gamma_l^- = \gamma_l^+$

For Case II, the limits are computed respect to  $\alpha$

$$P(\mathbf{A}_l) = 1 - \int_{\alpha}^{\infty} p_1(y) dy \quad (\text{A.14})$$

$$= \frac{1}{\Delta} \sqrt{\frac{\gamma_l^-}{2\pi}} \int_{\alpha}^{\infty} e^{-\frac{\gamma_l^-(y-\Delta)^2}{2\Delta^2}} dy \quad (\text{A.15})$$

$$= \frac{1}{\sqrt{\pi}} \int_{g_l^\alpha}^{\infty} e^{-t^2} dt \quad (\text{A.16})$$

and the probability of event  $\mathbf{A}_l$  for this case is expressed as

$$\boxed{P(\mathbf{A}_l) = \frac{1}{2}\text{erfc}(g_l^\alpha)} \quad (\text{A.17})$$

The probability of event  $\mathbf{B}_l$  also is computed respect to the constant  $\alpha$  in the next form

$$P(\mathbf{B}_l) = \int_{-\infty}^{\alpha} p_2(y) dy \quad (\text{A.18})$$

$$= \frac{1}{\Delta} \sqrt{\frac{\gamma_l^+}{2\pi}} \int_{-\alpha}^{\infty} e^{-\frac{\gamma_l^+ y^2}{2\Delta^2}} dy \quad (\text{A.19})$$

$$= \frac{1}{\sqrt{\pi}} \int_{-h_l^\alpha}^{\infty} e^{-t^2} dt \quad (\text{A.20})$$

and the probability of event  $\mathbf{B}_l$  for this case is expressed as

$$\boxed{P(\mathbf{B}_l) = 1 - \frac{1}{2}\text{erfc}(h_l^\alpha)} \quad (\text{A.21})$$

**Case III:**  $\sigma_x < \sigma_y$ ,  $\gamma_l^- > \gamma_l^+$

Following the procedure in **Case I**, the limits are defined and the gaussian function replaced:

$$P(\mathbf{B}_l) = \int_{\alpha}^{\beta} p_1(y) dy \quad (\text{A.22})$$

$$= \frac{1}{\Delta} \sqrt{\frac{\gamma_l^-}{2\pi}} \int_{\alpha}^{\beta} e^{-\frac{\gamma_l^- (y-\Delta)^2}{2\Delta^2}} dy. \quad (\text{A.23})$$

Using the limits defined to equation A.9  $g_l^\beta$  and  $g_l^\alpha$ , the probability  $P(\mathbf{B})$  can be expressed as

$$P(\mathbf{A}) = \frac{1}{\sqrt{\pi}} \int_{g_l^\alpha}^{g_l^\beta} e^{-t^2} dt \quad (\text{A.24})$$

and represented in terms of erf and erfc

$$\boxed{P(\mathbf{A}) = \frac{1}{2} [\text{erfc}(g_l^\alpha) - \text{erfc}(g_l^\beta)] = \frac{1}{2} [\text{erf}(g_l^\beta) - \text{erf}(g_l^\alpha)]} \quad (\text{A.25})$$

The probability of event  $P(\mathbf{B})$  for this case is deduced replacing the gaussian function and setting the limits according to values of  $\alpha$  and  $\beta$

$$P(\mathbf{B}_l) = 1 - \int_{\alpha}^{\beta} p_2(y) dy \quad (\text{A.26})$$

$$= 1 - \frac{1}{\Delta} \sqrt{\frac{\gamma_l^+}{2\pi}} \int_{\alpha}^{\beta} e^{-\frac{\gamma_l^+ y^2}{2\Delta^2}} dy \quad (\text{A.27})$$

using the limits as in the equation (A.12)  $h_l^\beta$  and  $h_l^\alpha$  it is obtained the next equation

---

$$P(\mathbf{B}) = \int_{h_l^\alpha}^{h_l^\beta} e^{-t^2} dt \quad (\text{A.28})$$

Finally, the probability of event  $\mathbf{B}$ ,  $P(\mathbf{B})$ , is represented in terms of erf and erfc

$$\boxed{P(\mathbf{A}) = 1 + \frac{1}{2}[\text{erfc}(h_l^\beta) - \text{erfc}(h_l^\alpha)] = 1 + \frac{1}{2}[\text{erf}(h_l^\alpha) - \text{erf}(h_l^\beta)]}. \quad (\text{A.29})$$

