



HAL
open science

Statistical Approaches for Segmentation : Application to Genome Annotation

Alice Cleynen

► **To cite this version:**

Alice Cleynen. Statistical Approaches for Segmentation : Application to Genome Annotation. General Mathematics [math.GM]. Université Paris Sud - Paris XI, 2013. English. NNT : 2013PA112258 . tel-00913851

HAL Id: tel-00913851

<https://theses.hal.science/tel-00913851v1>

Submitted on 4 Dec 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ PARIS-SUD

ECOLE DOCTORALE MATHÉMATIQUES
DE LA RÉGION PARIS-SUD

DISCIPLINE : MATHÉMATIQUES (STATISTIQUES)

THÈSE DE DOCTORAT

Soutenue le 15 novembre 2013 par

Alice Cleynen

**Approches statistiques en segmentation :
application à la ré-annotation de génome**

~

**Statistical Approaches for Segmentation :
Application to Genome Annotation**

Directeur de thèse : M. Stéphane Robin
Co-directrice : Mme. Sandrine Dudoit

Directeur de recherche, INRA
Professor, UC Berkeley

Composition du jury :

Président du jury :	Mr Jean-Philippe Vert	Directeur de recherche, MinesParisTech
Rapporteurs :	Mr Olivier Cappé	Directeur de recherche, CNRS
	Mr Simon Tavaré	Professor, Cambridge
Examineurs :	Mme. Elisabeth Gassiat	Professeur, UPSud
	Mme. Patricia Reynaud-Bouret	Chargée de recherche, CNRS
	Mr. Mark Van de Wiel	Professor, VU Amsterdam



Thèse préparée à
AgroParisTech / INRA et UC Berkeley
Laboratoire MIA Stat & Génome (UMR 518)
16 rue Claude Bernard
75 005 Paris



Abstract

We propose to model the output of transcriptome sequencing technologies (RNA-Seq) using the negative binomial distribution, as well as build segmentation models suited to their study at different biological scales, in the context of these technologies becoming a valuable tool for genome annotation, gene expression analysis, and new-transcript discovery. We develop a fast segmentation algorithm to analyze whole chromosomes series, and we propose two methods for estimating the number of segments, a key feature related to the number of genes expressed in the cell, should they be identified from previous experiments or discovered at this occasion.

Research on precise gene annotation, and in particular comparison of transcription boundaries for individuals, naturally leads us to the statistical comparison of change-points in independent series. To address our questions, we build tools, in a Bayesian segmentation framework, for which we are able to provide uncertainty measures. We illustrate our models, all implemented in R packages, on an RNA-Seq dataset from a study on yeast, and show for instance that the intron boundaries are conserved across conditions while the beginning and end of transcripts are subject to differential splicing.

Key words: Segmentation, negative binomial, algorithm, credibility intervals, model selection, RNA-Seq

Résumé

Nous proposons de modéliser les données issues des technologies de séquençage du transcriptome (RNA-Seq) à l'aide de la loi binomiale négative, et nous construisons des modèles de segmentation adaptés à leur étude à différentes échelles biologiques, dans le contexte où ces technologies sont devenues un outil précieux pour l'annotation de génome, l'analyse de l'expression des gènes, et la détection de nouveaux transcrits. Nous développons un algorithme de segmentation rapide pour analyser des séries à l'échelle du chromosome, et nous proposons deux méthodes pour l'estimation du nombre de segments, directement lié au nombre de gènes exprimés dans la cellule, qu'ils soient précédemment annotés ou détectés à cette même occasion.

L'objectif d'annotation précise des gènes, et plus particulièrement de comparaison des sites de début et fin de transcription entre individus, nous amène naturellement à nous intéresser à la comparaison des localisations de ruptures dans des séries indépendantes. Nous construisons ainsi dans un cadre de segmentation bayésienne des outils de réponse à nos questions pour lesquels nous sommes capable de fournir des mesures d'incertitude. Nous illustrons nos modèles, tous implémentés dans des packages R, sur des données RNA-Seq provenant d'expériences sur la levure, et montrons par exemple que les frontières des introns sont conservées entre conditions tandis que les débuts et fin de transcriptions sont soumis à l'épissage différentiel.

Mots Clés : Segmentation, binomiale négative, Sélection de modèle, algorithmes, intervalles de crédibilité, RNA-Seq

Remerciements

Stéphane, Sandrine, c'est un privilège d'ouvrir ces remerciements en m'adressant à vous. C'est aussi un exercice de style, tant il y aurait à dire et si difficile est-il d'écrire tout ce que vous m'aurez inspiré. J'ai commencé cette thèse un peu par hasard, et je te dois, Stéphane, tout le goût que j'y ai pris à mesure des années. Je ne compte plus les discussions dans ton bureau, et pourtant ton optimisme et ta confiance me surprennent encore. Je ne connais personne de plus généreux que toi, sur ton temps et ton savoir, tes conseils et tes encouragements, et je ne peux rêver meilleur directeur que toi (même s'il a été parfois frustrant que tu répondes à mes questions avant même que je n'arrive à les formuler). Je te suis extrêmement reconnaissante de tout ce que tu m'auras apporté, et je ne peux qu'espérer être en mesure, un jour, de transmettre à mon tour un peu de tout cela.

Sandrine, ces deux visites à Berkeley m'auront donné la chance de te connaître, d'apprécier le temps passé avec toi, tes qualités humaines exceptionnelles, et ton sourire ne te quittant jamais. Au-delà de nos intérêts scientifiques, ton soutien et ta compréhension dans les moments difficiles m'auront été d'une grande aide, et il est extraordinaire d'avoir tant pu partager avec toi malgré les milliers de kilomètres nous séparant. J'espère avoir encore de nombreuses occasions de profiter de tes conseils, de partager nos idées, et de visiter encore San Francisco !

C'est ensuite à Olivier Cappé et Simon Tavaré que je voudrais adresser mes remerciements, pour s'être intéressés à mon travail en acceptant de rapporter ma thèse. Le temps

que vous avez consacré à ce manuscrit et les nombreux commentaires dont vous m'avez fait part sont un grand honneur.

Merci aussi aux nombreux membres de mon jury, Elisabeth Gassiat, Patricia Reynaud-Bouret, Mark Van de Wiel, et Jean-Philippe Vert d'avoir accepté d'y participer, et pour le soutien que vous m'avez témoigné pendant ma thèse.

Parmi les nombreux membres de mon labo qui ont joué un rôle important dans la personne que je suis devenue, Emilie tient une place toute particulière. Je suis désolée de t'en avoir fait voir de toutes les couleurs avec ma binomiale négative, et je te remercie pour les heures consacrées à nos majorations, les week-ends et les voyages en train entachés de contrôles du χ^2 (comment ça, il fallait majorer et pas minorer?), et les interminables relectures de tentatives de rédaction. Il faut absolument que tu t'attèles à cette HDR, car je suis certaine que tu seras une directrice de thèse formidable. Merci Emilie, pour ta bonne humeur et ta générosité...

Merci ensuite à mes co-bureau, Antoine, pour avoir partagé Stéphane avec moi, Aurore, pour avoir dompté Jean-Benoist, et Jean-Benoist, pour ton infinie patience toutes les fois où j'ai planté les serveurs. Au fond j'ai bien souvent pesté contre votre insupportable bordel, vos mauvaises blagues, les lancers d'ustensiles à travers le bureau, l'opéra à toutes pompes dans les écouteurs et j'en passe, mais vous avez indéniablement contribué à la joyeuse troupe de notre équipe et je garderai d'excellents souvenirs de nos moments partagés.

Merci aussi au BDDDB pour leur imagination incroyable, à Loïc et Souhil qu'on ne voyait pas beaucoup (plus de Stéphane pour nous!), Eleana, ou encore Marie, Sarah et Anna que je n'aurai pas eu la chance de connaître bien longtemps... Un merci particulier au bureau d'à côté, toujours prêt à aider... Tristan (tu vas pouvoir récupérer le chauffage quand je serai partie), Marie (le labo ne serait pas le même sans toi, ton sacré caractère et ta bonne humeur sans faille) et Pierre (et sa galanterie légendaire), merci pour votre bonne humeur, et tous vos conseils et encouragements.

Il serait trop long d'écrire ici à chacun d'entre vous l'intégralité des sentiments que vous

m'inspirez. Mais je ne peux que rendre hommage à tous les membres de l'équipe, Julien, Maud, Liliane, Eric, Gabriel, Julie, Marie-Laure, Xiao, Céline, Michel, Jean-Jacques, et ceux qui en sont partis, Caroline, Stevonn, Nathalie, Ophélie, Florence... Vous formez le labo le plus généreux que je connaisse. Enfin, et c'est aussi grâce à vous que ce labo fonctionne si bien, un grand merci à Odile, Sophie, Francine et Valérie pour votre disponibilité et votre aide dans toutes les démarches.

Merci aussi à mes collaborateurs extérieurs, et tout particulièrement à Guillem sans qui je ne serais pas arrivée là aujourd'hui. J'ai toujours pu compter sur toi, et tu n'as jamais renoncé à m'expliquer encore et encore les mêmes choses ; merci pour ton temps. C'est sans oublier Gregory et The Minh, ou encore toutes les personnes rencontrées à Berkeley, Davide, Abha, Anne-Claire, Laurent et les autres. Enfin à Mahendra, merci de t'être intéressé à moi depuis notre rencontre et tout au long de ma thèse, merci aussi pour tes idées et tes conseils (tu remarqueras comme je les ai souvent suivis).

Un merci encore à mes amis, les plus vieux pour me supporter encore, les plus récents pour ne pas réaliser qu'il est un jour où ils ne me supporteront plus ! Jess (pour ta recette du gâteau au chocolat, ta gentillesse et ton caractère), Cyril (pour avoir été le premier à me faire comprendre que les choses n'ont pas besoin d'être compliquées pour être belles), Nicole et Christophe (et leur bonne grâce à perdre à tous les jeux que nous faisons !), merci pour votre amitié au delà de Toulouse.

À la joyeuse bande cachanaise, Sarah, Christophe, Jules, Florent, Xan, Pierre, Jean et bien sûr Sandrine et Nicolas, merci pour tous les supers moments passés, et pour ceux qui ne manqueront pas de suivre.

A ma famille enfin, si présente et encourageante. Merci à mon père, pour être plus fier encore de mes rêves que de mes accomplissements, et à mon grand frère, pour tout ce qu'il implique d'être mon grand frère. A mon incroyable grand-mère, à ma super cousine, et à ma tante, pour son inconditionnel soutien et sa présence, encore plus cette dernière année ;

merci. A ma mère surtout, pour avoir été tout ce dont j'ai pu rêver, et pour m'avoir construit un monde si fabuleux qu'il m'est possible de continuer à avancer aujourd'hui.

A Hoel enfin, par ce que tu t'appelles Hoel et qu'il n'y en a qu'un comme toi! Pour avoir supporté mes histoires de fonctions convexes et puis concaves, mes développements de Taylor en dimension d , je te dois bien un morceau de cette thèse. Mais surtout pour ton infinie patience, ton amour inconditionnel ; merci Hoel, d'être mon pilier sur cette terre, de me faire rire toujours, et pour tout ce qu'il nous reste à partager.

Contents

Abstract	3
1 Introduction	17
1.1 Biological framework	18
1.2 Negative binomial distribution and change-point analysis	36
1.3 Contribution	58
2 Segmentation methods for whole genome analysis using RNA-Seq data	69
2.1 An efficient algorithm for the segmentation of RNA-Seq data	71
2.2 Model selection	90
2.3 Constrained HMM approach	127
2.4 Results on the yeast data-set	153
3 Segmentation methods for gene annotation using RNA-Seq data	161
3.1 Method comparison	163
3.2 Profile comparison	206
3.3 EBS: R package for Exact Bayesian Segmentation	230
3.4 Results on the yeast dataset	237

List of Figures

1.1	Example of a sequence of nucleotides forming a DNA fragment.	19
1.2	The Central Dogma of molecular biology.	20
1.3	The Central Dogma detailed in eukaryotes.	21
1.4	Map of the yeast genome.	24
1.5	Evolution of the sequencing cost in the last decade.	26
1.6	Technical process for the determination of a sequence of nucleotides.	27
1.7	Steps of a sequencing technology.	29
1.8	Overview of the benchmark dataset.	33
1.9	Data overview.	34
1.10	Example of change-point analysis of a signal.	40
1.11	Classification of change-point approaches.	41
1.12	Graphical model of the exact Bayesian Segmentation approach.	56
1.13	Original and modified graphical models for the comparison of change-point location.	66
2.1	Data overview.	81
2.2	Run-time analysis for segmentation with negative binomial distribution.	84
2.3	Rand Index for the quality of the segmentation.	85
2.4	Segmentation of the yeast chromosome 1 using the negative binomial loss.	86
2.5	Estimation of K on resampled datasets.	106
2.6	Segmentation of the yeast chromosome 1 using Poisson loss.	108
2.7	Segmentation of the yeast chromosome 1 using the negative binomial loss.	109
2.8	Rand Index for the comparison of initialization methods.	138
2.9	Performance of our method on small datasets.	140
2.10	Performance of our method on large datasets.	141
2.11	Algorithm comparison on short and regular dataset.	144
2.12	Algorithm comparison on short but irregular dataset.	145
2.13	Algorithm comparison on medium length and irregular dataset.	146
2.14	Algorithm comparison on long and irregular dataset.	147

2.15	Segmentation of yeast dataset.	149
2.16	Subset of the segmentation of chromosome 3 (negative strand).	155
2.17	Distribution of UTR lengths. From NAGALAKSHMI <i>et al.</i> (2008). Reprinted with permission from AAAS.	156
2.18	Distribution of UTR lengths using the PDPA algorithm.	156
2.19	Run time of PDPA on bench-mark dataset.	157
2.20	Power curve of the Wald test.	159
2.21	pruned DP algorithm and coverage.	160
3.1	Segmentation of gene YAL030W.	178
3.2	Global fit.	179
3.3	Local fit.	180
3.4	ROC curves.	181
3.5	Global fit and estimation of K	184
3.6	ROC curves and estimation of K	185
3.7	EBS segmentation of five yeast genes.	188
3.8	Yeast dataset.	190
3.9	Global fit for NB and MU simulations.	191
3.10	Posterior probabilities of change-point location.	192
3.11	ROC curves for RS simulation.	193
3.12	Estimation of K for NB simulation.	194
3.13	Estimation of K for MU simulation.	195
3.14	Estimation of K for RS simulation.	196
3.15	Estimation of K for RS simulation, transcript YAL030W.	197
3.16	BIC and ICL criterion for postCP and EBS.	198
3.17	<i>D. melanogaster</i> 's Inr-a gene.	199
3.18	Estimation of K and ROC curves for Drosophila-like simulations.	200
3.19	ROC curves for EBS with Negative Binomial and Gaussian distributions.	201
3.20	Impact of the priors on the ICL and BIC values.	205
3.21	RNA-Seq data for a two-exon gene in three growth conditions.	207
3.22	Graphical model.	216

3.23	Simulation design.	218
3.24	Impact of estimating the dispersion parameter.	220
3.25	Posterior distribution of change-point location.	222
3.26	Distribution of change-point location and 95% credibility intervals.	224
3.27	Boxplot of posterior probabilities of E_0 for Poisson.	226
3.28	Boxplot of posterior probabilities of E_0 for negative Binomial, with $p_0 = 0.8$	227
3.29	Boxplot of posterior probabilities of E_0 for negative Binomial, with $p_0 = 0.5$	228
3.30	Posterior distribution of change-points location.	233
3.31	Posterior distribution of change-points location of three profiles.	235
3.32	Credibility interval of difference in change-point location.	236
3.33	Posterior probability of E_0	237
3.34	Example of alternative isoform.	238
3.35	Posterior probability of E_0 with informative priors.	239

List of tables

1.1	Element of matrix A.	57
1.2	Overview of the segmentation issues addressed in this Thesis.	61
1.3	Overview of Thesis contribution.	68
2.1	Properties of segmentation algorithms.	105
2.2	Distributions from the exponential family and characteristics.	126
2.3	Output of the pruned dynamic programming algorithm.	154
3.1	Properties of segmentation algorithms.	173
3.2	Estimates of model parameters for each of the five yeast genes.	174
3.3	Credibility intervals.	182
3.4	Values of parameters used in the simulation study.	219
3.5	Posterior probability of a common change point across conditions for gene YAL013W.	225

Introduction

1.1	Biological framework	18
1.1.1	The central dogma of molecular biology	18
1.1.2	Next-Generation Sequencing technology	24
1.1.3	Benchmark dataset	31
1.2	Negative binomial distribution and change-point analysis	36
1.2.1	Negative binomial distribution	36
1.2.2	Change-point analysis	39
1.2.3	Segmentation issues in this framework	44
1.3	Contribution	58
1.3.1	Introduction	58
1.3.2	Whole genome analysis	60
1.3.3	Gene annotation	64

1.1 Biological framework

1.1.1 The central dogma of molecular biology

Discovering the role of DNA: When one thinks about the laws of genetic information transmission, the first thing that comes to mind is experiments with peas: characteristics of a strain (color of the flowers, shape of the peas, etc.) are observed in the offspring with proportions resulting from the transmission of half of each parent's information. Yet these experiments led by Johann Gregor Mendel throughout his years in his monastery (?) did not encounter at the time the enthusiasm one could expect. Even though he is now recognized as the pioneer of molecular biology, and the father of hereditary genetic information, it is not until the early 20th century that credit was given to his work. Even then, when in 1944 the biological specificity of the Deoxyribonucleic Acid (DNA) was discovered (?), the work was not well accepted and diffused in the scientific community. It is only when its structure was discovered by ? that DNA turned fundamental in the comprehension of living organisms.

DNA and genetic information: DNA is a molecule made of two long sequences of nucleotides called strands, and is present under the form of chromosomes in each cell of an organism: in the cytoplasm in prokaryotes, and in the nucleus in eukaryotes. Nucleotides are made of a five-carbon sugar, one or more phosphate groups and one of 4 possible nucleobases: Adenine (*A*), Guanine (*G*), Thymine (*T*) and Cytosine (*C*). The structure of nucleotides (most commonly referred to as bases) gives an orientation, said 5' to 3', to the strand of DNA: the numbers refer to the direction of the 3rd and 5th carbon atoms of the sugar molecule. The couples of nucleotides *A-T* and *C-G* are complementary, meaning that they can hybridize, *i.e.* bound together, through hydrogen bounds in such a way that their orientation is opposite. DNA strands, the Watson (or positive) strand, and the Crick (or negative) strand, are themselves said complementary as the sequence of nucleotides they are made of are hybridized. Figure 1.1 illustrates a possible fragment of DNA.

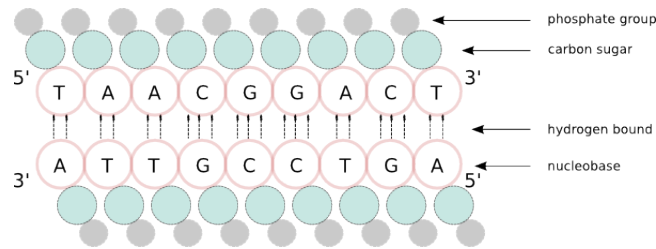


Figure 1.1: **Example of a sequence of nucleotides forming a DNA fragment.**

DNA is made of two oriented strands: sequences of nucleotides that are hybridized through hydrogen bounds.

This hybridization property is the base of natural perpetuation and propagation of the genetic information. It is also the foundation of all sequencing technologies (see Section 1.1.2). When two complementary sequences of nucleotides are present in a medium, they will naturally tend to hybridize to form a double-stranded molecule. Moreover, some specific enzymes, the DNA polymerases, are responsible for DNA replication: reading a strand of DNA from the 5' to the 3' end, they create the complementary strand by associating a *T* to an *A*, an *A* to a *T*, a *C* to a *G* and a *G* to a *C*. These sequences of nucleotides are now known to encode for the expression of the phenotype (observable characteristics of organisms) by a process known as the Central Dogma illustrated in Figure 1.2.

Central dogma: In its simplest form, the central dogma can be described as follows. Some regions of the DNA, called genes, contain the inherited genetic information; they are separated by regions said 'non-coding'. Genes are transcribed into Ribonucleic Acid (RNA). RNA is a molecule very similar to DNA with the two main following differences: it is usually single-stranded, and Thymine (*T*) is replaced by a very similar Uracil (*U*) nucleobase. This RNA molecule is then itself translated into sequences of amino acids named proteins, essential components responsible for most regulating activities of the organism.

Depending on the complexity of organisms, the non-coding regions represent from 2% (as in bacteria) to 98% of the DNA (as in humans). They not only separate genes from

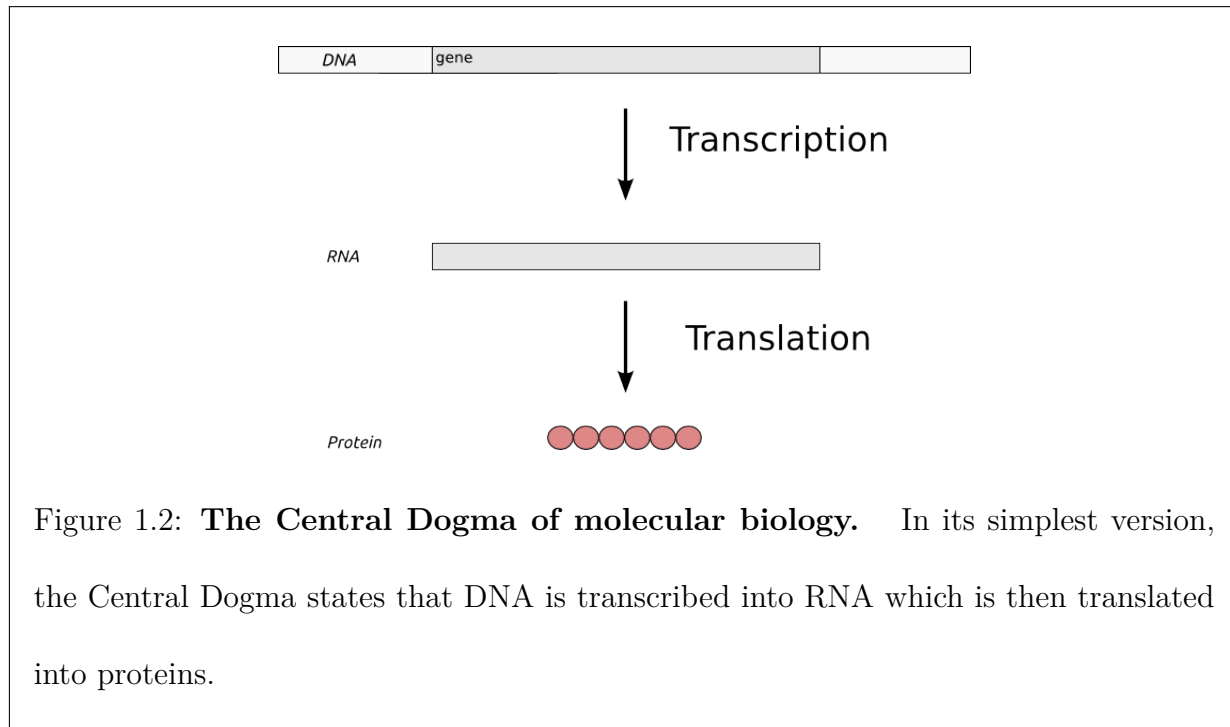


Figure 1.2: **The Central Dogma of molecular biology.** In its simplest version, the Central Dogma states that DNA is transcribed into RNA which is then translated into proteins.

one-another, but can also be present inside a gene: the latter is then made of a succession of coding sequences, called exons, and non-coding sequences, called introns. For instance, in the human genome, a gene is on average made of 9 exons thus separated by 8 introns.

In eukaryote organisms, on which we will focus from now on, the central dogma can be detailed as follows (see Figure 1.3). In the nucleus, genes (both exons and introns) present on DNA are transcribed into pre-RNA, which will be subject to a series of processes before reaching maturity. Among these processes, we find the splicing and removal of transcribed introns, the migration of RNA from the nucleus to the cytoplasm, the addition of a 5' cap (a sequence of a few nucleotides) to the 5' end, and the addition of a poly-A tail, a sequence of *As* with average length varying between species (50-70 nucleotides in yeast, about 250 nucleotides in mammalian cells) to the 3' end. Most importantly, this tail is the last of the transformations undergone by the RNA, and its presence thus characterizes a mature RNA (also called 'messenger', and denoted mRNA). Part of this mRNA will then be translated into proteins through the 'Genetic Code': to each triplet of nucleotides corresponds one amino acid.

This last 5' to 3' directed translation process only concerns parts of mRNA: on both

ends, sequences of nucleotides are not translated. These include the cap and tail, as well as sequences which were present on DNA, thus which have been transcribed. These sections are called UnTranslated Regions (UTR) on which we will be focusing.

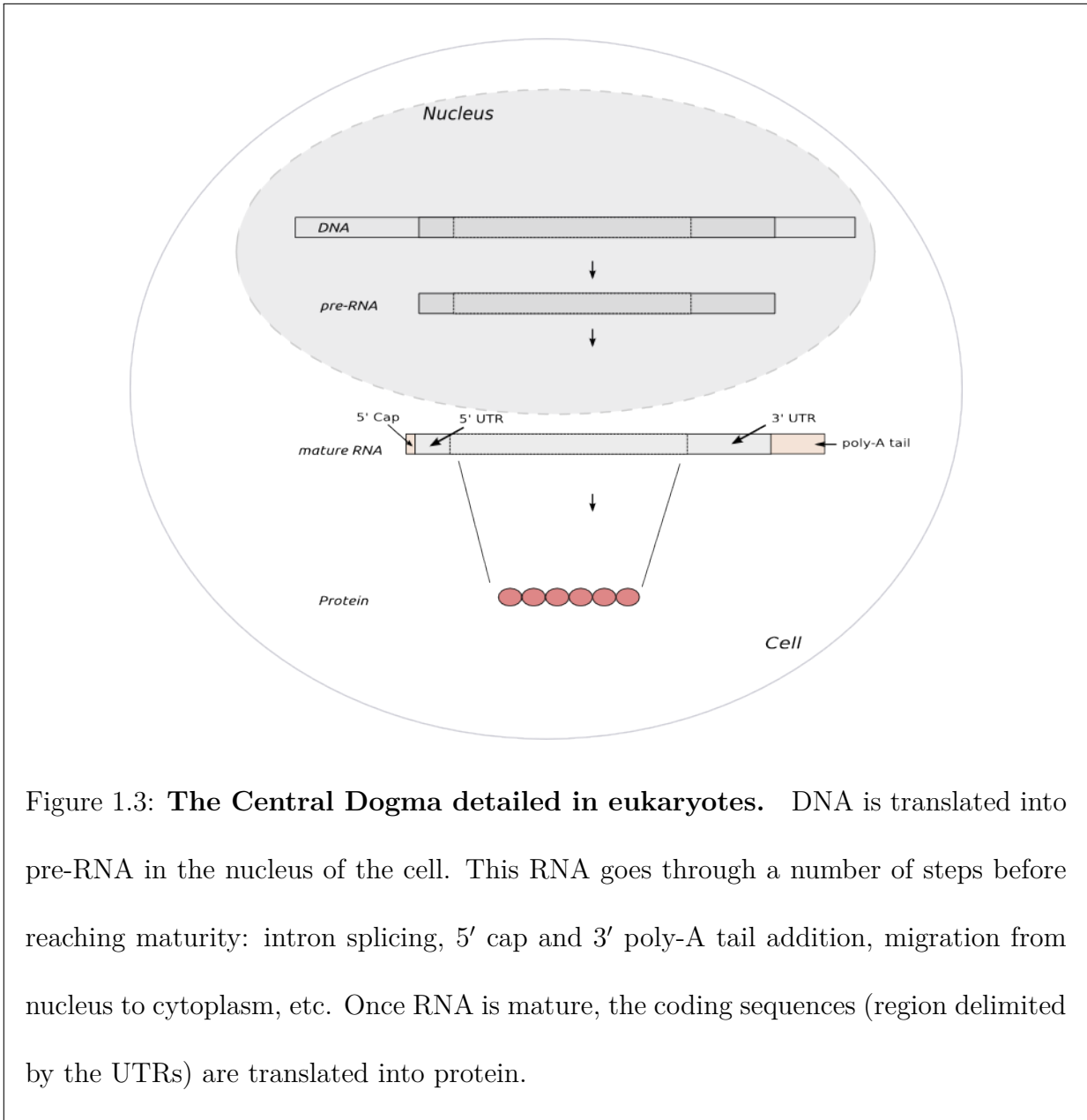


Figure 1.3: **The Central Dogma detailed in eukaryotes.** DNA is translated into pre-RNA in the nucleus of the cell. This RNA goes through a number of steps before reaching maturity: intron splicing, 5' cap and 3' poly-A tail addition, migration from nucleus to cytoplasm, etc. Once RNA is mature, the coding sequences (region delimited by the UTRs) are translated into protein.

UnTranslated Regions: The central dogma, even in the detailed version described above, is often read in one direction: DNA \rightarrow RNA \rightarrow proteins. Because the sequence of UTRs is not directly responsible for protein composition, initially little attention was

paid to them. In the last decades however, it was shown that proteins and RNA in turn regulate DNA, and micro-RNA and their role were discovered. It was then realized that UTRs do have an influence over the functionality of organisms. Recent studies have evidenced that they play a number of important roles: for instance, they are binding sites for proteins responsible for translation (??), they promote the initiation of translation (?), they are involved in translational regulation (?) and in the location of the translated protein in the cell (?). Moreover, mutations (change of a nucleotide in the DNA sequence) occurring in UTRs may be responsible for genetic diseases (??), for instance by preventing the expression of the gene.

UTRs of a given gene may vary in size depending on the environment condition. In almost all organisms, a large proportion of genes —40 to 50% in mouse and humans (?), about 72% in yeast (?)— have more than one polyadenylation sites (position of the genome where the poly-A tail will be added), and thus different possible UTR length. Even though 5' UTRs have been less studied, genes may also allow different 5' UTR length, and for instance, ? show that they are longer when genes are up-regulated (*i.e.* are more expressed than in a normal environment).

While each cell of an organism has the exact same genetic information, their specificity is determined by which genes they express. For instance, a gene coding for eye color might be expressed in eye cells but not in heart cells, or the gene coding for cell proliferation might be more expressed in an individual affected by cancer than in another individual.

This cell specificity and gene regulatory role call for methods to assess both genotype and gene expression with the goal of better understanding organism functionality, a key element in the study of pathologies such as cancer. To this end, it is necessary to have the annotation of the genome of the species studied, *i.e.* the knowledge of the boundaries between coding and non-coding regions, in order to determine the variations between different individuals of the same species. The studies cited above and many others agree about the importance of UTRs and the need to annotate them, study their mutations, or compare their length in different environment.

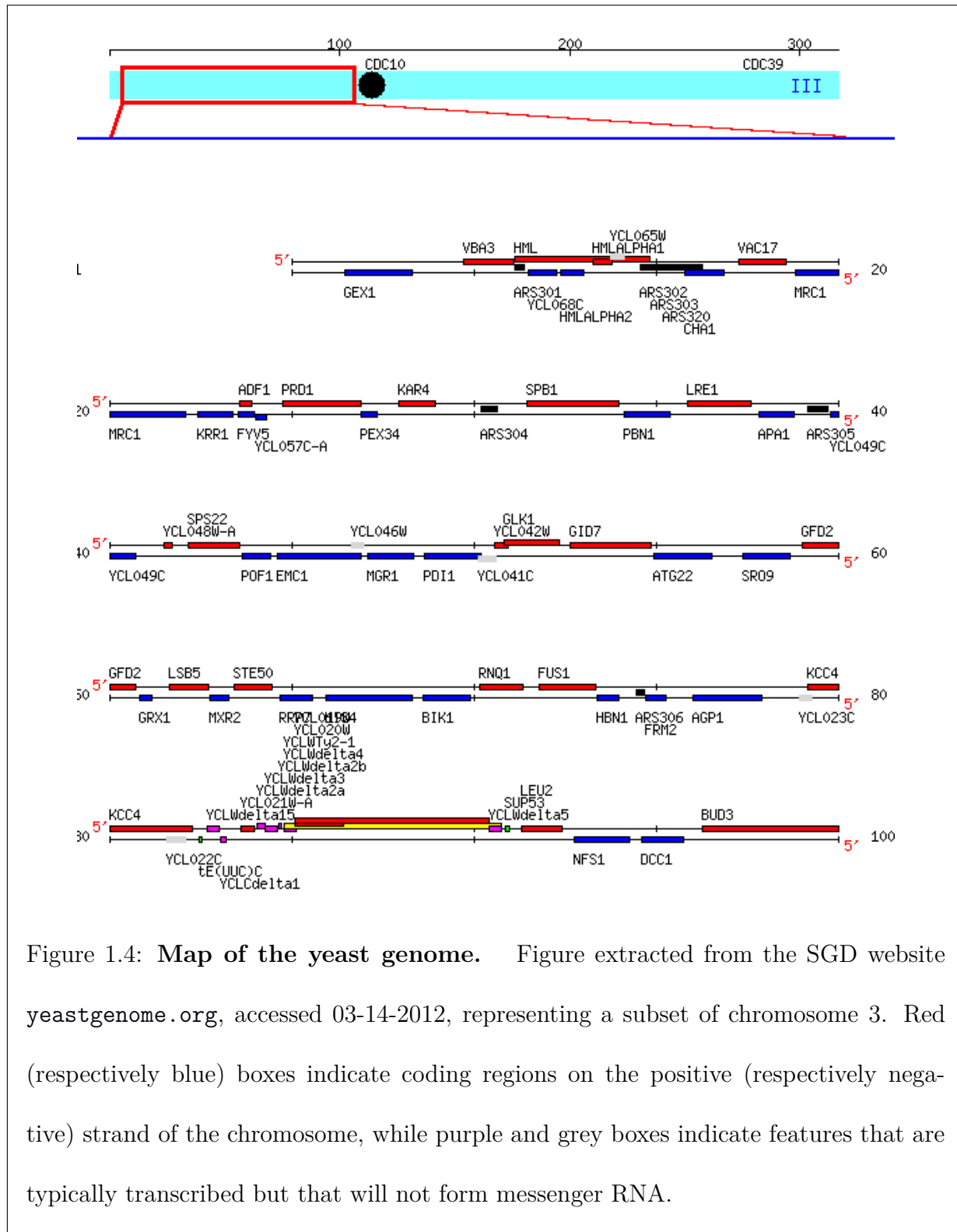
Now remembering that UTRs are present both on DNA and mature RNA molecules

(they are transcribed but untranslated sequences), sequencing the latter (*i.e.* the reconstructing its sequence of nucleotides) is an appropriate approach to their study. Section 1.1.2 will briefly recall the history of genome sequencing (be it DNA or RNA) and present a recent technology called 'Next-Generation Sequencing' (NGS).

Yeast genome: Yeast is a unicellular eukaryote family of about 1500 known species, among which *Saccharomyces Cerevisiae* is the most famous, mostly due to its use as baking powder. As is commonly the case, when there is no ambiguity we will use the term 'yeast' to refer to this particular species.

The yeast genome is composed of about 12 million nucleotides divided into 16 chromosomes. Approximately 6300 genes, with an average length of 1450 base-pairs (bp), have been annotated, and an official annotation is available on the Saccharomyces Genome Database (SGD) website: www.yeastgenome.org. Those genes have a rate of 0.007 intron per gene and account for 72% of the genome. Figure 1.4 presents a portion of the yeast genome: even though genes represent a large percentage of the total DNA, they are usually well separated from one-another, rarely located on both strands at the same time, and very few have introns.

The SGD annotation is that of coding sequences: it consists of a description of transcript boundaries (introns and internal UTR delimitations). However, up until recently no annotation of the UTRs was available, and though information are included with time, it still needs completion. A few studies have annotated the UTR boundaries of a fraction of genes, with heuristic methods: the use of in vitro experiments during which the impact of imposing different 5' UTR boundaries is studied (?), detection of a shift in the signal of sequencing experiments (NAGALAKSHMI *et al.*, 2008), peak calling, development of specific sequencing methods to target the 3' UTR (??), etc. While this work is not complete and might be improved, some useful general trends can still be used: the median lengths are respectively 50 and 100 bases for the 5' and 3' UTRs, but UTR's length of a gene are not correlated.



1.1.2 Next-Generation Sequencing technology

Brief overview of the history: The first genome ever fully sequenced, bacteriophage ϕ X174, had a length of 5386 nucleotides and the sequencing earned its publisher a Nobel

Prize (?). Less than 30 years later, the cost of sequencing a human genome (3.2×10^9 nucleotides) is less than 10 thousand dollars. Figure 1.5 compares the evolution of this cost to Moore's Law, which describes the observed trend that the number of transistors in computers doubles every two years. This comparison shows that we are facing a scientific field which improvement and performance translate into costs trends similar to computer power, and this calls for the development of new methodologies to analyze the tremendous, continually-growing resulting data sets.

This might also lead to wonder whether statistical methods proposed to deal with these data will not be outdated as soon as they are developed. In fact, if a continued decrease of sequencing costs can be expected, as will be described in the next paragraph with a brief history of genome sequencing, the evolution of the resolution has reached its maximum as we are now able to obtain information at the nucleotide scale. Methods developed for these technologies will therefore hopefully be improved, but never outdated.

Genome sequencing began in 1975 and 1977 as Sanger and Gilbert proposed almost simultaneously two methods for sequencing DNA: an "enzyme approach" (?), and a "chemistry approach" (?). While the second method was at first more popular, the surge for gene sequencing really started when Sanger sequencing imposed itself with the development in 1984 of Polymerase Chain Reaction (PCR), a technique for the amplification of DNA. This method, still at the core of all sequencing technologies today, is based on creating complementary strands of the target, as illustrated in Figure 1.6.

Sanger sequencing: A polymerase enzyme is introduced in the medium along with a large amount of nucleotides, some of which are colored with a fluorescent nucleobase-specific marker, and tied to a reversible terminator, a molecule which prevents further ligation to other nucleotides. The polymerase reads the target fragment (from 5' to 3') and associates to each base a complementary nucleotide from those available in the medium. When this nucleotide is tied to a terminator, the fragment of DNA thus created and completed is freed in the medium, and the enzyme starts over again. When all nucleotides are used,

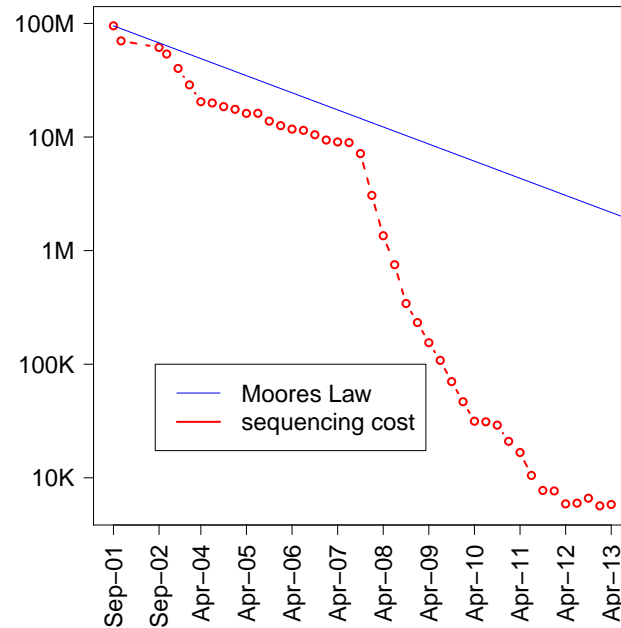


Figure 1.5: **Evolution of the sequencing cost in the last decade.** Comparison of the cost of sequencing the human genome (in dollars) to Moore's law. Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts. Accessed 07-20-2013.

the fragments obtained are analyzed: their length is assessed by chromatography, and their last base is read by fluorescence. Because of the large amount of fragments, there is a very large probability that each possible fragment-length will be observed, and then combining the information provided by length and last base composition allows to obtain the whole sequence of the initial target fragment. In the context of NGS sequencing, this fragment is called a read.

In the two decades following the development of PCR, over 20 sequencing methods were developed, and many genomes were at least partly sequenced. In parallel, detecting and determining the relative abundance of transcripts in RNA samples became a central theme in a number of biological studies. Microarrays, which had been introduced in 1983 (?),

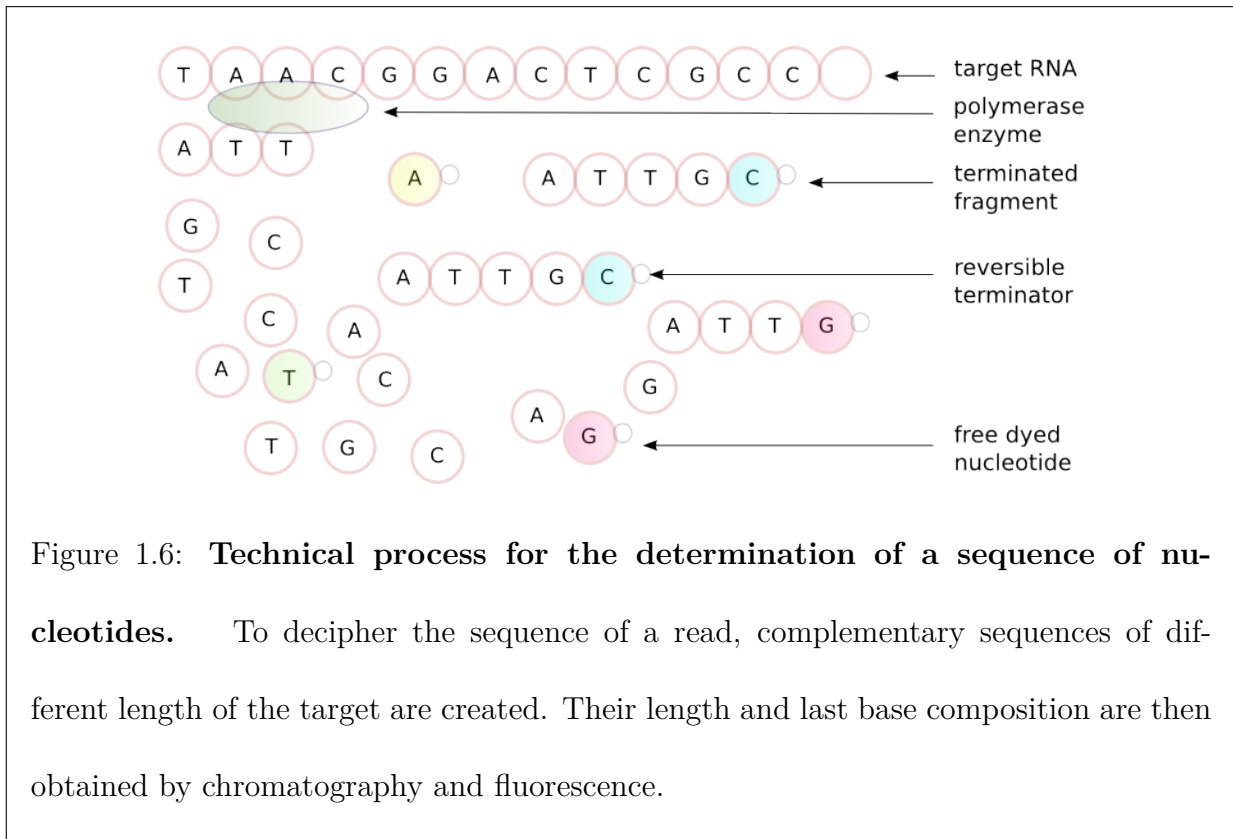


Figure 1.6: **Technical process for the determination of a sequence of nucleotides.** To decipher the sequence of a read, complementary sequences of different length of the target are created. Their length and last base composition are then obtained by chromatography and fluorescence.

became the largely preferred method in the 90's with the establishment of Affymetrix and Illumina companies to address such questions.