# Appendix B

## B.1 Skew Laplace Distribution

The probability density function of the skew Laplace distribution is proposed in [59] and defined in continuous time as:

$$f(x) = \frac{1}{\sigma} \frac{\kappa}{1 + \kappa^2} \begin{cases} e^{-\frac{\kappa}{\sigma}x}, & x \geq 0, \\ e^{-\frac{1}{\kappa\sigma}x}, & x \leq 0, \end{cases} \quad (\text{B.1})$$

where  $\sigma > 0$  is a scale parameter and skewness parameter  $\kappa$ . In the symmetric case ( $\kappa = 1$ ) this leads to a discrete analog of the classical Laplace distribution

$$f(x) = \frac{1}{2\sigma} e^{-\frac{|x|}{\sigma}}. \quad (\text{B.2})$$

The discrete distribution of equation (B.1) takes on an explicit form in terms of the parameters  $d = e^{\kappa/\sigma}$  and  $q = e^{1/\kappa\sigma}$  leading to the following definition.

**Definition.** A random variable  $Y$  has the discrete Laplace distribution with parameters  $d \in (0, 1)$  and  $q \in (0, 1)$ , if

$$f(k|d, q) = \mathbb{P}(Y = k) = \frac{(1 - d_l)(1 - q_l)}{1 - d_l q_l} \begin{cases} d_l^k, & k \geq 0, \\ q_l^{|k|}, & k \leq 0. \end{cases} \quad (\text{B.3})$$

Now, in order to find  $\kappa$  and  $\nu$ , it is needed to change the variable  $\sigma$  by  $\nu$  and based on the relationship  $0 < d_l = e^{-\frac{\kappa_l}{\nu_l}} = P(B_l)^{-1} - 1 < 1$ ,  $0 < q_l = e^{-\frac{1}{\kappa_l \nu_l}} = P(A_l)^{-1} - 1 < 1$  the equation (B.3) can be represented as:

$$p(k) = \frac{1}{\phi} \begin{cases} [P^{-1}(B) - 1]^k & , k > 0, \\ 1 & , k = 0, \\ [P^{-1}(A) - 1]^{|k|} & , k < 0, \end{cases} \quad (\text{B.4})$$

where  $\phi$  is a parameter of normalization.

Using (B.4) at  $k = -1$  and  $k = 1$ , it is obtained the next expressions

$$\frac{(1 - d_l)(1 - q_l)d_l}{1 - d_l q_l} = \frac{1}{\phi} \left( \frac{1}{P(B)} - 1 \right), k = 1 \quad (\text{B.5})$$

$$\frac{(1 - d_l)(1 - q_l)q_l}{1 - d_l q_l} = \frac{1}{\phi} \left( \frac{1}{P(A)} - 1 \right), k = -1. \quad (\text{B.6})$$

which are used to establish a ratio  $s$

$$\frac{d_l}{q_l} = \frac{P(A)(1 - P(B))}{P(B)(1 - P(A))} = s \quad (\text{B.7})$$

isolating the variable  $d_l$  and replacing the variable  $s$

$$d_l = s q_l \rightarrow e^{-\frac{\kappa_l}{\nu_l}} = s e^{-\frac{1}{\kappa_l \nu_l}} \quad (\text{B.8})$$

then, based on (B.8)

$$s = e^{\frac{1}{\kappa_l \nu_l} - \frac{\kappa_l}{\nu_l}} \rightarrow \ln s = \frac{1}{\kappa_l \nu_l} - \frac{\kappa_l}{\nu_l} = \frac{1 - \kappa_l^2}{\kappa_l \nu_l} \quad (\text{B.9})$$

so it is obtained an initial representation of  $\nu_l$

$$\boxed{\nu_l = \frac{1 - \kappa_l^2}{\ln s \kappa_l}}. \quad (\text{B.10})$$

Substituting the equation (B.10) and setting

$$\frac{(1-d_l)(1-q_l)}{1-d_l q_l} = \frac{1}{\phi_l} k = 0 \quad (\text{B.11})$$

$$\begin{aligned} 1-d_l q_l &= \phi_l(1-q_l-d_l-q_l d_l) \\ 1-e^{-\frac{1}{\nu_l} \frac{\kappa_l^2+1}{\kappa_l}} &= \phi_l \left( 1-e^{-\frac{1}{\kappa_l \nu_l}} - e^{-\frac{\kappa_l}{\nu_l}} + e^{-\frac{1}{\nu_l} \frac{\kappa_l^2+1}{\kappa_l}} \right) \\ 1-s^{-\frac{1+\kappa_l^2+1}{1-\kappa_l^2}} &= \phi_l \left( 1-s^{-\frac{1}{1-\kappa_l^2}} - s^{-\frac{\kappa_l^2}{1-\kappa_l^2}} + s^{-\frac{1+\kappa_l^2}{1-\kappa_l^2}} \right) \\ 1-\phi_l &= (1+\phi_l) s^{-1} s^{-2\frac{\kappa_l^2}{1-\kappa_l^2}} - \phi_l s^{-1} s^{-\frac{\kappa_l^2}{1-\kappa_l^2}} - \phi_l s^{-\frac{\kappa_l^2}{1-\kappa_l^2}} \end{aligned} \quad (\text{B.12})$$

To simplify, it is needed to use the variable  $x_l$  defined as  $x_l = s^{-\frac{\kappa_l^2}{1-\kappa_l^2}}$ , and rewriting the equation (B.11)

$$\begin{aligned} 1-\phi_l &= (1+\phi_l) \frac{1}{s} x_l^2 - \phi_l \frac{1}{d} x_l - \phi_l x_l \\ x_l^2 - \frac{\phi_l(1+s)}{1+\phi_l} x_l - \frac{1-\phi_l}{1+\phi_l} s &= 0 \end{aligned} \quad (\text{B.13})$$

So, the equation (B.13) is solved applying the standard solution of a quadratic function to find  $x_l$ :

$$\boxed{x_l = \frac{\phi_l(1+\mu_l)}{2(1+\phi_l)} \left( 1 - \sqrt{1 + \frac{4\mu_l(1-\phi_l^2)}{\phi_l^2(1+\mu_l)^2}} \right)}. \quad (\text{B.14})$$

If  $x_l = s^{-\frac{\kappa_l^2}{1-\kappa_l^2}}$  then it can be assumed that

$$\begin{aligned} \ln x_l &= \frac{\kappa_l^2}{1-\kappa_l^2} \ln s \\ \frac{\ln x_l}{\ln s} - \kappa_l^2 \frac{\ln x_l}{\ln s} + \kappa_l^2 &= 0 \end{aligned} \quad (\text{B.15})$$

and the variable  $\kappa_l$  is computed as

$$\boxed{\kappa_l = \sqrt{\frac{\ln x_l}{\ln \left( \frac{x_l}{s} \right)}}} \quad (\text{B.16})$$

Finally, the variable  $\nu_l$  can be obtained assuming that  $\frac{1-\kappa_l^2}{\kappa_l} = -\kappa_l \frac{\ln s}{\ln x_l}$

$$\boxed{\nu_l = -\frac{\kappa_l}{\ln x_l}} \quad (\text{B.17})$$

# Appendix C

## C.1 Computational Algorithm

The algorithm for computing the UB  $\mathcal{B}_n^U$  and LB  $\mathcal{B}_n^L$  bounds is developed in Table C.1 and Table C.2 [71]. Its inputs are the detected CNAs  $y_n$ , breakpoint locations  $\hat{\eta}_l$ , bound wideness  $\vartheta$ , number  $L$  of the breakpoints, and number of the probes  $M$ . At the output, it has two confidence masks  $\mathcal{B}_n^U$  and  $\mathcal{B}_n^L$ . The first algorithmic block (3–6) computes the segmental statistics  $\hat{a}_j$  and  $\sigma_j$  on intervals between neighboring breakpoints. The second block (7) is a function, described in table C.2, that employs equations (3.29, 3.30, 3.10, 3.37, 3.40) to compute the right jitter  $k_l^R$  and left jitter  $k_l^L$ . The third block (8–12) finds the jitter boundaries  $\mathcal{I}_l$  and  $\mathcal{E}_l$  for the UB and LB masks. The fourth block (13–16) make corrections to jitter boundaries in the cases when some boundaries merge or overlap. The fifth block (17–23) skips some points in the case when the UB mask or LB mask occurs to be uniform for several breakpoints. The masks  $\mathcal{B}_n^U$  and  $\mathcal{B}_n^L$  finally go to the output. Note that 3.39 approximates jitter in the breakpoints of CNVs in the lower bound sense. That means that wide jitter boundaries detected by the algorithm may be wider in practice.

Table C.1: Algorithm for computing the UB mask  $\mathcal{B}_n^U$  and LB mask  $\mathcal{B}_n^L$  via SNP array CNVs measurements  $y_n$  and the breakpoint locations estimates  $\hat{\eta}_l$ . Given: bound wideness ( $\vartheta$ -sigma).