Microarray approaches rely on the hybridization of the target fragmented single-stranded RNA to sequences of complementary DNA of known composition, often called probes, previously attached to a glass-slide. The sequence of target RNA and its abundance can then be inferred using laser fluorescence to assess the extent of hybridization to the probes. Probes are designed in such a way that quantities such as transcript amount, chromosome copy number or allele specificity can be assessed. The resolution of microarray methods is then determined by considering both the probe size (*i.e.* the length of the fragment of DNA attached to the glass-slide) and the genomic distance between the probes. In the last decade, most arrays had a resolution of about 5 thousand nucleotides.

Up until early in the last decade, it seemed like the sequencing technology would follow a growth rate equal to that of Moore's law: progress could be made on the chromatogram and chemistry materials but the techniques remained limited by the amount of space and

human expertise needed. In 2005 however, a technique to optimize the Sanger protocol was developed, and by 2008 all industries turned to what was called 'second-generation sequencing' technology. This technology can be used to sequence DNA, with the possible aims of creating reference genomes or assessing chromosomal aberrations (for instance extra copies of a chromosome), and RNA, with the aim of determining gene expression in particular cells. In the first case, we talk of DNA-Seq experiments, in the second, of RNA-Seq experiments. The next paragraph details a particular sequencing process, from the 'fire' technology, which was used for the benchmark dataset of this thesis. The protocol concerns RNA sequencing, but in the case of DNA-Seq, it would be identical except from the first step which would be skipped.

The sequencing process, summarized in Figure (1.7), can be described in five steps:

1. *Mature RNA is extracted and reverse-transcribed.* This is usually done using enzymes which target the poly-A tail by, for instance, hybridizing oligo-(dT) cellulose (long sequences of *T*s) to the tail. Extracted (and further purified) RNA is then reverse-transcribed into (a single-stranded) complementary DNA. To perform this step, nucleotides and enzymes are added to the RNA medium. The enzymes will run through the RNA from the 5' to the 3' end creating a first strand of DNA by completion: a *T* for an *A*, an *A* for a *U*, a *C* for a *G* and a *G* for a *C*. Once this first strand of cDNA is created, the procedure is repeated, this time producing the second DNA strand from the first. The original RNA is then removed and we are left with double-stranded DNA molecules.
2. *Double-stranded DNA is amplified and sheared into 200-300 base-pair-long fragments.* At this point, one DNA version of each initial molecule (be it RNA in RNA-Seq experiments, or the actual initial DNA molecule in DNA-Seq experiments) is present in the medium. Even the most recent version of NGS technologies cannot sequence fragments longer than 1000 bases (and most often will not exceed 200bp), therefore to obtain the sequence of a whole RNA molecule, it needs to be sheared into pieces. Because in the shearing process, and later on in the sequencing step, information on nucleotides is likely to be lost, the size of the library (set of pieces of DNA which will

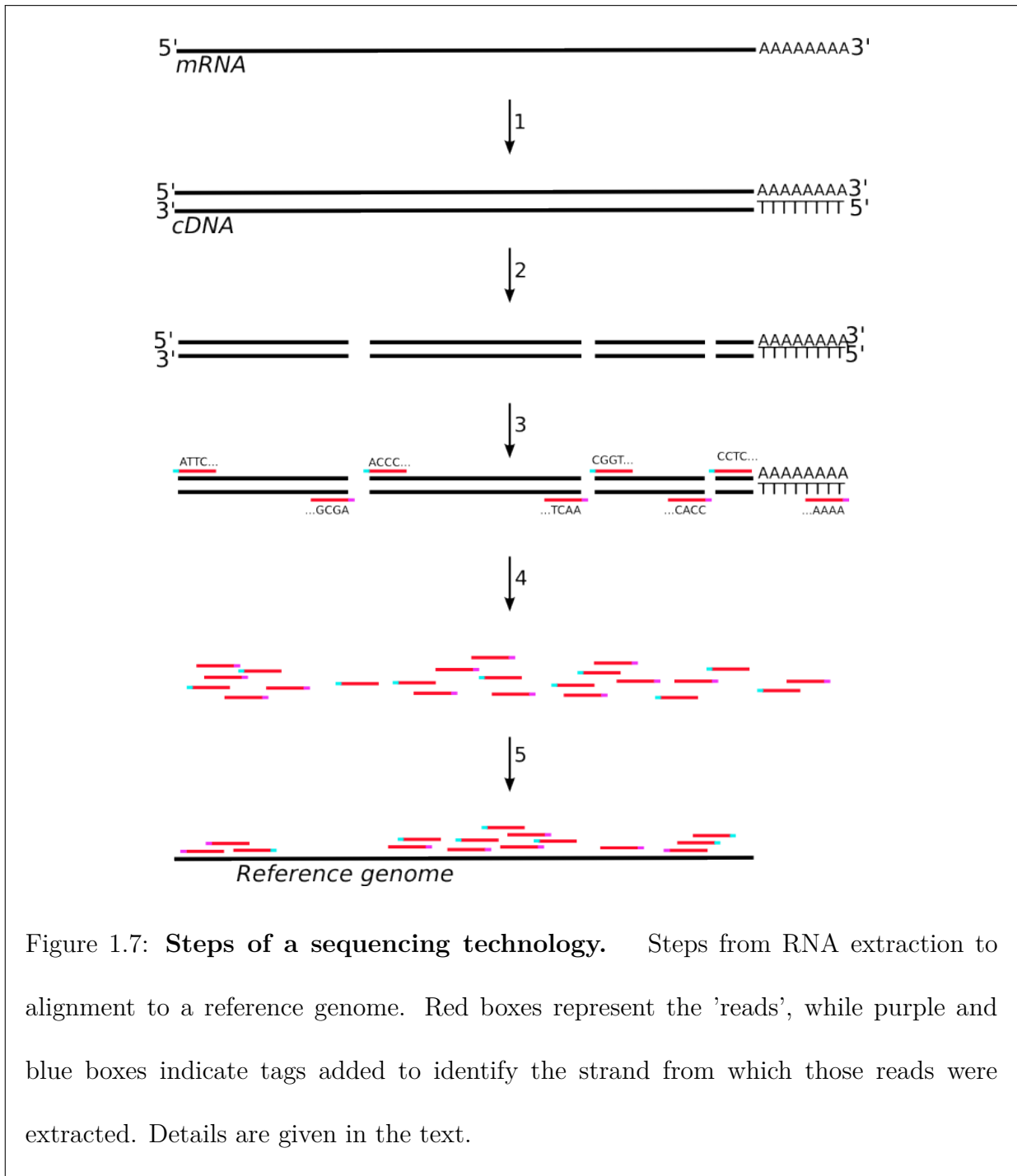


Figure 1.7: **Steps of a sequencing technology.** Steps from RNA extraction to alignment to a reference genome. Red boxes represent the 'reads', while purple and blue boxes indicate tags added to identify the strand from which those reads were extracted. Details are given in the text.

be sequenced) is usually not large enough to ensure uniform and sufficient coverage of the target. Most sequencing technologies therefore use an amplification step such as PCR prior to sequencing. After amplification, the DNA is sheared randomly (using enzymes) into fragments of size varying between 200 and 300 bases.

3. *Sequencing step.* The first 36 bases of the 5' end of both strands of each fragments is then sequenced. In the first-generation sequencing technologies, the protocol to perform this step was Sanger's (see Figure 1.6). The huge improvement of second-generation sequencing comes from the ability to read fragment-length and fluorescence at the same time. This is done by fixing the target fragments to a surface and moving a laser with extreme precision. The laser reads the fluorescence of the last base of the created fragment, and its position informs of the length of the fragment. A new enzyme is introduced in the medium which cleaves the terminator from the new piece of cDNA once the fluorescence is read. Instead of freeing the sequence in the medium and starting over, the polymerase enzyme resumes its task where it left it. This not only implies using less chemical materials, but also less time and space. It is also during this step that in strand-specific protocols (protocols which can identify the strand from which reads were sequenced) such as that used in the benchmark dataset, a strand-specific tag is added to each fragment prior to sequencing. This is represented by the blue and purple boxes in steps 3 to 5.
4. *Reads extraction.* Once all fragments are sequenced, the reads (red boxes in Figure 1.7) are extracted and only those of length 36 are kept.
5. *Alignment.* If no reference genome is available, or in the case of *de novo* assembly, specific software assemble reads together so as to form contigs, *i.e.* contiguous sequences of nucleotides representing fragments of the genome. When a reference genome is available, reads are aligned using developed software which compare the read sequences to the genome, allowing for a user-defined number of mismatches between the sequences. In our analysis, we used Bowtie (LANGMEAD *et al.*, 2008) allowing for two mismatches in each read, and kept only those that uniquely aligned. In strand-specific protocols, comparing the tags and directions in which the reads align provides information on their origin.

The protocols vary slightly depending on sequencing companies. The main difference is the length of the reads, 36 bases being among the shortest at the time of writing. Using longer read-length gives more information and facilitates the construction of reference

genome, at the price of decreased sequencing quality (more errors are observed in the output) and, in the case of RNA-sequencing, missing the smallest transcripts.

In studies examining the abundance of transcript, that the resolution is of the order of the base is a major improvement compared to the microarrays technologies, for which information on regions of at least thousands of nucleotides were summarized into a single number. This necessarily results in longer signals, but the resolution shall never be increased, and developing statistical methods efficient enough for the analysis of such signals will remain a crucial issue in the upcoming years.

1.1.3 Benchmark dataset

The benchmark dataset that will be used to illustrate the contributions all along this manuscript comes from a study performed by the Sherlock lab in Stanford. Published by RISSO *et al.* (2011), it is publicly available in the Sequence Read Archive (SRA) repository, <http://www.ncbi.nlm.nih.gov/sra>, with the accession number SRA048710.

The study aimed at comparing different yeast species grown in different media both in terms of gene expression (study not presented here) and UTR length. Out of the four studied species (*S. cerevisiae*, *S. mikatae*, *S. paradoxus* and *S. bayanus*), only *S. cerevisiae* has entirely been sequenced, and only contigs are available for the other species. These contigs are sufficient for the assessment of gene expression, but they usually are not long enough to allow the wider exploration of the genome which is needed for re-annotation. For this reason, this manuscript is only illustrated with examples from *S. cerevisiae* (from now on referred to as yeast).

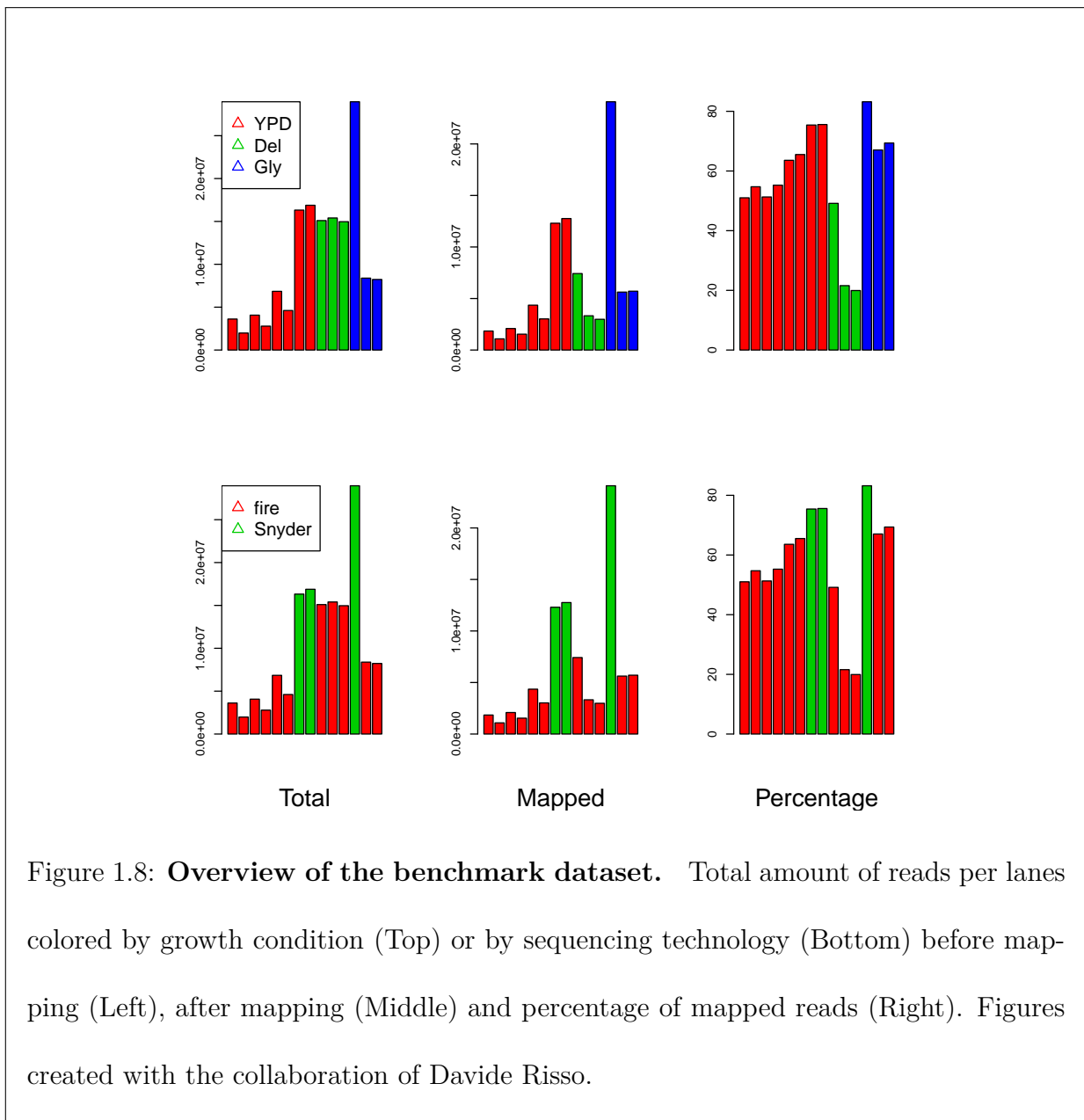
The yeast strains were grown for the same amount of time in three different media: ypd, delft and glycerol. Ypd is a rich medium made of YP glucose which is the standard growing condition of yeast, and delft is a similar but poorer medium. In both conditions, yeast cells ferment. In glycerol however, yeast respire, and we thus expect to observe more differences in gene expression and UTR sizes between this medium and any other than between ypd and delft. Details on the biological experiments are available in RISSO *et al.* (2011).

Samples of the strains were then sequenced using two different (but similar) protocols, Fire and Snyder, both of them being strand-specific. Their output, sequences of nucleotides of length 36, are similar, but the bias introduced in the sequencing step might differ. 8 samples, called lanes, were sequenced for ypd (2 Snyder and 8 Fire), 3 for delft (all Fire) and 3 for glycerol (1 Snyder and 2 Fire). Figure 1.8 represents barplot of the quantities related to the number of reads obtained for each lane: the total number of reads, the number that were mapped using Bowtie, and the percentage of mapped reads. On top, lanes were colored by growth condition, on bottom, by sequencing technology.

In most studies where a genome of reference or contigs are available, the output of RNA-Seq experiments is summarized in terms of number of reads per gene. Moreover, in studies comparing gene expression between conditions or species, normalization procedures are necessary to correct for library size or technology effects. As had been the case for microarrays, a vast literature of available methods has been proposed in the last decade.

In the context of genome re-annotation, the question of interest is the location of transcripts on the genome. The library size must be large enough for all expressed transcripts to be sequenced, but normalization procedure should not be required as they will not influence the delimitation of regions with signal. Rather than summarizing the signal in terms of number of reads per gene, we will use read counts per position, *i.e.* the number of reads which first (or last) nucleotide corresponds to position t of the genome. The higher the amount of reads, the easier the delimitations between coding and non-coding sequences will be to assess. For this reason, read counts for lanes corresponding to the same medium were summed.

The presence of the poly-*A* tail in mature RNA leads to a number of reads terminating in a long sequence of *As* (or *Ts* depending on the strand of complementary DNA from which they were issued). Since the poly-*A* tail is not present on the initial DNA, those reads will fail to align to the genome of reference, despite their useful information on UTR boundaries. In our benchmark dataset, after performing a first run of Bowtie, we sorted the unaligned reads ending with sequences of *As* (respectively *Ts*) longer than 3, trimmed the



As (resp Ts) and realigned these shorter semi-artificial reads with a second run of Bowtie. This added over 9000 reads to the alignment. Our final output signal was then, at each position t of the yeast genome, the number of reads whose last base ($3'$ end) aligned to position t . Figure 1.9 illustrates the data for the positive strand of chromosome 1 of yeast grown in ypd, using different scales.

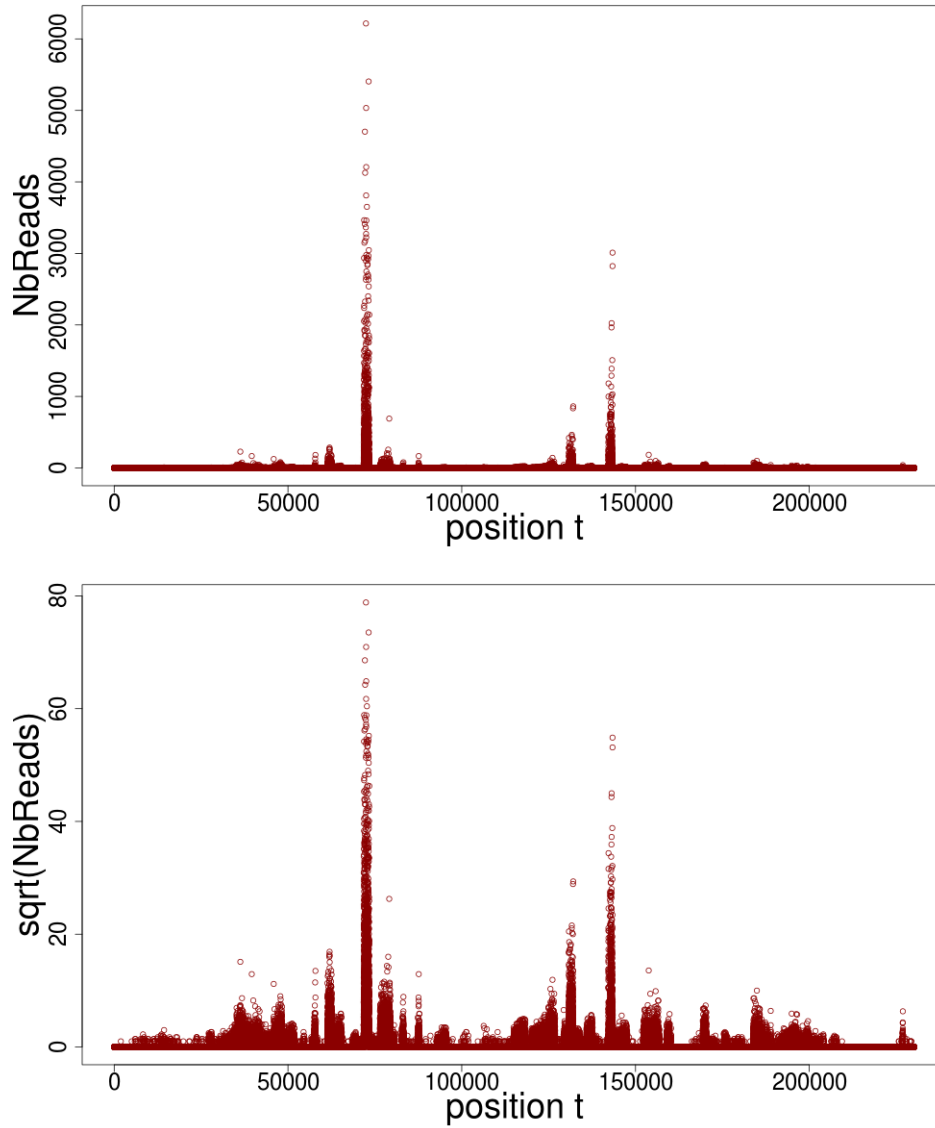


Figure 1.9: **Data overview.** Number of reads ending at each position of the genome. Regions with signal should correspond to coding regions of the genome. On the second graph, the data are plotted using a square-root scale for more visibility.

1.2 Negative binomial distribution and change-point analysis

1.2.1 Negative binomial distribution

The negative binomial distribution (\mathcal{NB}) is a two-parameter discrete probability distribution widely used to model dispersed count data, especially in biological literature, due to its many possible interpretations. Throughout this manuscript, the parametrization $\mathcal{NB}(p, \phi)$ will be used, with $0 \leq p \leq 1$, and $\phi > 0$.

Definition and interpretation: When ϕ is an integer, the negative binomial distribution is that of the number of successes Y in a sequence of Bernoulli trials of parameter p before ϕ failures occur. This corresponds to defining the negative binomial as the distribution of the sum of ϕ random variables of geometric distribution with parameter p . The probability mass function is then, for any positive integer y ,

$$P(Y = y) = \binom{y + \phi - 1}{y} p^\phi (1 - p)^y.$$

This definition is then easily extended to the case where ϕ is a positive real using the Gamma function Γ , in which case the probability mass function becomes

$$P(Y = y) = \frac{\Gamma(y + \phi)}{\Gamma(\phi)y!} p^\phi (1 - p)^y.$$

The negative binomial distribution is often referred to as the overdispersed Poisson distribution, as its variance is greater than its mean. This dispersion is introduced in the Poisson distribution by considering its mean parameter λ as a random variable with gamma distribution, resulting in a Gamma-Poisson mixture corresponding exactly to the negative binomial distribution. Specifically, let

$$\begin{aligned} \lambda &\sim \mathcal{Gam}\left(\phi, \frac{1-p}{p}\right) \\ Y &\sim \mathcal{P}(\lambda) \end{aligned}$$

Then

$$\begin{aligned}
 P(Y = y) &= \int_0^{+\infty} \frac{e^{-\lambda} \lambda^y}{y!} \lambda^{\phi-1} \frac{e^{-\lambda \frac{p}{1-p}}}{((1-p)/p)^\phi \Gamma(\phi)} d\lambda \\
 &= \frac{(\frac{p}{1-p})^\phi}{\Gamma(\phi) y!} (1-p)^{y+\phi} \Gamma(y+\phi) \\
 &= \frac{\Gamma(y+\phi)}{\Gamma(\phi) y!} (1-p)^y p^\phi.
 \end{aligned}$$

In fact, the negative binomial distribution converges to the Poisson distribution as the parameter ϕ tends to infinity: choosing p such that $\phi \frac{1-p}{p} = \lambda$ remains constant,

$$f_{NB(\phi \frac{1-p}{p}, \phi)}(y) \rightarrow f_{\mathcal{P}(\lambda)}(y).$$

Properties: Many properties and computations on the negative binomial distribution are recalled in JOHNSON *et al.* (2005); we do however detail a few of them here which will be useful in all further analysis:

- the mean and variance of a random variable Y with distribution $\mathcal{NB}(p, \phi)$ can be expressed as

$$\mathbf{E}(Y) = \phi \frac{1-p}{p} \quad \text{and} \quad \text{Var}(Y) = \phi \frac{1-p}{p^2}.$$

This corresponds to the definition of the overdispersion as

$$\text{Var}(Y) = \mathbf{E}(Y) + \phi^{-1} \mathbf{E}(Y)^2.$$

- the sum of two independent negative binomial distribution with same probability parameter p is a negative binomial distribution with parameters p and the sum of their dispersion parameters, in other words

$$\mathcal{NB}(p, \phi_1) + \mathcal{NB}(p, \phi_2) = \mathcal{NB}(p, \phi_1 + \phi_2).$$

- the transformation

$$\tilde{Y} = \sqrt{\phi} \sinh^{-1} \sqrt{\frac{Y}{\phi}}$$

approximately normalizes and variance-stabilizes the data, so that \tilde{Y} is approximately standard-normal distributed.

Estimation of the parameters: Given the value of ϕ , it is easy to estimate the probability parameter p from the observation $\{y_1, \dots, y_n\}$ of a sequence of independent random variables $\{Y_1, \dots, Y_n\}$ identically distributed with distribution $\mathcal{NB}(p, \phi)$. Let us denote $\bar{y} = \frac{1}{n} \sum_{t=1}^n y_t$ and $S_n^2 = \frac{1}{n-1} \sum_{t=1}^n (y_t - \bar{y})^2$. Then the moment and maximum likelihood estimators of this parameter coincide and have the explicit expression

$$\hat{p} = \frac{\phi}{\phi + \bar{y}}.$$

On the other hand, the maximum likelihood estimator of ϕ is obtained by solving equation

$$\sum_{i=1}^n \psi(y_i + \phi) - n\psi(\phi) + n \ln \left(\frac{\phi}{\phi + \bar{y}} \right) = 0,$$

where ψ is the digamma function. Since there is no explicit solution to this equation, $\hat{\phi}$ is typically obtained by using iterative algorithm such as Newton Raphson's. Its moment estimator however, can be computed explicitly as

$$\hat{\phi} = \frac{\bar{y}^2}{S_n^2 - \bar{y}}.$$

Note that all those estimators are biased. An explicit expression for the minimum variance unbiased estimator of p is given by

$$p^o = \frac{n\phi - 1}{n(\phi + \bar{y}) - 1},$$

however no explicit expression of an unbiased estimator for ϕ is available.

In the context of I datasets sharing their dispersion parameter, JOHNSON *et al.* (2005) propose an estimator of the latter based on a weighted average of each individual moment estimator of ϕ . Specifically, they obtain

$$\hat{\phi} = \frac{\sum_{\ell=1}^I w_{\ell} \hat{\phi}_{\ell}}{\sum_{\ell=1}^I w_{\ell}},$$

where $\hat{\phi}_{\ell}$ is the moment estimator of ϕ on profile ℓ and w_{ℓ} are weights given by

$$w_{\ell} = n_{\ell} \frac{S_{n_{\ell}, \ell}^2 - \bar{y}_{\ell}}{S_{n_{\ell}, \ell}^2}.$$

Inspired by this result, we will consider in our framework an estimator of ϕ for a single dataset based on sliding windows of length h . Our procedure is the following:

1. set $h = 15$;
2. For each sliding window L of size h , compute $\hat{\Phi}_L$ the moment estimator of Φ_L ;
3. $\hat{\Phi} = \text{median}\{\hat{\Phi}_L\}$;
4. while $\hat{\Phi} < 0$ set $h = 2.h$ and go back to 2.

While this estimator does not have theoretical guarantees, it allows to deal with variations in the intensity of the signal in a profile which is assumed to have constant overdispersion. Perspectives on improving the quality of this estimator should include allowing windows of different sizes to deal with large bands of zeros, and the inclusion of weights as in Johnson, Kotz and Kemp's estimator which we would define so as to robustify the estimator.

1.2.2 Change-point analysis

Definition of segmentation

Change-point analysis is an area of statistics that relates to the analysis of time series in order to identify instants, called change-points or break-points, where statistical properties before and after these instants are different. Typically, the distribution of the data is supposed to be piece-wise constant, with abrupt changes at locations τ_1, τ_2 , etc.

We define the change-point analysis of a signal of length n as the union of a partition m of $\{1, \dots, n\}$ and a set of distributions which are segment specific.

Let us denote

- K the number of segments of m ,
- τ_k the k^{th} change-point, with $0 \leq k \leq K$ and conventions $\tau_0 = 1, \tau_K = n + 1$,
- $m = (\tau_1, \dots, \tau_{K-1})$ and
- f_k the distribution in the k^{th} segment $[[\tau_{k-1}, \tau_k[$.

Then we are interested in the set

$$\{m, \{f_k\}_{1 \leq k \leq K}\}.$$

One simple example is the change-point analysis of a piece-wise constant signal S taking values 0 if $1 \leq t < t_1$, 1 if $t_1 \leq t < t_2$ and 0 if $t_2 \leq t \leq n + 1$ (cf. Figure 1.10). Then a possible summary of S is $\{(t_1, t_2), \{\delta_0, \delta_1, \delta_0\}\}$. In this example, we can see that the distribution δ_0 is used twice. This introduces the need to distinguish between approaches that will suppose that some distributions can be used in more than one segment from those that will assume one distribution per segment. In the first case, the inference is typically performed using Hidden Markov Models (HMM). The second, on which we will focus in this Thesis, is an area of statistics which we will refer to as *segmentation*. Very often in the literature and in what follows, the vocabulary of HMMs is borrowed in change-point analysis, and for instance an observation which is the realization of a random variable with distribution f_k will be said to belong to *state* k .

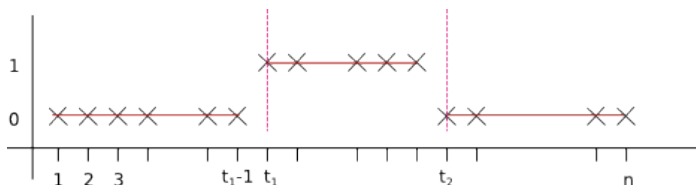


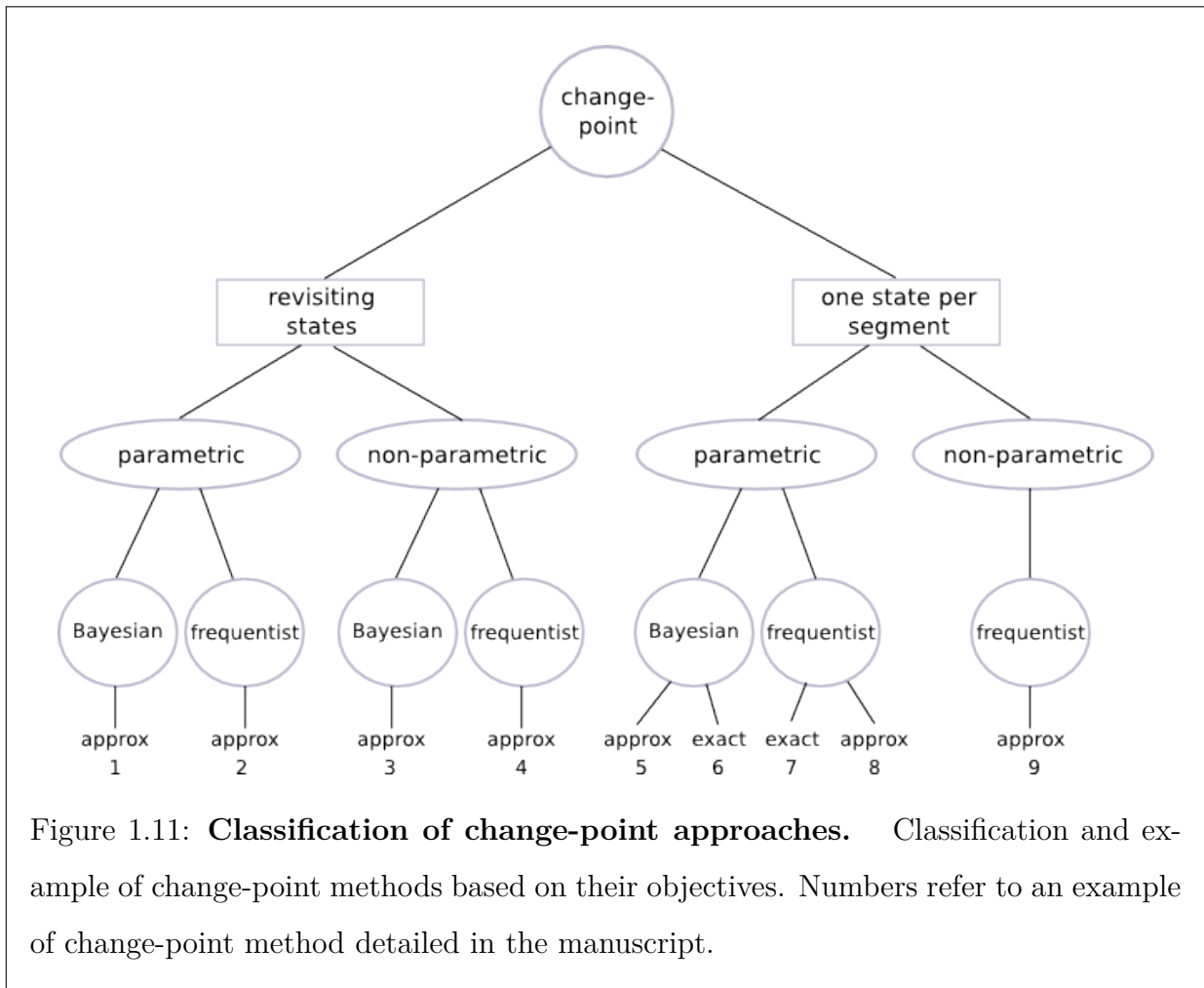
Figure 1.10: **Example of change-point analysis of a signal.** Three segments are considered.

By definition, change-point methods aim at proposing partitions of signal which verify some statistical properties. They consist in the combination of three issues:

- (i) a modeling issue: defining the family of distributions used to describe the signal in a relevant manner,
- (ii) an inference issue: estimating the parameters of the distribution, the change-points location, the number of segments, etc.,
- (iii) and a computational issue: developing algorithms to perform the inference in an efficient manner.

Overview of change-point approaches

There is a plethora of change-point methods, and any attempt at classifying them reveals more than one duality in their approaches. This paragraph tries to clarify intrinsic differences between them. The tree presented in Figure 1.11 illustrates a possible stratification in the objectives and approaches of change-point methods.



Probably, as suggested before, the most essential difference lies in the possibility of revisiting states compared to assuming one state per segment. In the first class of methods, the number r of distributions is less than the number K of segments, and this number might, or not, belong to the set of parameters to estimate. Those methods are very useful when each state is associated with a particular event, as for instance in un-supervised classification.

One simple example is the segmentation of a signal measuring rainfall abundance over a period of time long enough to overlap both sunny and rainy periods. In this case, each of the sunny and rainy states can be modeled with its specific distribution and the signal is expected to be divided in segments according to the weather. This is of course easily generalized to more than two-state signals. The second class of methods, the segmentation methods, is more adapted to signals where changes are related to irreversible events, for example the occurrence of successive renewal or breakdown in measuring devices affecting the signals.

The most famous approaches of the first class of methods rely on hidden Markov models where latent variables, assumed to follow a Markov chain, are introduced to describe the membership of each observation. The second class of methods generally rely on the exploration of the (whole or partial) segmentation space, *i.e.* the set of all possible partitions combined with the set of distributions considered in the modeling issue (*i*).

For each of those classes, we can then differentiate between parametric and non-parametric approaches. In the first case, the distribution of each segment is generally assumed to come from the same family, of which some or all parameters, θ , change abruptly at each change-point. This drastically simplifies the inference issue since classic inference techniques (such as maximum likelihood) can be used once the partition is identified. However, parametric approaches sometimes fail at capturing specific shapes in the distribution, and non-parametric methods can improve the estimation issue.

Bayesian and frequentist approaches are another form of duality in segmentation frameworks. In the parametric case, Bayesian and frequentist differ only in the traditional opposition of 'true value of the parameter' versus parameter as a random variable. In the non-parametric case however the difference is more subtle. The frequentist approach assumes almost no constraint on the shape of the distribution except for its belonging to a certain set of functions (typically Holder classes, translation distribution, etc.). On the other hand, the Bayesian approach assumes that the distribution is an infinite mixture of parametric distributions (for example Gaussian) and uses a prior (for example Dirichlet

process) on the parameters of the latter such that the inference on this infinite mixture is made possible.

Finally, a last discrimination between methods is their ability to exactly optimize the statistical criterion defined in the inference issue (*ii*). The approximation of inexact methods might come from two levels: the intractability of the inference for the criterion (issue (*ii*)) or the complexity of the computational issue (*iii*). The first case usually results in the definition of alternative target criterion (for instance optimization of composite likelihood, convexification of the likelihood, etc.), or the use of iterative algorithm (Expectation-Maximization (EM), Monte Carlo Markov Chain (MCMC), etc.) which one hopes to stop when convergence to the optimal solution is reached. The second case results in computational tricks (for instance reduction of the dimension of the segmentation space explored).

For each leaf of our tree, we give one example of segmentation approach, even though there are plenty to be found in the literature.

1. ? propose a statistical framework for the analysis of autoregressive series subject to changes in their regime. The change-point model relies on a Bayesian HMM where the inference is obtained by Monte Carlo sampling.
2. ? propose to use an HMM with six pre-defined states which emission distributions are mixtures of uniforms, normal and Dirac laws. This method is applied to the classification of regions of the genome depending on allelic proportion and number of copies.
3. ? use a non-parametric Bayesian HMM for the change-point identification of multivariate inhomogeneous space-time Poisson process. The non-parametric approach allows to assume no knowledge on the number of components in the mixture emission distribution of the Poisson process intensities. Their model is applied to the localization and intensity of crime occurrence in American cities.
4. ? propose the use of a non-parametric HMM where the emission distributions are estimated using wavelets transforms. They apply their method to the analysis of electrocardiograms for the identification of the heart state of patients.

5. ? propose a Bayesian segmentation model where the change-points are estimated one at a time as new datapoints are considered in the profile, thus reducing the dimension of the segmentation space. Their model is applied with normal distributions for the identification of chromosome copy-number variations in populations.
6. RIGAILL *et al.* (2012) propose an exact Bayesian segmentation framework using conjugate priors for exact computations, and exploring the entire segmentation space. They apply their method with the normal distribution for the segmentation of copy-number profiles.
7. ? propose to use the Dynamic Programming algorithm for the exact segmentation of profiles. Their method is illustrated for the least-square criterion on the detection of changes in the hydrophobic index imputed to changes in the structure of proteins.
8. OLSHEN *et al.* (2004) use an optimized version of binary segmentation which is based on the reduction of the complexity of the segmentation space which is explored. They apply their method to the identification of changes in chromosome copy-number.
9. ? use a wavelet-based approach for the segmentation of a signal issued from a piecewise smooth regression function in a white Gaussian noise. Their method is illustrated on the analysis of yearly temperature series in Prague.

1.2.3 Segmentation issues in this framework

In this thesis we focus on exact (*i.e.* not relying on iterative algorithms or computational tricks) parametric methods. This paragraph aims at precisely identifying the difficulties associated with each of the change-point issues (*i*) to (*iii*).

Modeling issues

This issue is the combination of the choice of segmentation approaches versus change-point methods which allow to re-visit states, and the choice of the distribution to use to fit the data.

In our context, the change-point models are introduced for the labelization of segments depending on their biological status, which differentiate coding regions of the genome which are transcribed in the cell at the time of the experiment from either non-coding or non-expressed regions. In an ideal situation where the technologies do not induce bias in the profiles, and if the DNA was transcribed at a global rate into RNA, that is independently of the localization of the coding regions, a model with only two possible segment labels (coding or non-coding) corresponding to only two intensity values could fit the data and address our biological goal. However, since for instance different cells will not express the same genes in equal proportions, we have reason to expect that each gene will have its own intensity and a change-point framework based on the ability to re-visit states will not be appropriate. Therefore, to circumvent this problem, we will consider throughout this Thesis segmentation models as introduced in Section 1.2.2.

To tackle the choice of the distribution, let us recall that our data are the read counts y_t associated to each position t of a portion of the genome of length n . In our parametric setting, we will assume that the y_t are the realization of random variables Y_t which are independent and follow a negative binomial distribution with known dispersion ϕ .

Our intuition is that the probability parameter p will be constant over each region (exon, intron, non-coding sequence, etc.) of the genome. For ease of interpretation, let us use the Poisson-Gamma mixture view for the negative binomial distribution. Then the mean parameter λ of the Poisson distribution represents the expected number of reads that will be aligned at a nucleotide of a region. Neighboring positions are likely to share the same expected value, but the latter is assumed to vary from region to region (for instance since different genes are not expressed at the same level). Now the λ parameter is Gamma distributed with shape parameter ϕ , which is common to all regions, and scale parameter $(1 - p)/p$ which will model the variability between regions.

This model corresponds exactly to a segmentation in the parameter p of the negative binomial distribution, and can be written as $\{m, \{\mathcal{NB}(p_k, \phi)\}_{1 \leq k \leq K(m)}\}$, where $K(m)$ is the number of segments of the partition m . We will often use an equivalent writing of this

model under the form

$$\forall J \in m, \forall t \in J, Y_t \sim \mathcal{NB}(p_J, \phi),$$

where J denotes a segment of m .

However, at some points in this Thesis we will be interested in comparing our approach to two other distributions for modeling the data. The first is the Poisson distribution, commonly used to model count data, for which we will assume that the mean parameter λ is segment-specific. The other is the normal homoscedastic distribution which will require transforming the data, and for which we will also assume that the mean parameter μ is segment-specific. Still, which-ever the choice of distribution, it will always be possible to summarize the segmentation as the union of the partition m and the set of parameters $\theta_1, \dots, \theta_{K(m)}$. In most cases, the estimation of those parameters will be easily performed by maximum likelihood inference, and in this case the terms segmentation and partition will often be taken one for the other, so that a partition m will be considered as a model.

Inference issues

In our context, the inference issue is itself the combination of three items:

- the choice of the number of segments $\{K\}$,
- the choice of the partition m into K segments,
- and the estimation of the parameters θ corresponding to partition m .

There are different possible criteria to choose the best model and perform the inference of the parameters (mean square, likelihood, etc). The criterion adopted in this Thesis is the maximum likelihood, with the two associated difficulties:

- The adequacy to the data always increases with the number of segments. To choose a more parsimonious, and thus more meaningful segmentation, an appropriate penalty term on K has to be defined.
- The change-point are discrete parameters, so that the likelihood function is not smooth and cannot be optimized directly. This therefore implies that the whole segmentation space has to be explored.

As stated above, in practice the difficulty is the choice of K and the estimation of the τ_k . Depending on segmentation methods, these two estimations are performed together or separately.

Major approaches for choosing the number of segments:

There are three main model-selection approaches encountered in the change-point literature. Note that although cross-validation methods have been proposed in the context of segmentation (ARLOT and CELISSE, 2011), the interpretation of cross-validation is problematic due to the spatial structure and hence dependence of the data. For this reason and because this approach is typically time-consuming, we only consider the approaches that we classify as follows:

Asymptotic considerations, that is the use of a model selection criterion developed so as to verify asymptotic properties. These include for instance the BIC criterion (YAO, 1988) and its derivatives such as the modified BIC (mBIC, ZHANG and SIEGMUND, 2007) proposed in order to avoid un-proper assumptions made by the former. In the segmentation context however these criteria are known to over-estimate the number of segments.