---

**Input:**  $y_n, \hat{\eta}_l, \vartheta$

1:  $\xi = \text{erfc}(\frac{\vartheta}{\sqrt{2}})$ ,  $L = \text{length}(\hat{\eta}_l)$ ,  $M = \text{length}(y_n)$

---



---

```

2:   $N_{L+1} = M - \hat{n}_L, \quad \hat{n}_0 = 0$ 
3:  for  $j = 1 : L + 1$  do
4:     $N_j = \hat{n}_j - \hat{n}_{j-1}, \quad \hat{a}_j = \frac{1}{N_j} \sum_{v=\hat{n}_{j-1}}^{\hat{n}_j-1} y_v$ 
5:     $\sigma_j = \sqrt{\frac{1}{N_j} \sum_{v=\hat{n}_{j-1}}^{\hat{n}_j-1} (y_v - \hat{a}_j)^2}$ 
6:  end for
7:   $[k_l^R, k_l^L] = k_l^R k_l^L \text{-jitter}(\hat{a}_j, \hat{\sigma}_j, L) \quad \triangleright \text{right jitter}$ 
                                          $\triangleright \text{left jitter}$ 
8:   $\mathcal{I}_{L+1} = M - 1, \mathcal{E}_{L+1} = M - 1$ 
9:  for  $l = 1 : L$  do
10:    $\mathcal{I}_l = \begin{cases} \hat{n}_l - k_l^R & \text{if } \Delta_l > 0 \\ \hat{n}_l + k_l^L & \text{if } \Delta_l < 0 \end{cases}$ 
11:    $\mathcal{E}_l = \begin{cases} \hat{n}_l + k_l^L & \text{if } \Delta_l > 0 \\ \hat{n}_l - k_l^R & \text{if } \Delta_l < 0 \end{cases}$ 
12: end for
13: for  $l = 1 : L$  do
14:    $\mathcal{I}_l = \begin{cases} \mathcal{I}_l & \text{if } \text{Im } \mathcal{I}_l = 0 \\ \mathcal{I}_{l-1} & \text{if } \Delta_l \geq 0 \wedge \text{Im } \mathcal{I}_l \neq 0 \\ \mathcal{I}_{l+1} & \text{if } \Delta_l < 0 \wedge \text{Im } \mathcal{I}_l \neq 0 \\ & \text{elseif } (\text{Im } \mathcal{I}_{l+1}) \rightarrow \mathcal{I}_{l+2} \end{cases}$ 
15:    $\mathcal{E}_l = \begin{cases} \mathcal{E}_l & \text{if } \text{Im } \mathcal{E}_l = 0 \\ \mathcal{E}_{l+1} & \text{if } \Delta_l \geq 0 \wedge \text{Im } \mathcal{E}_l \neq 0 \\ & \text{elseif } (\text{Im } \mathcal{E}_l + 1) \rightarrow \mathcal{E}_{l+2} \\ \mathcal{E}_{l-1} & \text{if } \Delta_l < 0 \wedge \text{Im } \mathcal{E}_l \neq 0 \end{cases}$ 
16: end for
17:  $l = 1, k = 1$ 
18: for  $n = 0 : M - 1$  do
19:    $l = \begin{cases} l & \text{if } n < \mathcal{I}_l \\ l + 1 & \text{if } n \geq \mathcal{I}_l \wedge \mathcal{I}_{l+1} > C_l \\ l + 2 & \text{if } n \geq \mathcal{I}_l \wedge \mathcal{I}_{l+1} \leq C_l \end{cases}$ 

```

---

---

```

20:    $k = \begin{cases} k & \text{if } n < \mathcal{E}_l \\ k + 1 & \text{if } n \geq \mathcal{E}_l \wedge \mathcal{E}_{l+1} > C_l \\ k + 2 & \text{if } n \geq \mathcal{E}_l \wedge \mathcal{E}_{l+1} \leq C_l \end{cases}$ 
21:    $\mathcal{B}_n^U = \hat{a}_l + \vartheta \sqrt{\frac{\sigma_l^2}{N_l}} \quad \triangleright \text{UB mask}$ 
22:    $\mathcal{B}_n^L = \hat{a}_k - \vartheta \sqrt{\frac{\sigma_k^2}{N_k}} \quad \triangleright \text{LB mask}$ 
23: end for
Output:  $\mathcal{B}_n^U, \mathcal{B}_n^L$ 

```

---

Table C.2: Algorithm for computing the KR jitter  $k_l^R$  and KL jitter  $k_l^L$ . Given:  $\hat{a}_j, \hat{\sigma}_j$  and number  $L$  of breakpoints.