Birgé-Massart approaches, which are based on non-asymptotic properties of the risk of the models. In our segmentation context, this area of model selection can be explained as follows. Simplifying the set of models to the set of all partitions m as we have explained before, and with our choice of likelihood as the criterion to optimize (which gives us for each model m the best estimator \hat{s}_m), it is natural to want to choose among these \hat{s}_m that which will minimize the Kullback-Leibler risk to the true distribution s . But of course, obtaining this estimator $\hat{s}_{m(s)}$ requires the knowledge of s , which is why we call it the *oracle*. Since in practice it cannot be reached, the goal is to try to do almost as well, *i.e.* to choose one estimator $\hat{s}_{\hat{m}}$ satisfying an *oracle inequality* of the form

$$R(s, \hat{s}_{\hat{m}}) \leq CR(s, \hat{s}_{m(s)})$$

where $R(s, u)$ is the Kullback-Leibler risk between s and u , and C is a constant that we hope as close as possible to 1. To this effect, \hat{m} is chosen so as to minimize the likelihood

penalized by a function pen depending on the model dimension, and which needs to be well defined. Now writing, for all m ,

$$KL(s, \hat{s}_{\hat{m}}) \leq KL(s, \bar{s}_m) + \bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{\hat{m}}) - pen(\hat{m}) + pen(m).$$

(where KL is the Kullback-Leibler divergence, \bar{s}_m is the projection of s on partition m and $\bar{\gamma}$ is the centered likelihood $\gamma - \mathbf{E}(\gamma)$), we can see that the penalty needs to be chosen large enough for $pen(\hat{m})$ to compensate the fluctuations of $\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{\hat{m}})$, but not too large for $pen(m)$ not to penalize the difference between $KL(s, \bar{s}_m)$ and $KL(s, \hat{s}_{\hat{m}})$.

Note that the distinction between asymptotic and non-asymptotic approaches is made unusual here since n can be very large. Yet, unlike most asymptotic settings where the size of the models does not depend on the number of observations, in the segmentation area the size of each model as well as the size of the list of models will increase with n . These kinds of approaches are becoming more and more popular with the enthusiasm in 'Big Data' related problems. In particular, various papers were proposed to analyze Gaussian datasets or mixtures (??), categorical variables (AKAKPO, 2011) and Poisson processes (REYNAUD-BOURET, 2003; BIRGÉ, 2007; BARAUD and BIRGÉ, 2009).

Classification-based approaches, such as the Integrated Complete Likelihood (ICL, BIERNACKI *et al.*, 2000) which are inspired from the missing data framework. The original expression of the ICL was given by

$$\log \mathbb{P}(y, S | \mathcal{M}_K)_{|S=\hat{S}}$$

where \mathcal{M}_K is the set of all possible models with K labels, y are the observations and \hat{S} is the estimator of S , the corresponding (unknown) clustering membership. ? then proposed to replace \hat{S} by its conditional expectation given the observations, so that the definition of the ICL which is currently widely used can be written as

$$\begin{aligned} ICL(K) &= \mathbb{E} [\log \mathbb{P}(y, S | \mathcal{M}_K) | y] \\ &= \log \mathbb{P}(y | \mathcal{M}_K) + \mathbb{E} [\log \mathbb{P}(S | y, \mathcal{M}_K) | y] \end{aligned}$$

The most-right term, or more precisely its opposite, is called the negative *entropy* of the classification and is commonly denoted as $\mathcal{H}(\mathcal{M}_K)$. This expression reveals the classification purpose of the ICL, since it is equivalent to using the integrated likelihood criterion

penalized by a term which measures the uncertainty of the classification. Indeed, the entropy will be highest when all models from \mathcal{M}_K will be equiprobable given the data, which, in the mixture framework, is equivalent to saying that the components of the mixture are poorly separated. On the contrary, a value of K for which one model outperforms all others will yield an entropy close to zero. In the segmentation framework, RIGAILL *et al.* (2012) propose an exact computation of the ICL which they use as an effective method for the choice of the number of segments K . Their idea is to consider a partition m as the set of clustering membership S . Indeed, the aim of segmentation is exactly the labeling of each observation into segment numbers. From this observation, and using the usual shortcut K for \mathcal{M}_K , the ICL criterion becomes

$$ICL(K) = -\log \mathbb{P}(y, K) - \sum_{m \in \mathcal{M}_K} \mathbb{P}(m|y, K) \log \mathbb{P}(m|y, K).$$

Confidence in the change-point location:

A few techniques have been proposed in the literature to evaluate the uncertainty in the change-point location, such as the derivation of asymptotic properties of the estimators (FEDER, 1975) and of likelihood-ratio tests (MUGGEO, 2003). Other approaches imply using Bootstrap techniques (HUŠKOVÁ and KIRCH, 2008). In a non-asymptotic framework, the constrained HMM approach of LUONG *et al.* (2013) which we will introduce in Section 1.2.3 computes, for a given number of segments K , the posterior distribution of change-point locations conditional on the set of parameters Θ_K . A more complete approach has been proposed by RIGAILL *et al.* (2012) in a Bayesian framework to derive the posterior distributions of various quantities of interest – including change-point locations – in the context of exponential family distributions with conjugate prior. Applying their model to the negative binomial distribution results in considering the model

$$\begin{cases} \forall J \in m, & p_J \sim \mathcal{Beta}(a, b) \\ \forall J \in m, \forall t \in J, & Y_t \sim \mathcal{NB}(p_J, \phi) \end{cases}$$

Implementing their algorithm however leads to a quadratic complexity and restricts its use to small dataset (see Section 1.2.3).

As for the comparison of change-point location in different profiles, we know of no method having ever addressed this question. Indeed, in the literature, two approaches can typically be considered for the analysis of multiple profiles. The first consists in the simultaneous segmentation of all series, looking for changes that are common to all of them. This approach amounts to the segmentation of one single multivariate series but might permit the detection of break-points in series with too low a signal to allow their independent analysis. The second approach consists in the joint segmentation of all the series, each having its specific number and location of changes. This allows to account for dependence between the series without imposing that the changes occur simultaneously. Still, none of these techniques can deal with the statistical problem of comparing the change-point locations in series that have been segmented separately.

Computational issues

Although the effectiveness of the algorithm is not directly involved in the resulting segmentation quality, it can be a limiting factor for the analysis of large signals. Indeed, the inference issue (ii) requires that for a given number of segments K , the whole segmentation space in K segments be explored, so that there are $\binom{n-1}{K-1}$ possibilities. This exploration can therefore not be performed in a naive manner. Most segmentation algorithms proceed in two steps:

- (a) the identification of the optimal segmentation in each given number of segments k from 1 to a user-defined K_{max} , followed by
- (b) the choice of the optimal K among those explored based on some criterion defined in the inference issue.

Typically, the computational difficulty results from the (a) step, and we can therefore assume that the choice of K is not an issue (for instance we can consider that $K = K_{max}$). Up-until 2010, there was really only one algorithm addressing step (a) in an exact manner: dynamic programming (DP). DP was introduced by ? but was first used in the context of segmentation by ? more than 25 years later, under the form of the segment neighborhood algorithm, which is still widely used today.

This algorithm reduces the complexity of the exhaustive exploration from $\mathcal{O}(n^K)$ to $\mathcal{O}(Kn^2)$. To do so, the algorithm relies on the segment additivity property. Indeed, the best partition of $\{1, \dots, n\}$ in k segments can be obtained as a minimizer amongst only $n - k$ possibilities indexed by $k - 1 < t \leq n$ which are the union of the best partition of $\{1, \dots, t\}$ into $k - 1$ segments and the last segment $\llbracket t, n + 1 \llbracket$. Specifically, denoting $C_{k,t}$ the cost of the optimal segmentation of $\{1, \dots, t\}$ in k segments it suffices at step k to obtain

$$C_{k,n} = \min_{\{k-1 < t \leq n\}} \left\{ C_{k-1,t} + \min_{\mu} \{c(\llbracket y_{t+1}, y_{n+1} \llbracket, \mu)\} \right\}$$

where c is the segment cost function associated to the *loss function* γ (typically γ is the log-likelihood and c its sum over each point in the segment) and μ is its parameter. At each step k , a first minimization on μ is performed for all t , and it is followed by a minimization on t . Implemented in the R package `changepoint` for the Gaussian distribution, the DP algorithm still cannot be used for values of n as large as those that we will want to consider in our analysis.

Very recently, two faster and still exact algorithms were proposed: the pruned DP algorithm (RIGAILL, 2010), and the pruned exact linear time (PELT) algorithm (KILLICK *et al.*, 2012). As the DP algorithm, the former recovers each optimal segmentation in 1 to K_{max} segments, but at each step a pruning process discards most suboptimal segmentations hence decreases the number of comparisons to be made, and results in an empirical complexity faster than $\mathcal{O}(Kn \log(n))$. Implemented in the R package `cghseg` for the Gaussian distribution, the PDPA remains valid as soon as the following conditions are verified:

- the cost function c is point-additive (*i.e.* $c(\llbracket y_1, y_2 \llbracket, \mu) = \sum_{y_t \in \llbracket y_1, y_2 \llbracket} \gamma(y_t, \mu)$)
- γ is a one-parameter loss function (*i.e.* $\mu \in \mathbb{R}$)
- γ is convex with respect to its parameter μ .

Indeed, while the DP relies on segment-additivity, the pruned DP algorithm is based on point-additivity. Specifically, if it computes the same quantity $C_{k,n}$, it exchanges the order on which the minimization are performed, so that

$$C_{k,n} = \min_{\{k-1 < t \leq n\}} \left\{ C_{k-1,t} + \min_{\mu} \left\{ \sum_{i=t}^{n+1} \gamma(y_i, \mu) \right\} \right\}$$

becomes

$$C_{k,n} = \min_{\mu} \left\{ \min_{\{k-1 < t \leq n\}} \left\{ C_{k-1,t} + \sum_{i=t}^{n+1} \gamma(y_i, \mu) \right\} \right\}.$$

Now at each step k , a first minimization on t is performed for each value of μ . Each t being a candidate for the last change-point of the segmentation in k segments up to point n , if we store, for each of them the values of μ for which it is optimal, we can discard t when the set of μ s becomes empty. Then when adding new datapoints (remember that, for instance, performing step $k + 1$ requires that each $C_{k,i}$ has been computed for $k \leq i \leq n$), the number of those candidates has been reduced, resulting in an empirical almost linear complexity in most applications. Note that PDPA and the original DP algorithm give the same results. Part of our contribution is its implementation for the Poisson and negative binomial distributions in the package `Segmentor3IsBack` (see Section 2.1).

The second algorithm, PELT, is an exact algorithm for the optimization of a penalized version of the likelihood which results in the simultaneous choice of segments number K and optimal segmentation in this K segments (*i.e.* steps (a) and (b) are performed simultaneously). This, under some specific assumptions, drastically reduces the size of the exploration space and leads to a $\mathcal{O}(Kn)$ complexity.

Specifically, PELT optimizes the cost function

$$\sum_{J \in m} \sum_{t \in J} -\log g(y_t; \theta_J, \phi) + \beta |m|$$

over all possible partitions m , where g is some parametric probability density function with parameters θ_J which is segment-specific and ϕ which is global, and β is a constant to be chosen independently of segment number and location. Even though the authors propose a generalization to concave penalty function (at the price of higher complexity), this algorithm is dedicated to penalties which are proportional to the number of segments, such as BIC.

Similarly to the PDPA, PELT relies on the pruning of candidates for the last change-point location based on the principle that once this candidate has been beaten by another

location, it can never become optimal again as more datapoints are considered. Under some specific conditions, the authors prove that the time complexity of the algorithm is $\mathcal{O}(n)$ (the space complexity being linear too). If most of those conditions are verified for the log-likelihood criterion, one of them is most restrictive which requires that the expected number of segments increases linearly with n , as might indeed be expected in domains such as economy. In practice, this algorithm is extremely fast, but in some contexts it suffers from the inability to choose the number of segments K . It is implemented in the R package `changept` for a set of parametric distributions including Gaussian and Poisson, but not the negative binomial.

A completely different approach is that of postCP (LUONG *et al.*, 2013). Let us denote S a segmentation in K segments (instead of defining the segmentation m as the sequence of change-points, its equivalent definition in terms of sequence of segment labels is used here and denoted S , so that $S_i \in \{1, \dots, K\}$ is the index of the segment at position i), $g_{\theta_{S_i}}(\cdot)$ a parametric distribution with parameter θ_{S_i} , and $\Theta_K = (\theta_1, \dots, \theta_K)$ the global parameter.

The idea is to notice that for a segmentation S in K segments, the likelihood of the data is

$$\mathbb{P}(y|S; \Theta_K) = \prod_{i=1}^n g_{\theta_{S_i}}(y_i) = \prod_{k=1}^K \prod_{i:S_i=k} g_{\theta_k}(y_i),$$

and that this formulation is equivalent to that obtained when assuming that S is a heterogeneous Markov chain over $\{1, 2, \dots, K, K+1\}$ with constraints imposed so that states cannot be revisited. This is particularly pleasant since the posterior distribution $\mathbb{P}(S|y, K)$ can be computed efficiently using classical algorithms such as Forward-Backward. The authors therefore impose that S verifies

$$\begin{aligned} \mathbb{P}(S_1 = 1) &= 1 \\ \forall 2 \leq i \leq n, \quad \forall 1 \leq k \leq K, &\begin{cases} \mathbb{P}(S_i = k | S_{i-1} = k) &= 1 - \eta \\ \mathbb{P}(S_i = k + 1 | S_{i-1} = k) &= \eta, \end{cases} \\ \mathbb{P}(S_i = K + 1 | S_{i-1} = K + 1) &= 1, \end{aligned}$$

and η is a constant with values between 0 and 1.

Now for a given K , due to the sparse transition matrix induced by the constraints, the complexity of the Forward-Backward algorithm is $\mathcal{O}(Kn)$ in time and $\mathcal{O}(n)$ in space. This fast algorithm, implemented in the R package `postCP` for parametric distributions including Poisson and Gaussian, allows the computation of quantities such as posterior distribution of the change-points, conditional on the parameters Θ_K . Part of our contribution consisted in its extension to the negative binomial distribution.

Note that if this approach is HMM-based, contrary to most change-point methods based on HMMs such as those cited in the Introduction (see 1.2.2), this model corresponds to a *segmentation* approach, resulting in one state per segment being specified. Moreover, since no iterative algorithm is required, this approach is still considered *exact*.

As illustrated before, some approaches choose to reduce their complexity by using tricks such as exploration of a subset of models, or modification of the criterion to optimize. This is the case of the CART algorithm (BREIMAN *et al.*, 1984), a heuristic procedure based on binary segmentation (BS, SCOTT and KNOTT, 1974) to approach the best segmentations in 1 to K_{max} segments. It is implemented in the R package `changepoint` for a few parametric distributions including Poisson and Gaussian, but not the negative binomial.

The idea of binary segmentation is to split a segment into two segments at each step by minimizing a criterion (for instance the log-likelihood in our context), and to keep the best partition for the next step. Specifically, it runs the following way: one has an interval, say I , to be split into k pieces but the splits are computed and decided one by one. At the first step one choses the best way to split I in two parts, I_1^1 and I_2^1 such that $I_1^1 \sqcup I_2^1 = I$. At the second step one has to compute the best splitting of I_1^1 and I_2^1 and keep the best of them. Thus one splits one of the two intervals I_1^1 and I_2^1 and obtains I_1^2 , I_2^2 and I_3^2 and so on... The algorithm is typically stopped when a fixed number K_{max} of segments is reached. This drastically reduces the number of partitions to explore, and with it the complexity of the algorithm. Indeed, if implemented using a heap, the complexity of CART is no worse than $\mathcal{O}(n \log n)$ in time, and empirically in $\mathcal{O}(n \log(K_{max}))$, and $\mathcal{O}(n)$ in space.

This complexity can be refined by computing the number of operations which are required: at each new step, in order to find the best division of a segment of length l into 2 pieces, CART requires $7l - 3$ elementary operations. Thus the worst number of elementary operations is bounded by $K_{max}(7n - 3) + \sum_1^{K_{max}} \log_2(k)$. (The term $\log_2(k)$ is related to the comparisons to perform and comes from the use of the heap.) If the splits are quite regularly distributed, one obtains $(7n - 3) \log_2(K_{max}) + \sum_1^{K_{max}} \log_2(k)$, and may therefore expect a $n \log_2(K_{max})$ time complexity.

Note, for comparison, that PDPA needs at each step to find the roots and minimum of cost functions for each candidate, for which it requires 13 elementary operations. Intersecting two intervals requires 2 more elementary operations. If the pruning is perfect, *i.e.* if only one candidate is left at each step, then PDPA will require $15K_{max}n$ elementary operations. Therefore the best configuration for PDPA is still worse than the worst configuration for CART.

Finally, at the price of larger space and time complexity, some algorithms are able to generate additional relevant information, such as exact confidence intervals for the parameters. This is the case of the Bayesian segmentation algorithm proposed in RIGAILL *et al.* (2012), which is valid for models where the distribution is from the exponential family and which verify the factorability assumption, that is,

$$P(Y, m) = C \prod_{J \in m} a_J P(Y_J | J),$$

where

$$P(Y_J | J) = \int P(Y_J | \theta_J) P(\theta_J) d\theta_J.$$

The model is specified in Figure 1.12 and is the following:

- the number of segments K is drawn from the prior distribution $P(K)$;
- m is drawn conditional on K in $P(m|K)$;
- the parameters θ_J for each segment J are supposed to be independent and are drawn from the same distribution $P(\theta_J)$;
- finally, the observed data $Y = (Y_t)$ are independent conditional on m and (θ_J) and

have a distribution which depends on the segment:

$$(Y_t | m, J \in m, \theta_J, t \in J) \sim \mathcal{G}(\theta_J, \phi).$$

where ϕ is a global parameter that is supposed to be known, and \mathcal{G} is a parametric distribution which possesses a conjugate prior on θ .

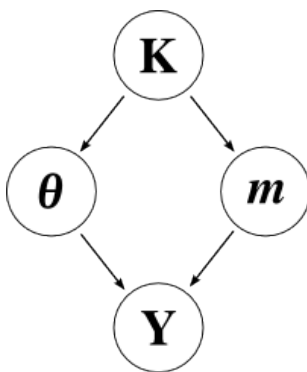


Figure 1.12: **Graphical model of the exact Bayesian Segmentation approach.** Hierarchical model of the exact Bayesian segmentation approach proposed in RIGAILL *et al.* (2012).

Part of our contribution was to implement this algorithm in an R package, **EBS**, which we fully describe in Section 3.3. The key element in this algorithm is the triangular probability matrix A defined by $\forall 1 \leq i < j \leq n+1, [A]_{i,j} = a_j P(Y_{[i,j]} | [i, j])$ (and 0 elsewhere). Indeed, the authors show that all quantities of interest can be computed by simple operations on the lines and columns of this matrix, so that the resulting complexity of the algorithm is $\mathcal{O}(Kn^2)$ both in time and space.

The generic element $[A]_{i,j}/a_j$ of this matrix are given under the form of the product of three terms in Table 1.1 for the distributions included in the package:

- the Poisson distribution with conjugate prior $\mathcal{Gam}(\alpha, \beta)$,
- the Gaussian homoscedastic distribution with known variance σ^2 and with conjugate prior $\mathcal{N}(\mu_0, \sigma^2/\sigma_0^2)$ on the mean,

- the negative binomial distribution with known dispersion ϕ and with conjugate prior $\mathcal{B}eta(\alpha, \beta)$ on p , and
- the Gaussian heteroscedastic with conjugate priors $\mathcal{IG}(\nu/2, s/2)$ on σ_0^2 and $\mathcal{N}(\mu_0, \sigma_0^2/n_0)$ on $\mu|\sigma_0^2$.

distribution	term 1	term 2	term 3
Poisson	$\prod_{t=i}^{j-1} \frac{1}{Y_t!}$	$\frac{\beta^\alpha}{\Gamma(\alpha)}$	$\frac{\Gamma(\alpha+Y_J)}{(\beta+n_J)^{\alpha+Y_J}}$
Gaussian homoscedastic	$(2\pi\sigma^2)^{-n_J/2}$	$\sqrt{\sigma_0^2}$	$\exp\left[-\frac{(\sigma_0^2+n_J)(\sum_{t=i}^{j-1} Y_t^2 + \sigma_0^2\mu_0^2) - (Y_J + \sigma_0^2\mu_0)^2}{2\sigma^2(\sigma_0^2+n_J)}\right] (n_J + \sigma_0^2)^{-\frac{1}{2}}$
Negative binomial	$\prod_{t=i}^{j-1} \frac{\Gamma(\phi+Y_t)}{\Gamma(\phi)Y_t!}$	$\frac{1}{\beta(\alpha,\beta)}$	$\beta(\alpha + Y_J, n_J\phi + \beta)$
Gaussian heteroscedastic	$\pi^{-\frac{n_J}{2}}$	$\frac{\sqrt{n_0 s^{\nu_0}}}{\Gamma(\nu/2)}$	$\frac{(n_J+n_0)^{\frac{\nu+n_J-1}{2}} \Gamma(\frac{n_J+n_0}{2})}{[(n_J+n_0)[\sum_{t=i}^{j-1} Y_t^2 - \bar{Y}_J + s] + n_J n_0 (\bar{Y}_J - \mu_0)^2]^{\frac{\nu+n_J}{2}}}$

Table 1.1: **Element of matrix A.** The generic term $[A]_{i,j}$ is the product of 4 components: the factor a_J which is induced by the prior on the distribution, and 3 terms which we discuss in Section 3.3.

1.3 Contribution

1.3.1 Introduction

Our work is primarily dedicated to segmentation methods for the general biological framework of genome annotation, using transcriptome sequencing (RNA-Seq) data. We have developed three different algorithms with different complexities in order to go from a comprehensive analysis of the genome to progressively more local scales for which we obtain more precise statistical results. Our manuscript is naturally organized around this scale, and we present the successive constructions and the links between each proposed method.

As suggested in the previous sections, two issues are characteristic of RNA-Seq data: their discrete nature, since they are directly related to the number of reads associated to each nucleotide, and their size, due to the base-resolution and the length of some chromosomes.

Discrete nature of the data. This first point could be circumvented by applying transformation techniques as it has always been done for the analysis of microarray data. Indeed, multiple existing segmentation algorithms are dedicated to data modeled by the normal distribution, and many normalization and transformation methods for NGS data have been proposed, such as that which has resulted in the publication of our benchmark dataset (RISSE *et al.*, 2011). A comparison of the main methods can for example be found in ?. However, applying normalization methods was justified in the context of microarray analysis since they were always based on the comparison of the profile from the target individual to that of a reference. Our framework is different because we want to identify significant regions on the sole basis of the profile of interest. Even when it will come to comparing the transcripts locations for different profiles in the last section of this manuscript, each of them will be segmented independently thus normalization will not be required. We therefore wanted to keep the raw data and model them using discrete distributions. A later study (partly presented in Section 3.1) comparing our model and its equivalent for the same data but transformed and modeled by the normal distribution proved us right. Indeed, even

though in most cases the results are similar, we were able to highlight situations in which our model was able to identify regions undetected by the transformed approach.

The natural question is then the choice of the distribution to use to model these data. The simplest model, which corresponds to the Poisson distribution, fails to take into account the large and intrinsic dispersion in the data, and that is what we will observe throughout the comparisons performed between methods using this distribution and our approaches (for example in Chapters 2.2 and 3.1). Two more flexible other laws require only one additional parameter : the Zero-Inflated Poisson, which is a mixture between a Dirac distribution at zero and a Poisson distribution, and the negative binomial, which we described in the introduction, and which can be interpreted as the over-dispersed Poisson distribution. Although none of them belongs to the exponential family, the second does on the condition that the dispersion parameter ϕ is fixed. Moreover, data in segments corresponding to highly expressed genes show large dispersion but their distribution does not always present a large peak in zero. For this reason and because the negative binomial distribution has become the consensus in most Seq-data modeling (especially in differential expression analysis (ROBINSON *et al.*, 2010; RISSO *et al.*, 2011)), we will therefore use it all along our contribution.

This lead us to choose for the modeling issue (*i*) the negative binomial with probability parameter p_J which is segment-specific, and with global dispersion parameter ϕ .

It is then necessary to estimate the dispersion parameter ϕ in the data so as to use its estimate as a known parameter in our segmentation models. This cannot be done in a naive manner as by definition we expect the signal to be fragmented into segments with different distributions, but with common dispersion. Ideally we would wish to use a robust estimator based on sliding windows, as the Median Average Deviation estimator (MAD, HAMPEL, 1974; DONOHO, 1995) or that of HALL *et al.* (1990) which have been proposed for estimating the variance. However, the complicated expression of the bias of classical estimators (such as the maximum likelihood or the moment estimators), and the discrete nature of the data did not allow us to propose a satisfactory estimator. We therefore chose to use a simple moment estimator on sliding windows of which we keep the median, as was

described in Section 1.2.1. Although we can show that its impact is minimal on the quality of our results, it remains the weak point of our contribution. If, however, a better estimator were to be proposed, it would be very easy to integrate it to our entire contribution which would remain unchanged.

Length of the data. This second point is more challenging than the first because it will greatly limit the range of possible algorithms. Indeed, since the resolution of the data is that of the nucleotide, we have series whose length can be as large as that of chromosomes, that is, up to n of the order of 10^8 . In such a case, algorithms able to handle the profile need linear time complexity, provided both that the space complexity is not too large (we will discuss these concepts later), and secondly that the multiplicative constants are not prohibitive.

As in many areas of statistics, the question that arises is 'what information can we get from our algorithms, and at what price?' For example, we want to get the optimal change-points in the series, but we would also wish to assess the quality of these breaks, that is, the uncertainty that is associated with their location. However, we know of no linear time algorithm that can both deal with the parameter uncertainty and that related to the location of change-points. When the goal is the precise gene annotation for which the series are much shorter, this might not be a restriction. Our contribution is the development of various methods to address the segmentation issues described in Section 1.2.3, in two biological complementary framework : whole-genome analysis, and precise gene re-annotation. Table 1.2 is an overview of the structure and content of our contribution.

1.3.2 Whole genome analysis

This study is the subject of Chapter 2 of our manuscript. We are here typically in a context where the length of the series is a limiting factor. Note though that in the case of our reference data, we are at worst only in the order of the million, whereas in the general case the length can be up to 10^8 .

	modeling (<i>i</i>)	inference (<i>ii</i>)	computational (<i>iii</i>)
Whole-genome analysis		2.2 ; 2.3	2.1
Gene annotation	3.1	3.2	3.2 ;3.3

Table 1.2: **Overview of the segmentation issues addressed in this Thesis.** Our contribution is organized by biological scale: whole-genome analysis or local, gene scale. For each of them, we recall which segmentation issue is addressed in which chapters of our manuscript.

Even though all methods developed in this Chapter hold for various distributions, we will suppose the modeling issue (*i*) resolved by the choice of the negative binomial distribution with global dispersion parameter. Indeed, the same approaches performed with the Poisson distribution resulted in an oversegmentation of the data due to the inability to capture the dispersion, while the use of the Gaussian distribution performed almost as well only when the data was transformed using the \sinh^{-1} transformation. The latter implies the estimation of the dispersion ϕ , sole real drawback in the use of the negative binomial.

An algorithm for the segmentation of long RNA-Seq profiles. Performing the inference requires the estimation of the number of segments K and the exploration of the segmentation space \mathcal{M}_K . Supposing that K is known, there are $\binom{n-1}{K-1}$ distinct partitions of $\{1, \dots, n\}$ in K segments, and here we have a very large n . The computational issue (*iii*) is therefore crucial, as using an exploration approach that is naive will result in a complexity in $\mathcal{O}(n^K)$, intractable in our context. It is in this spirit that the dynamic programming algorithm (DP, ?) presented in Section 1.2.3 was developed; unfortunately, it remains of quadratic time complexity, thus it is inappropriate for our context.

We propose to adapt the *pruned dynamic programming algorithm* (PDPA) proposed by RIGAILL (2010) to the negative binomial distribution. This choice is motivated both by the algorithmic aspect, since its time complexity, at worst in $\mathcal{O}(Kn^2)$, is empirically less, and by the assumptions necessary for its implementation. Indeed, the requirement of the PELT algorithm (KILLICK *et al.*, 2012) that the expected number of segments increases

linearly with n does not seem appropriate in our context, as explained in Section 2.1, and its inability to choose the number of segments is very restrictive. In this latter section, we describe the PDPA algorithm and its performance, detailing the conditions required for it to be valid and how the negative binomial needed to be parametrized to satisfy those requirements. Although PDPA's time complexity depends on the signal, we show that it is almost linear (in $\mathcal{O}(Kn)$) with reasonable constant in our RNA-Seq data. With space complexity in $\mathcal{O}(Kn)$, this algorithm can thus widely be used in our context. Its C++ implementation, in collaboration with Michel Koskas, from AgroParisTech, and Guillem Rigaille, from the university of Evry, has represented a topic of this thesis, as well as its distribution as an R package, `Segmentor3IsBack`, for several distributions including the negative binomial.

However, the algorithm has two major limitations, the first of which being its inability to assess the quality of the proposed segmentation. Indeed, we obtain the best segmentation in K segments, but it does not tell us about the second best. Is it radically different, or otherwise very close? We propose a first answer to this question by calculating, for each change-point, the cost of the best segmentation depending on its position. In cases where the resulting curves (examples are given in Figure 2.1) each have a clear minimum, we can have confidence in the optimal segmentation. Otherwise it means that the location of the breaks is uncertain. This is thus a rapid descriptive criterion for change-point quality assessment, but we aimed at developing statistical criteria to quantify this uncertainty.

The second limitation is that it does not propose to choose the number of segments K ; on the contrary, it obtains for each value of k between 1 and a user-specified K_{max} the optimal segmentation in k segments. This issue is a recurrent difficulty in segmentation frameworks.

Criteria for choosing the number of segments. In this context, the development of a model selection procedure, directly related to the inference issue (*ii*), that would complete the algorithm was a natural problem for which we propose two solutions in the following chapters. Indeed, the two families of approaches described in Section 1.2.3, the Birgé-

Massart approaches, based on non-asymptotic considerations on the risk associated to the likelihood criterion, and the classification-based approaches, seemed equally interesting in our context.

Thus in Chapter 2.2 we develop a penalized likelihood approach inspired by the Birgé and Massart literature (BIRGÉ and MASSART, 1997; BARRON *et al.*, 1999; BIRGÉ and MASSART, 2001, 2007) in which we seek to determine the form of the penalty in order to obtain an oracle inequality for our estimator of the distribution.

Our framework differs from previous approaches in that we are dealing with data that is both discrete and unbounded. Thus, traditional concentration inequalities cannot be applied directly and finer decomposition as well as large deviations results are required. In collaboration with Emilie Lebarbier, we showed that a penalty of the form

$$\text{pen}(m) = \beta|m| \left(1 + 4\sqrt{1.1 + \log\left(\frac{n}{|m|}\right)} \right)^2$$

satisfies our problem, and we discuss this result in Section 2.2. This approach (whose application is of complexity $\mathcal{O}(K)$) combined with the previous algorithm can thus provide a complete procedure (with final empirically linear complexity) to segment RNA-Seq data corresponding to whole chromosomes.

In Chapter 2.3 however, we are interested in the Integrated Completed Likelihood (ICL BIERNACKI *et al.*, 2000) as a criterion for the choice of K .

Our work draws on two major works in the field of segmentation which we detailed in Section 1.2.3. The first is the Bayesian segmentation model, published in RIGAILL *et al.* (2012) in which an exact computation of the ICL is proposed in quadratic time. The second is the constrained HMM approach proposed by LUONG *et al.* (2013). Our contribution consisted in implementing, in collaboration with The Minh Luong and Gregory Nuel, from the University of Paris Descartes and Guillem Rigaiill, from the University of Evry, this second algorithm for the negative binomial distribution, and propose a computation of the ICL criterion conditional on the model parameters. Using it to choose K requires its calculation for all values of k between 1 and a (user-defined) K_{max} and results in making the global process complexity in $\mathcal{O}(K^2n)$, thus limited to profiles of intermediate size.

However, the segmentation approach by constrained HMM provides more information than the dynamic programming algorithm as we can for example obtain confidence intervals for the location of change-points, still conditional on the parameter values. In our context of genome annotation, examples of such intermediate size series can be obtained by dividing a chromosome into two pieces at the centromere. Indeed, this region is known for not including coding regions, so that we don't expect any change-point to be located there. Now if we were to divide n and K by two, we would gain a factor 8 and 4 respectively in time and space.

We conclude these chapters with an illustration of the PDPA algorithm on our benchmark dataset in Chapter 2.4.

1.3.3 Gene annotation

In Chapter 3, our ultimate goal is the comparison of the location of transcribed regions of the genome of a species which has been grown in different environments. In the segmentation context, this problem is equivalent to comparing the change-point locations of independent profiles.

An algorithm for the computation of change-point location uncertainty. We are here typically in a framework where the complexity of the computational issue (*iii*) is less crucial than before, for instance the approaches presented in the previous sections, namely the exact Bayesian segmentation and the constrained HMM, can be applied. Indeed, here n is the order of 10^3 as we consider genes, and K is of the order of at most ten, since we want to separate the coding regions (*i.e.* exons) from non-coding regions (*i.e.* introns) within a same gene. However, dealing with change-point comparison requires the ability to compute quantities such as the uncertainty of their location. To this effect, we have proved that the negative binomial distribution verifies the requirements of the approach of RIGAILL *et al.* (2012) and have implemented it, as well as for diverse other distributions, in an R package called EBS (for Exact Bayesian Segmentation). We describe in Chapter 3.3 how this package operates, as well as the computational tricks we used.

Assessing the quality of the models. Because the goal is the comparison of independent segmentations, it seemed natural to check the relevance of our modeling (*i*) contributions beginning with a state of the art on segmentation methods that can take into account both the discrete nature of our data and the absence of a reference profile. We show (in Chapter 3.1), in collaboration with Sandrine Dudoit and Stéphane Robin, that algorithms are more effective when K is known, an assumption that is not absurd in contexts where we already have an approximate genome annotation that we seek to refine. The PDPA and EBS algorithms then have optimal performances, while the constrained HMM approach is faster than EBS but has slightly worse results and therefore does not represent a gain in this context of 'small data'. The other considered algorithms, which were only implemented for the Poisson distribution, failed to match any of our three approaches.

Methods for the comparison of change-point location. We have subsequently retained the exact Bayesian segmentation model to perform our location comparison (associated with the inference issue (*ii*)), and proposed, in collaboration with Stéphane Robin, two approaches which are presented in Chapter 3.2, the first one dedicated to the comparison of two profiles, while the other applies regardless of the number of profiles considered.

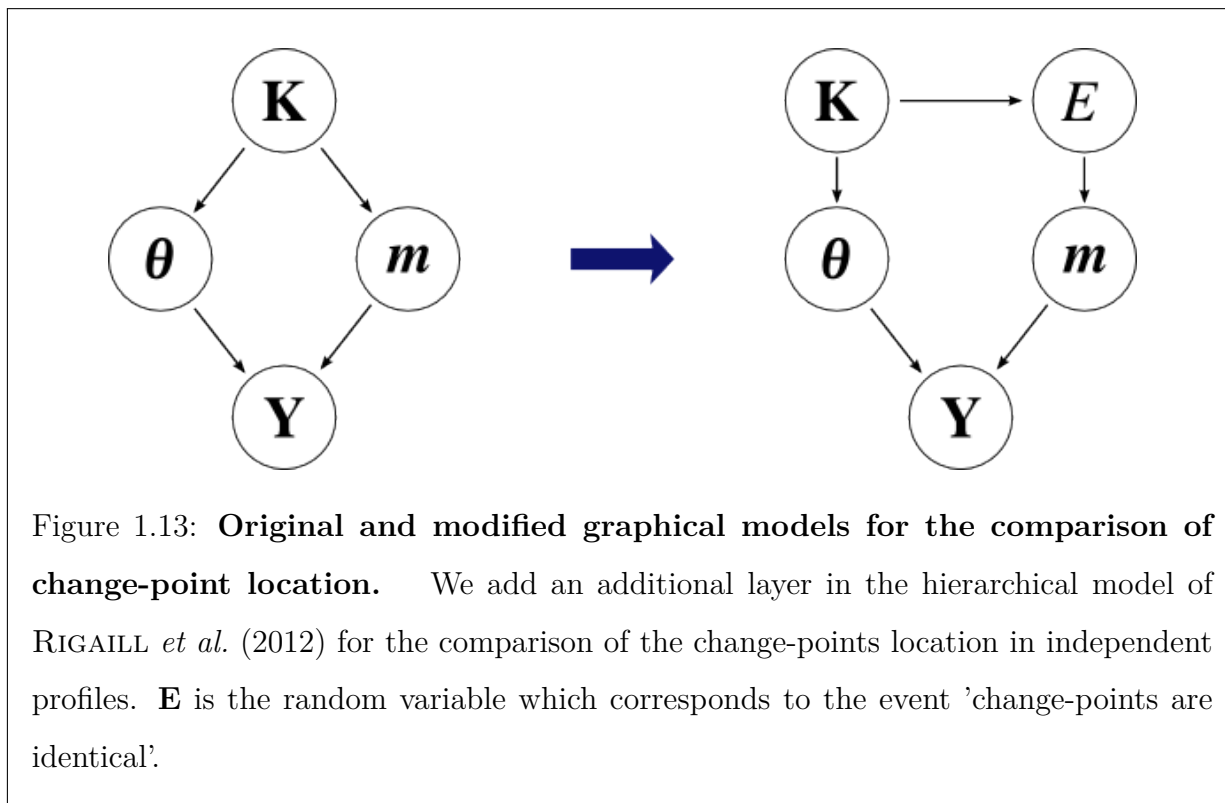
In the case of two profiles, the posterior distribution of the shift in locations can be computed by simple convolution as

$$\delta_{k_1, k_2}(d; K^1, K^2) = \sum_t p_{k_1}(t; Y^1; K^1) p_{k_2}(t - d; Y^2; K^2).$$

where $p_{k_\ell}(t, Y^\ell, K^\ell)$ is the posterior distribution of change-point τ_{k_ℓ} from the segmentation in K^ℓ segments of profile Y^ℓ .

This does not hold as soon as we have more than two series. It is then natural to compute the posterior probability of the event $E_0 = \{\tau_{k_1}^1 = \dots = \tau_{k_I}^I\}$ to decide on the equality of the change-points. This led us to introduce an additional layer in the graphical model as illustrated in Figure 1.13.

This new model allows to set the prior p_0 of event E_0 according to our expertise, and we



can then exactly compute the posterior probability of this event. Both frameworks provide natural decision rules for the equality of change-points in the profiles.

We return to our benchmark dataset in Chapter 3.4 in which we apply these rules to a subset of yeast genes with two exons. We illustrate the expected result which is that the boundaries of introns are not dependent on growth condition, while the beginning and end of transcription are subject to changes according to their environment.

Conclusion. We have developed in this thesis several segmentation methods for the general framework of genome annotation which we have illustrated on the same dataset throughout the manuscript. This has allowed us to highlight their richness when studying biological phenomena such as differential splicing. Table 1.3 summarizes the majority of our contribution. Depending on the depth of the analysis performed (from whole-genome to single gene), each of our three methods, namely the pruned Dynamic Programming algorithm with the non-asymptotic penalty, the constrained HMM with the ICL penalty, and the exact Bayesian approach, can be applied to determine the localization of the change-points

and assess their credibility. Moreover, they all meet the following three requirements:

- they are suitable for modeling count data, especially with the negative binomial, but can however be extended to many other distributions,
- they solve the criterion they seek to optimize in an exact manner, and
- they are implemented in R packages and freely available to the public.

In the next two chapters, the sections correspond to papers submitted during the PhD completed by some discussion. Depending on the papers, those discussions are either small remarks or illustrations (as in Section 3.1), or more global extensions of the work (as in Section 2.2).

Biological framework and <i>examples</i>	Max values and complexity	Computational (<i>iii</i>) package	Inference (<i>ii</i>)	uncertainty
Whole genome <i>e.g. expressed genes</i> <i>e.g. new transcripts</i>	$n: 10^8$ $\mathcal{O}(Kn)$ $K: 10^3$	pruned Dynamic Programming Segmentor3IsBack	optimal segmentation oracle inequality	qualitative
	$n: 10^5$ $\mathcal{O}(K^2n)$ $K: 10^2$	constrained HMM postCP	ICL	conditional
Genes <i>e.g. confident annotation</i> <i>e.g. profile comparison</i>	$n: 10^4$ $\mathcal{O}(Kn^2)$ $K: 10^1$	Exact Bayesian Segmentation EBS	optimal segmentation ICL decision rules	exact

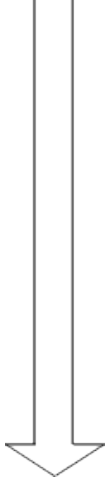


Table 1.3: **Overview of Thesis contribution.** Our contribution is organized by scale of the profiles (from whole genomes to single genes) for which we give potential biological applications and the tools developed for their analysis. For each of them, we recall their complexity and the maximum values of the parameters n (length of the data) and K (number of segments) and some examples of information provided by the methods.

Segmentation methods for whole genome analysis using RNA-Seq data

2.1	An efficient algorithm for the segmentation of RNA-Seq data	71
2.1.1	Background	73
2.1.2	Segmentation model and algorithm	75
2.1.3	Contribution	77
2.1.4	Results and discussion	82
2.2	Model selection	90
2.2.1	Introduction	91
2.2.2	Model Selection Procedure	94
2.2.3	Exponential bounds	99
2.2.4	Simulations and application	103
2.2.5	Proof of Theorem 2.2.3	107
2.2.6	Appendices	111
2.2.7	Discussion and perspectives	120
2.3	Constrained HMM approach	127
2.3.1	Introduction	128
2.3.2	Integrated Completed Likelihood criterion estimation using a constrained HMM	131
2.3.3	Validation	137
2.3.4	Discussion	142
2.4	Results on the yeast data-set	153

In this chapter we are interested in the segmentation of series corresponding to whole chromosomes. This question is usually introduced with one of two possible purposes: the discovery of new transcripts, which may arise when a region is labeled as transcribed while it had not previously been annotated as a gene, and the identification of genes specifically expressed in the target cells. In this context, we will address the modeling issue (*i*) by using the negative binomial distribution with segment-specific probability parameter p to fit the data. Still, the length of the data-set is expected to be very large (typically from 10^6 to 10^8 datapoints), so that the other two aspects, inference (*ii*) and computational (*iii*), remain difficult issues which are the at the heart of the contribution of this section.

2.1 An efficient algorithm for the segmentation of RNA-Seq data

Up to this contribution, no segmentation algorithm optimizing the likelihood criterion could allow modeling the data using the negative binomial distribution, and very few actually proposed to deal with count data-sets. In this section, we propose to adapt the pruned Dynamic Programming algorithm to our framework, assuming that the choice of K is not an issue.

Our procedure and proof that the negative binomial distribution fulfills the required conditions are described in a paper which is given (in its submitted version available at <http://arxiv.org/abs/1204.5564>) in the next paragraph.

One question remains open is that of this algorithm's complexity. In his initial paper, Rigaiil shows that it is at worst equivalent to that of the DP ($\mathcal{O}(Kn^2)$) and obtains, in a simulation study a complexity close to $\mathcal{O}(Kn \log n)$. In practice, our analysis on RNA-Seq data indicated a complexity faster than $n^{3/2}$ and slightly slower than linear in the size of our profiles. This is for instance illustrated in Figure 2.19 in the last section of this chapter, where we present the output of our whole procedure on each chromosome of the yeast species.

Segmentor3IsBack: an *R* package for the fast and exact segmentation of Seq-data

Alice Cleynen, Michel Koskas, Emilie Lebarbier, Guillem Rigail, Stéphane Robin

abstract

Background: Genome annotation is an important issue in biology which has long been addressed with gene prediction methods and manual experiments requiring biological expertise. The expanding Next Generation Sequencing technologies and their enhanced precision allow a new approach to the domain: the segmentation of RNA-Seq data to determine gene boundaries.

Results: Because of its almost linear complexity, we propose to use the Pruned Dynamic Programming Algorithm, which performances had been acknowledged for CGH arrays, for Seq-experiment outputs. This requires the adaptation of the algorithm to the negative binomial distribution with which we model the data. We show that if the dispersion in the signal is known, the PDP algorithm can be used and we provide an estimator for this dispersion. We then propose to estimate the number of segments, which can be associated to coding or non-coding regions of the genome, using an oracle penalty.

Conclusions: We illustrate the results of our approach on a real data-set and show its good performance. Our algorithm is available as an *R* package on the CRAN repository.

Keywords

segmentation algorithm; exact; fast; RNA-Seq data; count data

2.1.1 Background

Change-point detection methods have long been used in the analysis of genetic data, for instance they proved a useful tool in the study of DNA sequences with various purposes. BRAUN and MULLER (1998); DUROT *et al.* (2009) have developed segmentation methods for categorical variables with the aim of identifying patterns for gene predictions, while BOCKHORST and JOJIC (2012) uses the sequence segmentation for the detection of SNPs. In the last two decades, with the large spread of micro-arrays, change-point methods have been widely used for the analysis of DNA copy number variations and the identification of amplification or deletion of genomic regions in pathologies such as cancers (SHEN and ZHANG, 2012; ERDMAN and EMERSON, 2008; OLSHEN *et al.*, 2004; PICARD *et al.*, 2005, 2011).

The recent development of Next-Generation Sequencing technologies gives rise to new applications along with new difficulties: (a) the increased size of profiles (up to 10^8 data-points when micro-arrays signals were closer to 10^5), and (b) the discrete nature of the output (number of reads starting at each position of the genome). Yet applying segmentation methods to DNA-Seq data and its greater resolution should lead to the analysis of copy-number variation with a much improved precision than CGH arrays. Moreover, in the case of poly-(A) RNA-Seq data on lower organisms, since coding regions of the genome are well separated from non-coding regions with lower activity, segmentation methods should allow the identification of transcribed genes as well as address the issue of new transcript discovery. Our objective is therefore to develop a segmentation method to tackle both (a) and (b) with some specific requirements: the amount of reads falling in a segment should be representative of the biological information associated (relative copy-number of the region, relative level of expression of the gene) and comparison to neighboring regions should be sufficient to label the segment (for instance normal or deleted region of the chromosome in DNA-Seq data, exon or non-coding region in RNA-Seq), so that no comparison profile should be needed. This also suppresses the need for normalization, and thus we wish to analyze the raw count-profile.

So far, most methods addressing the analysis of these datasets require some normalization process to allow the use of algorithms relying on Gaussian-distributed data or previously developed for micro-arrays (CHIANG *et al.*, 2008; XIE and TAMMI, 2009; YOON *et al.*, 2009; BOEVA *et al.*, 2011). Indeed, methods adapted to count data-sets are not many, and highly focused on Poisson distribution. SHEN and ZHANG (2012) proposes a method based on the comparison of Poisson processes associated with the read counts of a case and a control sample, allowing for the detection of alteration of genomic sequences but not for expressed genes in a normal condition. RIVERA and WALTHER (2012) developed a likelihood ratio statistic for the localization of a shift in the intensity of a Poisson process while FRANKE *et al.* (2012) developed a test statistic for the existence of a change-point in the Poisson autoregression of order 1. Those two latter methods do not require a comparison profile but they only allow for the detection of a single change-point and have too high a time-complexity to be applied to RNA-Seq profiles. Binary Segmentation, a fast heuristic (OLSHEN *et al.*, 2004) and Pruned Exact Linear Time (PELT), (KILLICK *et al.*, 2012) an exact algorithm for optimal segmentation with respect to the likelihood, are both implemented for the Poisson distribution in package **changepoint**. Even though both are extremely fast, do not require a comparison profile and analyse count-data, the Poisson distribution is ill-adapted to our kind of data-sets.

A recent study of HOCKING *et al.* (2013) has compared 13 segmentation methods for the analysis of chromosomal copy number profiles and has shown the excellent performances of the Pruned Dynamic Programming (PDP) algorithm proposed by RIGAILL (2010) in its initial implementation for the analysis of Gaussian data in the *R* package **cghseg**. We propose to use the PDP algorithm which we have implemented for the Poisson and negative binomial distributions.

In the next section we recall the general segmentation framework and the definition and requirements of the PDP algorithm. Our contributions are given in the third section where we define the negative binomial model and show that it satisfies the PDP algorithm requirements. We also give a model selection criterion with theoretical guarantees, which makes the whole approach complete. We conclude with a simulation study, which illustrates

the performances of the proposed method.

2.1.2 Segmentation model and algorithm

General segmentation model

The general segmentation problem consists in partitioning a signal of n data-points $\{y_t\}_{t \in [1, n]}$ into K pieces or segments. The model can be written as follows: the observed data $\{y_t\}_{t=1, \dots, n}$ are supposed to be a realization of an independent random process $Y = \{Y_t\}_{t=1, \dots, n}$. This process is drawn from a probability distribution \mathcal{G} which depends on a set of parameters among which one parameter θ is assumed to be affected by $K - 1$ abrupt changes, called change-points, so that

$$Y_t \sim \mathcal{G}(\theta_r, \phi) \quad \text{if } t \in r \quad \text{and} \quad r \in m$$

where m is a partition of $[1, n]$ into segments r , θ_r stands for the parameter of segment r and ϕ is constant. The objective is to estimate the change-points or the positions of the segments and the parameters θ_r both resulting from the segmentation. More precisely, we define $\mathcal{M}_{k,t}$ the set of all possible partitions in $k > 0$ regions of the sequence up to point t . We recall that the number of possible partitions is

$$\text{card}(\mathcal{M}_{K,t}) = \binom{t-1}{K-1}.$$

We aim at choosing the partition in $\mathcal{M}_{K,n}$ of minimal loss γ , where the loss is usually taken as the negative log-likelihood of the model. We define the loss of a segment with given parameter θ as $c(r, \theta) = \sum_{i \in r} \gamma(y_i, \theta)$, so its optimal cost is $c(r) = \min_{\theta} \{c(r, \theta)\}$. This allows us to define the cost of a segmentation m as $\sum_{r \in m} c(r)$ and our goal is to recover the optimal segmentation $M_{K,n}$ and its cost $C_{K,n}$ where :

$$M_{k,t} = \arg \min_{\{m \in \mathcal{M}_{k,t}\}} \left\{ \sum_{r \in m} c(r) \right\}$$

and $C_{k,t} = \min_{\{m \in \mathcal{M}_{k,t}\}} \left\{ \sum_{r \in m} c(r) \right\}.$

Quick overview of the pruned DPA

The pruned DPA relies on the function $H_{k,t}(\theta)$ which is the cost of the best partition in k regions up to t , the parameter of the last segment being θ :

$$H_{k,t}(\theta) = \min_{k-1 \leq \tau \leq t} \{ C_{k-1,\tau} + c([\tau + 1, t], \theta) \},$$

and from there gets $C_{k,t}$ as $\min_{\theta} \{ H_{k,t}(\theta) \}$. More precisely, for each total number of regions, k , from 2 to K , the pruned DPA works on a list of last change-point candidates: $ListCandidate_k$. For each of these candidate change-points, τ , the algorithm stores a cost function and a set of optimal-cost intervals. To be more specific, we define:

- $H_{k,t}^{\tau}(\theta) = C_{k,\tau} + \sum_{j=\tau+1}^t \gamma(y_j, \theta)$: the optimal cost if the last change is τ ;
- $S_{k,t}^{\tau} = \{ \theta \mid H_{k,t}^{\tau}(\theta) \leq H_{k,t}(\theta) \}$: the set of θ such that τ is optimal;
- $I_{k,t}^{\tau} = \{ \theta \mid H_{k,n}^{\tau}(\theta) \leq H_{k,n}^t(\theta) \}$: the set of θ such that τ is better than t in terms of cost, with $\tau < t$.

We have $H_{k,t}(\theta) = \min_{\tau \leq t} \{ H_{k,t}^{\tau}(\theta) \}$.

The PDP algorithm rely on four basic properties of these quantities:

- (i) if all $\sum_{j=\tau+1}^{t+1} \gamma(y_j, \theta)$ are unimodal in θ then $I_{k,t}^{\tau}$ are intervals;
- (ii) $H_{k,t+1}^{\tau}(\theta)$ is obtained from $H_{k,t}^{\tau}(\theta)$ using:

$$H_{k,t+1}^{\tau}(\theta) = H_{k,t}^{\tau}(\theta) + \gamma(y_{t+1}, \theta);$$

- (iii) it is easy to update $S_{k,t+1}^{\tau}$ using:

$$\begin{aligned} S_{k,t+1}^{\tau} &= S_{k,t}^{\tau} \cap I_{k,t+1}^{\tau} \\ S_{k,t}^t &= \mathbb{G}_{\mathbb{R}}(\cup_{\tau \in \llbracket k-1, t-1 \rrbracket} I_{k,t}^{\tau}); \end{aligned}$$

- (iv) once it has been determined that $S_{k,t}^{\tau}$ is empty, the region-border τ can be discarded from the list of candidates $ListCandidate_k$:

$$S_{k,t}^{\tau} = \emptyset \Rightarrow \forall t' \geq t \quad S_{k,t'}^{\tau} = \emptyset.$$

Requirements of the pruned dynamic programming algorithm.

Proposition 2.1.1. *Properties (i) to (iv) are satisfied as soon as the following conditions on the loss $c(r, \theta)$ are met:*

- (a) *it is point additive,*
- (b) *it is convex with respect to its parameter θ ,*
- (c) *it can be stored and updated efficiently.*

It is possible to include an additional penalty term in the loss function. For example, in the case of RNA-seq data one could add a lasso ($\lambda|\theta|$) or ridge penalty ($\lambda\theta^2$) to encode that a priori the coverage in most regions should be close to 0. Our C++ implementation of the pruned DPA includes the possibility to add such a penalty term, however we do not provide an R interface to this functionality in our R package. One of the reason for this choice is that choosing an appropriate value for λ is not a simple problem.

2.1.3 Contribution

Pruned dynamic programming algorithm for count data

We now show that the PDP algorithm can be applied to the segmentation of RNA-Seq data using a negative binomial model, and propose a criterion for the choice of K . Though not discussed here, our results also hold for the Poisson segmentation model.

Negative binomial model. We consider that in each segment r all y_t are the realization of random variables Y_t which are independent and follow the same negative binomial distribution. Assuming the dispersion parameter ϕ to be known, we will use the natural parametrization from the exponential family (also classically used in R) so that parameter θ_r will be the probability of success. In this framework, θ_r is specific to segment r whereas ϕ is common to all segments.

We have $E(Y_t) = \phi(1 - \theta)/\theta$ and $Var(Y_t) = \phi(1 - \theta)/\theta^2$. We choose the loss as the negative log-likelihood associated to data-point t belonging to segment r : $-\phi \log(\theta_r) - y_t \log(1 - \theta_r) + A(\phi, y_t)$, or more simply $\gamma(y_t, \theta_r) = -\phi \log(\theta_r) - y_t \log(1 - \theta_r)$ since A is a function that does not depend on θ_r .

Validity of the pruned dynamic programming algorithm for the negative binomial model

Proposition 2.1.2. *Assuming parameter ϕ to be known, the negative binomial model satisfies (a), (b) and (c):*

- (a) As we assume that Y_t are independent we indeed have that the loss is point additive :

$$c(r, \theta) = \sum_{t \in r} \gamma(y_t, \theta).$$
- (b) As $\gamma(y_t, \theta) = -\phi \log(\theta) - y_t \log(1 - \theta)$ is convex with respect to θ , $c(r, \theta)$ is also convex as the sum of convex functions.
- (c) Finally, we have $c(r, \theta) = -n_r \phi \log(\theta) + \sum_{t \in r} y_t \log(1 - \theta)$. This function can be stored and updated using only two doubles: one for $-n_r \phi$, and the other for $\sum_{t \in r} y_t$.

Estimation of the overdispersion parameter. We propose to estimate ϕ using a modified version of the estimator proposed by JOHNSON *et al.* (2005): compute the moment estimator of ϕ on each sliding window of size h using the formulae $\phi = \mathbb{E}(Y)^2 / (Var(Y) - \mathbb{E}(Y))$ and keep the median $\hat{\phi}$.

C++ implementation of the pruned DPA

We implemented the pruned DPA in C++ with in mind the possibility of adding new loss functions in potential future applications. The difficulties we had to come through were the versatility of the program to design and the design of the objects it could work on. Indeed, the use of full templates implied that we used stable sets of objects for the operations that were to be performed on.

Namely:

- The sets were to be chosen in a *tribe*. This means that they all belong to a set \mathcal{S} of sets such that any set $s \in \mathcal{S}$ can be conveniently handled and stored into the computer.

A set of sets \mathcal{S} is said *acceptable* if it satisfies:

1. if s belongs to \mathcal{S} , $\mathbb{R} \setminus s \in \mathcal{S}$
 2. if $s_1, s_2 \in \mathcal{S}$, $s_1 \cap s_2 \in \mathcal{S}$
 3. if $s_1, s_2 \in \mathcal{S}$, $s_1 \cup s_2 \in \mathcal{S}$
- The cost functions were chosen in a set \mathcal{F} such that
 1. each function may be conveniently handled and stored by the software
 2. for any $f \in \mathcal{F}$, $f(x) = 0$ can be easily solved and the set of solutions belongs to an acceptable set of sets
 3. for any $f \in \mathcal{F}$ and any constant c , $f(x) \leq c$ can be easily solved and the set of solutions belongs to an acceptable set of sets
 4. for any $f, g \in \mathcal{F}$, $f + g \in \mathcal{F}$.

Thus we defined two collections for the sets of sets, intervals and parallelepipeds, and implemented the loss functions corresponding to negative binomial, Poisson or normal distributions. The program is thus designed in a way that any user can add his own cost function or acceptable set of probability function and use it without rewriting a line in the code.

Model Selection

The last issue concerns the estimate of the number of segments K . This model selection issue can be solved using penalized log-likelihood criterion where the choice of a good penalty function is crucial. This kind of procedure requires the visit of the optimal segmentations in $k = 1, \dots, K_{\max}$ segments where K_{\max} is generally chosen smaller than n . The most popular criteria (AIC, AKAIKE (1973) and BIC, YAO (1988)) failed in the segmentation context due to the discrete nature of the segmentation parameter. In a non-asymptotic point of view and for the negative binomial model, CLEYNEN and LEBARBIER

(2013) proposed to choose the number of segments as follows: denoting \hat{m}_K the optimal segmentation of the data in K segments,

$$\hat{K} = \arg \min_{K \in 1:K_{\max}} \left\{ \sum_{r \in \hat{m}_K} \sum_{t \in r} \left[-\phi \log \frac{\phi}{\phi + \bar{y}_r} - Y_t \log \left(1 - \frac{\phi}{\phi + \bar{y}_r} \right) \right] + \beta K \left(1 + 4 \sqrt{1.1 + \log \left(\frac{n}{K} \right)} \right)^2 \right\}, \quad (2.1)$$

where $\bar{y}_r = \frac{\sum_{t \in r} y_t}{\hat{n}_r}$ and \hat{n}_r is the size of segment r . The first term corresponds to the cost of the optimal segmentation while the second is a penalty term which depends on the dimension K and of a constant β that has to be tuned according to the data (see the next section). With this choice of penalty, so-called oracle penalty, the resulting estimator satisfies an oracle-type inequality. A more complete performance study is done in CLEYNEN and LEBARBIER (2013) and showed that the proposed criterion outperforms the existing ones.

R package

The Pruned Dynamic Programming algorithm is available in the function `Segmentor` of the *R* package **Segmentor3IsBack**. The user can choose the distribution with the slot `model` (1 for Poisson, 2 for Gaussian homoscedastic, 3 for negative binomial and 4 for segmentation of the variance). It returns an S4 object of class `Segmentor` which can later be processed for other purposes. The function `SelectModel` provides four criteria for choosing the optimal number of segments: AIC (AKAIKE, 1973), BIC (YAO, 1988), the modified BIC (ZHANG and SIEGMUND, 2007) (available for Gaussian and Poisson distribution) and oracle penalties (available for the Gaussian distribution (LEBARBIER, 2005) and for the Poisson and negative binomial (CLEYNEN and LEBARBIER, 2013) as described previously). This latter kind of penalties require tuning a constant according to the data, which is done using the slope heuristic (ARLOT and MASSART, 2009).

Figure 2.4 (which is detailed in the Results and discussion section) was obtained with the following 4 lines of code (assuming the data was contained in vector \mathbf{x}):

```
Seg <- Segmentor(x, model=3, Kmax=200)
```

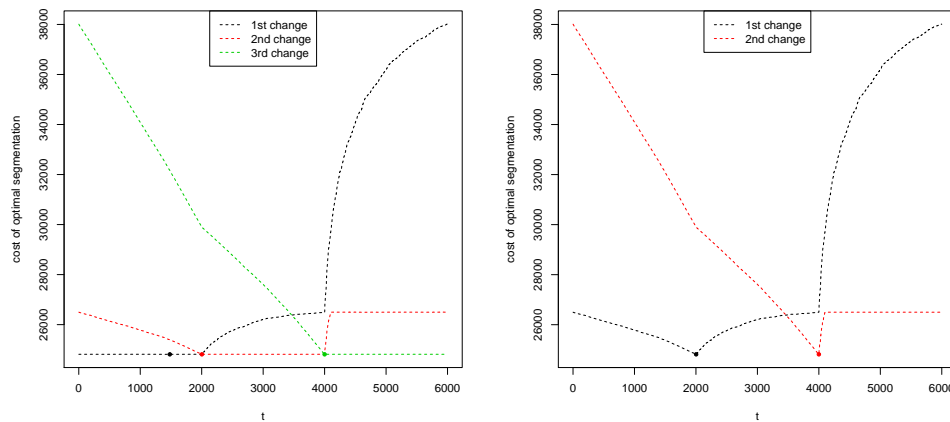


Figure 2.1: **Cost of optimal segmentation in 4 and 3 segments.** Cost of optimal segmentation depending on the location of the j^{th} change-point when the number of segments is 4 (right) and 3 (left) and the signal was simulated with 3 segments. Illustration of the output of function `BestSegmentation`.

```
Kchoose<-SelectModel(Seg, penalty="oracle")
plot(sqrt(x),col='dark red')
abline(v=getBreaks(Seg)[Kchoose, 1:Kchoose],col='blue')
```

The function `BestSegmentation` allows, for a given K , to find the optimal segmentation with a change-point at location t (slot `$bestSeg`). It also provides, through the slot `$bestCost`, the cost of the optimal segmentation with t for j^{th} change-point. The right side of Figure 2.1 illustrates this result for the optimal segmentations in 4 segments of a signal simulated with only 3 segments. We can see for instance that any choice of first change-point location between 1 and 2000 yields almost the same cost (the minimum is obtained for $t = 1481$), thus the optimal segmentation is not clearly better than the second or third. On the contrary, the same function with 3 segments shows that the optimal segmentation outperforms all other segmentations in 3 segments (left side of Figure 2.1).

2.1.4 Results and discussion

Performance study

We designed a simulation study on the negative binomial distribution to assess the performance of the PDP algorithm in terms of computational efficiency, while studying the impact of the overdispersion parameter ϕ by comparing the results for two different values of this parameter. After running different estimators (median on sliding windows of maximum, quasi-maximum likelihood and moment estimators) on several real RNA-Seq data (whole chromosome and genes of various sizes) we fixed $\phi_1 = 0.3$ as a typical value for highly dispersed data as observed in real RNA-Seq data, and chose $\phi_2 = 2.3$ for comparison with a reasonably dispersed data-set. For each value, we simulated data-sets of size n with various densities of number of segments K , and only two possible values for the parameter p_j : 0.8 on even segments (corresponding to low signal) and 0.2 on odd segments for a higher signal. We had n vary on a logarithmic scale between 10^3 and 10^6 and K between $\sqrt{n}/6$ and $\sqrt{n}/3$. For each configuration, we segmented the signal up to $K_{\max} = \sqrt{n}$ twice: once with the known value of ϕ and once with our estimator $\hat{\phi}$ as described above. We started with a window width $h = 15$. When the estimate was negative, we doubled h and repeated the experience until the median is positive.

Each configuration was simulated 100 times.

For our analysis we checked the run-time on a standard laptop, and assessed the quality of the segmentation using the Rand Index \mathcal{I} . Specifically, let C_t be the true index of the segment to which base t belongs and let \hat{C}_t be the index estimated by the method, then

$$\mathcal{I} = \frac{2 \sum_{t>s} [\mathbf{1}_{C_t=C_s} \mathbf{1}_{\hat{C}_t=\hat{C}_s} + \mathbf{1}_{C_t \neq C_s} \mathbf{1}_{\hat{C}_t \neq \hat{C}_s}]}{(n-1)(n-2)}.$$

Figure 2.2 shows, for the particular case of $K = \sqrt{n}/3$, the almost linear complexity of the algorithm in the size n of the signal. As the maximal number of segments K_{\max} considered increased with n , we normalized the run-time to allow comparison. This underlines an empirical complexity smaller than $\mathcal{O}(K_{\max} n \log n)$, and independent on the value of ϕ

or its knowledge. Moreover, the algorithm, and therefore the pruning, is faster when the overdispersion is high, phenomenon already encountered with the L^2 loss when the distribution of the errors is Cauchy. However, the knowledge of ϕ does not affect the run-time of the algorithm. Figure 2.3 illustrates through the Rand Index the quality of the proposed segmentation for a few values of n . Even though the indexes are slightly lower for ϕ_1 than for ϕ_2 (see left panel), they range between 0.94 and 1 showing a great quality in the results. Moreover, the knowledge of ϕ does not increase the quality (see right panel), which validates the use of our estimator.

Yeast RNAseq experiment

We applied our algorithm to the segmentation of chromosome 1 of the *S. Cerevisiae* (yeast) using RNA-Seq data from the Sherlock Laboratory at Stanford University (Risso *et al.*, 2011) and publicly available from the NCBI's Sequence Read Archive (SRA, <http://www.ncbi.nlm.nih.gov/sra>, accession number SRA048710). We selected the number of segments using our oracle penalty described in the previous section. An existing annotation is available on the Saccharomyces Genome Database (SGD) at <http://www.yeastgenome.org>, which allows us to validate our results.

With a run-time of 25 minutes (for a signal length of 230218), we selected 103 segments with the negative binomial distribution, most of which (all but 3) were found to surround known genes from the SGD. Figure 2.4 illustrates the result.

Conclusion

Segmentation has been a useful tool for the analysis of biological data-sets for a few decades. We propose to extend its application with the use of the Pruned Dynamic Programming algorithm for count data-sets such as outputs of sequencing experiments. We show that the negative binomial distribution can be used to model such data-sets on the

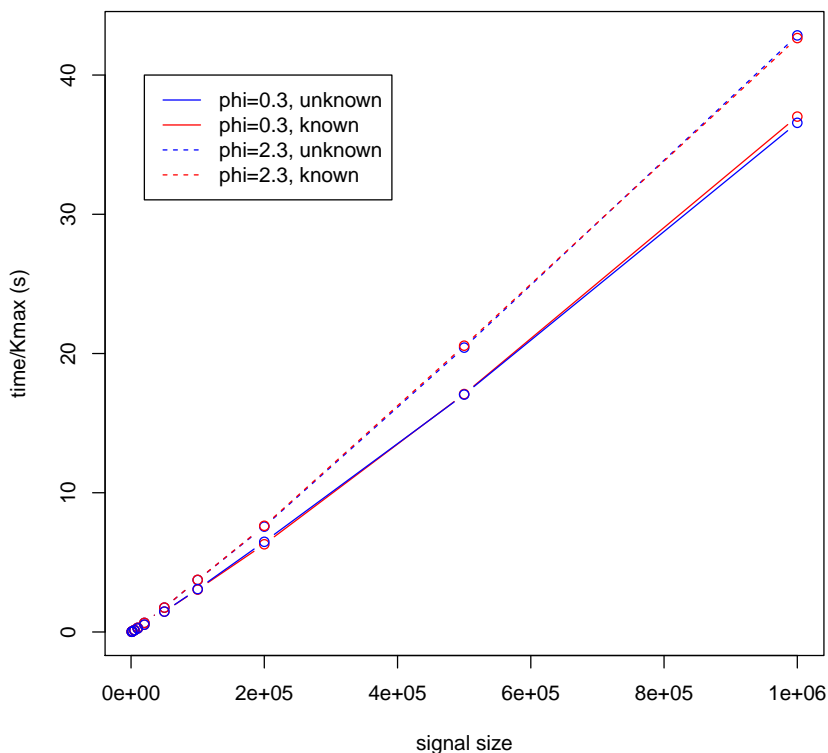


Figure 2.2: **Run-time analysis for segmentation with negative binomial distribution.** This figure displays the normalized (by K_{\max}) run-time in seconds of the **Segmentor3IsBack** package for the segmentation of signals with increasing length n , for two values of the dispersion ϕ , and with separate analysis when its value is known or estimated. While the algorithm is faster for more over-dispersed data, the estimation of the parameter does not slow the processing.

condition that the overdispersion parameter is known, and proposed an estimator of this parameter that performs well in our segmentation framework.

We propose to choose the number of segments using our oracle penalty criterion, which makes the package fully operational. This package also allows the use of other criteria such as AIC or BIC. Similarly, the algorithm is not restricted to the negative binomial distribution but also allows the use of Poisson and Gaussian losses for instance, and could

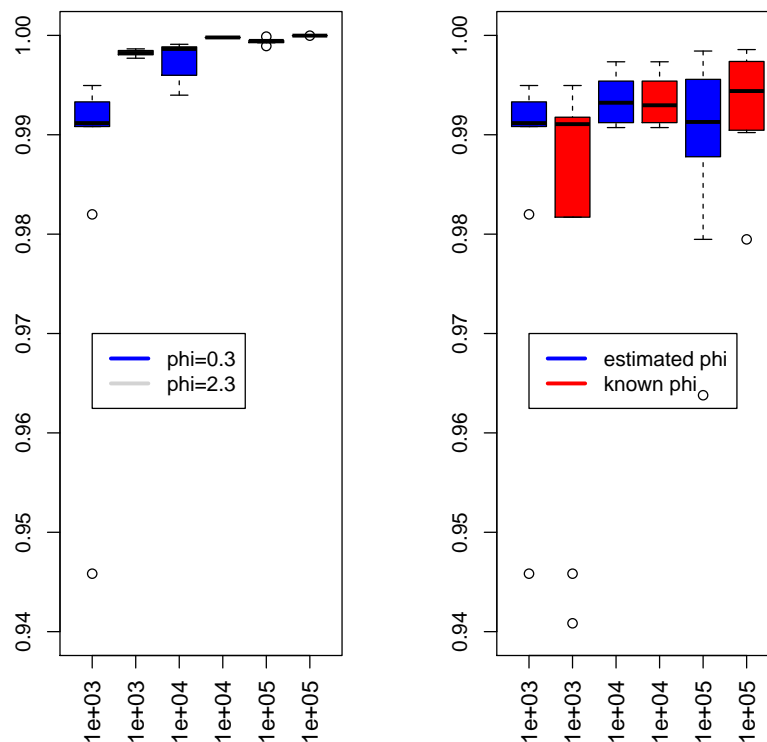
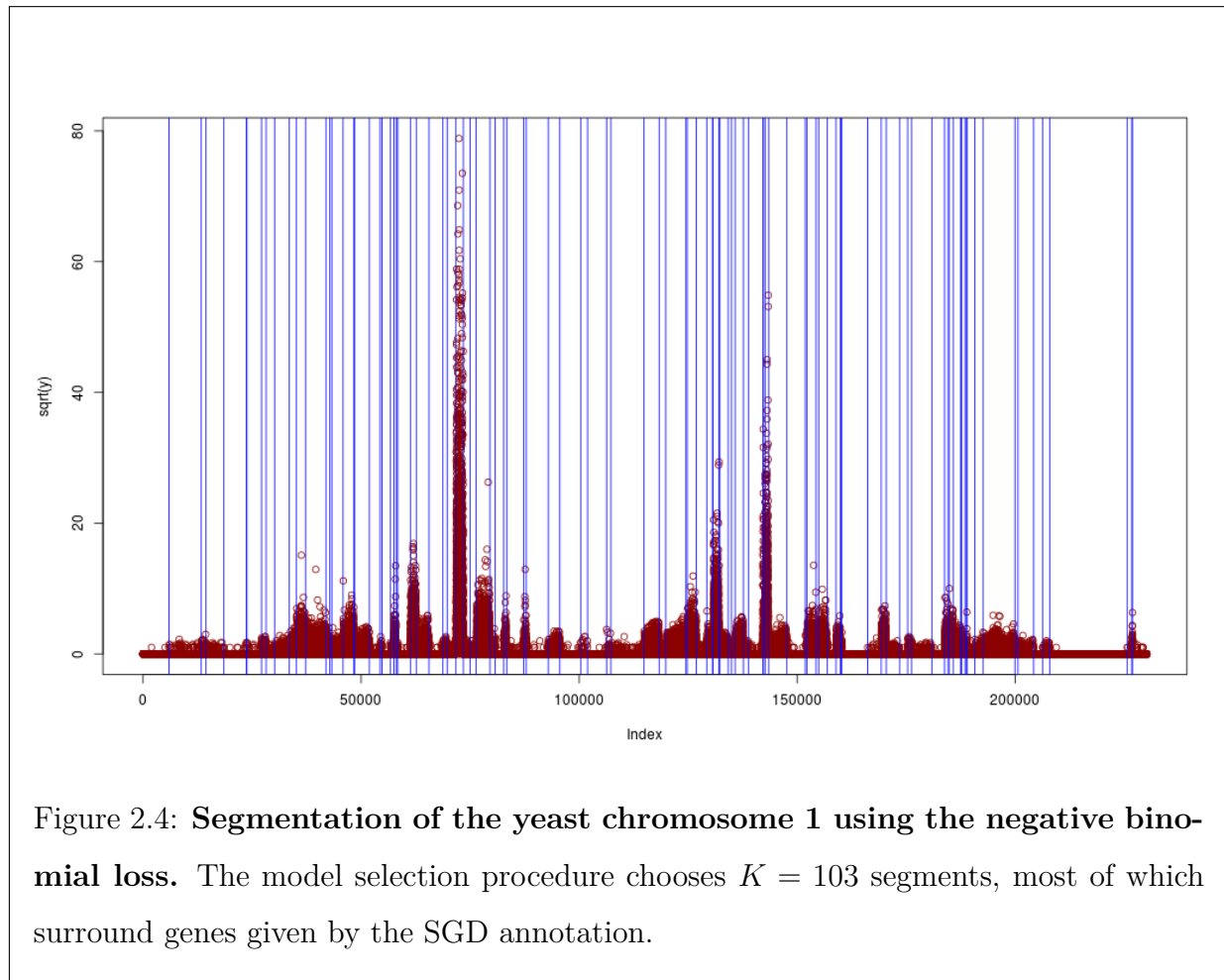


Figure 2.3: **Rand Index for the quality of the segmentation.** This figure displays the boxplot of the Rand Index computed for each of the hundred simulations performed in the following situations: comparing the values with ϕ_1 and ϕ_2 when estimated (left figure), and comparing the impact of estimating ϕ_1 (right figure). While the estimation does not decrease the quality of the segmentation, the value of the dispersion affects the recovery of the true change-points.

easily be adapted to other convex one-parameter losses.

With its empirical complexity of $\mathcal{O}(K_{\max}n \log n)$, it can be applied to large signals such as read-alignment of whole chromosomes, and we illustrated its result on a real-data sets from the yeast genomes. Moreover, this algorithm can be used as a base for further analysis. For example, LUONG *et al.* (2013) use it to initialize their Hidden Markov Model to compute change-point location probabilities.



Availability and requirements

- Project name: Segmentor3IsBack
- Project home page: <http://cran.r-project.org/web/packages/Segmentor3IsBack/index.html>
- Operating systems: Platform independent
- Programming language: C++ code embedded in *R* package
- License: GNU GPL
- Any restrictions to use by non-academics: none

List of abbreviations used

- PELT: Pruned Exact Linear Time
- PDP: Pruned Dynamic Programming

- AIC: Akaike Information Criterion
- BIC: Bayesian Information Criterion
- NCBI: National Center for Biotechnology Information
- SGD: Saccharomyces Genome Database

Competing interests The authors have no competing interest to declare.

Author’s contributions AC co-wrote the C++ code, wrote the R-package, performed data analysis and co-wrote the manuscript. MK co-wrote the C++ code. EL co-supervised the work and co-wrote the manuscript. GR co-wrote the C++ code, and co-wrote the manuscript. SR co-wrote the manuscript and co-supervised the work.

References

- H. Akaike. Information theory and extension of the maximum likelihood principle. *Second international symposium on information theory*, pages 267–281, 1973.
- Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *The Journal of Machine Learning Research*, 10:245–279, 2009.
- Joseph Bockhorst and Nebojsa Jojic. Discovering patterns in biological sequences by optimal segmentation. *arXiv preprint arXiv:1206.5256*, 2012.
- Valentina Boeva, Andrei Zinovyev, Kevin Bleakley, Jean-Philippe Vert, Isabelle Janoueix-Lerosey, Olivier Delattre, and Emmanuel Barillot. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics*, 27 (2):268–269, 2011.
- Jerome V Braun and Hans-Georg Muller. Statistical methods for DNA sequence segmentation. *Statistical Science*, pages 142–162, 1998.
- Derek Y Chiang, Gad Getz, David B Jaffe, Michael JT O’Kelly, Xiaojun Zhao, Scott L Carter, Carsten Russ, Chad Nusbaum, Matthew Meyerson, and Eric S Lander. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature methods*, 6(1):99–103, 2008.
- Alice Cleynen and Emilie Lebarbier. Segmentation of the Poisson and negative binomial rate models: a penalized estimator. *arXiv preprint arXiv:1301.2534*, 2013.
- C. Durot, E. Lebarbier, and AS Tocquet. Estimating the joint distribution of independent categorical variables via model selection. *Bernoulli*, 15(2):475–507, 2009.

Chandra Erdman and John W Emerson. A fast Bayesian change point analysis for the segmentation of microarray data. *Bioinformatics*, 24(19):2143–2148, 2008.

Jurgen Franke, Claudia Kirch, and Joseph Tadjuidje Kamgaing. Changepoints in times series of counts. *Journal of Time Series Analysis*, 33(5):757–770, 2012.

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.

N. Johnson, A.W. Kemp, and S. Kotz. Univariate discrete distributions. John Wiley & Sons, Inc., 2005.

Rebecca Killick, Paul Fearnhead, and IA Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

E. Lebarbier. Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, 85(4):717–736, April 2005. ISSN 0165-1684.

T.M. Luong, Y. Rozenholc, and G. Nuel. Fast estimation of posterior probabilities in change-point models through a constrained hidden Markov model. *Arxiv preprint arXiv:1203.4394*, 2012.

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.

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.

Franck Picard, Emilie Lebarbier, Mark Hoebeke, Guillem Rigaiil, Baba Thiam, and Stéphane Robin. Joint segmentation, calling, and normalization of multiple CGH profiles. *Biostatistics*, 12(3):413–428, 2011.

G. Rigaiil. Pruned dynamic programming for optimal multiple change-point detection. *Arxiv:1004.0887*, April 2010. URL <http://arxiv.org/abs/1004.0887>.

Davide Risso, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, 12(1):480, 2011.

Camilo Rivera and Guenther Walther. Optimal detection of a jump in the intensity of a Poisson process or in a density with likelihood ratio statistics. *arXiv preprint arXiv:1211.2859*, 2012.

Jeremy J Shen and Nancy R Zhang. Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next-generation DNA sequencing. *The Annals of Applied Statistics*, 6(2):476–496, 2012.

Chao Xie and Martti Tammi. CNV-seq, a new method to detect copy number variation

using high-throughput sequencing. *BMC bioinformatics*, 10(1):80, 2009.

Y.-C. Yao. Estimating the number of change-points via schwarz' criterion. *Statistics & Probability Letters*, 6(3):181–189, February 1988.

Seungtai Yoon, Zhenyu Xuan, Vladimir Makarov, Kenny Ye, and Jonathan Sebat. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome research*, 19(9):1586–1592, 2009.

Nancy R Zhang and David O Siegmund. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32, 2007.

2.2 Model selection

We are now interested in the choice of the best number of segments K . This model selection problem is one of the most difficult issue of all segmentation approaches, and there are multiple likelihood-based methods to address it, most of which are adapted to one specific context or dataset type. However, up to this contribution, no proposition had yet been made for the analysis of count (and possibly unbounded) data.

In the next paragraph, we propose, in a joint work with Emilie Lebarbier, a first approach based on the Birgé and Massart literature (see Section 1.2.3). This paragraph differs slightly from the submitted paper (available at <http://arxiv.org/abs/1301.2534>) in that we corrected two errors which we identified while working on a generalization of this approach, itself presented in Section 2.2.7.

Segmentation of the Poisson and negative binomial rate models: a penalized estimator

Alice Cleynen and Emilie Lebarbier

abstract

We consider the segmentation problem of Poisson and negative binomial (i.e. overdispersed Poisson) rate distributions. In segmentation, an important issue remains the choice of the number of segments. To this end, we propose a penalized log-likelihood estimator where the penalty function is constructed in a non-asymptotic context following the works of L. Birgé and P. Massart. The resulting estimator is proved to satisfy an oracle inequality. The performances of our criterion is assessed using simulated and real datasets in the RNA-seq data analysis context.

Keywords

Distribution estimation; Change-points detection; Count data (RNA-seq); Poisson and negative binomial distributions; Model selection.

2.2.1 Introduction

We consider a multiple change-point detection setting for count datasets, which can be written as follows: we observe a finite sequence $\{y_t\}_{t \in \{1, \dots, n\}}$ realisation of independent variables Y_t . These variables are supposed to be drawn from a probability distribution \mathcal{G} which depends on a set of parameters. Here two types of parameters are distinguished:

$$Y_t \sim \mathcal{G}(\theta_t, \phi) = s(t), \quad 1 \leq t \leq n,$$

where ϕ is a constant parameter while the θ s are point-specific. In many contexts, we might want to consider that the θ s are piece-wise constant and so subject to an unknown

number $K - 1$ of abrupt changes (for instance with climatic or financial data). Thus, we want to assume the existence of partition of $\{1, \dots, n\}$ into K segments within which the observations follow the same distribution and between which observations have different distributions, i.e. θ is constant within a segment and differ from a segment to another. A motivating example is sequencing data analysis. For instance, the output of RNA-seq experiments is the number of reads (i.e. short portions of the genome) which first position maps to each location of a genome of reference. Supposing that we dispose of such a sequence, we expect to observe a stationarity in the amount of reads falling in different areas of the genome: expressed genes, intronic regions, etc. We wish to localize those regions that are biologically significant. In our context, we consider for \mathcal{G} the Poisson and negative binomial distributions, adapted to RNA-seq experiment analysis (RISSE *et al.*, 2011).

Change-point detection problems are not new and many methods have been proposed in the literature. For count data-sets, BRAUN and MULLER (1998) provide a detailed bibliography of methods in the particular case of the segmentation of the DNA sequences that includes Bayesian approaches, scan statistics, likelihood-ratio tests, binary segmentation and numerous other methods such as penalized contrast estimation procedures. In a Bayesian framework, BIERNACKI *et al.* (2000) proposes to use an exact "ICL" criterion for the choice of K , while its approximation is computed in the constrained HMM approach of LUONG *et al.* (2013). In this paper, we consider a penalized contrast estimation method which consists first, for every fixed K , in finding the best segmentation in K segments by minimizing the contrast over all the partitions with K segments, and then in selecting a convenient number of segments K by penalizing the contrast. Choosing the number of segments, i.e. choosing a "good" penalty, is a crucial issue and not so easy. The most basic examples of penalty are the Akaike Information Criterion (AIC, AKAIKE, 1973) and the Bayes Information Criterion (BIC, YAO, 1988) but these criteria are not well adapted in the segmentation context and tend to overestimate the number of change-points (see BIRGÉ and MASSART (2007); ZHANG and SIEGMUND (2007) for theoretical explanations). In this particular context, some modified versions of these criteria have been proposed. For instance, ZHANG and SIEGMUND (2007); BRAUN *et al.* (2000) have proposed modified versions of the BIC criterion (shown to be consistent) in the segmentation of Gaussian pro-

cesses and DNA sequences respectively. However, these criteria are based on asymptotic considerations. In the last years there has been an extensive literature influenced by BIRGÉ and MASSART (1997); BARRON *et al.* (1999) introducing non-asymptotic model selection procedures, in the sense that the size of the models as well as the size of the list of models are allowed to be large when n is large. This penalized contrast procedure consists in selecting a model amongst a collection such that its performance is as close as possible to that of the best but unreachable model in terms of risk. This approach has been now considered in various function estimation contexts. In particular, AKAKPO (2011) proposed a penalty for estimating the density of independent categorical variables in a least-squares framework, while REYNAUD-BOURET (2003); BIRGÉ (2007), or BARAUD and BIRGÉ (2009), focused on the estimation of the density of a Poisson process.

When the number of models is large, as in the case of an exhaustive search in segmentation problem, it can be shown that penalties which only depend on the number of parameters of each model, as for the classical criteria, are theoretically (and also practically) not adapted. This was suggested by LEBARBIER (2005) and BIRGÉ and MASSART (2007) who show that the penalty term needs to be well defined, and in particular needs to depend on the complexity of the list of models, i.e. the number of models having the same dimension. For this reason, following the work of BIRGÉ and MASSART (1997) and in particular CASTELLAN (1999) in the density estimation framework, we consider a penalized log-likelihood procedure to estimate the true distribution s of a Poisson or negative binomial-distributed sequence \mathbf{y} . We prove that, up to a $\log n$ factor, the resulting estimator satisfies an oracle inequality.