---

```

Function  $k_l^R, k_l^L$  jitter, Input:  $\hat{a}_j, \hat{\sigma}_j, L$ 
1: for  $l = 1 : L$  do
2:    $\Delta_l = \hat{a}_{l+1} - \hat{a}_l, \quad \gamma_l^- = \frac{\Delta_l^2}{\sigma_l^2}, \quad \gamma_l^+ = \frac{\Delta_l^2}{\sigma_{l+1}^2}$ 
3:    $\alpha_l$  by (3.10) with “-” and  $a_l = \hat{a}_l$ 
4:    $\beta_l$  by (3.10) with “+” and  $a_l = \hat{a}_l$ 
5:    $g_l^\beta = \frac{\beta_l - \Delta_l}{|\Delta_l|} \sqrt{\frac{\gamma_l^-}{2}}, \quad g_l^\alpha = \frac{\alpha_l - \Delta_l}{|\Delta_l|} \sqrt{\frac{\gamma_l^-}{2}}$ 
6:    $h_l^\beta = \frac{\beta_l}{|\Delta_l|} \sqrt{\frac{\gamma_l^+}{2}}, \quad h_l^\alpha = \frac{\alpha_l}{|\Delta_l|} \sqrt{\frac{\gamma_l^+}{2}}$ 
7:    $P_l^A$  by (3.29),  $P_l^B$  by (3.30),  $\phi_l$  by (3.37)
8:    $\mu_l = \frac{P_l^A(1-P_l^B)}{P_l^B(1-P_l^A)}, \quad x_l$  by (3.40),  $\kappa_l = \sqrt{\frac{\ln x_l}{\ln(x_l/\mu_l)}}$ 
9:    $\nu_l = -\frac{\kappa_l}{\ln(x_l)}, \quad d_l = e^{-\frac{\kappa_l}{\nu_l}}, \quad q_l = e^{-\frac{1}{\kappa_l \nu_l}}$ 
10:   $k_l^R = \left\lfloor \frac{\nu_l \ln \frac{(1-d_l)(1-q_l)}{\xi(1-d_l q_l)}}{\kappa_l} \right\rfloor \quad \triangleright \text{right jitter}$ 
11:   $k_l^L = \left\lfloor \nu_l \kappa_l \ln \frac{(1-d_l)(1-q_l)}{\xi(1-d_l q_l)} \right\rfloor \quad \triangleright \text{left jitter}$ 
12: end for
Output:  $k_l^R, k_l^L$ 

```

---

# Appendix D

## D.1 Comparison of Approximations

The (*mse*) computed to each proposed approximation is summarized in tables D.1 and D.2. Following a particular methodology -Heuristic, Parametrization of Laplace or Asymmetric Exponential Power distribution- the functions obtained are compared with the measurements generated at three levels of simulation: *slow*, *fast* and *detailed*. So, in tables D.1 and D.2 are exposed the best performance of each function proposed.

Table D.1: Typical MSEs produced by all the approximations proposed for different values of Signal to Noise Ratio  $\gamma = \gamma_i^- = \gamma_i^+$

Section I						
$\gamma$	Slow Algorithm		Fast Algorithm			
	pdf (3.39)	MBA (4.3)	pdf (3.39)	(3.39) with (4.9)	(3.39) with (4.10)	(3.39) with (4.11)
0.1	$7.6 \times 10^{-5}$	$4.2 \times 10^{-6}$	$8.6 \times 10^{-5}$	$1.4 \times 10^{-6}$	$1.5 \times 10^{-6}$	$1.8 \times 10^{-7}$
0.2	$7.7 \times 10^{-5}$	$1.8 \times 10^{-6}$	$7.8 \times 10^{-5}$	$4.1 \times 10^{-6}$	$3.8 \times 10^{-6}$	$1.1 \times 10^{-7}$
0.3	$7.5 \times 10^{-5}$	$1.1 \times 10^{-6}$	$7.4 \times 10^{-5}$	$5.6 \times 10^{-6}$	$5.3 \times 10^{-6}$	$7.9 \times 10^{-8}$
0.4	$7.3 \times 10^{-5}$	$8.0 \times 10^{-7}$	$7.0 \times 10^{-5}$	$6.9 \times 10^{-6}$	$6.3 \times 10^{-6}$	$8.5 \times 10^{-8}$
0.5	$6.6 \times 10^{-5}$	$6.1 \times 10^{-7}$	$6.6 \times 10^{-5}$	$8.1 \times 10^{-6}$	$7.3 \times 10^{-6}$	$1.1 \times 10^{-7}$
0.6	$6.3 \times 10^{-5}$	$4.9 \times 10^{-7}$	$5.9 \times 10^{-5}$	$1.0 \times 10^{-5}$	$9.0 \times 10^{-6}$	$1.2 \times 10^{-7}$