The paper is organized as follows. The general framework is described in Section 2.2.2. More precisely, we present our proposed penalized maximum-likelihood estimator, the form of the penalty and give some non-asymptotic risk bounds for the resulting estimator. The studies of the two considered models (Poisson and negative binomial) are done in parallel along the paper. Some exponential bounds are derived in Section 2.2.3. A simulation study is performed to compare our proposed criterion with others and an application to the segmentation of RNA-seq data illustrates the procedure in Section 2.2.4. The proof of the main result is given in Section 2.2.5 for which the proofs of some intermediate results

are given in the Appendix 2.2.6.

2.2.2 Model Selection Procedure

Penalized maximum-likelihood estimator

Let us denote by m a partition of $\llbracket 1, n \rrbracket$, $m = \{\llbracket 1, \tau_1 \rrbracket, \llbracket \tau_1, \tau_2 \rrbracket, \dots, \llbracket \tau_k, n \rrbracket\}$ and by \mathcal{M}_n a set of partitions of $\llbracket 1, n \rrbracket$. In our framework we want to estimate the distribution s defined by $s(t) = \mathcal{G}(\theta_t, \phi)$, $1 \leq t \leq n$, and we consider the two following models:

$$\mathcal{G}(\theta_t, \phi) = \mathcal{P}(\lambda_t) \quad (\mathcal{P})$$

$$\mathcal{G}(\theta_t, \phi) = \mathcal{NB}(p_t, \phi) \quad (\mathcal{NB})$$

In the (\mathcal{NB}) case, we suppose that the over-dispersion parameter ϕ is known. We define the collection of models :

Definition 2.2.1. *The collection of models associated to partition m is \mathcal{S}_m the set of distribution of sequences of length n such that for each element s_m of \mathcal{S}_m , for each segment J of m , and for each t in J , $s_m(t) = \mathcal{G}(\theta_J, \phi)$:*

$$\mathcal{S}_m = \{s_m \mid \forall J \in m, \forall t \in J, s_m(t) = \mathcal{G}(\theta_J, \phi)\}.$$

We shall denote by $|m|$ the number of segments in partition m , and by $|J|$ the length of segment J .

We consider the log-likelihood contrast $\gamma(u) = \sum_{t=1}^n -\log \mathbf{P}_u(Y_t)$, namely respectively for $u(t) = \mathcal{P}(\mu_t)$ and $u(t) = \mathcal{NB}(q_t, \phi)$,

$$\gamma(u) = \sum_{t=1}^n \mu_t - Y_t \log(\mu_t) + \log(Y_t!), \quad (\mathcal{P})$$

$$\gamma(u) = \sum_{t=1}^n -\phi \log q_t - Y_t \log(1 - q_t) - \log \left(\frac{\Gamma(\phi + Y_t)}{\Gamma(\phi) Y_t!} \right). \quad (\mathcal{NB})$$

Then the minimal contrast estimator \hat{s}_m of s on the collection \mathcal{S}_m is

$$\hat{s}_m = \arg \min_{u \in \mathcal{S}_m} \gamma(u), \quad (2.1)$$

so that, noting $\bar{Y}_J = \frac{\sum_{t \in J} Y_t}{|J|}$, for all $J \in m$ and $t \in J$

$$\hat{s}_m(t) = \mathcal{P}(\bar{Y}_J) \text{ for } (\mathcal{P}) \quad \text{and} \quad \hat{s}_m(t) = \mathcal{NB} \left(\frac{\phi}{\phi + \bar{Y}_J}, \phi \right) \text{ for } (\mathcal{NB}). \quad (2.2)$$

Therefore, for each partition m of \mathcal{M}_n we can obtain the best estimator \hat{s}_m as in equation (2.2), and thus define a collection of estimators $\{(\hat{s}_m)_{m \in \mathcal{M}_n}\}$. Ideally, we would wish to select the estimator $\hat{s}_{m(s)}$ amongst this collection with the minimum given risk. In the log-likelihood framework, it is natural to consider the Kullback-Leibler risk, with $K(s, u) = \mathbf{E}[\gamma(u) - \gamma(s)]$. In the following we note \mathbf{E} and \mathbf{P} the expectation and the probability under the true distribution s respectively (otherwise the underlying distribution is mentioned). In our models, the Kullback-Leibler between distributions s and u can be developed into

$$K(s, u) = \sum_{t=1}^n \left(\mu_t - \lambda_t - \lambda_t \log \frac{\mu_t}{\lambda_t} \right), \quad (\mathcal{P})$$

$$K(s, u) = \phi \sum_{t=1}^n \log \left(\frac{p_t}{q_t} \right) + \frac{1 - p_t}{p_t} \log \left(\frac{1 - p_t}{1 - q_t} \right). \quad (\mathcal{NB})$$

Unfortunately, minimizing this risk requires the knowledge of the true distribution s , and is unreachable. We will therefore want to consider the estimator $\hat{s}_{\hat{m}}$ where \hat{m} minimizes $\gamma(\hat{s}_m) + \text{pen}(m)$ for a well-chosen function pen (depending on the data). By doing so, we hope to select an estimator $\hat{s}_{\hat{m}}$ whose risk is as close as possible to the risk of $\hat{s}_{m(s)} = \arg \min_{m \in \mathcal{M}_n} \mathbf{E}_s[K(s, \hat{s}_m)]$ in the sense that

$$\mathbf{E}[K(s, \hat{s}_{\hat{m}})] \leq C \mathbf{E}[K(s, \hat{s}_{m(s)})],$$

where C is a nonnegative constant hopefully close to 1. We therefore introduce the following definition:

Definition 2.2.2. *Let \mathcal{M}_n be a collection of partitions of $\llbracket 1, n \rrbracket$ constructed on a partition m_f (i.e. m_f is a refinement of every m in \mathcal{M}_n). Given a nonnegative, increasing in the size of m penalty function $\text{pen}: \mathcal{M}_n \rightarrow \mathbf{R}_+$, and choosing*

$$\hat{m} = \arg \min_{m \in \mathcal{M}_n} \{\gamma(\hat{s}_m) + \text{pen}(m)\},$$

we define the penalized maximum-likelihood estimator as $\hat{s}_{\hat{m}}$.

In the following paragraph we provide a choice of penalty function, and show that the resulting estimator satisfies an oracle inequality.

Choice of the penalty function

Main result

The following result shows that for an appropriate choice of the penalty function, we have a non-asymptotic risk bound for the penalized maximum-likelihood estimator.

Theorem 2.2.3. *Let \mathcal{M}_n be a collection of partitions constructed on a partition m_f such that there exist absolute positive constants ρ_{min} , ρ_{max} and Γ satisfying:*

- $\forall t, \rho_{min} \leq \theta_t \leq \rho_{max}$ and
- $\forall J \in m_f, |J| \geq \Gamma(\log(n))^2$.

Let $(L_m)_{m \in \mathcal{M}_n}$ be some family of positive weights satisfying

$$\Sigma = \sum_{m \in \mathcal{M}_n} \exp(-L_m |m|) < +\infty. \quad (2.3)$$

Let $\beta > 1/2$ in the Poisson case, $\beta > 1/2\rho_{min}$ in the negative binomial case. If for every $m \in \mathcal{M}_n$

$$pen(m) \geq \beta |m| \left(1 + 4\sqrt{L_m}\right)^2, \quad (2.4)$$

then

$$\mathbf{E} \left[h^2(s, \hat{s}_{\hat{m}}) \right] \leq C_\beta \inf_{m \in \mathcal{M}_n} \{K(s, \bar{s}_m) + pen(m)\} + C(\phi, \Gamma, \rho_{min}, \rho_{max}, \beta, \Sigma),$$

with $C_\beta = \frac{(16\beta)^{1/3}}{(2\beta)^{1/3} - 1}$ in model (\mathcal{P}) and $C_\beta = \frac{(2\rho_{min}\beta)^{1/3}}{(2\rho_{min}\beta)^{1/3} - 1}$ in model (\mathcal{NB}) .

We note $h^2(s, u)$ the squared Hellinger distance between distribution s and u and \bar{s}_m is the projection of s onto the collection \mathcal{S}_m according to the Kullback-Leibler distance. The proof of this Theorem is given in Section 2.2.5.

Denoting $\bar{s}_m = \arg \min_{u \in \mathcal{S}_m} K(s, u)$, we have for $J \in m$ and $t \in J$,

$$\begin{aligned} \bar{s}_m(t) &= \mathcal{P}(\bar{\lambda}_J) & \text{where } \bar{\lambda}_J &= \frac{\sum_{t \in J} \lambda_t}{|J|} & (\mathcal{P}) \\ \bar{s}_m(t) &= \mathcal{NB}(p_J, \phi) & \text{where } p_J &= \frac{|J|}{\sum_{t \in J} 1/p_t}. & (\mathcal{NB}) \end{aligned} \quad (2.5)$$

We remark that the risk of the penalized estimator $\hat{s}_{\hat{m}}$ is treated in terms of Hellinger distance instead of the Kullback-Leibler information. This is due to the fact that the Kullback-Leibler is possibly infinite, and so difficult to control. It is possible to obtain a risk bound in term of Kullback-Leibler if we have a uniform control of $\|\log(s/\bar{s}_m)\|_\infty$ (see MASSART, 2007, for more explanation).

Choice of the weights $\{L_m, m \in \mathcal{M}_n\}$.

The penalty function depends on the family \mathcal{M}_n through the choice of the weights L_m which satisfy (2.3). We consider for \mathcal{M}_n the set of all possible partitions of $\llbracket 1, n \rrbracket$ constructed on a partition m_f which satisfies, for all segment J in m_f , $|J| \geq \Gamma(\log n)^2$. Classically (see BIRGÉ and MASSART, 2001) the weights are chosen as a function of the dimension of the model s , which is here $|m|$. The number of partitions of \mathcal{M}_n having dimension D being bounded by $\binom{n}{D}$, we have

$$\begin{aligned} \Sigma &= \sum_{m \in \mathcal{M}_n} e^{L_m |m|} = \sum_{D=1}^n e^{-L_D D} \text{Card}\{m \in \mathcal{M}_n, |m| = D\} \\ &\leq \sum_{D=1}^n \binom{n}{D} e^{-L_D D} \leq \sum_{D=1}^n \left(\frac{en}{D}\right)^D e^{-L_D D} \\ &\leq \sum_{D=1}^n e^{-D \left(L_D - 1 - \log\left(\frac{n}{D}\right)\right)}. \end{aligned}$$

So with the choice $L_D = 1 + \kappa + \log\left(\frac{n}{D}\right)$ with $\kappa > 0$, condition (2.3) is satisfied. Choosing, say $\kappa = 0.1$, the penalty function can be chosen of the form

$$\text{pen}(m) = \beta |m| \left(1 + 4 \sqrt{1.1 + \log\left(\frac{n}{|m|}\right)} \right)^2, \quad (2.6)$$

where β is a constant to be calibrated.

Integrating this penalty in Theorem 2.2.3 leads to the following control:

$$\mathbf{E} \left[h^2(s, \hat{s}_{\hat{m}}) \right] \leq C_1 \inf_{m \in \mathcal{M}_n} \left\{ K(s, \bar{s}_m) + \beta |m| \left(1 + 4 \sqrt{1.1 + \log \left(\frac{n}{|m|} \right)} \right)^2 \right\} + C(\phi, \Gamma, \rho_{min}, \rho_{max}, \beta, \Sigma). \quad (2.7)$$

The following proposition gives a bound on the Kullback-Leibler risk associated to \hat{s}_m :

Proposition 2.2.4. *Let m be a partition of \mathcal{M}_n , \hat{s}_m be the minimum contrast estimator and \bar{s}_m be the projection of s given by equations (2.2) and (2.5) respectively. Assume that there exists some positive absolute constants ρ_{min} , ρ_{max} and Γ such that $\forall t, \rho_{min} \leq \theta_t \leq \rho_{max}$ and $|J| \geq \Gamma(\log n)^2$. Then $\forall \varepsilon > 0, \forall a > 2$*

$$K(s, \bar{s}_m) - \frac{C_1(\phi, \Gamma, \rho_{min}, \rho_{max}, \varepsilon, a)}{n^{a/2-\alpha}} + C_2(\varepsilon)|m| \leq \mathbf{E}[K(s, \hat{s}_m)],$$

where $\alpha < 1$ is a constant that can be expressed according to n , $C_2(\varepsilon) = \frac{1}{2} \frac{1-\varepsilon}{(1+\varepsilon)^2}$ in the Poisson model (\mathcal{P}) and $C_2(\varepsilon) = \rho_{min}^2 \frac{(1-\varepsilon)^2}{(1+\varepsilon)^4}$ in the negative binomial model (\mathcal{NB}).

The proof is given in appendix 2.2.6.

Combining proposition 2.2.4 and equation (2.7), we obtain the following oracle-type inequality:

Corollary 2.2.5. *Let \mathcal{M}_n be a collection of partitions constructed on a partition m_f such that there exist absolute positive constants ρ_{min} , ρ_{max} and Γ verifying:*

- $\forall t, \rho_{min} \leq \theta_t \leq \rho_{max}$ and
- $\forall J \in m_f, |J| \geq \Gamma(\log n)^2$.

There exists some constant C such that

$$\mathbf{E} \left[h^2(s, \hat{s}_{\hat{m}}) \right] \leq C \log(n) \inf_{m \in \mathcal{M}_n} \{ \mathbf{E}[K(s, \hat{s}_m)] \} + C(\phi, \Gamma, \rho_{min}, \rho_{max}, \beta, \Sigma).$$

2.2.3 Exponential bounds

In order to prove Theorem 2.2.3, the general procedure in this model selection framework (see for example BIRGÉ and MASSART, 2001) is the following: by definitions of \hat{m} and \hat{s}_m (see definition 2.2.2 and equation (2.1)), we have, $\forall m \in \mathcal{M}_n$

$$\gamma(\hat{s}_{\hat{m}}) + \text{pen}(\hat{m}) \leq \gamma(\hat{s}_m) + \text{pen}(m) \leq \gamma(\bar{s}_m) + \text{pen}(m).$$

Then, with $\bar{\gamma}(u) = \gamma(u) - \mathbf{E}[\gamma(u)]$,

$$K(s, \hat{s}_{\hat{m}}) \leq K(s, \bar{s}_m) + \bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{\hat{m}}) - \text{pen}(\hat{m}) + \text{pen}(m).$$

The idea is therefore to control $\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{m'})$ uniformly over $m' \in \mathcal{M}_n$. This is more complicated when dealing with different models m and m' . Thus, following the work of CASTELLAN (1999) (see proof of Theorem 3.2, also recalled in MASSART, 2007), we propose the following decomposition

$$\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{m'}) = (\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'})) + (\bar{\gamma}(s) - \bar{\gamma}(\bar{s}_{m'})) + (\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s)), \quad (2.8)$$

and control each term separately. The first term is the most delicate to handle, and requires the introduction and the control of a chi-square statistic. The main difficulty here is the non-bounded characteristic of the objects we are dealing with. Indeed, in the classic density estimation context such as that of CASTELLAN (1999), the objects are probabilities which are bounded and so facilitate the direct use of concentration inequalities.

In our case, the chi-square statistic we introduce is denoted χ_m^2 and defined by

$$\chi_m^2 = \chi^2(\bar{s}_m, \hat{s}_m) = \sum_{J \in \mathcal{m}} |J| \frac{(\bar{Y}_J - \bar{E}_J)^2}{\bar{E}_J}, \quad (2.9)$$

where we recall that $\bar{Y}_J = \frac{\sum_{t \in J} Y_t}{|J|}$ and use the notation $\bar{E}_J = \frac{E_J}{|J|}$ with $E_J = \sum_{t \in J} E_t$. Respectively for (\mathcal{P}) and (\mathcal{NB}) , we have $E_t = \lambda_t$ and $E_t = \phi^{\frac{1-p_t}{p_t}}$. The purpose is thus to control χ_m^2 uniformly over \mathcal{M}_n . To this effect, we need to obtain an exponential bound of $Y_J = \sum_{t \in J} Y_t$ around its expectation. In the next paragraph, we recall a result of BARAUD and BIRGÉ (2009) that we use to derive an exponential bound for χ_m^2 .

Control of Y_J

First we recall a large deviation results established by BARAUD and BIRGÉ (2009) (lemma 3) that we apply in the Poisson and negative binomial frameworks.

Lemma 2.2.6. *Let Y_1, \dots, Y_n be n independent centered random variables.*

If $\log(\mathbf{E}[e^{zY_i}]) \leq \kappa \frac{z^2 \theta_i}{2(1-z\tau)}$ for all $z \in [0, 1/\tau[$, and $1 \leq i \leq n$, then

$$\mathbf{P} \left[\sum_{i=1}^n Y_i \geq \left(2\kappa x \sum_{i=1}^n \theta_i \right)^{1/2} + \tau x \right] \leq e^{-x} \text{ for all } x > 0.$$

If for $1 \leq i \leq n$ and all $z > 0$ $\log(\mathbf{E}[e^{-zY_i}]) \leq \kappa z^2 \theta_i / 2$, then

$$\mathbf{P} \left[\sum_{i=1}^n Y_i \leq - \left(2\kappa x \sum_{i=1}^n \theta_i \right)^{1/2} \right] \leq e^{-x} \text{ for all } x > 0.$$

To apply this lemma we therefore need a majoration of $\log \mathbf{E} \left[e^{z(Y_t - E_t)} \right]$ and $\log \mathbf{E} \left[e^{-z(Y_t - E_t)} \right]$ for $z > 0$.

Poisson case. With $E_t = \lambda_t$, we have:

$$\log \mathbf{E} \left[e^{z(Y_t - \lambda_t)} \right] = -z\lambda_t + \log \mathbf{E} \left[e^{zY_J} \right] = -z\lambda_t + \log e^{(\lambda_t(e^z - 1))} = \lambda_t(e^z - z - 1).$$

Using $e^z - z - 1 \leq \frac{z^2}{2(1-z)}$ for $0 < z < 1$ and $e^z - z - 1 \leq \frac{z^2}{2}$ for $z < 0$, we have

$$\log \mathbf{E} \left[e^{z(Y_t - E_t)} \right] \leq E_t \frac{z^2}{2(1-z)} \quad \text{and} \quad \log \mathbf{E} \left[e^{-z(Y_t - E_t)} \right] \leq E_t \frac{z^2}{2}$$

Negative binomial case. In this case $E_t = \phi \frac{1-p_t}{p_t}$ and we have

$$\begin{aligned} \log \mathbf{E} \left(e^{z \left(Y_t - \phi \frac{1-p_t}{p_t} \right)} \right) &= \frac{z^2}{2} \sum_{k \geq 0} \frac{2\kappa_{k+2}}{(k+2)!} z^k \text{ for } z \leq -\log(1-p_t) \\ &\leq E_t \frac{z^2}{2} \frac{2}{p_t} \sum_{k \geq 0} \left(\frac{z}{p_t} \right)^k \end{aligned}$$

where the κ_k are the cumulants of the negative binomial distribution.

Then

$$\begin{aligned} \log \mathbf{E} \left(e^{z \left(Y_t - \phi \frac{1-pt}{pt} \right)} \right) &\leq E_t \frac{z^2}{2} \frac{2}{\rho_{min}} \frac{1}{1 - \frac{z}{\rho_{min}}} \quad \text{for } z \leq \rho_{min} \\ &\leq E_t \frac{z^2}{2} \frac{2}{\rho_{min}} \quad \text{for } -1 \leq 0 \leq z \end{aligned}$$

Finally, with $\kappa = 1$ in the Poisson case and $\kappa = 2/\rho_{min}$ in the negative binomial case, we get

$$P \left[Y_J - E_J \geq \sqrt{2\kappa x E_J} + \kappa x \right] \leq e^{-x},$$

leading to

$$P [Y_J - E_J \geq x] \leq e^{-\frac{x^2}{2\kappa(E_J+x)}} \quad \text{and} \quad P [|Y_J - E_J| \geq x] \leq 2e^{-\frac{x^2}{2\kappa(E_J+x)}} \quad (2.10)$$

Exponential bound for χ_m^2

We first introduce the following set Ω_m defined by:

$$\Omega_m(\varepsilon) = \bigcap_{J \in m} \left\{ \left| \frac{Y_J}{E_J} - 1 \right| \leq \varepsilon \right\}, \quad (2.11)$$

for all $\varepsilon \in]0, 1[$ and all segmentations m such that each segment J verifies $|J| \geq \Gamma(\log(n))^2$.

This set has a large probability since we obtain

$$\begin{aligned} \mathbf{P}(\Omega_m(\varepsilon)^C) &\leq \sum_{J \in m} \mathbf{P}(|Y_J - E_J| > \varepsilon E_J) \leq 2 \sum_{J \in m} e^{-\frac{\varepsilon^2 E_J}{2\kappa(1+\varepsilon)}} \\ &\leq 2 \sum_{J \in m} e^{-|J| \varepsilon' f(\phi, \rho_{min})} \leq 2|m| \exp(-\varepsilon' \Gamma f(\phi, \rho_{min}) (\log(n))^2) \end{aligned}$$

by applying equation (2.10) with $x = \varepsilon E_J$ and where $\varepsilon' = \varepsilon^2 / (2(1 + \varepsilon))$ and $f(\phi, \rho_{min}) > 0$.

Thus

$$\mathbf{P}(\Omega_m(\varepsilon)^C) \leq \frac{C(\phi, \Gamma, \rho_{min}, \varepsilon, a)}{n^a}, \quad (2.12)$$

with $a > 2$.

The reason for introducing this set is double: in addition to enable the control of χ_m^2 given

by equation (2.9) on this restricted set, it allows us to link $K(\hat{s}_m, \bar{s}_m)$ to V_m^2 (see (2.18) for the control of the first term in the decomposition) and so to χ_m^2 , relation that we use to evaluate the risk of one model (see (2.20)).

Let m_f be a partition of \mathcal{M}_n such that $\forall J \in m_f, |J| \geq \Gamma(\log(n))^2$ and assume that all considered partitions in \mathcal{M}_n are constructed on this grid m_f . The following proposition gives an exponential bound for χ_m^2 on the restricted event $\Omega_{m_f}(\epsilon)$.

Proposition 2.2.7. *Let Y_1, \dots, Y_n be independent random variables with distribution \mathcal{G} (Poisson or negative binomial distribution). Let m be a partition of \mathcal{M}_n with $|m|$ segments and χ_m^2 the statistic given by (2.9). For any positive x , we have*

$$\mathbf{P} \left[\chi_m^2 \mathbf{1}_{\Omega_{m_f}(\epsilon)} \geq C(\rho_{min}) \left(|m| + 8(1 + \epsilon) \sqrt{x|m|} + 4(1 + \epsilon)x \right) \right] \leq e^{-x}.$$

with $C(\rho_{min}) = 1$ in the Poisson case and $2/\rho_{min}$ in the negative binomial case.

Proof. As in the density estimation framework, this quantity can be controlled using the Bernstein inequality. In our context, noting $\chi_m^2 = \sum_{J \in m} Z_J$ where

$$Z_J = \frac{(Y_J - E_J)^2}{E_J},$$

we need

- the calculation (or bounds) of the expectation of χ_m^2 :

Poisson case Y_J is distributed according to a Poisson distribution with parameter λ_J so that

$$\mathbf{E} [\chi_m^2] = |m|. \tag{2.13}$$

Negative binomial case We have

$$\mathbf{E} [\chi_m^2] = \sum_{J \in m} \frac{1}{|J|} \frac{\sum_{t \in J} \text{Var}(Y_t)}{\phi^{\frac{1-p_J}{p_J}}} = \sum_{J \in m} \frac{1}{|J|} \frac{\sum_{t \in J} \phi^{\frac{1-pt}{p_t}}}{\phi^{\frac{1-p_J}{p_J}}},$$

and thus

$$|m| \leq \mathbf{E} [\chi_m^2] \leq \frac{1}{\rho_{min}} |m|. \tag{2.14}$$

- an upper bound of $\sum_{J \in m} \mathbf{E}[Z_J^p]$. For every $p \geq 2$ we have,

$$\begin{aligned} \mathbf{E} \left[Z_J^p \mathbf{1}_{\Omega_{m_f}(\epsilon)} \right] &= \frac{1}{E_J^p} \int_0^{+\infty} 2p x^{2p-1} P \left[\{|Y_J - E_J| \geq x\} \cap \Omega_{m_f}(\epsilon) \right] dx \\ &\leq \frac{1}{E_J^p} \int_0^{\varepsilon E_J} 2p x^{2p-1} P \left[|Y_J - E_J| \geq x \right] dx. \end{aligned}$$

Using equation (2.10) and since $x \leq \varepsilon E_J$, we obtain the exponential bound $P \left[|Y_J - E_J| \geq x \right] \leq 2e^{-\frac{x^2}{2\kappa E_J(1+\varepsilon)}}$.

Therefore

$$\begin{aligned} \mathbf{E} \left[Z_J^p \mathbf{1}_{\Omega_{m_f}(\epsilon)} \right] &\leq \frac{1}{E_J^p} \int_0^{\varepsilon E_J} 4p x^{2p-1} e^{-\frac{x^2}{2\kappa E_J(1+\varepsilon)}} dx \\ &\leq 4p\kappa^p (1+\varepsilon)^p \int_0^{+\infty} u^{2p-1} e^{-\frac{u^2}{2}} du \\ &\leq 4p\kappa^p (1+\varepsilon)^p \int_0^{+\infty} (2t)^{p-1} e^{-t} dt \\ &\leq 2^{p+1} p\kappa^p (1+\varepsilon)^p p!, \end{aligned}$$

and

$$\sum_{J \in m} \mathbf{E} \left[Z_J^p \mathbf{1}_{\Omega_{m_f}(\epsilon)} \right] \leq 2^{p+1} p\kappa^p (1+\varepsilon)^p p! |m|.$$

Since $p \leq 2^{p-1}$,

$$\sum_{J \in m} \mathbf{E} \left[Z_J^p \mathbf{1}_{\Omega_{m_f}(\epsilon)} \right] \leq \frac{p!}{2} \times \left[2^5 (\kappa(1+\varepsilon))^2 |m| \right] \times [4(\kappa(1+\varepsilon))]^{p-2}.$$

We conclude by taking $v = 2^5 (\kappa(1+\varepsilon))^2 |m|$ and $c = 4(\kappa(1+\varepsilon))$ (see proposition 2.9 of MASSART (2007) for the definition of the Bernstein's inequality).

□

2.2.4 Simulations and application

In the context of RNA-seq experiments, an important question is the (re)-annotation of the genome, that is, the precise localisation of the transcribed regions on the chromosomes.

In an ideal situation, when considering the number of reads starting at each position, one would expect to observe a uniform coverage over each gene (proportional to its expression level), separated by regions of null signal (corresponding to non-transcribed regions of the genome). In practice however, those experiments tend to return very noisy signals that are best modelled by the negative binomial distribution.

In this Section, we first study the performance of the proposed penalized criterion by comparing it with others model selection criteria on a resampling dataset and then we provide an application on real data. Since the penalty depends on the partition only through its size, the segmentation procedure is two-steps: first we estimate, for all number of segments K between 1 and K_{max} , the optimal partition with K segments (i.e. construct the collection of estimators $\{\hat{s}_K\}_{1 \leq K \leq K_{max}}$ where $\hat{s}_K = \arg \min_{\hat{s}_m, m \in \mathcal{M}_K} \{\gamma(\hat{s}_m)\}$). The optimal solution is obtained using a fast segmentation algorithm such as the Pruned Dynamic Programming Algorithm (PDPA, RIGAILL, 2010) implemented for the Poisson and negative binomial losses or contrasts in the R package `Segmentor3IsBack` (CLEYNEN *et al.*, under review). Then, we choose K using our penalty function which requires the calibration of the constant β that can be tuned according to the data by using the slope heuristic (see BIRGÉ and MASSART (2007); ARLOT and MASSART (2009)). Using the negative binomial distribution requires the knowledge of parameter ϕ . We propose to estimate it using a modified version of the Johnson and Kotz's estimator (JOHNSON *et al.*, 2005).

Simulation study

We have assessed the performances of the proposed method (called Penalized PDPA) on a simulation scenario by comparing to five other procedures both its choice in the number of segments and the quality of the obtained segmentation using the Rand Index \mathcal{I} . This index is defined as follows: let C_t be the true index of the segment to which base t belongs and let \hat{C}_t be the corresponding estimated index, then

$$\mathcal{I} = \frac{2 \sum_{t>s} [\mathbf{1}_{C_t=C_s} \mathbf{1}_{\hat{C}_t=\hat{C}_s} + \mathbf{1}_{C_t \neq C_s} \mathbf{1}_{\hat{C}_t \neq \hat{C}_s}]}{(n-1)(n-2)}.$$

The characteristics of the different algorithms are described in Table 2.1.

Algorithm	Dist	Complexity	Inference	Pen	Exact	Reference
Penalized PDPA	NB	$n \log n$	frequentist	external	exact	CLEYNEN <i>et al.</i> (under review)
PDPA with BIC	NB	$n \log n$	frequentist	external	exact	CLEYNEN <i>et al.</i> (under review)
Penalized PDPA	P	$n \log n$	frequentist	external	exact	CLEYNEN <i>et al.</i> (under review)
PDPA with BIC	P	$n \log n$	frequentist	external	exact	CLEYNEN <i>et al.</i> (under review)
PELT with BIC	P	n	frequentist	internal	exact	KILLICK and ECKLEY (2011)
CART with BIC	P	$n \log n$	frequentist	external	heuristic	BREIMAN <i>et al.</i> (1984)
postCP with ICL	NB	n	frequentist	external	exact	LUONG <i>et al.</i> (2013)
EBS with ICL	NB	n^2	Bayesian	external	exact	RIGAILL <i>et al.</i> (2012)

Table 2.1: **Properties of segmentation algorithms.** The first column indicates the name of the algorithm and the criterion used for the choice of K . In the second column, NB stands for the negative binomial distribution and P for Poisson. The time of each algorithm is given (column "Complexity") and column "Exact" states if the exact solution is reached.

The data we considered comes from a resampling procedure using real RNA-seq data. The original data, from a study by the Sherlock Genomics laboratory at Stanford University, is publicly available on the NCBI's Sequence Read Archive (SRA, url: <http://www.ncbi.nlm.nih.gov/sra>) with the accession number SRA048710. We created an artificial gene, inspired from the *Drosophila* *inr-a* gene, resulting in a 14-segment signal with irregular intensities mimicking a differentially transcribed gene. 100 datasets are thus created. Results are presented using boxplots in Figure 2.5. Because PELT's estimate of K averaged around 427 segments, we did not show its corresponding boxplot.

We can see that with the negative binomial distribution, not only do we perfectly recover the true number of segments, but our procedure outperforms all other approaches. Moreover, the impressive results in terms of Rand Index prove that our choice of number of segments also leads to the almost perfect recovery of the true segmentation. However, the use of the Poisson loss leads to a constant underestimation of the number of segments, which is reflected on the Rand Index values. This is due to the inappropriate choice of

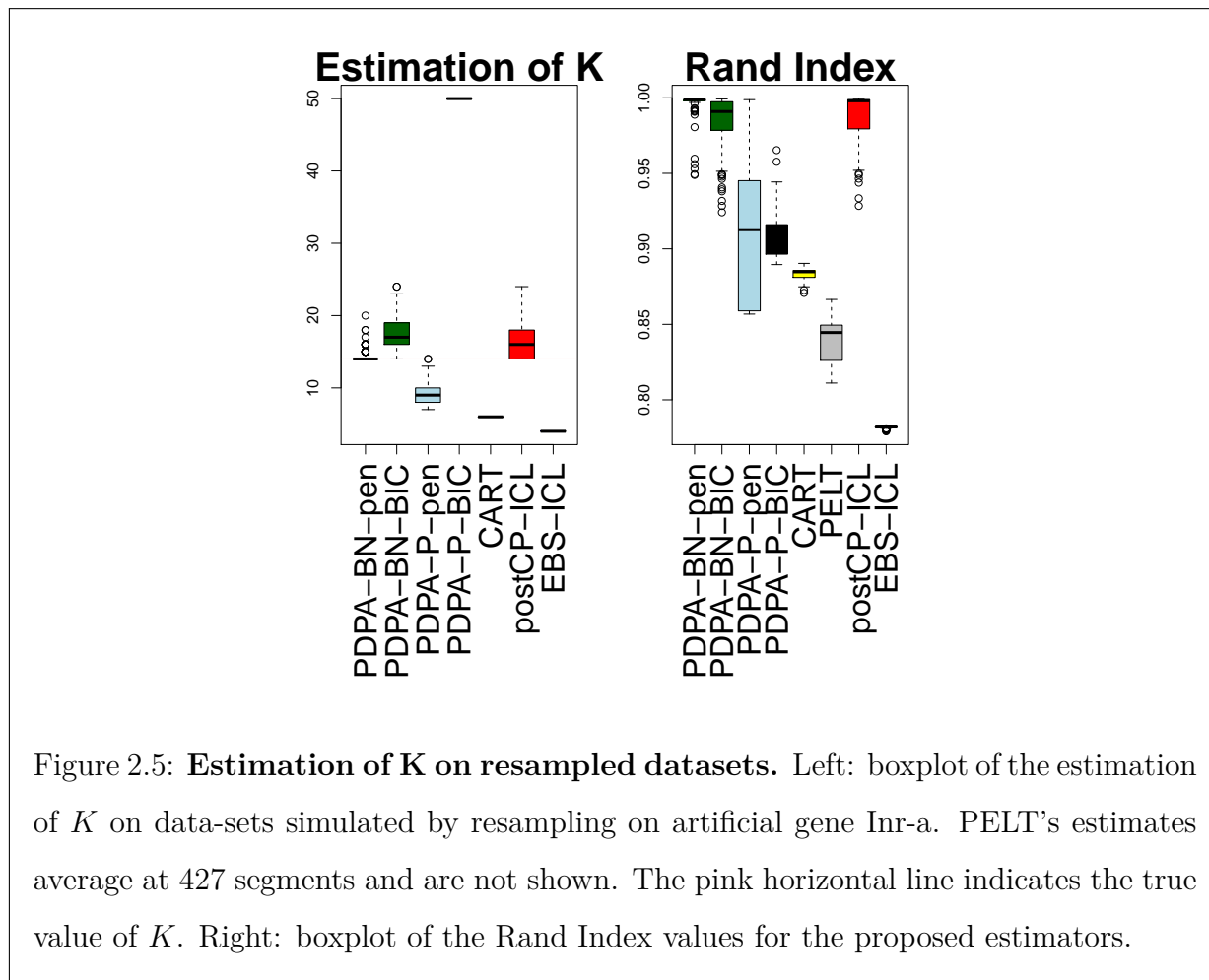


Figure 2.5: **Estimation of K on resampled datasets.** Left: boxplot of the estimation of K on data-sets simulated by resampling on artificial gene Inr-a. PELT’s estimates average at 427 segments and are not shown. The pink horizontal line indicates the true value of K . Right: boxplot of the Rand Index values for the proposed estimators.

distribution (confirmed by the other algorithms implemented for the Poisson loss which perform worse than the others). It however underlines the need for the development of methods for the negative binomial distribution. Moreover, in terms of computational time, the fast algorithm is in $\mathcal{O}(n \log n)$ (CLEYNEN *et al.*, under review), allowing its use on long signals (such as a whole-genome analysis), even though it is not as fast as CART or PELT.

Segmentation of RNA-Seq data

We apply our proposed procedure for segmenting chromosome 1 of the *S. Cerevisiae* (yeast) using RNA-Seq data from the Sherlock Laboratory at Stanford University (RISSE *et al.*, 2011) and publicly available from the NCBI’s Sequence Read Archive (SRA, url:<http://www.ncbi.nlm.nih.gov/sra>, accession number SRA048710). An

existing annotation is available on the Saccharomyces Genome Database (SGD) at url: <http://www.yeastgenome.org>, which allows us to validate our results. The two distributions (Poisson and negative binomial) are considered here to show the difference.

In the Poisson distribution case, we select 106 segments of which only 19 are related to the SGD annotation. Indeed, as illustrated by Figure 2.6, 36 of the segments have a length smaller than 10: the Poisson loss is not adapted to this kind of data with high variability and it tends to select outliers as segment. On the contrary, we select 103 segments in the negative binomial case most of which (all but 3) surround known genes from the SGD. Figure 2.7 illustrates the result. However, almost none of those change-points correspond exactly to annotated boundaries. Discussion with biologists has increased our belief in the need for genome (re-)annotation using RNA-seq data, and in the validity of our approach.

2.2.5 Proof of Theorem 2.2.3

Recall that we want to control the three terms in the decomposition given by (2.8). All the proofs of the different propositions are given in Section 2.2.6.

- The control of the term $\bar{\gamma}(\hat{s}_{m'}) - \bar{\gamma}(\bar{s}_{m'})$ is obtained with the following proposition where the set $\Omega_1(\xi)$ is defined by

$$\Omega_1(\xi) = \bigcap_{m' \in \mathcal{M}_n} \left\{ \chi_{m'}^2 \mathbf{1}_{\Omega_{m_f}(\epsilon)} \leq C(\rho_{min}) \left[|m'| + 8(1 + \epsilon) \sqrt{(L_{m'} |m'| + \xi) |m'|} + 4(1 + \epsilon)(L_{m'} |m'| + \xi) \right] \right\}.$$

Proposition 2.2.8. *Let m' be a partition of \mathcal{M}_n . Then*

$$\begin{aligned} (\bar{\gamma}(\hat{s}_{m'}) - \bar{\gamma}(\bar{s}_{m'})) \mathbf{1}_{\Omega_{m_f}(\epsilon) \cap \Omega_1(\xi)} &\leq C(\epsilon) \left[|m'| + 8(1 + \epsilon) \sqrt{(L_{m'} |m'| + \xi) |m'|} \right. \\ &\quad \left. + 4(1 + \epsilon)(L_{m'} |m'| + \xi) \right] + \frac{1}{1 + \epsilon} K(\bar{s}_{m'}, \hat{s}_{m'}), \end{aligned}$$

with $C(\epsilon) = \frac{1}{2} \left(\frac{1+\epsilon}{1-\epsilon} \right)$ in the Poisson case and $C(\epsilon) = \frac{1+\epsilon}{2\rho_{min}}$ in the negative binomial case.

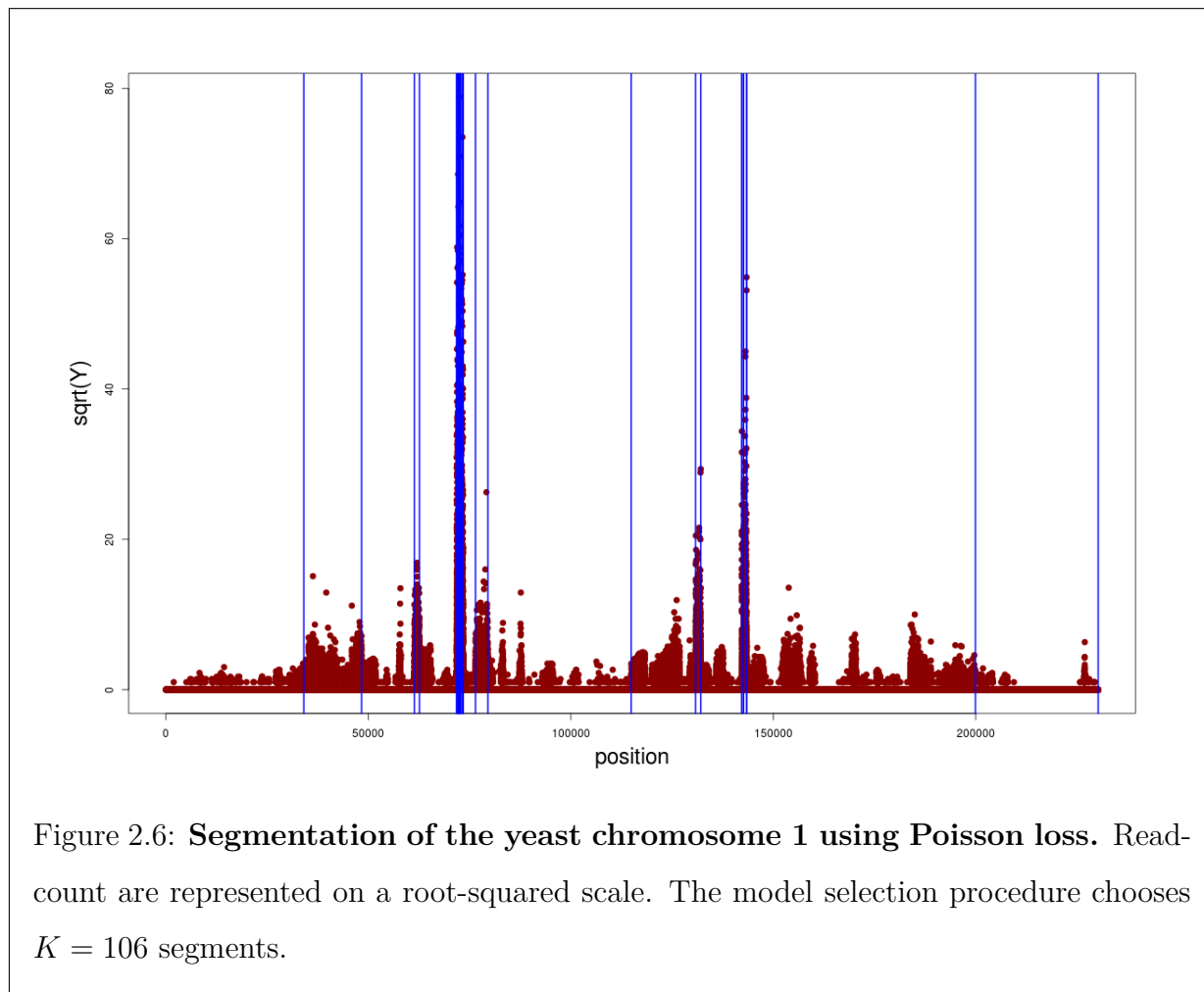


Figure 2.6: **Segmentation of the yeast chromosome 1 using Poisson loss.** Read-count are represented on a root-squared scale. The model selection procedure chooses $K = 106$ segments.