Section I						
$\gamma$	Slow Algorithm		Fast Algorithm			
	pdf	MBA	pdf	(3.39)	(3.39)	(3.39)
	(3.39)	(4.3)	(3.39)	with (4.9)	with (4.10)	with (4.11)
0.7	$5.9 \times 10^{-5}$	$4.1 \times 10^{-7}$	$5.9 \times 10^{-5}$	$1.0 \times 10^{-5}$	$9.0 \times 10^{-6}$	$1.2 \times 10^{-7}$
0.8	$5.5 \times 10^{-5}$	$3.5 \times 10^{-7}$	$5.3 \times 10^{-5}$	$1.2 \times 10^{-5}$	$1.0 \times 10^{-5}$	$9.2 \times 10^{-8}$
0.9	$5.3 \times 10^{-5}$	$3.0 \times 10^{-7}$	$5.3 \times 10^{-5}$	$1.2 \times 10^{-5}$	$1.0 \times 10^{-5}$	$9.2 \times 10^{-8}$
1.0	$5.1 \times 10^{-5}$	$2.7 \times 10^{-7}$	$5.0 \times 10^{-5}$	$2.2 \times 10^{-5}$	$2.7 \times 10^{-5}$	$1.7 \times 10^{-7}$
1.1	$4.9 \times 10^{-5}$	$2.4 \times 10^{-5}$	$2.7 \times 10^{-5}$	$2.5 \times 10^{-5}$	$3.2 \times 10^{-5}$	$1.7 \times 10^{-7}$
1.37	$4.1 \times 10^{-5}$	$1.8 \times 10^{-5}$	$4.0 \times 10^{-5}$	$2.7 \times 10^{-5}$	$3.4 \times 10^{-5}$	$2.6 \times 10^{-7}$

Table D.2: Typical MSEs produced by all the approximations proposed for different values of Signal to Noise Ratio  $\gamma = \gamma_l^- = \gamma_l^+$

Section II						
$\gamma$	Detailed Algorithm					
	pdf	MBA	(3.39)	(3.39)	(3.39)	AEP 4.11
	(3.39)	(4.3)	with (4.9)	with (4.10)	with (4.11)	
0.1	$7.3 \times 10^{-5}$	$2.18 \times 10^{-5}$	$1.3 \times 10^{-5}$	$1.2 \times 10^{-5}$	$9.5 \times 10^{-6}$	$1.11 \times 10^{-7}$
0.2	$7.6 \times 10^{-5}$	$2.8 \times 10^{-5}$	$9.9 \times 10^{-6}$	$8.5 \times 10^{-6}$	$1.3 \times 10^{-5}$	$7.09 \times 10^{-8}$
0.3	$7.5 \times 10^{-5}$	$3.3 \times 10^{-5}$	$9.5 \times 10^{-6}$	$7.8 \times 10^{-6}$	$1.7 \times 10^{-5}$	$1.16 \times 10^{-7}$
0.4	$7.2 \times 10^{-5}$	$3.5 \times 10^{-5}$	$8.17 \times 10^{-6}$	$6.2 \times 10^{-6}$	$1.8 \times 10^{-5}$	$1.03 \times 10^{-7}$
0.5	$6.8 \times 10^{-5}$	$3.5 \times 10^{-5}$	$7.12 \times 10^{-6}$	$5.0 \times 10^{-6}$	$1.9 \times 10^{-5}$	$1.24 \times 10^{-7}$
0.6	$6.5 \times 10^{-5}$	$3.4 \times 10^{-5}$	$5.9 \times 10^{-6}$	$5.1 \times 10^{-6}$	$2.0 \times 10^{-5}$	$1.04 \times 10^{-7}$
0.7	$5.9 \times 10^{-5}$	$3.3 \times 10^{-5}$	$4.7 \times 10^{-6}$	$8.2 \times 10^{-6}$	$2.0 \times 10^{-5}$	$8.33 \times 10^{-8}$
0.8	$5.7 \times 10^{-5}$	$3.2 \times 10^{-5}$	$3.6 \times 10^{-6}$	$1.5 \times 10^{-6}$	$1.8 \times 10^{-5}$	$9.99 \times 10^{-8}$
0.9	$5.2 \times 10^{-5}$	$3.0 \times 10^{-5}$	$2.4 \times 10^{-6}$	$6.9 \times 10^{-7}$	$1.6 \times 10^{-5}$	$1.13 \times 10^{-7}$
1.0	$5.0 \times 10^{-5}$	$2.9 \times 10^{-5}$	$2.0 \times 10^{-6}$	$4.4 \times 10^{-7}$	$1.6 \times 10^{-5}$	$1.21 \times 10^{-7}$
1.1	$4.7 \times 10^{-5}$	$2.8 \times 10^{-5}$	$1.4 \times 10^{-6}$	$2.2 \times 10^{-7}$	$1.4 \times 10^{-5}$	$1.13 \times 10^{-7}$
1.37	$3.9 \times 10^{-5}$	$2.37 \times 10^{-5}$	$7.2 \times 10^{-7}$	$7.9 \times 10^{-7}$	$1.0 \times 10^{-5}$	$6.80 \times 10^{-8}$