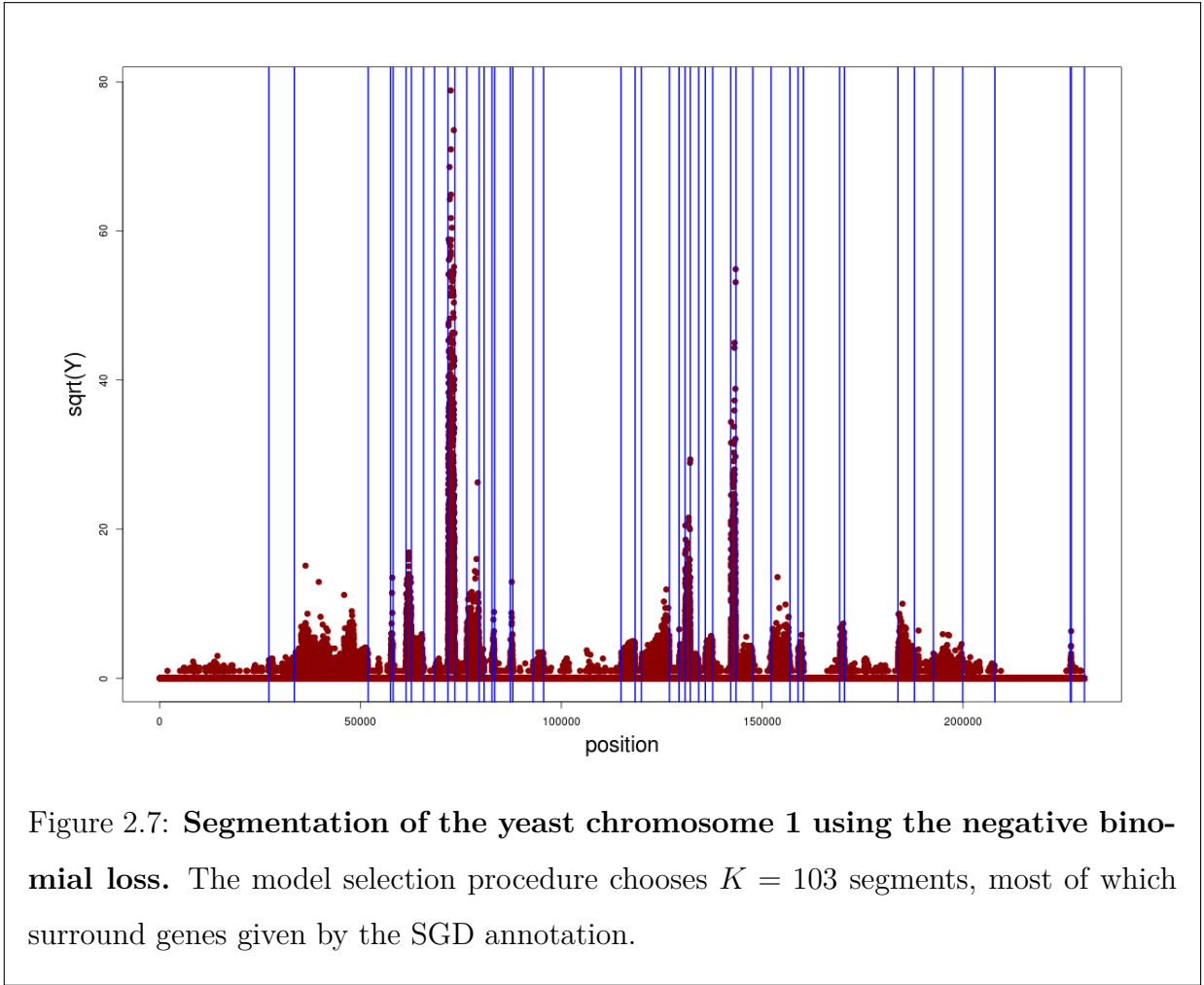
- The control of the term $\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s)$, or more precisely its expectation, is given by the following proposition:

Proposition 2.2.9.

$$|\mathbf{E}[(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))\mathbf{1}_{\Omega_{m_f}(\epsilon)}]| \leq \frac{C(\phi, \Gamma, \rho_{min}, \rho_{max}, \epsilon, a)}{n^{(a-1)/2}}. \quad (2.15)$$

- To control $\bar{\gamma}(s) - \bar{\gamma}(\bar{s}_{m'})$, we use the following proposition which gives an exponential bound for $\bar{\gamma}(s) - \bar{\gamma}(u)$.

Proposition 2.2.10. *Let s and u be two distributions of a sequence Y . Let γ be the log-likelihood contrast, $\bar{\gamma}(u) = \gamma(u) - \mathbf{E}[\gamma(u)]$, and $K(s, u)$ and $h^2(s, u)$ be respectively the Kullback-Leibler and the squared Hellinger distances between distributions s and*



u. Then $\forall x > 0$,

$$\mathbf{P} \left[\bar{\gamma}(s) - \bar{\gamma}(u) \geq K(s, u) - 2h^2(s, u) + 2x \right] \leq e^{-x}.$$

Applying it to $u = \bar{s}_{m'}$ yields:

$$\mathbf{P} \left[\bar{\gamma}(s) - \bar{\gamma}(\bar{s}_{m'}) \geq K(s, \bar{s}_{m'}) - 2h^2(s, \bar{s}_{m'}) + 2x \right] \leq e^{-x}. \quad (2.16)$$

We then define

$$\Omega_2(\xi) = \bigcap_{m' \in \mathcal{M}_n} \left\{ \bar{\gamma}(s) - \bar{\gamma}(\bar{s}_{m'}) \leq K(s, \bar{s}_{m'}) - 2h^2(s, \bar{s}_{m'}) + 2(L_{m'}|m'| + \xi) \right\}.$$

Let $\Omega(\varepsilon, \xi) = \Omega_{m_f}(\varepsilon) \cap \Omega_1(\xi) \cap \Omega_2(\xi)$. Then, combining equation (2.16) and proposition 2.2.8, we get for $m' = \hat{m}$,

$$\begin{aligned}
(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{\hat{m}}))\mathbf{1}_{\Omega(\epsilon, \xi)} &= (\bar{\gamma}(s) - \bar{\gamma}(\bar{s}_{\hat{m}}))\mathbf{1}_{\Omega(\epsilon, \xi)} + (\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))\mathbf{1}_{\Omega(\epsilon, \xi)} + (\bar{\gamma}(\bar{s}_{\hat{m}}) - \bar{\gamma}(\hat{s}_{\hat{m}}))\mathbf{1}_{\Omega(\epsilon, \xi)} \\
&\leq \left[K(s, \bar{s}_{\hat{m}}) - 2h^2(s, \bar{s}_{\hat{m}}) \right] \mathbf{1}_{\Omega(\epsilon, \xi)} + R\mathbf{1}_{\Omega(\epsilon, \xi)} + \frac{1}{1 + \epsilon} K(\bar{s}_{\hat{m}}, \hat{s}_{\hat{m}})\mathbf{1}_{\Omega(\epsilon, \xi)} \\
&\quad + C(\epsilon) \left[|\hat{m}| + 8(1 + \epsilon)\sqrt{(L_{\hat{m}}|\hat{m}| + \xi)|\hat{m}|} + 4(1 + \epsilon)(L_{\hat{m}}|\hat{m}| + \xi) \right] \\
&\quad + 2L_{\hat{m}}|\hat{m}| + 2\xi,
\end{aligned}$$

with $R = \bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s)$. So that

$$\begin{aligned}
K(s, \hat{s}_{\hat{m}})\mathbf{1}_{\Omega(\epsilon, \xi)} &\leq \left[K(s, \bar{s}_{\hat{m}}) - 2h^2(s, \bar{s}_{\hat{m}}) \right] \mathbf{1}_{\Omega(\epsilon, \xi)} + \frac{1}{1 + \epsilon} K(\bar{s}_{\hat{m}}, \hat{s}_{\hat{m}})\mathbf{1}_{\Omega(\epsilon, \xi)} \\
&\quad + C(\epsilon) \left[|\hat{m}| + 8(1 + \epsilon)\sqrt{(L_{\hat{m}}|\hat{m}| + \xi)|\hat{m}|} + 4(1 + \epsilon)(L_{\hat{m}}|\hat{m}| + \xi) \right] \\
&\quad + K(s, \bar{s}_m)\mathbf{1}_{\Omega(\epsilon, \xi)} + 2L_{\hat{m}}|\hat{m}| + 2\xi + R\mathbf{1}_{\Omega(\epsilon, \xi)} - \text{pen}(\hat{m}) + \text{pen}(m).
\end{aligned}$$

And since

- $K(s, \hat{s}_{\hat{m}}) = K(s, \bar{s}_{\hat{m}}) + K(\bar{s}_{\hat{m}}, \hat{s}_{\hat{m}})$ (see equation (2.17)),
- $K(s, u) \geq 2h^2(s, u)$ (see lemma 7.23 in MASSART, 2007),
- $h^2(s, \hat{s}_{\hat{m}}) \leq 2(h^2(s, \bar{s}_{\hat{m}}) + h^2(\bar{s}_{\hat{m}}, \hat{s}_{\hat{m}}))$ (using inequality $2ab \leq \kappa a^2 + \kappa^{-1}b^2$ with $\kappa = 1$),

$$\begin{aligned}
\frac{\epsilon}{1 + \epsilon} h^2(s, \hat{s}_{\hat{m}})\mathbf{1}_{\Omega(\epsilon, \xi)} &\leq K(s, \bar{s}_m)\mathbf{1}_{\Omega(\epsilon, \xi)} + R\mathbf{1}_{\Omega(\epsilon, \xi)} - \text{pen}(\hat{m}) + \text{pen}(m) \\
&\quad + |\hat{m}|C(\epsilon) \left[1 + (1 + \epsilon) \left(8\sqrt{L_{\hat{m}}} + \epsilon + 4L_{\hat{m}} \right) \right] + 2L_{\hat{m}}|\hat{m}| \\
&\quad + 2\xi \left[1 + C(\epsilon) \left(8(1 + \epsilon)\frac{2}{\epsilon} + 4(1 + \epsilon) \right) \right].
\end{aligned}$$

But

$$\begin{aligned}
C(\epsilon) \left[1 + (1 + \epsilon) \left(8\sqrt{L_{\hat{m}}} + \epsilon + 4L_{\hat{m}} \right) \right] + 2L_{\hat{m}} &\leq C(\epsilon) \left[1 + (1 + \epsilon) \left(\epsilon + 8\sqrt{L_{\hat{m}}} + 8L_{\hat{m}} \right) \right] \\
&\leq C_2(\epsilon) \left[1 + 8\sqrt{L_{\hat{m}}} + 8L_{\hat{m}} \right].
\end{aligned}$$

with $C_2(\epsilon) = \frac{1}{2} \left(\frac{1 + \epsilon}{1 - \epsilon} \right)^3$ for (\mathcal{P}) and $C_2(\epsilon) = \frac{1}{2\rho_{\min}} (1 + \epsilon)^3$ for (\mathcal{NB}) . So we have

$$\begin{aligned}
\frac{\epsilon}{1 + \epsilon} h^2(s, \hat{s}_{\hat{m}})\mathbf{1}_{\Omega(\epsilon, \xi)} &\leq K(s, \bar{s}_m)\mathbf{1}_{\Omega(\epsilon, \xi)} + R\mathbf{1}_{\Omega(\epsilon, \xi)} - \text{pen}(\hat{m}) + \text{pen}(m) \\
&\quad + |\hat{m}|C_2(\epsilon) \left(1 + 4\sqrt{L_{\hat{m}}} \right)^2 + 2\xi \left[1 + (1 + \epsilon)C(\epsilon) \left(\frac{8}{\epsilon} + 2 \right) \right].
\end{aligned}$$

By assumption, $pen(\hat{m}) \geq \beta|\hat{m}| \left(1 + 4\sqrt{L_{\hat{m}}}\right)^2$. Choosing $\beta = C_2(\varepsilon)$ yields

$$h^2(s, \hat{s}_{\hat{m}}) \mathbf{1}_{\Omega(\varepsilon, \xi)} \leq C_\beta \left[K(s, \bar{s}_m) \mathbf{1}_{\Omega(\varepsilon, \xi)} + R \mathbf{1}_{\Omega(\varepsilon, \xi)} + pen(m) \right] + \xi C(\beta).$$

Then, using propositions 2.2.9 and 2.2.8, we have $\mathbf{P} \left(\Omega_1(\xi)^C \right) \leq \sum_{m' \in \mathcal{M}_n} e^{-L_{m'}|m'|+\xi}$ and $\mathbf{P} \left(\Omega_2(\xi)^C \right) \leq \sum_{m' \in \mathcal{M}_n} e^{-L_{m'}|m'|+\xi}$. So that using hypothesis (2.3),

$$\mathbf{P} \left(\Omega_1(\xi)^C \cup \Omega_2(\xi)^C \right) \leq 2 \sum_{m' \in \mathcal{M}_n} e^{-L_{m'}|m'|+\xi} \leq 2\Sigma e^{-\xi},$$

and thus $\mathbf{P} \left(\Omega_1(\xi) \cap \Omega_2(\xi) \right) \geq 1 - 2\Sigma e^{-\xi}$. We now integrate over ξ , and using equation (2.15), we get with a probability larger than $1 - 2\Sigma e^{-\xi}$

$$\mathbf{E} \left[h^2(s, \hat{s}_{\hat{m}}) \mathbf{1}_{\Omega_{m_f}(\varepsilon)} \right] \leq C_\beta \left[K(s, \bar{s}_m) + \frac{C(\phi, \Gamma, \rho_{min}, \rho_{max}, \beta, a)}{n^{(a-1)/2}} + pen(m) \right] + \Sigma C(\beta).$$

And since $\mathbf{E} \left[h^2(s, \hat{s}_{\hat{m}}) \mathbf{1}_{\Omega_{m_f}(\varepsilon)^C} \right] \leq \frac{C(\phi, \Gamma, \rho_{min}, \rho_{max}, \beta, a)}{n^{a-1}}$, we have

$$\mathbf{E} \left[h^2(s, \hat{s}_{\hat{m}}) \right] \leq C_\beta \left[K(s, \bar{s}_m) + pen(m) \right] + C'(\phi, \Gamma, \rho_{min}, \rho_{max}, \beta, \Sigma).$$

Finally, by minimizing over $m \in \mathcal{M}_n$, we get

$$\mathbf{E} \left[h^2(s, \hat{s}_{\hat{m}}) \right] \leq C_\beta \inf_{m \in \mathcal{M}_n} \{ K(s, \bar{s}_m) + pen(m) \} + C'(\phi, \Gamma, \rho_{min}, \rho_{max}, \beta, \Sigma).$$

2.2.6 Appendices

Proof of proposition 2.2.4

Using Pythagorean-type identity, we obtain the following decomposition (see for example CASTELLAN, 1999):

$$K(s, \hat{s}_m) = K(s, \bar{s}_m) + K(\bar{s}_m, \hat{s}_m). \quad (2.17)$$

The objective is then to obtain a lower bound of $\mathbf{E}[K(\bar{s}_m, \hat{s}_m)]$ in the two considered distribution cases.

Poisson case We have

$$K(\bar{s}_m, \hat{s}_m) = \sum_{J \in m} |J| \left(\bar{Y}_J - \bar{\lambda}_J - \bar{\lambda}_J \log \frac{\bar{Y}_J}{\bar{\lambda}_J} \right) = \sum_{J \in m} |J| \bar{\lambda}_J \Phi \left(\log \frac{\bar{Y}_J}{\bar{\lambda}_J} \right).$$

where $\Phi(x) = e^x - 1 - x$. Since $\frac{1}{2}x^2(1 \wedge e^x) \leq \Phi(x) \leq \frac{1}{2}x^2(1 \vee e^x)$, then on $\Omega_{m_f}(\epsilon)$, we have

$$\begin{aligned} \frac{1}{2} \log^2 \frac{\bar{Y}_J}{\bar{\lambda}_J} \left(1 \wedge \frac{\bar{Y}_J}{\bar{\lambda}_J} \right) &\leq \Phi \left(\log \frac{\bar{Y}_J}{\bar{\lambda}_J} \right) \leq \frac{1}{2} \log^2 \frac{\bar{Y}_J}{\bar{\lambda}_J} \left(1 \vee \frac{\bar{Y}_J}{\bar{\lambda}_J} \right), \\ \frac{1-\epsilon}{2} \log^2 \frac{\bar{Y}_J}{\bar{\lambda}_J} &\leq \Phi \left(\log \frac{\bar{Y}_J}{\bar{\lambda}_J} \right) \leq \frac{1+\epsilon}{2} \log^2 \frac{\bar{Y}_J}{\bar{\lambda}_J}. \end{aligned}$$

So

$$\frac{1-\epsilon}{2} V_m^2 \leq K(\bar{s}_m, \hat{s}_m) \leq \frac{1+\epsilon}{2} V_m^2, \quad (2.18)$$

where

$$V_m^2 = V^2(\bar{s}_m, \hat{s}_m) = \sum_{J \in m} |J| \bar{\lambda}_J \log^2 \frac{\bar{Y}_J}{\bar{\lambda}_J} = \sum_{J \in m} |J| \frac{(\bar{Y}_J - \bar{\lambda}_J)^2}{\bar{\lambda}_J} \left(\frac{\log \frac{\bar{Y}_J}{\bar{\lambda}_J}}{\frac{\bar{Y}_J}{\bar{\lambda}_J} - 1} \right)^2. \quad (2.19)$$

And using, for $x > 0$, $\frac{1}{1 \vee x} \leq \frac{\log x}{x-1} \leq \frac{1}{1 \wedge x}$, we get, on $\Omega_{m_f}(\epsilon)$

$$\frac{1}{(1+\epsilon)^2} \chi_m^2 \leq V_m^2 \leq \frac{1}{(1-\epsilon)^2} \chi_m^2. \quad (2.20)$$

So

$$\frac{1-\epsilon}{2(1+\epsilon)^2} \chi_m^2 \mathbf{1}_{\Omega_{m_f}(\epsilon)} \leq K(\bar{s}_m, \hat{s}_m) \mathbf{1}_{\Omega_{m_f}(\epsilon)} \leq \frac{1+\epsilon}{2(1-\epsilon)^2} \chi_m^2 \mathbf{1}_{\Omega_{m_f}(\epsilon)}.$$

On one hand, $\mathbf{E}[\chi_m^2] = |m|$, and

$$\frac{1-\epsilon}{2(1+\epsilon)^2} |m| - \mathbf{E} \left[\chi_m^2 \mathbf{1}_{\Omega_{m_f}(\epsilon)^c} \right] \leq \mathbf{E} \left[K(\bar{s}_m, \hat{s}_m) \mathbf{1}_{\Omega_{m_f}(\epsilon)} \right] \leq \frac{1+\epsilon}{2(1-\epsilon)^2} |m|.$$

Since $\chi_m^2 \leq \frac{1}{\Gamma(\log(n))^2 \rho_{\min}} \sum_{J \in m} (Y_J - \lambda_J)^2 \leq \frac{1}{\Gamma(\log(n))^2 \rho_{\min}} (\sum_t Y_t - \sum_t \lambda_t)^2$, using Cauchy-Schwarz Inequality, we get

$$\begin{aligned}
\mathbf{E} \left[\chi_m^2 \mathbf{1}_{\Omega_{m_f}(\epsilon)^C} \right] &\leq \frac{1}{\Gamma(\log(n))^2 \rho_{min}} \left[3 \left(\sum_t \lambda_t \right)^2 + \sum_t \lambda_t \right]^{1/2} P(\Omega_{m_f}(\epsilon)^C)^{1/2} \\
&\leq C(\Gamma, \rho_{min}, \rho_{max}) \frac{n}{(\log(n))^2} P(\Omega_{m_f}(\epsilon)^C)^{1/2} \\
&\leq C(\Gamma, \rho_{min}, \rho_{max}) n^\alpha P(\Omega_{m_f}(\epsilon)^C)^{1/2} \\
&\leq \frac{C(\phi, \Gamma, \rho_{min}, \rho_{max}, \varepsilon, a)}{n^{\alpha/2-\alpha}},
\end{aligned}$$

where $\alpha = 1 - 2 \frac{\log(\log(n))}{\log(n)}$, $n \geq 2$. For example, $\alpha = 0.62$ for $n = 10^6$.

On the other hand, using $\log 1/x \geq 1 - x$ for all $x > 0$, $\mathbf{E} \left[K(\bar{s}_m, \hat{s}_m) \mathbf{1}_{\Omega_{m_f}(\epsilon)^C} \right] \geq 0$.

Finally, we have

$$K(s, \bar{s}_m) + \frac{1 - \varepsilon}{2(1 + \varepsilon)^2} |m| - \frac{C_1(\Gamma, \rho_{min}, \rho_{max}, \varepsilon, a)}{n^{\alpha/2-\alpha}} \leq \mathbf{E}[K(s, \hat{s}_m)],$$

Negative binomial case

We have

- $K(\bar{s}_m, \hat{s}_m) = \phi \sum_{J \in m} \frac{|J|}{p_J} h_{\frac{\phi}{\phi + \bar{Y}_J}}(p_J)$, and
- $\forall 0 < a < 1, \quad h_a(x) \geq \frac{1-x}{1-a} \log^2 \left(\frac{1-x}{1-a} \right)$.

Then on $\Omega_{m_f}(\epsilon)$

$$K(\bar{s}_m, \hat{s}_m) \geq \phi \sum_{J \in m} \frac{|J|}{p_J} \frac{1-p_J}{\frac{\bar{Y}_J}{\phi + \bar{Y}_J}} \log^2 \left(\frac{\frac{\bar{Y}_J}{\phi + \bar{Y}_J}}{1-p_J} \right).$$

Introducing

$$V_m^2 = \sum_{J \in m} \phi |J| \frac{1-p_J}{p_J} \log^2 \left(\frac{\frac{\bar{Y}_J}{\phi + \bar{Y}_J}}{1-p_J} \right), \quad (2.21)$$

we get

$$K(\bar{s}_m, \hat{s}_m) \geq V_m^2, \quad (2.22)$$

and since $\bar{Y}_J - \phi \frac{1-p_J}{p_J} = \frac{\phi + \bar{Y}_J}{p_J} \left(\frac{\bar{Y}_J}{\phi + \bar{Y}_J} - (1-p_J) \right)$, we have

$$V_m^2 = \sum_{J \in m} |J| \left(\frac{\phi}{\phi + \bar{Y}_J} \right)^2 \frac{\left(\bar{Y}_J - \phi \frac{1-p_J}{p_J} \right)^2}{\phi \frac{1-p_J}{p_J}} \left[\frac{\log \left(\frac{\frac{\bar{Y}_J}{\phi + \bar{Y}_J}}{1-p_J} \right)}{\frac{\frac{\bar{Y}_J}{\phi + \bar{Y}_J}}{1-p_J} - 1} \right]^2.$$

And finally,

$$K(\bar{s}_m, \hat{s}_m) \mathbf{1}_{\Omega_{m_f}(\epsilon)} \geq \rho_{min}^2 \frac{(1-\epsilon)^2}{(1+\epsilon)^4} \chi_m^2 \mathbf{1}_{\Omega_{m_f}(\epsilon)}.$$

Moreover, on one hand we have $|m| \leq \mathbf{E}[\chi_m^2] \leq \frac{1}{\rho_{min}} |m|$. On the other hand, since $\chi_m^2 \leq \frac{1}{\Gamma(\log(n))^2 \phi(1-\rho_{max})} (\sum_t Y_t - \sum_t E_t)^2$, using Cauchy-Schwarz Inequality, we get

$$\begin{aligned} \mathbf{E} \left[\chi_m^2 \mathbf{1}_{\Omega_{m_f}(\epsilon)^C} \right] &\leq \frac{\left[\sum_t \mathbf{E} (Y_t - E_t)^4 + 6\phi^2 \sum_{(t,l), l \neq t} \frac{1-p_t}{p_t^2} \frac{1-p_l}{p_l^2} \right]^{1/2}}{\Gamma(\log(n))^2 \phi(1-\rho_{max})} P(\Omega_{m_f}(\epsilon)^C)^{1/2}, \\ &\leq C(\Gamma, \rho_{min}, \rho_{max}) n^\alpha P(\Omega_{m_f}(\epsilon)^C)^{1/2}, \\ &\leq \frac{C(\phi, \Gamma, \rho_{min}, \rho_{max}, \epsilon, a)}{n^{\alpha/2 - \alpha}}, \end{aligned}$$

where $\alpha = 1 - 2 \frac{\log(\log(n))}{\log(n)}$, $n \geq 2$. Finally, we have

$$K(s, \bar{s}_m) + \rho_{min}^2 \frac{(1 - \varepsilon)^2}{(1 + \varepsilon)^4} |m| - \frac{C(\phi, \Gamma, \rho_{min}, \rho_{max}, \varepsilon, a)}{n^{a/2 - \alpha}} \leq \mathbf{E}[K(s, \hat{s}_m)].$$

Proof of proposition 2.2.8

Poisson case

The term to be controlled is $\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'}) = \sum_{J \in m'} |J| (\bar{Y}_J - \bar{\lambda}_J) \log \frac{\bar{Y}_J}{\bar{\lambda}_J}$.

Using Cauchy-Schwarz inequality, we have

$$\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'}) \leq \sqrt{\chi_{m'}^2} \sqrt{V_{m'}^2},$$

with $\chi_{m'}^2$ and $V_{m'}^2$ defined as in equations (2.9) and (2.19). Then, using equation (2.18)

$$(\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'})) \mathbf{1}_{\Omega_{m_f}(\varepsilon)} \leq \sqrt{\chi_{m'}^2} \sqrt{\frac{2}{1 - \varepsilon} K(\bar{s}_{m'}, \hat{s}_{m'})},$$

and using $2ab \leq \kappa a^2 + \kappa^{-1} b^2$ for all $\kappa > 0$, we get

$$(\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'})) \mathbf{1}_{\Omega_{m_f}(\varepsilon)} \leq \frac{\kappa}{2} \chi_{m'}^2 + \frac{\kappa^{-1}}{1 - \varepsilon} K(\bar{s}_{m'}, \hat{s}_{m'}). \quad (2.23)$$

And with proposition 2.2.7, we get, for $\kappa = \frac{1 + \varepsilon}{1 - \varepsilon} = 2C(\varepsilon)$,

$$\begin{aligned} & (\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'})) \mathbf{1}_{\Omega_{m_f}(\varepsilon) \cap \Omega_1(\xi)} \\ & \leq \frac{1 + \varepsilon}{2(1 - \varepsilon)} \left[|m'| + 8(1 + \varepsilon) \sqrt{(L_{m'} |m'| + \xi) |m'|} + 4(1 + \varepsilon)(L_{m'} |m'| + \xi) \right] \\ & \quad + \frac{1}{1 + \varepsilon} K(\bar{s}_{m'}, \hat{s}_{m'}). \end{aligned}$$

Negative binomial case

In this case we can write $\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'}) = \sum_{J \in m'} |J| (\bar{Y}_J - \bar{E}_J) \log \frac{\bar{Y}_J}{1 - p_J}$. Again, using Cauchy-Schwarz inequality, and with χ_m^2 and V_m^2 defined by equations (2.9) and (2.21), we

get

$$\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'}) \leq \sqrt{\chi_{m'}^2} \sqrt{V_{m'}^2},$$

so that with equation (2.22) and $2ab \leq \kappa a^2 + \kappa^{-1}b^2$ for all $\kappa > 0$

$$(\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'})) \mathbf{1}_{\Omega_{m_f}(\epsilon)} \leq \frac{\kappa}{2} \chi_{m'}^2 + \frac{\kappa^{-1}}{2} K(\bar{s}_{m'}, \hat{s}_{m'}). \quad (2.24)$$

Finally, with proposition 2.2.7 and $\kappa = \frac{1+\epsilon}{2} = 2C(\epsilon)$,

$$\begin{aligned} & (\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'})) \mathbf{1}_{\Omega_{m_f}(\epsilon) \cap \Omega_1(\xi)} \\ & \leq C(\rho_{min}) \frac{1+\epsilon}{4} \left[|m'| + 8(1+\epsilon) \sqrt{(L_{m'}|m'| + \xi)|m'|} + 4(1+\epsilon)(L_{m'}|m'| + \xi) \right] \\ & \quad + \frac{1}{1+\epsilon} K(\bar{s}_{m'}, \hat{s}_{m'}). \end{aligned}$$

Proof of proposition 2.2.9

Poisson case

Noting that $\mathbf{E}[(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s)) \mathbf{1}_{\Omega_{m_f}(\epsilon)}] = -\mathbf{E}[(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s)) \mathbf{1}_{\Omega_{m_f}(\epsilon)^C}]$,

we have

$$\begin{aligned} & |\mathbf{E}[(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s)) \mathbf{1}_{\Omega_{m_f}(\epsilon)}]| \\ & \leq |\mathbf{E}[(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s)) \mathbf{1}_{\Omega_{m_f}(\epsilon)^C}]| \leq \mathbf{E}[|(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))| \mathbf{1}_{\Omega_{m_f}(\epsilon)^C}] \\ & \leq \mathbf{E} \left[\left| \left(\sum_J \sum_t (Y_t - E_t) \log(\rho_{max}/\rho_{min}) \right) \right| \mathbf{1}_{\Omega_{m_f}(\epsilon)^C} \right] \\ & \leq \log(\rho_{max}/\rho_{min}) \times \mathbf{E} \left[\left| \sum_t (Y_t - E_t) \right| \mathbf{1}_{\Omega_{m_f}(\epsilon)^C} \right] \\ & \leq \log(\rho_{max}/\rho_{min}) \times \left(\left[\mathbf{E} \left(\sum_t (Y_t - E_t) \right)^2 \right]^{1/2} \times (P(\Omega_{m_f}(\epsilon)^C))^{1/2} \right) \\ & \leq (n\rho_{max})^{1/2} \times \log(\rho_{max}/\rho_{min}) \times (P(\Omega_{m_f}(\epsilon)^C))^{1/2}, \end{aligned}$$

which concludes the proof.

Negative binomial case.

Once again, $\mathbf{E}[(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))\mathbf{1}_{\Omega_{m_f}(\epsilon)}] = -\mathbf{E}[(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))\mathbf{1}_{\Omega_{m_f}(\epsilon)^C}]$, and

$$\begin{aligned} & |\mathbf{E}[(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))\mathbf{1}_{\Omega_{m_f}(\epsilon)}]| \\ & \leq |\mathbf{E}[(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))\mathbf{1}_{\Omega_{m_f}(\epsilon)^C}]| \leq \mathbf{E}[|(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))|\mathbf{1}_{\Omega_{m_f}(\epsilon)^C}] \\ & \leq \mathbf{E}\left[\left|\left(\sum_J \sum_t \left(Y_t - \phi \frac{1-p_t}{p_t}\right) \log(1/(1-\rho_{min}))\right)\right|\mathbf{1}_{\Omega_{m_f}(\epsilon)^C}\right] \\ & \leq \log(1/(1-\rho_{min})) \times \mathbf{E}\left[\left|\sum_t (Y_t - E_t)\right|\mathbf{1}_{\Omega_{m_f}(\epsilon)^C}\right] \\ & \leq \left(n\phi \frac{1}{\rho_{min}^2}\right)^{1/2} \times \log \frac{1}{1-\rho_{min}} \times (P(\Omega_{m_f}(\epsilon)^C))^{1/2} \end{aligned}$$

which concludes the proof.

Proof of proposition 2.2.10

Using the Markov inequality $\mathbf{P}[\bar{\gamma}(s) - \bar{\gamma}(u) \geq b] \leq \inf_a \left[e^{-ab} \mathbf{E} \left(e^{a(\bar{\gamma}(s) - \bar{\gamma}(u))} \right) \right]$ with $a = \frac{1}{2}$, we get

$$\begin{aligned} & \mathbf{P}[\bar{\gamma}(s) - \bar{\gamma}(u) \geq b] \leq \\ & \leq \exp \left[-\frac{b}{2} + \log \mathbf{E} \left[\exp \left(\frac{1}{2} (\gamma(s) - \gamma(u)) + \frac{1}{2} \mathbf{E}[\gamma(u) - \gamma(s)] \right) \right] \right] \\ & \leq \exp \left[-\frac{b}{2} + \frac{1}{2} K(s, u) + \log \mathbf{E} \left[\exp \left(-\frac{1}{2} \sum_t \log \mathbf{P}_s(X_t = Y_t) + \log \mathbf{P}_u(X_t = Y_t) \right) \right] \right] \\ & \leq \exp \left[-\frac{b}{2} + \frac{1}{2} K(s, u) + \sum_t \log \mathbf{E} \sqrt{\frac{\mathbf{P}_u(X_t = Y_t)}{\mathbf{P}_s(X_t = Y_t)}} \right] \\ & \leq \exp \left[-\frac{b}{2} + \frac{1}{2} K(s, u) + \sum_t \mathbf{E} \sqrt{\frac{\mathbf{P}_u(X_t = Y_t)}{\mathbf{P}_s(X_t = Y_t)}} - n \right] \\ & \leq \exp \left[-\frac{b}{2} + \frac{1}{2} K(s, u) - h^2(s, u) \right] \end{aligned}$$

where $\mathbf{P}_s = \mathbf{P}$ denote the probability under the distribution s . Thus

$$\mathbf{P}[\bar{\gamma}(s) - \bar{\gamma}(u) \geq K(s, u) - 2h^2(s, u) + 2x] \leq e^{-x}.$$

Acknowledgement The authors wish to thank Stéphane Robin for more than helpful discussions on the statistical aspect and Gavin Sherlock for his insight on the biological applications.

References

H. Akaike. Information theory and extension of the maximum likelihood principle. *Second international symposium on information theory*, pages 267–281, 1973.

N. Akakpo. Estimating a discrete distribution via histogram selection. *ESAIM Probab. Statist.*, To appear, 2009.

Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *The Journal of Machine Learning Research*, 10:245–279, 2009.

Y. Baraud and L. Birgé. Estimating the intensity of a random measure by histogram type estimators. *Probab. Theory Related Fields*, 143(1-2):239–284, 2009. ISSN 0178-8051.

A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999. ISSN 0178-8051.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *EEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, 2000. ISSN 01628828.

L. Birgé. Model selection for Poisson processes. In *Asymptotics: particles, processes and inverse problems*, volume 55 of IMS Lecture Notes Monogr. Ser., pages 32–64. Inst. Math. Statist., Beachwood, OH, 2007.

L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001. ISSN 1435-9855.

L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007. ISSN 0178-8051.

Jerome V Braun and Hans-Georg Muller. Statistical methods for DNA sequence segmentation. *Statistical Science*, pages 142–162, 1998.

Jerome V Braun, RK Braun, and Hans-Georg Müller. Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika*, 87(2):301–314, 2000.

Breiman, Friedman, Olshen, and Stone. Classification and regression trees. *Wadsworth and Brooks*, 1984.

G Castellan. Modified Akaike’s criterion for histogram density estimation. *preprint*, 99,

1999.

Alice Cleynen, Michel Koskas, and Guillem Rigai. A generic implementation of the pruned dynamic programming algorithm. *Arxiv preprint arXiv:1204.5564*, under review.

N. Johnson, A.W. Kemp, and S. Kotz. Univariate discrete distributions. *John Wiley & Sons, Inc.*, 2005.

Rebecca Killick and Idris A Eckley. Changepoint: an R package for changepoint analysis. Lancaster University, 2011.

E. Lebarbier. Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, 85(4):717–736, April 2005. ISSN 0165-1684.

T.M. Luong, Y. Rozenholc, and G. Nuel. Fast estimation of posterior probabilities in change-point models through a constrained hidden Markov model. *Arxiv preprint arXiv:1203.4394*, 2012.

Pascal Massart. Concentration inequalities and model selection. 2007.

P. Reynaud-Bouret. Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Related Fields*, 126(1):103–153, 2003. ISSN 0178-8051.

G. Rigai. Pruned dynamic programming for optimal multiple change-point detection. *Arxiv:1004.0887*, April 2010. URL <http://arxiv.org/abs/1004.0887>.

G Rigai, E Lebarbier, and S Robin. Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and Computing*, 22(4):917–929, 2012.

Davide Risso, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, 12(1):480, 2011.

Y.-C. Yao. Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters*, 6(3):181–189, February 1988.

Nancy R Zhang and David O Siegmund. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32, 2007.

2.2.7 Discussion and perspectives

Our work shows that a penalty of the form

$$\text{pen}(m) = \beta K \left(1 + 4\sqrt{1.1 + \log\left(\frac{n}{K}\right)} \right)^2,$$

gives the following oracle inequality:

$$\mathbf{E} \left[h^2(s, \hat{s}_m) \right] \leq C_1 \log(n) \inf_{m \in \mathcal{M}_n} \{ \mathbf{E}[K(s, \hat{s}_m)] \} + C_2.$$

This calls for a few remarks.

- First, the penalty term can be interpreted as the sum of two terms. The first is a term proportional to K which is common to all penalized likelihood approach. The second however, is a term in $K \log \frac{n}{K}$, which is specific to and recurrently encountered in non-asymptotic approaches. This term is induced by the size of the collection of models to explore and was identified in BIRGÉ and MASSART (2001) as a minimizer of the minimax risk in the case of Gaussian-distributed data. Its appearance in the penalty function is therefore reassuring: indeed, it was shown (see for instance LEBARBIER (2005); BIRGÉ and MASSART (2007)) that the penalty term needs to depend on the number of models having the same dimension to provide adequate results.
- Second, the oracle inequality is one on the Hellinger distance and not on the Kullback-Leibler divergence as we would ideally wish to obtain. This is because in our framework we have no warranty that the quantities s and \bar{s}_m will remain bounded unless we further specify constraints on the true distribution s . Controlling the Hellinger distance is a common trick to avoid such unpleasant restrictions (CASTELLAN, 1999; MASSART, 2007), and it is easily related to the Kullback divergence when the later is finite.
- Third, the exact same penalty shape and oracle inequalities are obtained for Poisson and negative binomial distributed random variables. This result allows us to think

that it could be generalized to the larger exponential family of distributions. We describe here the essential steps towards such generalization for one-parameter distributions from the exponential family which sufficient statistic is the identity, and underline the aspects which still require some refinement. The left column of Table 2.2 gives a list of distributions which fit in our generalization.

In this context, the true distribution s is of the form

$$s(t) = h(Y_t) \exp [\theta_t Y_t - A(\theta_t)],$$

and the loss is

$$\gamma(s) = \sum_{t=1}^n -\theta_t Y_t + A(\theta_t).$$

The optimal estimator and the projection on a partition m are respectively, for $t \in J$,

$$\begin{aligned} \hat{s}_m(t) &= h(Y_t) \exp \left[A'^{-1}(\bar{Y}_J) Y_t - A \left[A'^{-1}(\bar{Y}_J) \right] \right], \text{ and} \\ \bar{s}_m(t) &= h(Y_t) \exp \left[A'^{-1}(\bar{E}_J) Y_t - A \left[A'^{-1}(\bar{E}_J) \right] \right] \end{aligned}$$

where $E_t = A'(\theta_t)$, $E_J = \sum_{t \in J} E_t$ and $\bar{E}_J = E_J/|J|$.

Exactly as in the Poisson and negative binomial cases, the following assumptions are made:

- $\forall t, \theta_{min} \leq \theta_t \leq \theta_{max}$, and
- $\forall J \in m_f, |J| \geq \Gamma(\log(n))^2$

Note that imposing a constraint on θ_t naturally implies that the expectation and variance of the variables are bounded. Indeed, A' is a non-decreasing function and A'' is continuous, so that $A'(\theta_{min}) \leq E_t \leq A'(\theta_{max})$ and $Var(Y_t) = A''(\theta_t)$ is bounded on one side by $A''(\theta_{min})$ and on the other side by $A''(\theta_{max})$.

Applying the same technique as for the negative binomial and Poisson distributions, we would have to control

1. $K(\bar{s}_m, \hat{s}_m)$,
2. $\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'})$, and

$$3. \bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s),$$

(since the last term $\bar{\gamma}(s) - \bar{\gamma}(\bar{s}_{m'})$ is controlled independently of the choice of distribution for s). However, as in the previous case, an essential tool should be an exponential bound of Y_J around its expectation. If we want to apply the Baraud and Birgé Lemma (BARAUD and BIRGÉ, 2009) as in our paper, we need to control the Laplace transform of Y . By definition of the cumulant-generating function of Y_t , we have :

$$\begin{aligned} \log \mathbf{E} \left(e^{z(Y_t - E_t)} \right) &= -zE_t + \sum_{n \geq 1} \frac{\kappa_n}{n!} z^n \\ &= \frac{z^2}{2} \sum_{n \geq 0} \frac{2\kappa_{n+2}}{(n+2)!} z^n. \end{aligned}$$

Provided that the sequence $2\kappa_n/n!A'(\theta_t)$, $n \geq 2$ is bounded by a universal constant $\kappa_{max} \geq 1$ for all $\theta_{min} \leq \theta_t \leq \theta_{max}$, we get, on $-1 \leq z < 1$,

$$\log \mathbf{E} \left(e^{z(Y_t - E_t)} \right) \leq E_t \frac{z^2}{2} \frac{\kappa_{max}}{1-z}.$$

We finally obtain the controls required for the Baraud and Birgé Lemma, namely

$$\log \mathbf{E} \left(e^{z(Y_t - E_t)} \right) \leq \begin{cases} E_t \frac{z^2}{2(1-z)} \kappa_{max} & \text{if } 1 > z \geq 0 \\ E_t \frac{z^2}{2} \kappa_{max} & \text{if } -1 < z \leq 0 \end{cases}.$$

Note that constraint $\kappa_n/n!A'(\theta_t) \leq \kappa_{max}$ can seem quite strong and could probably be refined. In practice, some classic distributions lack this property, such as the geometric for which $\kappa_r < r! \frac{(1-p)}{p^r}$, and some constraints on the range of possible z have to be added (we consider for instance $z < p$ for the geometric and negative binomial distributions). This is not an issue as long as one can define a set $-c < z < c$ with $c > 0$ for which the bounds required by the Baraud and Birgé lemma are verified. We give in the last column of Table 2.2 the values of c and κ_{max} obtained for the classical distributions of the exponential family.

We can now apply the Baraud and Birgé Lemma to obtain

$$\mathbb{P} \left[Y_J - E_J \geq \sqrt{2x\kappa_{max}E_J + \kappa_{max}x} \right] \leq e^{-x}, \text{ and}$$

$$\mathbb{P}[|Y_J - E_J| \geq x] \leq 2e^{-\frac{x^2}{2\kappa_{max}(E_J+x)}}.$$

From there we can define $\Omega_{m_f}(\epsilon) = \bigcap_{J \in m_f} \left\{ \left| \frac{Y_J}{E_J} - 1 \right| \leq \epsilon \right\}$ which is of large probability.

We now consider the three entities which we want to control:

1. The ultimate goal is to obtain a lower bound on the risk of one model $\mathbb{E}[K(s, \hat{s}_m)]$. For this purpose, we use the classical decomposition $K(s, \bar{s}_m) + K(\bar{s}_m, \hat{s}_m)$, and by using Taylor's decomposition we can write

$$\begin{aligned} K(\bar{s}_m, \hat{s}_m) &= \sum_J |J| \left[\bar{E}_J \left(A'^{-1}(\bar{E}_J) - A'^{-1}(\bar{Y}_J) \right) + A \left(A'^{-1}(\bar{Y}_J) \right) - A \left(A'^{-1}(\bar{E}_J) \right) \right] \\ &= \sum_J |J| \bar{E}_J \left[\left(A'^{-1}(\bar{E}_J) - A'^{-1}(\bar{Y}_J) \right)^2 \frac{A''(z)}{2\bar{E}_J} \right] \end{aligned}$$

where z is a real number between $A'^{-1}(\bar{E}_J)$ and $A'^{-1}(\bar{Y}_J)$.

Now because A' is non-decreasing, A'' is continuous, and considering the set $\Omega_{m_f}(\epsilon)$, it naturally comes

$$\frac{A''(\tau_{min}(\epsilon))}{2 A'(\theta_{max})} V_m^2 \mathbf{1}_{\Omega_{m_f}(\epsilon)} \leq K(\bar{s}_m, \hat{s}_m) \mathbf{1}_{\Omega_{m_f}(\epsilon)} \leq \frac{A''(\tau_{max}(\epsilon))}{2 A'(\theta_{min})} V_m^2 \mathbf{1}_{\Omega_{m_f}(\epsilon)},$$

where we denote

- $V_m^2 = \sum_J |J| \bar{E}_J \left[A'^{-1}(\bar{E}_J) - A'^{-1}(\bar{Y}_J) \right]^2$,
- $\tau_{min}(\epsilon) = \arg \min \{ A''(A'^{-1}[(1 - \epsilon)A'(\theta_{min})]), A''(A'^{-1}[(1 + \epsilon)A'(\theta_{max})]) \}$, and
- $\tau_{max}(\epsilon) = \arg \max \{ A''(A'^{-1}[(1 - \epsilon)A'(\theta_{min})]), A''(A'^{-1}[(1 + \epsilon)A'(\theta_{max})]) \}$.

The goal is then to link $K(\bar{s}_m, \hat{s}_m)$ to the χ_m^2 term, which is defined as in the Poisson case by $\chi_m^2 = \sum_{J \in m} |J| \left(\bar{Y}_J - \bar{E}_J \right)^2 / \bar{E}_J$.

Applying Taylor's decomposition to A'^{-1} , we get

$$K(\bar{s}_m, \hat{s}_m) = \sum_{J \in m} |J| \frac{(\bar{Y}_J - \bar{E}_J)^2}{\bar{E}_J} \left[\left[(A'^{-1})'(c) \right]^2 \bar{E}_J A''(z) / 2 \right]$$

for a real number c between \bar{Y}_J and \bar{E}_J . Considering once more the set $\Omega_{m_f}(\epsilon)$ yields

$$\frac{A'(\theta_{min})}{2} \frac{A''(\tau_{min}(\epsilon))}{A''(\tau_{max}(\epsilon))^2} \chi_m^2 \mathbf{1}_{\Omega_{m_f}(\epsilon)} \leq K(\bar{s}_m, \hat{s}_m) \mathbf{1}_{\Omega_{m_f}(\epsilon)} \leq \frac{A'(\theta_{max})}{2} \frac{A''(\tau_{max}(\epsilon))}{A''(\tau_{min}(\epsilon))^2} \chi_m^2 \mathbf{1}_{\Omega_{m_f}(\epsilon)}.$$

Now we have

$$|m| \frac{A''(\tau_{min}(0))}{A'(\theta_{max})} \leq \mathbf{E}[\chi_m^2] \leq |m| \frac{A''(\tau_{max}(0))}{A'(\theta_{min})}, \text{ and thus}$$

$$\frac{A'(\theta_{min})}{A'(\theta_{max})} \frac{A''(\tau_{min}(\varepsilon))A''(\tau_{min}(0))}{2A''(\tau_{max}(\varepsilon))^2} |m| - \mathbf{E} \left[\chi_m^2 \mathbf{1}_{\Omega_{m_f}(\varepsilon)^C} \right] \leq \mathbf{E} \left[K(\bar{s}_m, \hat{s}_m) \mathbf{1}_{\Omega_{m_f}(\varepsilon)} \right].$$

It then remains to control $\mathbf{E} \left[\chi_m^2 \mathbf{1}_{\Omega_{m_f}(\varepsilon)^C} \right]$. Since $\chi_m^2 \leq \frac{[\sum_t (Y_t - E_t)]^2}{\Gamma(\log n)^2 A'(\theta_{min})}$, using Cauchy-Schwarz inequality we get

$$\begin{aligned} \mathbf{E} \left[\chi_m^2 \mathbf{1}_{\Omega_{m_f}(\varepsilon)^C} \right] &\leq \frac{\left[\sum_t \mathbf{E} [Y_t - E_t]^4 + 6 \sum_{(t,l), l \neq t} \mathbf{E} [Y_t - E_t]^2 \mathbf{E} [Y_l - E_l]^2 \right]^{1/2}}{\Gamma(\log n)^2 A'(\theta_{min})} P(\Omega_{m_f}(\varepsilon)^C)^{1/2} \\ &\leq \frac{\left[\sum_t (\kappa_{4,t} + 3\text{Var}(Y_t)^2) + 6 \sum_{(t,l), l \neq t} \text{Var}(Y_t) \text{Var}(Y_l) \right]^{1/2}}{\Gamma(\log n)^2 A'(\theta_{min})} P(\Omega_{m_f}(\varepsilon)^C)^{1/2} \\ &\leq \frac{\left[12n\kappa_{max} A'(\theta_{max}) + 6n^2 A''(\tau_{max}(0))^2 \right]^{1/2}}{\Gamma(\log n)^2 A'(\theta_{min})} P(\Omega_{m_f}(\varepsilon)^C)^{1/2} \\ &\leq \frac{C(\Gamma, \theta_{min}, \theta_{max}, \kappa_{max}, \varepsilon, a)}{n^{a/2-\alpha}} \end{aligned}$$

with $0 < \alpha < 1$. Finally, since $\mathbf{E} \left[K(\bar{s}_m, \hat{s}_m) \mathbf{1}_{\Omega_{m_f}(\varepsilon)^C} \right] \geq 0$, we can conclude that

$$K(s, \bar{s}_m) + \frac{A'(\theta_{min})}{A'(\theta_{max})} \frac{A''(\tau_{min}(\varepsilon))A''(\tau_{min}(0))}{2A''(\tau_{max}(\varepsilon))^2} |m| - \frac{C(\Gamma, \theta_{min}, \theta_{max}, \kappa_{max}, \varepsilon, a)}{n^{a/2-\alpha}} \leq \mathbf{E} \left[K(\bar{s}_m, \hat{s}_m) \right].$$

2. We have $\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'}) = \sum_{J \in m'} |J| (E_J - Y_J) (A'^{-1}(\bar{E}_J) - A'^{-1}(\bar{Y}_J))$ which can be bounded by $\sqrt{\chi_{m'}^2} \sqrt{V_{m'}^2}$. By the above, we get, for any positive κ ,

$$\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'}) \leq \frac{\kappa}{2} \chi_{m'}^2 + \frac{\kappa^{-1} A'(\theta_{max})}{A''(\tau_{min}(\varepsilon))} K(\bar{s}_{m'}, \hat{s}_{m'}).$$

We can control the χ^2 term using Bernstein's inequality combined with the exponential bound on Y_J , to obtain:

$$\mathbb{P} \left[\chi_m^2 \mathbf{1}_{\Omega_{m_f}(\varepsilon)} \geq |m| \frac{A''(\tau_{max}(0))}{A'(\theta_{min})} + 8\kappa_{max} (1 + \varepsilon) \sqrt{x|m|} + 4\kappa_{max} (1 + \varepsilon) x \right] \leq e^{-x}.$$

3. Finally, the term $\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s) = \sum_{J \in m} \sum_{t \in J} (Y_t - E_t) (\theta_t - A'^{-1}(\bar{E}_J))$ can be controlled using Cauchy-Schwarz inequality as in the paper, yielding

$$|\mathbf{E}[(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s)) \mathbf{1}_{\Omega_{m_f}(\varepsilon)}]| \leq \left(n A''(\tau_{max}(0)) \right)^{1/2} \cdot (\theta_{max} - \theta_{min}) \cdot (P(\Omega_{m_f}(\varepsilon)^C))^{1/2}.$$

Now integrating each term together in the inequality

$$\begin{aligned} K(s, \hat{s}_{m'}) &\leq K(s, \bar{s}_m) - \text{pen}(m') + \text{pen}(m) \\ &\quad + (\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'})) + (\bar{\gamma}(s) - \bar{\gamma}(\bar{s}_{m'})) + (\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s)) \end{aligned}$$

gives, for $m' = \hat{m}$,

$$\begin{aligned} \frac{\varepsilon}{1+\varepsilon} h^2(s, \hat{s}_{\hat{m}}) \mathbf{1}_{\Omega(\varepsilon, \xi)} &\leq K(s, \bar{s}_m) \mathbf{1}_{\Omega(\varepsilon, \xi)} + R \mathbf{1}_{\Omega(\varepsilon, \xi)} - \text{pen}(m') + \text{pen}(m) \\ &\quad + |\hat{m}| C_a (1 + \varepsilon) \left[1 + \frac{A'(\theta_{\min}) \kappa_{\max}}{A''(\tau_{\max}(0))} (1 + \varepsilon) \left(8\sqrt{L_{\hat{m}}} + \varepsilon + 4L_{\hat{m}} \right) \right] + 2L_{\hat{m}} |\hat{m}| \\ &\quad + 2\xi \left[1 + C_a (1 + \varepsilon) \frac{A'(\theta_{\min}) \kappa_{\max}}{A''(\tau_{\max}(0))} (1 + \varepsilon) \left(8\frac{2}{\varepsilon} + 4 \right) \right], \end{aligned}$$

with $C_a = \frac{1}{2} \frac{A''(\tau_{\max}(0))}{A''(\tau_{\min}(\varepsilon))} \frac{A'(\theta_{\max})}{A'(\theta_{\min})}$.

Then by some simple bounding we get

$$C_a (1 + \varepsilon) \left[1 + \frac{A'(\theta_{\min}) \kappa_{\max}}{A''(\tau_{\max}(0))} (1 + \varepsilon) \left(8\sqrt{L_{\hat{m}}} + \varepsilon + 4L_{\hat{m}} \right) \right] + 2L_{\hat{m}} \leq \beta \left(1 + 4\sqrt{L_{\hat{m}}} \right)^2$$

and conclude taking $\text{pen}(m) \geq \beta \left(1 + 4\sqrt{L_{\hat{m}}} \right)^2$. As always in such approaches, the penalty function depends on the family \mathcal{M}_n through the choice of the weights L_m . In the segmentation context where we conduct an exhaustive search of the possible models, choosing the the classical $L_K = 1 + \kappa + \log \left(\frac{n}{K} \right)$ for all partition m in K segments results in obtaining the exact same penalty shape as in the Poisson and negative binomial cases, namely

$$\text{pen}(m) = \beta K \left(1 + 4\sqrt{1.1 + \log \left(\frac{n}{K} \right)} \right)^2.$$

The end of the argument follows as in the negative binomial and Poisson cases, and we obtain successively

$$\begin{aligned} \mathbf{E} \left[h^2(s, \hat{s}_{\hat{m}}) \right] &\leq C_\beta^1 \inf_{m \in \mathcal{M}_n} \{ K(s, \bar{s}_m) + \text{pen}(m) \} + C^2(\Gamma, \theta_{\min}, \theta_{\max}, \kappa_{\max}, \beta, \Sigma), \quad \text{and} \\ \mathbf{E} \left[h^2(s, \hat{s}_{\hat{m}}) \right] &\leq C_\beta \log(n) \inf_{m \in \mathcal{M}_n} \{ \mathbf{E}[K(s, \hat{s}_m)] \} + C(\Gamma, \theta_{\min}, \theta_{\max}, \kappa_{\max}, \beta, \Sigma). \end{aligned}$$

Note that in this context the constant β depends on the constraints imposed on the distribution s , such as θ_{min} , θ_{max} , κ_{max} , etc. In practice this is not a major drawback since β is tuned according to the data using the slope heuristic, so that these constraints are taken into account. Moreover the constants C_a and $\frac{A'(\theta_{min})}{A''(\tau_{max}(0))}$ can be computed explicitly when the distribution is specified. However in some cases performing the computations directly might give more precise results. For instance in the Poisson case, one will obtain $\mathbf{E}[\chi_m^2] = |m|$ directly instead of using the bounds $\frac{\lambda_{min}}{\lambda_{max}}|m| \leq \mathbf{E}[\chi_m^2] = \frac{\lambda_{max}}{\lambda_{min}}|m|$. For this reason in Table 2.2 we give examples of distributions which fit in our framework as well as the constraint κ_{max} defined earlier, but not the values of the other constant obtained.

distribution	fixed parameter	θ	c	κ_{max}
$\mathcal{N}(\mu, \sigma^2)$	σ^2	μ/σ^2	1	σ^2/μ_{min}
$\mathcal{E}(\lambda)$		$-\lambda$	1	1*
			λ_{min}	λ_{min}^{-1}
$\mathcal{P}(\lambda)$		$\log(\lambda)$	1	1
$\mathcal{B}(p)$		$\log\left(\frac{p}{1-p}\right)$	1	1
$\mathcal{B}(m, p)$	m	$\log\left(\frac{p}{1-p}\right)$	1	1
$\mathcal{G}(p)$		$\log(1-p)$	p_{min}	$2/p_{min}$
$\mathcal{NB}(p, \phi)$	ϕ	$\log(1-p)$	p_{min}	$2/p_{min}$
$\mathcal{Gam}(\alpha, \beta)$	α	$-\beta$	1	1*
			β_{min}	β_{min}^{-1}

Table 2.2: **Distributions from the exponential family and characteristics.** For the exponential and gamma distribution, we distinguish the cases where λ_{min} (resp. β_{min}) is larger than 1 (denoted by *) from the general case.

2.3 Constrained HMM approach

We now propose a second method for the selection of the number of segments which is based on classification-based approaches (see Section 1.2.3). Indeed, as stated in this Section, an Integrated Completed Likelihood (ICL) criterion has been proposed by RIGAILL *et al.* (2012) in a Bayesian segmentation context to select the number of segments. The goal of this work, in collaboration with Dr Rigail, Dr Luong and Pr Nuel, was to propose an approximation of the ICL criterion, which we call the *conditional ICL*, which could be computed in linear time.

Details of these computations are given in the paper (available at <http://arxiv.org/abs/1211.3210>) presented below.

**Fast estimation of the Integrated Completed Likelihood criterion
for change-point detection problems with applications
to Next-Generation Sequencing data**

Alice Cleynen, The Minh Luong, Guillem Rigaiil and Gregory Nuel

abstract

In this paper, we consider the Integrated Completed Likelihood (ICL) as a useful criterion for estimating the number of changes in the underlying distribution of data, specifically in problems where detecting the precise location of these changes is the main goal. The exact computation of the ICL requires $\mathcal{O}(Kn^2)$ operations (with K the number of segments and n the number of data-points) which is prohibitive in many practical situations with large sequences of data. We describe a framework to estimate the ICL with $\mathcal{O}(Kn)$ complexity. Our approach is general in the sense that it can accommodate any given model distribution. We checked the run-time and validity of our approach on simulated data and demonstrate its good performance when analyzing real Next-Generation Sequencing (NGS) data using a negative binomial model. Our method is implemented in the R package `postCP` and available on the CRAN repository.

Keywords

Hidden Markov Model; Integrated Completed Likelihood; Model Selection; Negative Binomial; Segmentation

2.3.1 Introduction

The estimation of the number of segments is a central aspect in change-point methodology. For instance, in the context of CGH-array or Next-Generation Sequencing experiments, identifying the number and corresponding location of segments is crucial as the segments

may relate to a biological event of interest. This theoretically complex problem can be handled in the more general context of model selection, leading to the use of *ad hoc* procedures in practical situations.

Among the procedures are the use of classical criteria based on penalized likelihoods such as the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC or SIC, YAO, 1988). However, when choosing the number of segments, the BIC criterion uses a Laplace approximation requiring differentiability conditions not satisfied by the model, which thus may not be appropriate when the number of observations in each segment are unequal and unknown. These criteria also tend to overestimate the number of segments as the clustering within segments tends to be ignored, as shown by ZHANG and SIEGMUND (2007) who proposed a modified BIC criterion using a Brownian motion model with changing drift for the specific case of normal data.

For this reason, there has been an extensive literature influenced by BIRGÉ and MAS-SART (2001) which proposes new penalty shapes and constants in order to select a lower number of segments in the profile. The idea is to choose the model that, within a set of models, performs closest to the true value by deriving a tight upper bound on the variance term. This leads to penalties that generally depend only on the number of segments K , and whose constants can be chosen adaptively to the data (LAVIELLE, 2005; LEBARBIER, 2005). However, a large proportion of those methods focused on normal data, and are not applicable to count datasets modeled by the Poisson or the negative binomial distributions.

Other approaches for model selection appearing in the literature include sequential likelihood ratio tests (HACCOU and MEELIS, 1988) and Bayesian approaches through estimating model posterior probabilities by various MCMC methods (GREEN, 1995; CHIB, 1998; ANDRIEU *et al.*, 2001; FEARNHEAD, 2005). However, the Bayesian approaches are often computationally intensive as they require re-sampling.

In the context of incomplete data models (e.g. mixture model for clustering) BIERNACKI *et al.* (2000) proposed a model selection criterion accounting for both observed and unobserved variables based on the Integrated Completed Likelihood (ICL): $\sum_S \mathbb{P}(S|X) \log \mathbb{P}(S|X)$ where X are the observations and S are the corresponding (un-

known) clustering membership.

RIGAILL *et al.* (2012) proposed the use of the ICL criterion in the multiple change-point detection context. Hence, the segmentation S can be considered as a set of unobserved variables in the sense that the segment-labels of each datapoint are not known. In this context, we can select the number of segments as:

$$\hat{K} = \arg \min_K \text{ICL}(K) \quad \text{where} \quad \text{ICL}(K) = -\log \mathbb{P}(X, K) + \mathcal{H}(K), \quad (2.1)$$

with $\mathcal{H}(K) = -\sum_{S \in \mathcal{M}_K} \mathbb{P}(S|X, K) \log \mathbb{P}(S|X, K)$, and \mathcal{M}_K representing the set of all segmentations of the signal in K segments.

The entropy term $\mathcal{H}(K)$ can be viewed as an intrinsic penalty to quantify the reliability of a given model with K segments by characterizing the separation of the observations in different segments. In other words, for fixed K segments, the entropy $\mathcal{H}(K)$ thus will be lower when the best segmentation provides a much better fit compared to other segmentations with the same number of segments, hence favoring models which provide the most evidence of similarity within the detected segments. While other penalized likelihood approaches are designed to select the most likely number of segments by relying on approximation of posterior probabilities or oracle inequalities, the ICL criterion aims at selecting the number of segments with the lowest uncertainty.

In the context of Hidden Markov Models (HMMs), it is well known that the posterior distribution $\mathbb{P}(S|X, K, \Theta_K)$ can be efficiently computed using standard forward-backward recursions with $\mathcal{O}(K^2n)$ complexity (MARTIN and ASTON, 2012). However, the HMM requires that emission parameters take their values in a limited set of levels which are recurrently visited by the underlying hidden process.

In the segmentation context, where each segment has its own specific level, an exact algorithm with $\mathcal{O}(Kn^2)$ complexity computes the ICL in a Bayesian framework. In a simulation study, RIGAILL *et al.* (2012) showed that the ICL performed better than standard model selection criteria such as BIC or Deviance Information Criterion (DIC). However the quadratic complexity and numerical precision restrict the use of this Bayesian ICL to relatively small profiles.

In this paper we suggest a computation of the ICL conditionally on the segment parameters and we propose a fast two-step procedure to compute this conditional ICL criterion with linear complexity in order to select the number of segments within a set of change-point data. First, we specify a range of possible K number of change-points, from one to a user-defined K_{\max} . We estimate the parameters of the segmentation in K segments, and given these estimates, we compute the ICL for each value of K in the range. Second, we select the K which minimizes the ICL criterion. In essence, our conditional ICL explores only one aspect of the segmentation uncertainty, the position of the change-points, and ignores the uncertainty due to the segment parameters.

Section 2.3.2 describes the ICL estimation procedure, through the use of a constrained hidden Markov model and Section 2.3.3 validates the approach by presenting the results of different simulations for detecting the correct number of change-points. Finally, Section 2.3.4 is a discussion of our method supported by a comparison with a few segmentation algorithms on data-sets simulated by re-sampling real RNA-Seq data, and an illustration on the original dataset from an experiment on a chromosome from the yeast species from the same study.

2.3.2 Integrated Completed Likelihood criterion estimation using a constrained HMM

In this paper we use the following *segment-based model* for the distribution of X given a segmentation $S \in \mathcal{M}_K$:

$$\mathbb{P}(X|S; \Theta_K) = \prod_{i=1}^n g_{\theta_{S_i}}(X_i) = \prod_{k=1}^K \prod_{i:S_i=k} g_{\theta_k}(X_i) \quad (2.2)$$

where $g_{\theta_{S_i}}(\cdot)$ is the parametric distribution (ex: normal, Poisson, negative binomial, etc.) with parameter θ_{S_i} , $\Theta_K = (\theta_1, \dots, \theta_K)$ is the global parameter, $S_i \in \{1, \dots, K\}$ is the index of the segment at position i (ex: $S_{1:5} = 11222$ corresponds to a segmentation of $n = 5$ points into $K = 2$ segments with a change-point occurring between positions 2 and 3), and \mathcal{M}_K is the set of all possible partitions of S_1, \dots, S_n with a fixed K number of segments, such that $S_1 = 1$ and $S_n = K$, and $S_i - S_{i-1} \in \{0, 1\}$ for all $i = 2, \dots, n$.

One should note that although this model has the same emission probabilities as its HMM counterpart, the constraints on the sequence S correspond *exactly* to the segmentation model where each segment has its own level, and *not* to any HMM where levels take their value in a recurring set.

Fast estimation of posterior quantities in ICL criterion

Our goal is to compute the conditional ICL given by the following equation :

$$\hat{K} = \arg \min_K \text{ICL}(K|\Theta_K)$$

$$\text{where } \text{ICL}(K|\Theta_K) = -\log \mathbb{P}(X, K|\Theta_K) + \mathcal{H}(K|\Theta_K). \quad (2.3)$$

The objective of this conditional ICL is to reproduce the performance of the non-conditional ICL (in Equation 2.1). The conditional ICL criterion is well defined given a prior distribution on the segmentations: $\mathbb{P}(S, K)$. We will only consider priors that can be decomposed as: $\mathbb{P}(S, K) = \mathbb{P}(S|K)\mathbb{P}(K)$; this choice is discussed in a later section. In both the conditional and non-conditional ICL, the first term, $\log \mathbb{P}(X, K)$ (and respectively $\log \mathbb{P}(X, K|\Theta_K)$), depends on both $\mathbb{P}(S|K)$ and $\mathbb{P}(K)$, however, the entropy term only depends on $\mathbb{P}(S|K)$.

To estimate this entropy term, we consider a specific hidden Markov model with constraints chosen specifically to correspond to a segmentation model (LUONG *et al.*, 2013) where the change-points separate segments consisting of contiguous observations with the same distribution. Introducing a prior distribution $\mathbb{P}(S, K)$ on any $S \in \mathcal{M}_K$, yields the posterior distribution of the segmentation:

$$\mathbb{P}(S, K|X; \Theta_K) = \frac{\mathbb{P}(X|S, K; \Theta_K)\mathbb{P}(S, K)}{\sum_R \mathbb{P}(X|R, K; \Theta_K)\mathbb{P}(R, K)}. \quad (2.4)$$

Considering the prior $\mathbb{P}(S, K) = \mathbb{P}(S|K)\mathbb{P}(K)$ and fixing the value of K , let us assume that S is a heterogeneous Markov chain over $\{1, 2, \dots, K, K + 1\}$. We only allow for

transitions of 0 or +1 by constraining the chain with:

$$\mathbb{P}(S_1 = 1) = 1$$

$$\forall 2 \leq i \leq n, \quad \forall 1 \leq k \leq K, \quad \begin{cases} \mathbb{P}(S_i = k | S_{i-1} = k) & = 1 - \eta_k(i) \\ \mathbb{P}(S_i = k + 1 | S_{i-1} = k) & = \eta_k(i), \end{cases}$$

where $\eta_k(i)$ is the transition probability from the k^{th} segment to $k + 1$ for observation i .

In the general case where K is not fixed, the choice of prior on S is known to be a critical point. However previous methods include the use of non-informative priors (ZHANG and SIEGMUND, 2007) when K is fixed. For that reason, we focus on the uniform prior by setting $\eta_k(i) = \eta$ for all k and i . Note that this particular case corresponds to the uniform prior $\mathbb{P}(S|K) = 1/\binom{n-1}{K-1} = 1/|\mathcal{M}_K|$ which is used in RIGAILL *et al.* (2012).

To estimate the properties of the K^{th} state we introduce a ‘junk’ state $K + 1$, and for consistency we choose $\mathbb{P}(S_i = K + 1 | S_{i-1} = K + 1) = 1$. We then estimate the emission distribution by using the maximum likelihood estimate $g_{\hat{\theta}_k}(x_i)$, or alternatively the E-M algorithm.

We define the forward and backward quantities as follows for observation i and state k :
For $1 \leq i \leq n - 1$:

$$F_i(k) = \mathbb{P}(X_{1:i} = x_{1:i}, S_i = k | \hat{\Theta}_k)$$

$$B_i(k) = \mathbb{P}(X_{i+1:n} = x_{i+1:n}, S_n = k | S_i = k, \hat{\Theta}_k).$$

We may use the following recursions to estimate the forward and backward quantities:

$$F_1(k) = \begin{cases} g_{\hat{\theta}_1}(x_1) & \text{if } k = 1 \\ 0 & \text{else} \end{cases}$$

$$\begin{aligned}
F_{i+1}(k) &= [F_i(k)(1 - \eta_k(i+1)) + \mathbf{1}_{k>1}F_i(k-1)\eta_k(i+1)]g_{\hat{\theta}_k}(x_{i+1}) \\
B_{n-1}(k) &= \begin{cases} \eta_K(n)g_{\hat{\theta}_K}(x_n) & \text{if } k = K-1 \\ (1 - \eta_K(n))g_{\hat{\theta}_K}(x_n) & \text{if } k = K \\ 0 & \text{else} \end{cases} \\
B_{i-1}(k) &= (1 - \eta_k(i))g_{\hat{\theta}_k}(x_i)B_i(k) + \mathbf{1}_{k<K}\eta_{k+1}(i)g_{\hat{\theta}_{k+1}}(x_i)B_i(k+1)
\end{aligned}$$

These quantities can then be used to obtain the marginal distributions μ_i and the transition π_i , being terms needed for the calculation of the entropy $\mathcal{H}(K|\hat{\Theta}_K)$ with:

$$\mu_i(k) = \frac{F_i(k)B_i(k)}{F_1(1)B_1(1)} \quad (2.5)$$

$$\pi_i(k, k') = \frac{\mathbb{P}(S_i = k'|S_{i-1} = k)g_{\hat{\theta}_k}(x_i)B_i(k')}{B_{i-1}(k)}. \quad (2.6)$$

where

$$\mathbb{P}(S_i = k'|S_{i-1} = k) = \begin{cases} 1 - \eta & \text{if } k' = k \\ \eta & \text{if } k' = k + 1 \\ 0 & \text{else} \end{cases}.$$

Calculation of $\log \mathbb{P}(X, K|\Theta_K)$

The non-conditional term $\mathbb{P}(X, K)$ can be written as $\sum_{S \in \mathcal{M}_K} \mathbb{P}(S, K, X) = \sum_{S \in \mathcal{M}_K} \mathbb{P}(X|S, K)\mathbb{P}(S|K)\mathbb{P}(K)$. In our constrained HMM approach we will therefore compute, for a given parameter $\hat{\Theta}_K$ for which the choice will be discussed later on, $\mathbb{P}(X, K|\hat{\Theta}_K)$ as $\sum_{S \in \mathcal{M}_K} \mathbb{P}(X|S, K, \hat{\Theta}_K)\mathbb{P}(S|K)\mathbb{P}(K)$, using the classic priors $\mathbb{P}(K) = \alpha$ and the previously discussed uniform prior $\mathbb{P}(S|K) = 1/\binom{n-1}{K-1}$. The remaining term $\sum_{S \in \mathcal{M}_K} \mathbb{P}(X|S, K, \hat{\Theta}_K)$ is then obtained directly using forward-backward recursions. Specifically, we obtain:

$$\begin{aligned}
\sum_{S \in \mathcal{M}_K} \mathbb{P}(X, S \in \mathcal{M}_K|K, \hat{\Theta}_K) &= F_1(1)B_1(1) \quad \text{and} \\
\mathbb{P}(S \in \mathcal{M}_K|K, \hat{\Theta}_K) &= F_1^0(1)B_1^0(1)
\end{aligned}$$

where $F_i(k)$ and $B_i(k)$ are the HMM forward and backward recursions as described, and

$F_i^0(k)$ and $B_i^0(k)$ are forward and backward terms obtained with the usual recursions where each emission probability is replaced by 1.

The likelihood term is finally obtained as

$$\mathbb{P}(X, K | \hat{\Theta}_K) = \frac{\mathbb{P}(K) F_1(1) B_1(1)}{\binom{n-1}{K-1} F_1^0(1) B_1^0(1)}. \quad (2.7)$$

Estimation of $\mathcal{H}(K)$

The ICL can be expressed, with the entropy term $\mathcal{H}(K)$ estimated by $\mathcal{H}(K | \hat{\Theta}_K) = -\sum_S \mathbb{P}(S | X, K, \hat{\Theta}_K) \log \mathbb{P}(S | X, K, \hat{\Theta}_K)$, as (BIERNACKI *et al.*, 2000):

$$\text{ICL}(K | \hat{\Theta}_K) = \mathbb{P}(X, K | \hat{\Theta}_K) + \mathcal{H}(K | \hat{\Theta}_K), \quad (2.8)$$

with K being the number of segments.

For a fixed K , the constrained HMM is an efficient way to estimate the posterior segmentation distribution $\mathbb{P}(S | X, K, \hat{\Theta}_K)$ for a given set of parameters $\hat{\Theta}_K$. This model consists of a heterogeneous Markov chain (HMC) with marginal distribution $\mu_i(S_i) = \mathbb{P}(S_i | X, K, \hat{\Theta}_K)$ and heterogeneous transition $\pi_i(S_{i-1}, S_i) = \mathbb{P}(S_i | S_{i-1}, X, K, \hat{\Theta}_K)$. Those quantities can be computed with the recursive formulas as described above.

It is hence easy (HERNANDO *et al.*, 2005) to derive the following expression for the entropy term:

$$\mathcal{H}(K | \hat{\Theta}_K) = - \left[\sum_{S_1} \mu_1(S_1) \log \mu_1(S_1) + \sum_{i=2}^n \sum_{S_{i-1}, S_i} \mu_{i-1}(S_{i-1}) \pi_i(S_{i-1}, S_i) \log \pi_i(S_{i-1}, S_i) \right] \quad (2.9)$$

Note that information theory ensures that we have $0 \leq \mathcal{H}(K | \hat{\Theta}_K) \leq \log \binom{n-1}{K-1}$.

The original entropy term, $\mathcal{H}(K)$ has an expression including posterior probabilities, thus requiring the estimation of the posterior distribution of S as detailed in Section 2.3.2. While it can be computed with quadratic complexity $O(Kn^2)$ (RIGAILL *et al.*, 2012) and intensive operations on probability matrices, its exact computation is usually intractable for large datasets of tens of thousands points or more. The forward-backward recursions of the

HMM and Equation (2.9) allow its estimation with linear complexity $O(Kn)$. One should note that the key point for fast computation lies in the fact that we work conditionally on $\hat{\Theta}_K$ rather than considering the whole parameter space.

Model selection procedure using ICL

For any given K , using our constrained HMM method requires a set of initial parameters $\Theta_K = \{\hat{\theta}_k\}_{1 \leq k \leq K}$. Because the quality of the results depends on the choice of those initial values, we propose the use an effective segmentation algorithm to obtain a set of $K - 1$ change-points, which can in turn be used to estimate the parameters Θ_k through maximum likelihood estimation.

We considered several options for the initialization algorithm: for normally distributed data we considered a K-means algorithm (HARTIGAN and WONG, 1979; COMTE and ROZENHOLC, 2004), which is a greedy method that minimizes the least-squares criterion, as well as binary segmentation (SCOTT and KNOTT, 1974), a fast heuristic to optimize the log-likelihood criterion. We also used the pruned dynamic programming algorithm (RIGAILL, 2010), a fast algorithm to compute the optimal segmentation according to loss functions including Poisson, negative binomial or normal losses. We then use the Viterbi algorithm (VITERBI, 1967) to obtain the *a posteriori* most probable set of change-points.

To estimate the ICL of a change-point model with K segments, we compute the posterior probabilities of interest through the forward-backward algorithm as previously described, which is implemented in the postCP package (available on the CRAN : <http://cran.r-project.org/web/packages/postCP>).

This procedure is repeated for a range of possible values of K : $K_{\text{range}} = \{1, \dots, K_{\text{max}}\}$. We finally choose the number of segments by minimizing the ICL criterion upon all values of K in K_{range} , *i.e.*

$$\hat{K}_{\text{ICL}} = \arg \min_{K \in K_{\text{range}}} \text{ICL}(K | \hat{\Theta}_K). \quad (2.10)$$

2.3.3 Validation

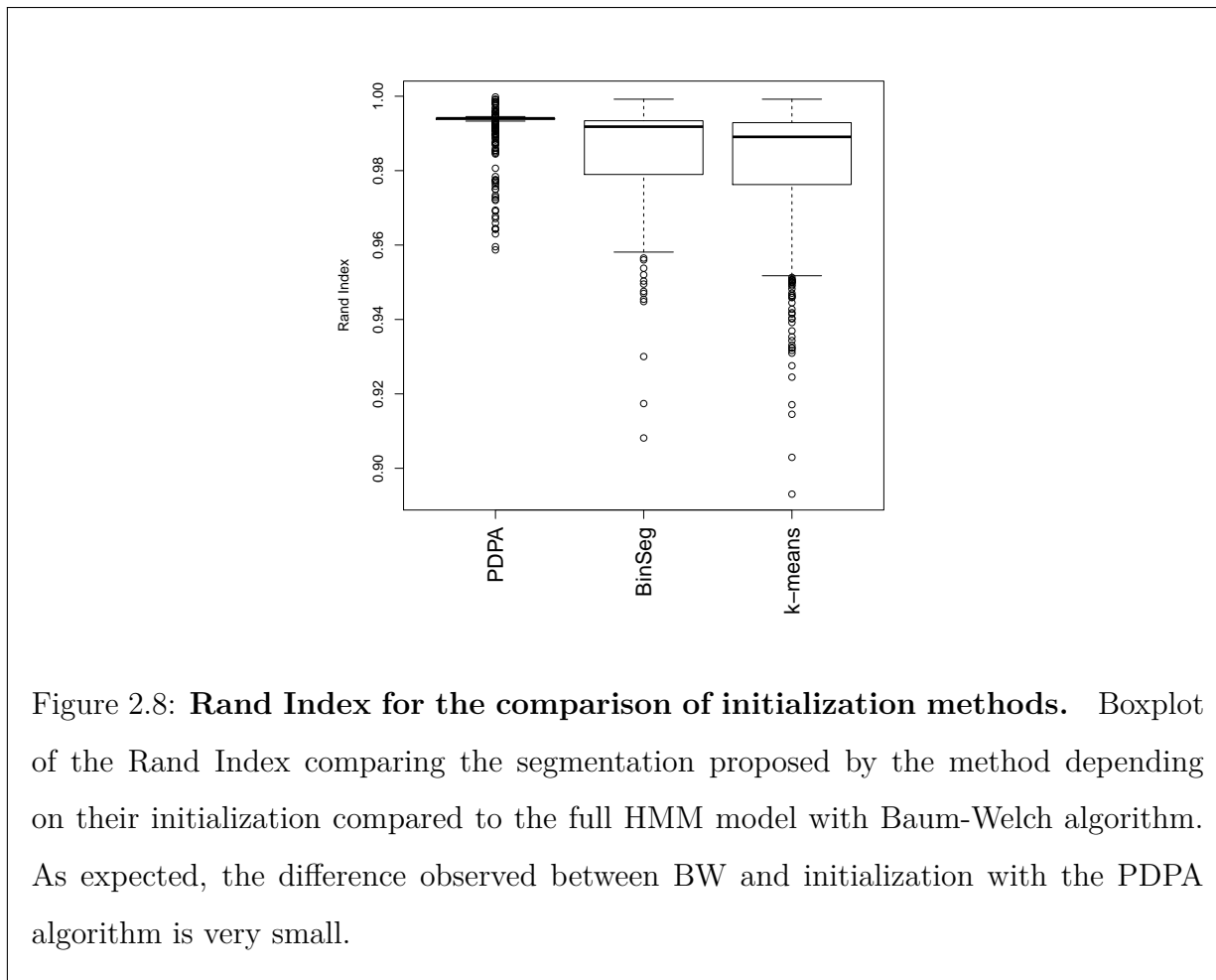
To validate the quality of our approach we first evaluated the impact of the initialization parameters. We implemented the Baum-Welch algorithm (BAUM *et al.*, 1970) for use as a reference, and computed the Rand Index between the segmentation resulting from the Baum-Welch approach to those resulting from our algorithm with different other initialization methods. The Rand Index compares the adequacy between different segmentations by computing the proportion of concordant pairs of data-points, including the proportion of pairs that either belong to the same segment in the two competitive segmentations, or that are in different segments in both segmentations. In a second step, we evaluated the results of our algorithm in terms of model selection on two sets of simulations.

Impact of initialization parameters

Because of the long run-time of the Baum-Welch (BW) algorithm, we considered a small simulation study where the data of size $n = 1,000$ is simulated from the Poisson distribution with parameter λ subject to 9 change-points (at locations 100, 130, 200, 475, 500, 600, 630, 800 and 975) and taking the successive values 1, 4.3, 1.15, 6 and 4.2 repeated twice. On each of the 1,000 replications, we ran our constrained HMM segmentation approach considering the number of segments to be known, but with different initialization parameters: those obtained by the Baum-Welch algorithm, those obtained by the pruned dynamic programming algorithm (PDPA), those obtained with a k-means approach and those obtained by running the Binary Segmentation (BinSeg) (SCOTT and KNOTT, 1974) for the Poisson distribution.

The results are illustrated in Figure 2.8. As expected, the Rand Index between the estimation by the Baum-Welch algorithm and the PDPA algorithm is very close to 1, and it decreases with other initialization methods that are not exact. Moreover, on average the Baum-Welch algorithm required 15.2 iterations when itself initialized with the PDPA output, while the run-time for the initialization by PDPA requires 0.24 seconds and an iteration of BW, 0.04 seconds. This finding suggests that the combination of PDPA and

postCP is advantageous in terms of run-time with a negligible difference in results, especially since the number of iterations of BW grows as n and the number of segments increase (not shown here).



Validation of the ICL approach

Our first simulation study consisted of relatively small signals ($n = 500$ points) where we compared our approach to the quadratic non-conditional algorithm. In our second simulation study, with larger signals ($n = 50,000$), we only ran our fast ICL criterion due to computing limitations.

The simulation designs were as follows:

Small design. We used a similar simulation design suggested by RIGAILL *et al.* (2012): we simulated a sequence of 500 observations from a Poisson model (requiring the choice of only one parameter) affected by six change-points at the following positions: 22, 65, 108, 219, 252 and 435. Odd segments had a mean of 1, while even segments had a mean of $1 + \lambda$, with λ varying from 0 to 9. Thus, the true number of change-points were more easily identified with higher values of λ . For each configuration, we simulated 1,000 sequences.

Large design. We repeated the preceding procedure for large-scale datasets. We generated a sequence of 50,000 observations with $K = 40$ segments by randomly selecting 39 change-points whose locations were drawn from a uniform distribution (without replacement), with each segment needing to be at least of length 25. For this sample size, we focus on the results from our approximated ICL as the non-conditional ICL implementation is not fast enough to be practical in this situation. For each configuration, we simulated 100 sequences.

We compared the performances of three different criteria:

- The conditional ICL *greedy* (C-ICL-g) criterion where initial parameters are obtained by the greedy algorithm using least-squares, and using the criterion described in the previous section and given by Equation (2.10) .
- The conditional ICL *exact* (C-ICL-e) criterion which corresponds to an initialization of the parameters using the pruned dynamic programming algorithm with Poisson loss.
- The non-conditional ICL (NC-ICL) criterion as described in RIGAILL *et al.* (2012). The hyper-parameters used for the prior on the data-distribution were set to 1. This choice is discussed in the previous paper. In this simple scenario, the results were robust to changes in the hyper-parameters (result not shown).

Figure 2.9 summarizes the results of the simulation study for simulations of length 500. While the non-conditional ICL criterion had the highest amount of correct estimates of number of segments \hat{K} , the faster ICL with pruned PDPA performed almost as well.

Of note, the average run-times of the methods were 4.2 seconds for the non-conditional approach, 0.001 and 0.12 seconds respectively for the initialization of postCP with the k-means and PDPA algorithms, and 0.46 seconds for the postCP algorithm.