# Appendix E

Table E.1: Part I. Left jitter  $k_l^-$  and right jitter  $k_l^+$  detected by different masks in the CNA breakpoints of the 13th chromosomal sample “BLC\_B1\_T37.txt” in the  $3\sigma$  sense with the confidence probability of  $P = 99.73\%$ . The chromosome is associated with breast cancer and all values are given for Log<sub>2</sub> Ratio. Here symbol “-” means that the jitter cannot be calculated by the masks.

$l$	SNR		MBA [71]		Laplace (3.39)		(3.39) with (4.9)		(3.39) with (4.10)		(3.39) with (4.11)		Hybrid		
	$\gamma_l^-$	$\gamma_l^+$	$ \gamma_l^- - \gamma_l^+ $	$k_l^-$	$k_l^+$	$k_l^-$	$k_l^+$	$k_l^-$	$k_l^+$	$k_l^-$	$k_l^+$	$k_l^-$	$k_l^+$	$k_l^-$	$k_l^+$
1	0.2101	0.3052	0.0950	10	12	-	-	-	-	-	-	-	-	10	12
2	0.4722	0.5412	0.0690	7	8	10	6	-	-	-	-	-	-	7	8
3	3.7920	3.9122	0.1201	2	2	3	3	3	3	3	3	2	2	3	3
4	0.0574	0.0536	0.0037	-	-	-	-	-	-	-	-	-	-	-	-
5	0.0221	0.0232	0.0011	-	-	-	-	-	-	-	-	-	-	-	-
6	13.6833	13.4657	0.2175	-	-	1	1	1	1	1	1	1	1	1	1
7	0.8156	0.8993	0.0837	6	6	7	5	10	8	10	8	6	5	10	8
8	8.7577	8.9276	0.1698	-	-	2	2	2	2	2	2	1	1	2	2
9	7.1710	5.0402	2.1308	-	-	2	2	2	3	2	3	1	1	2	2
10	1.1958	2.3267	1.1309	4	4	7	3	-	-	-	-	-	-	7	3
11	1.6459	1.1516	0.4942	4	4	4	7	-	-	-	-	-	-	4	7
12	0.3676	0.4120	0.0444	9	9	12	6	-	-	-	-	-	-	9	9
13	5.9443	4.8291	1.1151	2	2	2	3	3	3	3	3	2	2	3	3
14	5.6081	6.5830	0.9749	-	-	2	2	2	2	2	2	1	1	2	2
15	5.8848	3.9626	1.9222	2	2	2	3	3	3	3	3	2	2	2	3
16	3.5660	4.9407	1.3746	2	2	3	2	3	3	3	3	2	2	3	2
17	4.5575	4.2765	0.2810	2	2	3	3	3	3	3	3	2	2	3	3
18	7.5862	29.3433	21.7570	-	-	1	1	1	1	1	1	1	1	1	1
19	32.2283	8.9617	23.2666	-	-	1	1	1	1	1	1	0	1	1	1
20	5.4114	5.0298	0.3816	-	-	2	2	3	3	3	3	2	2	2	2

Table E.2: Part II. Left jitter  $k_l^-$  and right jitter  $k_l^+$  detected by different masks in the CNA breakpoints of the 13th chromosomal sample “BLC\_B1\_T37.txt” in the  $3\sigma$  sense with the confidence probability of  $P = 99.73\%$ . The chromosome is associated with breast cancer and all values are given for Log<sub>2</sub> Ratio. Here symbol “-” means that the jitter cannot be calculated by the masks.

$l$	SNR		MBA [71]		Laplace (3.39)		(3.39) with (4.9)		(3.39) with (4.10)		(3.39) with (4.11)		Hybrid	
	$\gamma_l^-$	$\gamma_l^+$	$k_l^-$	$k_l^+$	$k_l^-$	$k_l^+$	$k_l^-$	$k_l^+$	$k_l^-$	$k_l^+$	$k_l^-$	$k_l^+$	$k_l^-$	$k_l^+$
21	1.5649	1.4423	4	4	4	5	6	6	7	4	4	4	4	5
22	1.5162	1.6035	4	4	5	4	6	6	6	4	4	4	4	4
23	1.3162	1.3841	4	4	5	5	7	7	6	4	4	4	4	7
24	1.7423	1.9598	4	4	5	4	6	5	5	4	3	3	5	4
25	1.5106	1.9640	4	4	5	4	6	5	5	4	3	3	5	4
26	7.1659	5.6248	-	-	2	2	2	2	2	1	1	1	2	2
27	0.5958	0.6090	7	7	7	7	11	11	11	7	7	7	11	11
28	1.2939	1.2883	4	4	5	5	7	7	7	4	4	4	7	7
29	2.7600	2.0822	3	3	3	4	4	5	5	3	3	3	3	4
30	2.9136	3.2142	3	3	3	3	4	4	4	2	2	2	3	3
31	0.8566	0.7435	6	6	5	8	8	11	11	6	7	7	8	11
32	0.7246	1.1731	5	5	11	3	-	-	-	-	-	-	5	5
33	5.6282	3.5115	2	2	2	3	3	3	3	2	2	2	2	3
34	3.0722	4.1954	2	2	3	3	4	3	3	2	2	2	3	3
35	0.4690	0.3940	9	8	5	13	-	-	-	-	-	-	9	8
36	0.5150	0.5073	8	8	7	8	12	12	12	8	8	8	12	12
37	2.8922	2.5560	3	3	3	4	4	4	4	3	3	3	3	4
38	0.1076	0.1119	17	18	21	10	-	-	-	-	-	-	17	18
39	0.9593	0.9784	5	5	6	6	8	8	8	5	5	5	8	8
40	0.1841	0.1513	15	13	-	-	-	-	-	-	-	-	15	13

Table E.3: Part III. Left jitter  $k_l^-$  and right jitter  $k_l^+$  detected by different masks in the CNA breakpoints of the 13th chromosomal sample “BLC\_B1\_T37.txt” in the  $3\sigma$  sense with the confidence probability of  $P = 99.73\%$ . The chromosome is associated with breast cancer and all values are given for Log<sub>2</sub> Ratio. Here symbol “—” means that the jitter cannot be calculated by the masks.

$l$	SNR		$ \gamma_l^- - \gamma_l^+ $		MBA [71]		Laplace (3.39)		(3.39) with (4.9)		(3.39) with (4.10)		(3.39) with (4.11)		Hybrid	
	$\gamma_l^-$	$\gamma_l^+$	$k_l^-$	$k_l^+$	$k_l^-$	$k_l^+$	$k_l^-$	$k_l^+$	$k_l^-$	$k_l^+$	$k_l^-$	$k_l^+$	$k_l^-$	$k_l^+$	$k_l^-$	$k_l^+$
41	0.0828	0.0777	—	—	—	—	—	—	—	—	—	—	—	—	—	—
42	0.8173	0.6956	6	6	5	8	—	—	—	—	—	—	—	—	6	6
43	0.0286	0.0425	24	39	—	—	—	—	—	—	—	—	—	—	24	39
44	0.0597	0.0697	21	26	—	—	—	—	—	—	—	—	—	—	21	26
45	1.6864	0.9808	4	4	3	8	—	—	—	—	—	—	—	—	3	8
46	0.9639	1.7280	4	4	8	3	—	—	—	—	—	—	—	—	8	3
47	0.0837	0.0797	—	—	9	35	—	—	—	—	—	—	—	—	9	35
48	8.3634	9.7859	—	—	2	2	2	2	2	2	2	1	1	1	2	2
49	9.9867	9.0587	—	—	2	2	2	2	2	2	2	1	1	1	2	2
50	10.3886	16.6565	—	—	1	1	1	1	1	1	1	1	1	1	1	1
51	17.3197	10.4969	—	—	1	1	1	1	1	1	1	1	1	1	1	1
52	0.6826	0.6563	7	6	6	7	10	11	10	11	11	7	7	10	11	11
53	0.2242	0.2322	12	12	13	9	22	18	22	18	18	14	12	22	18	18
54	0.2228	0.2449	11	12	17	7	—	—	—	—	—	—	—	—	11	12
55	0.9405	0.5706	6	6	3	14	—	—	—	—	—	—	—	—	6	6
56	0.3219	0.3171	10	10	9	10	15	16	15	16	16	10	11	15	16	16
57	12.7085	16.4145	—	—	1	1	1	1	1	1	1	1	1	1	1	1
58	14.1556	10.3335	—	—	1	1	1	1	1	1	1	1	1	1	1	1