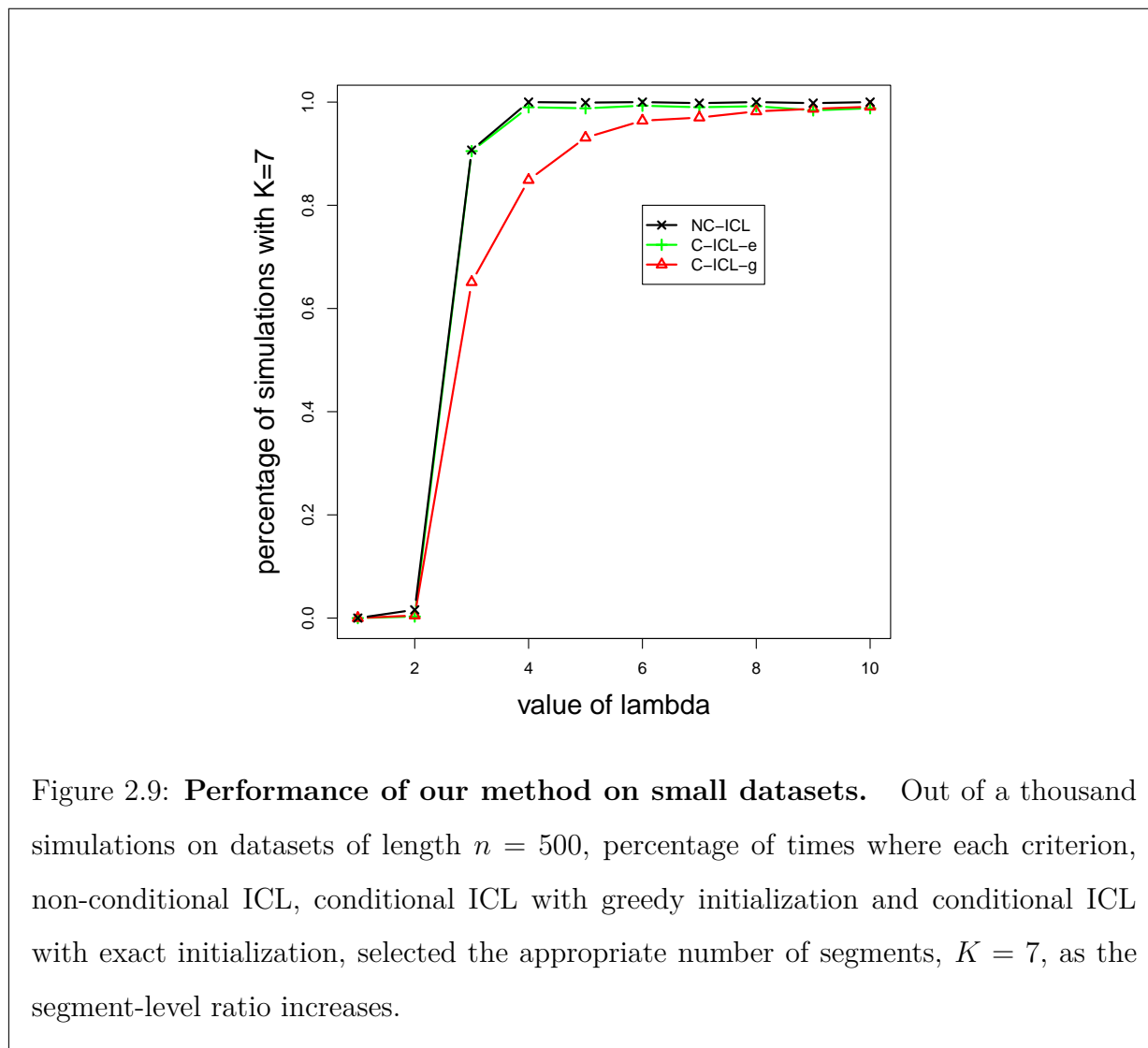
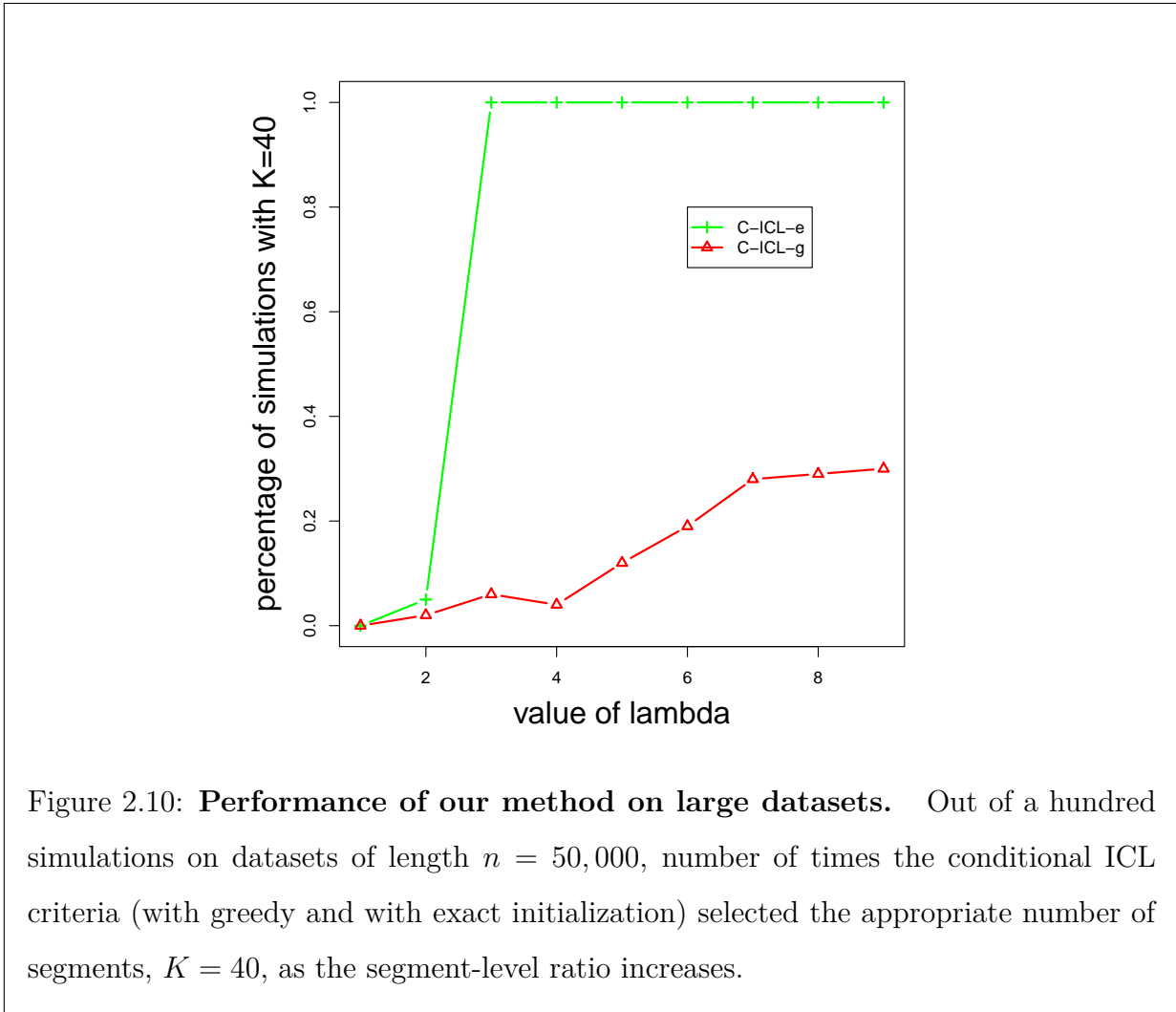


Figure 2.10 summarizes the results of the simulation study for simulations of length 50,000. For these larger sequences, the conditional ICL criteria performed much better when the initial change-point set was detected by PDPA than with the greedy algorithm. As the segmentation problem becomes more difficult with more segments, the greedy algorithm is less successful in providing accurate initial change-point location estimates. As a result,

less accurate values of $\hat{\Theta}_K$ are used and the conditional ICL is not as effective in predicting the number of segments as in the smaller sample size example.



On the other hand, the conditional ICL combined with PDPA detected the correct number of segments more than 80% of the time with larger inter-segmental differences of $\lambda > 2$. The average run-time for the initialization was 1.32 seconds for k-means and 142 seconds for PDPA, while the model selection procedure required on average 1,240 seconds (≈ 20 minutes). Despite the longer run-time, it is advised to use the PDPA for model selection in very long sequences as it provides a more accurate set of change-points than greedy methods.

2.3.4 Discussion

Choice of K_{\max}

Our strategy to compute the estimate of the ICL proceeds in two steps. First, we recover all the best segmentations in 1 to K_{\max} segments. Then, using the parameters from all these K_{\max} segmentations as an initialization, we run a forward-backward algorithm.

The initialization step takes on average an $\mathcal{O}(K_{\max}n \log n)$ complexity using the PDPA (see RIGAILL, 2010). The complexity of the second step is in $\mathcal{O}(K_{\max}n)$. Depending on the applications, it might be desirable or not to consider K_{\max} of the order of n , (see KILLICK *et al.*, 2012, for a discussion). In the second case, our strategy is efficient. On the other hand, in the first case the initialization step is on average in $\mathcal{O}(n^2 \log n)$ and at worst in $\mathcal{O}(n^3)$, while the second step is in $\mathcal{O}(n^2)$. The first step is thus the limiting factor.

When the goal is to initialize the HMM by recovering all the best segmentations of 1 to n segments, which we showed to be desirable for the quality of the procedure, there exists to our knowledge no faster algorithms to obtain an exact solution to this problem. Moreover, in any case, enumerating the $\sum_{k=1}^n k$ change-points of these n segmentations is already quadratic in n . An alternative is to use the binary segmentation heuristic (VENKATRAMAN and OLSHEN, 2007) which is on average in $\mathcal{O}(\log(K_{\max}n))$. In that case the limiting factor is the second step which still is quadratic in n .

Thus, we believe our strategy is most suited for the second case, when K_{\max} is much smaller than n . In the first case, when K_{\max} is of the order of n , our strategy is at least quadratic in n and its application is limited to medium size profiles.

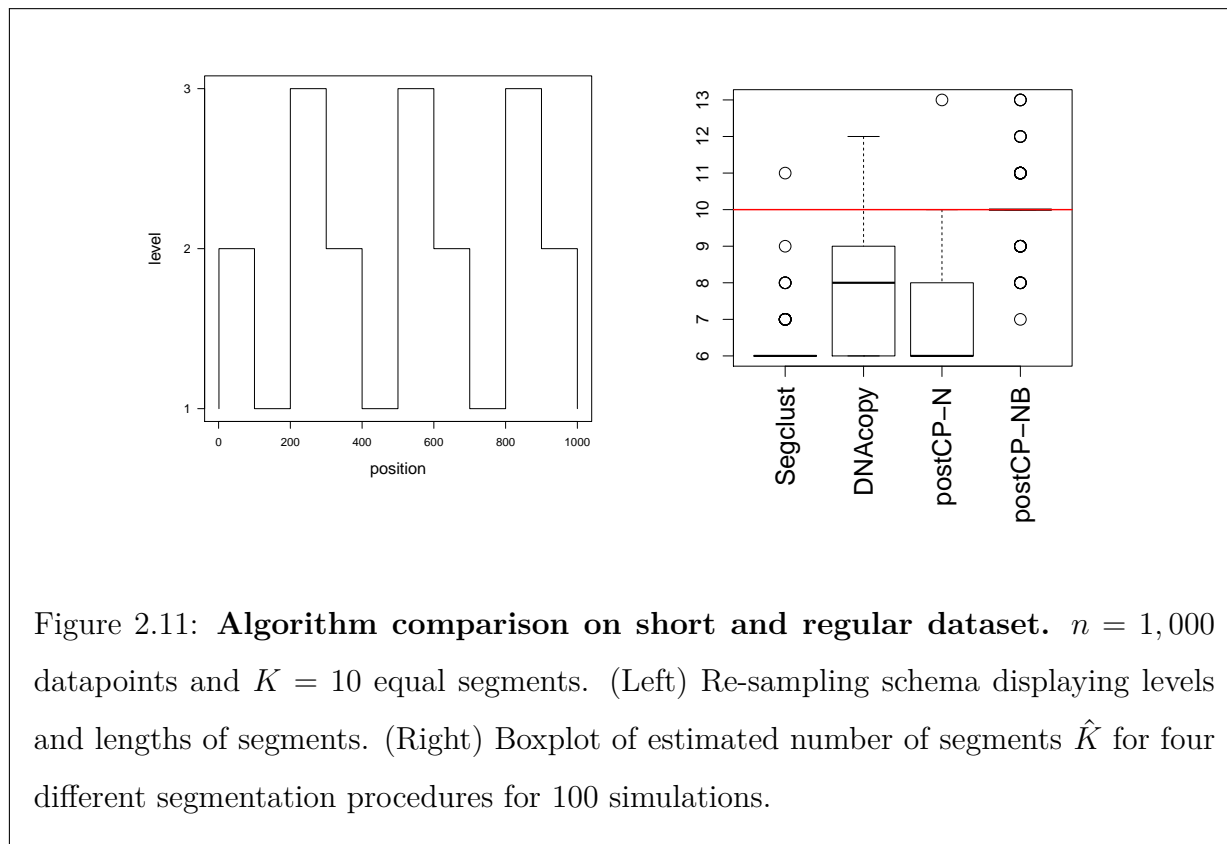
Re-sampling of yeast RNA-Seq data

To assess the quality of our criteria, we performed the following simulation study to compare two previously published packages on CRAN, `segclust` (PICARD *et al.*, 2007), which uses adaptive penalized likelihoods and `DNA copy`, an implementation of binary

segmentation for multiple change-points (VENKATRAMAN and OLSHEN, 2007), with our model selection method with the conditional ICL criterion. We performed the following re-sampling procedure using real RNA-seq data. The original data, from a study by the Sherlock Genomics laboratory at Stanford University, is publicly available on the NCBI's Sequence Read Archive (SRA, <http://www.ncbi.nlm.nih.gov/sra>) with the accession number SRA048710. We clustered the observed signal into the following classes: intronic region, low expressed, medium expressed and highly expressed genes that we will refer to as levels 1, 2, 3 and 4. We then designed four simulation studies, each repeated 100 times, by varying the number and level of segments as well as the signal and segment sizes, as described in the left Figures 2.11 through 2.14. On each segment, the data was obtained by re-sampling (with replacement) the observations in the classified clusters.

To assess the performance of our fast ICL approach in segmentation, we used three different distributions as the emission distribution $g_{\theta}(\cdot)$ a normal distribution (postCP-N), a Poisson distribution (postCP-P) and negative binomial (postCP-NB) and used PDPA to obtain the initial set of parameters. In all cases, we used homogeneous Markov chains with uniform priors; it is of note that the results of the constrained HMM methods may improve with informative priors (FEARNHEAD, 2005), for example those taken from *a posteriori* estimates. For segclust, DNACopy, and postCP-N, which assume a normal distribution, we applied the methods to the data after the widely used $\log(x + 1)$ transformation. In all our simulation studies, postCP-P grossly overestimated the number of segments, so the results are not displayed here.

In the simplest case, the left part of Figure 2.11 illustrates the resampling scheme for $n = 1,000$ and $K = 10$ evenly spaced segments, displaying the levels used for each segment and the change-point locations. The right part of the figure displays a boxplot of the number of segments found by each approach. In this quite unrealistic scenario, postCP-BN estimated the correct number of segments in 63 of 100 replicates. The next best algorithms were postCP-N and DNACopy, respectively, which both slightly underestimated the number of segments. The segclust procedure provided a consistent underestimate of the number of



segments.

Figure 2.12 illustrates the re-sampling schemes and boxplots for a slightly different and more realistic scenario of $n = 1,000$ and $K = 10$, with unevenly spaced segments this time. The results are comparable to the previous except that the methods performed slightly worse; the median postCP-NB estimate was still correct but missed 1 or 2 segments in 43 replicates. This suggests that postCP has more difficulties in detecting small segments.

We then replicated the methods for larger data sets and unevenly spaced segments. Figure 2.13 displays the methods and results for a $n = 5,000$ and $K = 10$ scenario. In this case, DNAcopy performs best, with the median number of estimated segments being correct. The postCP-NB method gave similar results but missed two change-points in 66 of the replicates. The segclust algorithm, once again, found consistent but overly conservative estimates of the number of segments, while postCP-N grossly overestimated the segments as the log-transformation was not adequate in this design.

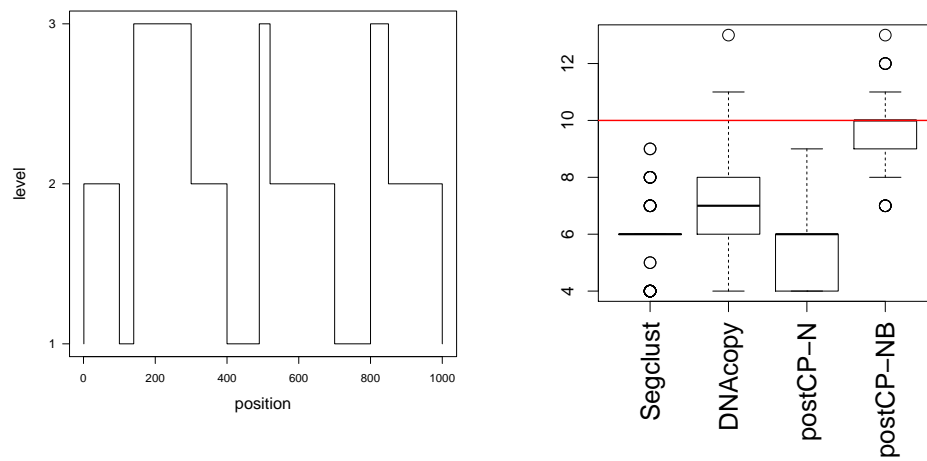


Figure 2.12: **Algorithm comparison on short but irregular dataset.** $n = 1,000$ datapoints and $K = 10$ uneven segments. (Left) Re-sampling schema displaying levels and lengths of segments. (Right) Boxplot of estimated number of segments \hat{K} for 4 different segmentation procedures for 100 simulations.

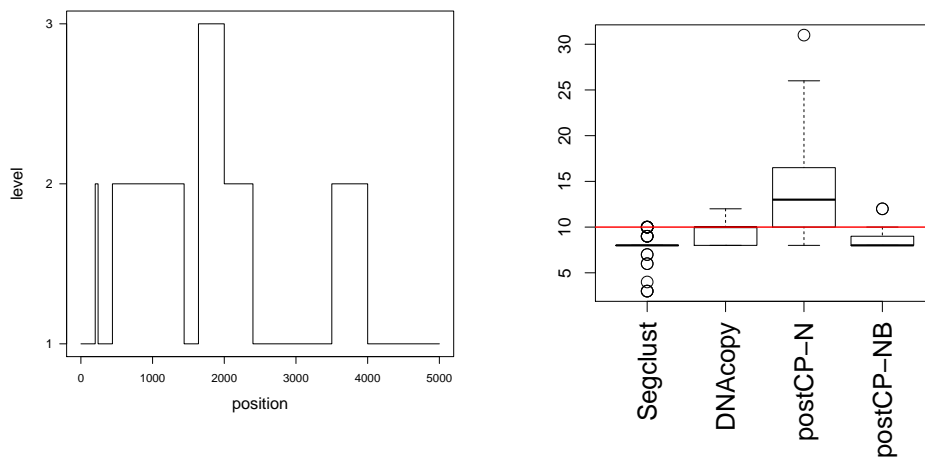


Figure 2.13: **Algorithm comparison on medium length and irregular dataset.** $n = 5,000$ datapoints and $K = 10$ uneven segments. (Left) Re-sampling schema displaying levels and lengths of segments. (Right) Boxplot of estimated number of segments \hat{K} for 4 different segmentation procedures for 100 simulations.

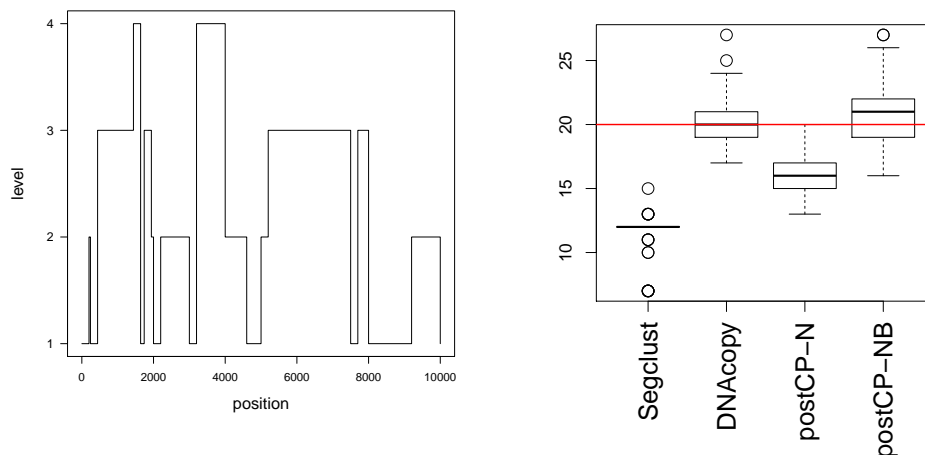


Figure 2.14: **Algorithm comparison on long and irregular dataset.** $n = 10,000$ datapoints and $K = 20$ uneven segments. (Left) Re-sampling schema displaying levels and lengths of segments. (Right) Boxplot of estimated number of segments \hat{K} for 4 different segmentation procedures for 100 simulations.

To understand the results, we ran the PDPA on the simulated datasets to obtain the optimal segmentations w.r.t. to negative binomial likelihood imposing $K = 10$ segments. We found that in 48 replicates out of 100, this segmentation did not include the second true segment but rather sheared other segments into more pieces. This finding suggests that, at least in these 48 replicates, precisely finding the position of the first two changes might be prohibitively difficult. Thus by selecting $K = 8$ change-points rather than 10, postCP-NB is coherent with the goal of the ICL (i.e. selecting a set of segments such that we are confident in the position of these changes).

In a $n = 10,000$ and $K = 20$ scenario with uneven segments (Figure 2.14), DNACopy was again best, with postCP-N and postCP-NB almost as effective, the former method slightly underestimating the number of segments and the latter approach slightly overestimating them.

In the investigated scenarios, we found postCP, when the correct negative binomial distribution was specified, provided accurate and precise results when segments were evenly spaced, but provided slightly less accurate results in more realistic scenarios where segment lengths were uneven. The results with postCP-N and postCP-P suggest that the postCP approach may be susceptible to misspecification of the emission distribution when there are very small segments present (Figure 2.13). Given the goal of the ICL this is to be expected. Indeed, it is reasonable to have high uncertainty in the identification of small segments when the emission distribution is misspecified.

On the other hand, DNACopy tended to underestimate segments in easier scenarios, where segments were even, but obtained more accurate results with more realistic uneven segments. The hybrid segmentation and clustering approach, segclust, generally was consistent but underestimated the number of segments.

Application to a real data-set

We finally illustrate the procedure on the real data-set from the Sherlock study described above, whose underlying characteristics are unknown. The signal corresponds to the positive strand of chromosome 1 from the yeast genome and has a length of 230,218.

We used a negative binomial model with global overdispersion parameters and initialized our procedure using the pruned dynamic programming algorithm (for a runtime of 25 minutes). The postCP algorithm then required 4 hours to analyze the profile, resulting in a choice of 79 segments.

We also compared these results to those proposed by the previously cited methods. However, we were not able to run the segclust algorithm on this long profile due to lack of memory capacity. With a similar runtime, the postCP algorithm with the normal distribution applied to the log-transformed data resulted in a choice of 80 segments, while DNACopy analyzed the signal in 47 seconds for a final choice of 465 segments. Figure 2.15 illustrates the segmentation proposed by each method. For clarity, we focus on a region of length 50,000 datapoints, and plotted the signal in a square-root scale. Even though the

constrained HMM approach chooses almost the same number of segments with different emission distributions, their corresponding resulting segmentations differ.

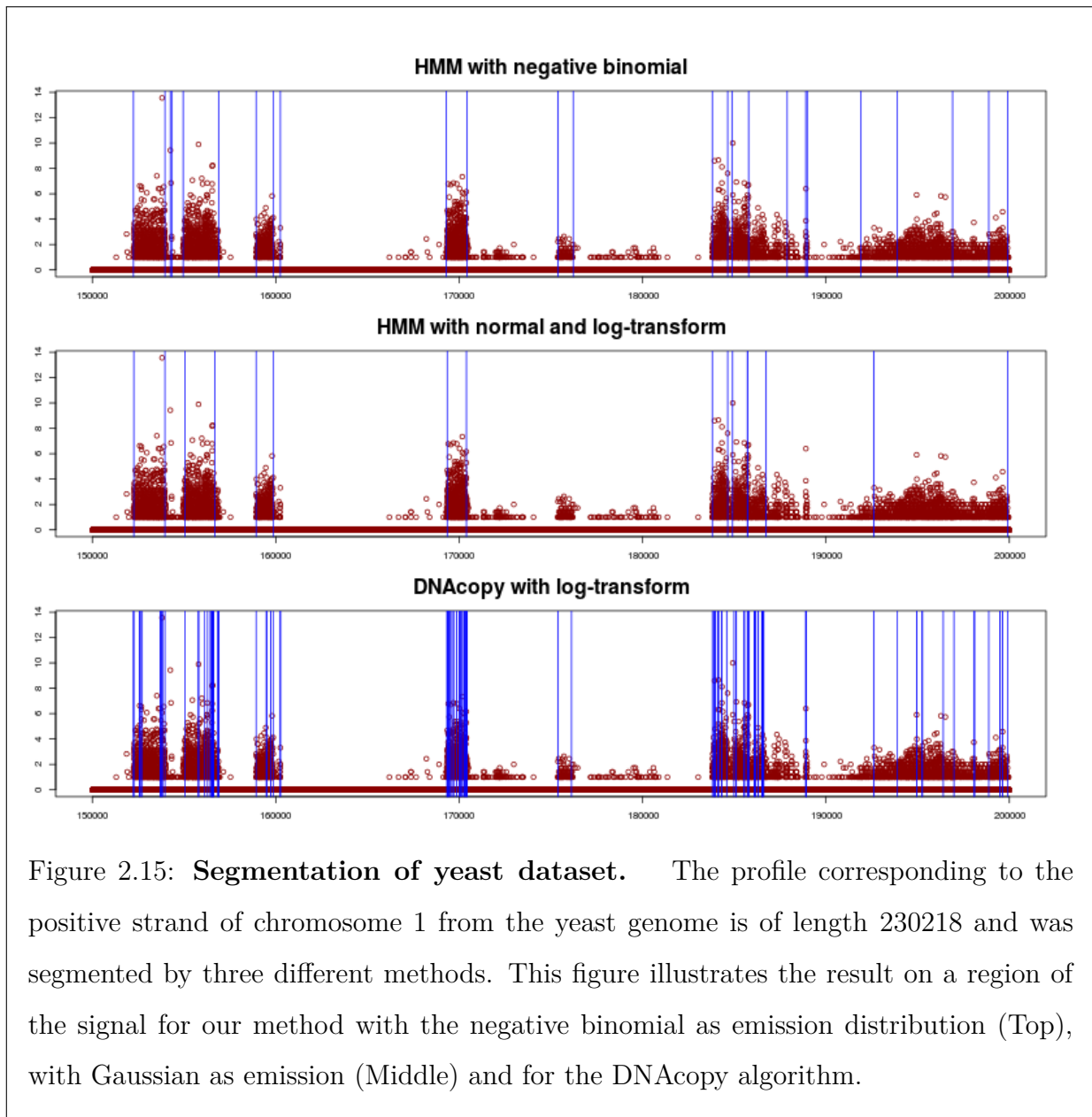


Figure 2.15: **Segmentation of yeast dataset.** The profile corresponding to the positive strand of chromosome 1 from the yeast genome is of length 230218 and was segmented by three different methods. This figure illustrates the result on a region of the signal for our method with the negative binomial as emission distribution (Top), with Gaussian as emission (Middle) and for the DNACopy algorithm.

Conclusion

We describe a fast procedure for estimating the ICL criterion in the context of model selection for segmentation. While simulations showed that the performance of the conditional

ICL approach was almost as good as that of the non-conditional approach, several features allow for its use in a wide range of applications. The described ICL algorithm is versatile as it can be applied to data of any model distribution when provided with an initialization for the HMM, through either maximum likelihood estimation or the expectation-maximization (E-M) algorithm. While there exists some model selection criteria that could be adapted to our problem such as the BIC or the MDL (DAVIS *et al.*, 2006) which provide a balance between data fitting and model complexity, the ICL also takes into account the entropy of the segmentation space. Given the very large collection of possible segmentations, we believe that the ICL is an interesting alternative to more standard model selection criteria.

Furthermore, our procedure can be applied to long signals due to its fast run-time. With its effective results in finding the number of segments, specifically those where the precise location of the change-points can be estimated, this paper shows the practicality of the conditional ICL procedure in a wide variety of segmentation problems.

Acknowledgments

The authors would like to thank Stéphane Robin for useful discussions. Alice Cleynen's research was supported by an Allocation Special Normalien at the Université Paris-Sud in Orsay and The Minh Luong's research was supported by an excellence postdoctoral grant at the Université Paris-Descartes.

References

- Christophe Andrieu, PM Djurić, and Arnaud Doucet. Model selection by MCMC computation. *Signal Processing*, 81(1):19–37, 2001.
- Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*, 41(1):164–171, 1970.
- C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, 2000. ISSN 01628828.
- L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):

203–268, 2001. ISSN 1435-9855.

Siddhartha Chib. Estimation and comparison of multiple change-point models. *Journal of econometrics*, 86(2):221–241, 1998.

Fabienne Comte and Yves Rozenholc. A new algorithm for fixed design regression and denoising. *Annals of the Institute of Statistical Mathematics*, 56(3):449–473, 2004.

Richard A Davis, Thomas C M Lee, and Gabriel A Rodriguez-Yam. Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101(473):223–239, 2006.

Paul Fearnhead. Exact Bayesian curve fitting and signal segmentation. *Signal Processing, IEEE Transactions on*, 53(6):2160–2166, 2005.

Peter J Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Patsy Haccou and Evert Meelis. Testing for the number of change points in a sequence of exponential random variables. *Journal of Statistical Computation and Simulation*, 30(4):285–298, 1988.

John A Hartigan and Manchek A Wong. A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

Diego Hernando, Valentino Crespi, and George Cybenko. Efficient computation of the hidden Markov model entropy for a given observation sequence. *Information Theory, IEEE Transactions on*, 51(7):2681–2685, 2005.

Rebecca Killick, Paul Fearnhead, and IA Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

Marc Lavielle. Using penalized contrasts for the change-point problem. *Signal Processing*, 85(8):1501–1510, 2005.

E. Lebarbier. Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, 85(4):717–736, April 2005. ISSN 0165-1684.

T.M. Luong, Y. Rozenholc, and G. Nuel. Fast estimation of posterior probabilities in change-point models through a constrained hidden Markov model. *Arxiv preprint arXiv:1203.4394*, 2012.

Donald EK Martin and John AD Aston. Distribution of statistics of hidden state sequences through the sum-product algorithm. *Methodology and Computing in Applied Probability*, pages 1–22, 2012.

Franck Picard, Stéphane Robin, E Lebarbier, and J-J Daudin. A segmentation/clustering model for the analysis of array cgh data. *Biometrics*, 63(3):758–766, 2007.

G. Rigaiil. Pruned dynamic programming for optimal multiple change-point detection. *Arxiv:1004.0887*, April 2010. URL <http://arxiv.org/abs/1004.0887>.

G Rigaille, E Lebarbier, and S Robin. Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and Computing*, 22(4): 917–929, 2012.

A.J. Scott and M. Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30:507–512, 1974.

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.

Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269, 1967.

Y.-C. Yao. Estimating the number of change-points via Schwarz’ criterion. *Statistics & Probability Letters*, 6(3):181–189, February 1988.

Nancy R Zhang and David O Siegmund. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32, 2007.

2.4 Results on the yeast data-set

In this last chapter, we propose to illustrate the pruned Dynamic Programming algorithm and its associated model selection criteria (see Chapters 2.1 and 2.2) on our benchmark dataset. We segmented all 16 chromosomes from the yeast genome, considering the positive and negative strands separately, so that 32 independent profiles are processed.

The annotation available on the SGD website (<http://yeastgenome.org>) allows to obtain an approximate idea of the expected number and location of segments for each chromosome simply by summing all known coding regions. Of course not all genes from the chromosome are expressed at the time of the experiments, and 'new' transcripts are not taken into account, so that this number remains approximate. There are multiple goals:

- identify expressed genes,
- identify new transcripts,
- obtain information on the length of the UTRs.

For this last objective, it might be useful to recall that the SGD annotation indicates the boundaries of translated regions (and therefore not that of the UTRs) while RNA-Seq data gives information on the mature transcribed regions. Comparing the boundaries of both quantities thus provides the UTR length.

In Table 2.3 are indicated, among other information detailed later, the expected number of segments (K_{ex}), and the number \hat{K} estimated by the PDPA, for each strand of each chromosome (indicated in column #).

It is interesting to note that in all cases where $n \leq 10^6$ and $K_{ex} \leq 600$ we obtain a \hat{K} larger than the expected numbers, while it is the contrary in other cases. This is easily explained by the use of the slope heuristic (ARLOT and MASSART, 2009) to choose the penalty constant β . Indeed, it implies exploring values of K much larger than the expected to estimate β_{min} . It is very likely that in these cases, our choice of K_{max} influences the resulting estimator.

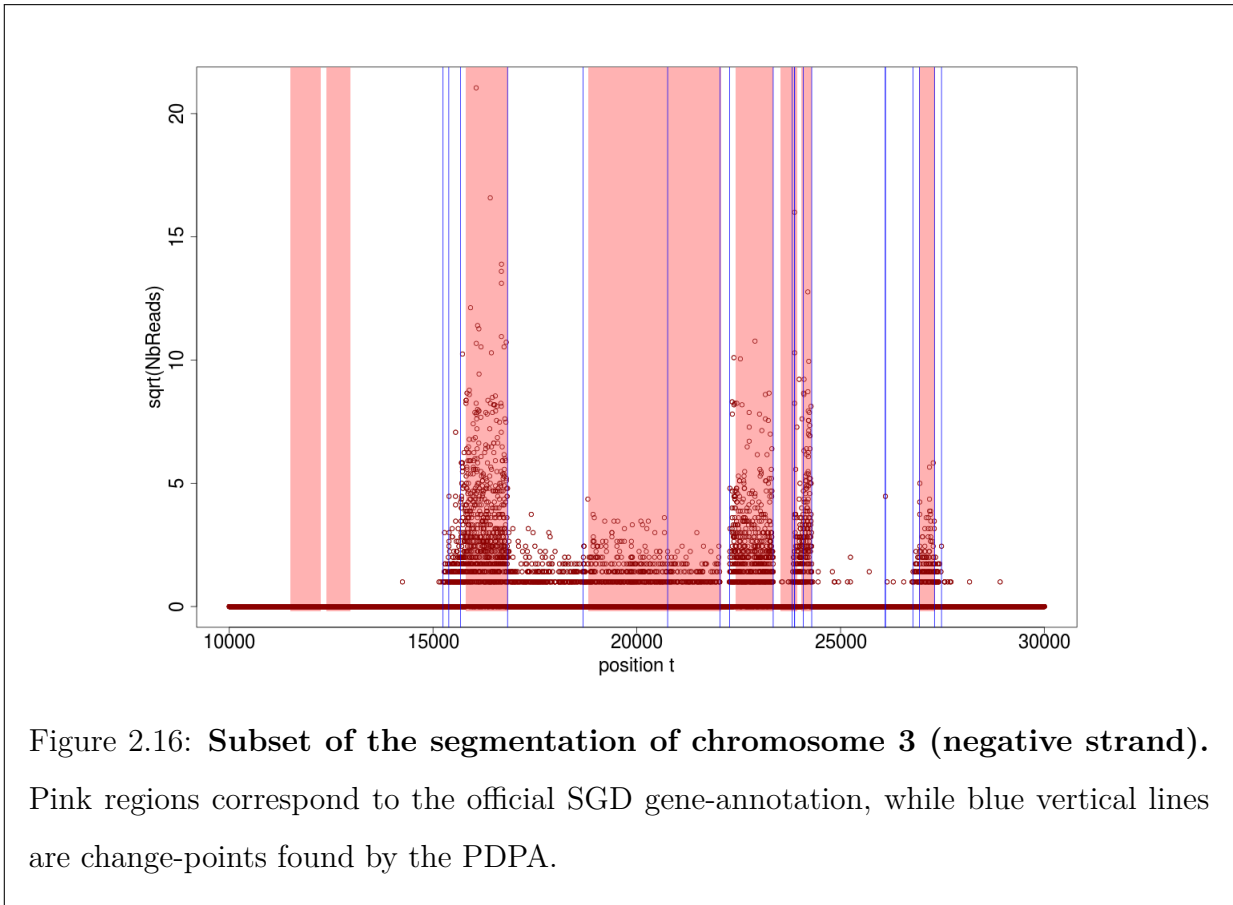
A closer look at the resulting segmentation shows that most of the difference between

#	length	Positive strand				Negative strand			
		K_{ex}	\hat{K}	\widehat{K}_f	rt	K_{ex}	\hat{K}	\widehat{K}_f	rt
1	230,218	137	201	142	85	123	198	148	83
2	813,184	497	530	426	433	551	584	476	491
3	316,620	191	302	220	132	223	288	202	146
4	1,531,933	919	713	604	1070	961	710	614	1087
5	576,874	361	480	350	316	365	449	338	299
6	270,161	163	219	160	103	167	269	196	106
7	1,090,940	671	645	542	678	637	627	502	677
8	562,643	385	458	346	292	353	423	334	287
9	439,888	273	345	262	195	283	365	278	208
10	745,751	467	531	416	389	419	518	400	405
11	666,816	405	493	384	368	355	468	362	361
12	1,078,177	637	608	478	669	671	631	510	668
13	924,431	589	604	506	541	567	593	478	545
14	784,333	519	580	472	436	451	529	426	440
15	1,091,291	667	613	508	652	621	626	530	681
16	948,066	609	637	510	555	559	572	458	545

Table 2.3: **Output of the pruned dynamic algorithm.** Expected number of segments (K_{ex}), number proposed by the algorithm (\hat{K}), final number after grouping regions (\widehat{K}_f) and runtime (rt, in minutes).

K_{ex} and \hat{K} is induced by extra change-points inside segments corresponding to expressed regions. This is for instance illustrated in Figure 2.16 which corresponds to a subset of the negative strand of chromosome 3. The first two genes of this region are not expressed at the time of the experiment, while others appear to have approximate official boundaries, and some are divided in sub-pieces by the algorithm. This last observation can be explained either by the technology biases introduced in the data which results in some regions being sequenced with deeper coverage than others, or by alternative splicing phenomena, *i.e.* subregions of genes not expressed in equal proportions in the cell. Columns \widehat{K}_f of Table 2.3 gives the final number of detected transcripts, *i.e.* after grouping over-segmented regions based on the value of the estimated parameter of each segments. We were then able to identify over a dozen regions of the genomes which had not been annotated as coding and for which the signal intensity gives support to their corresponding to new transcripts.

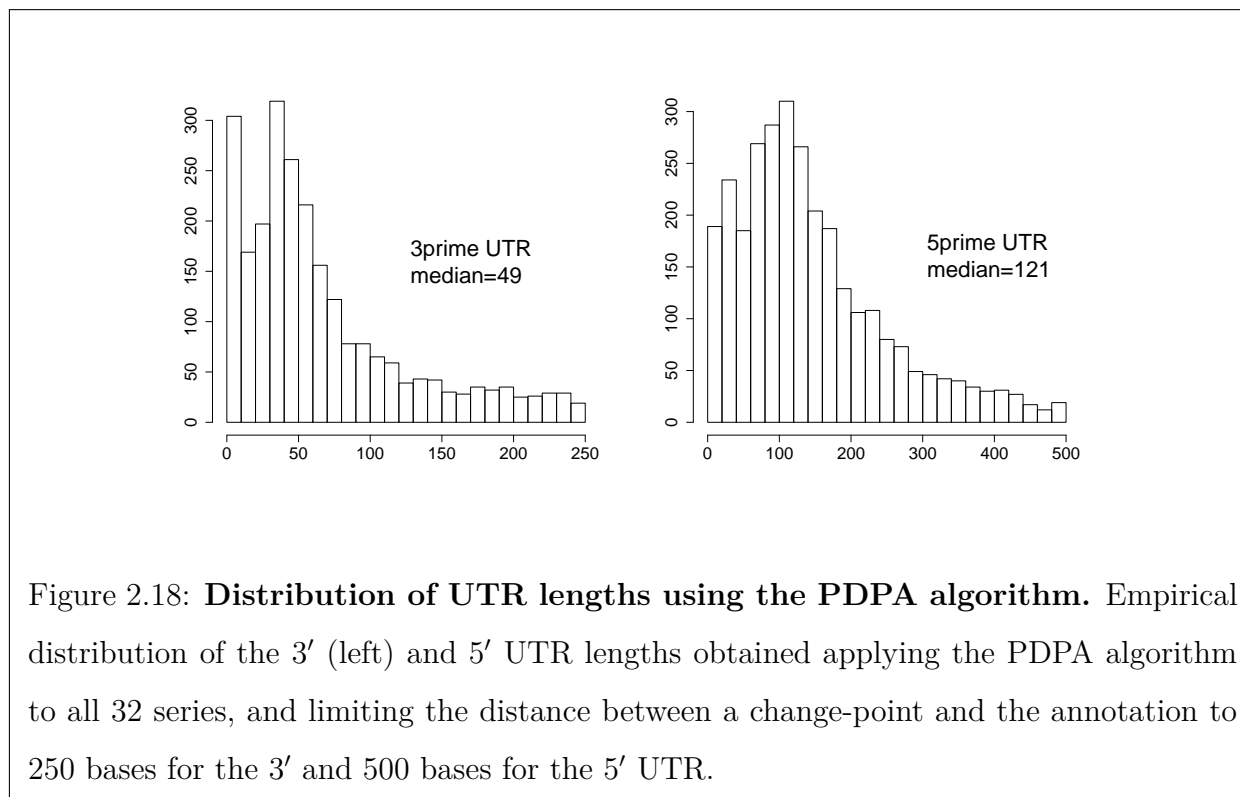
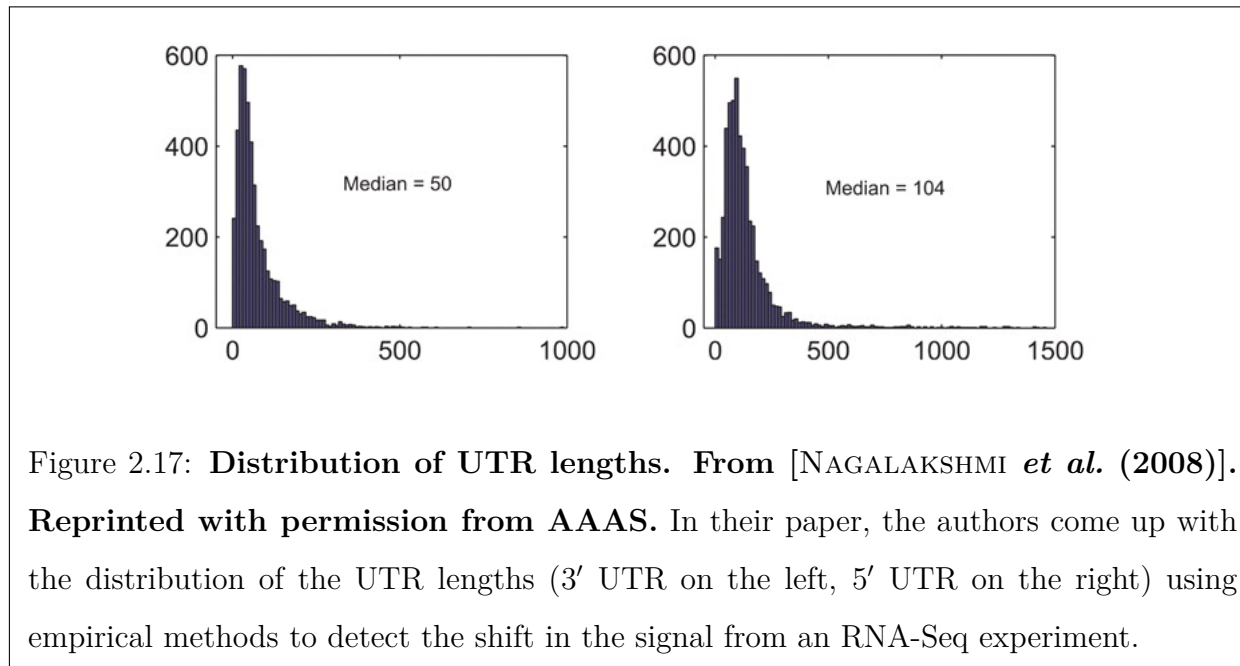
To address the question of UTR length, we used a paper from NAGALAKSHMI *et al.* (2008) from which we extracted Figure 2.17 giving the empirical distribution of both the



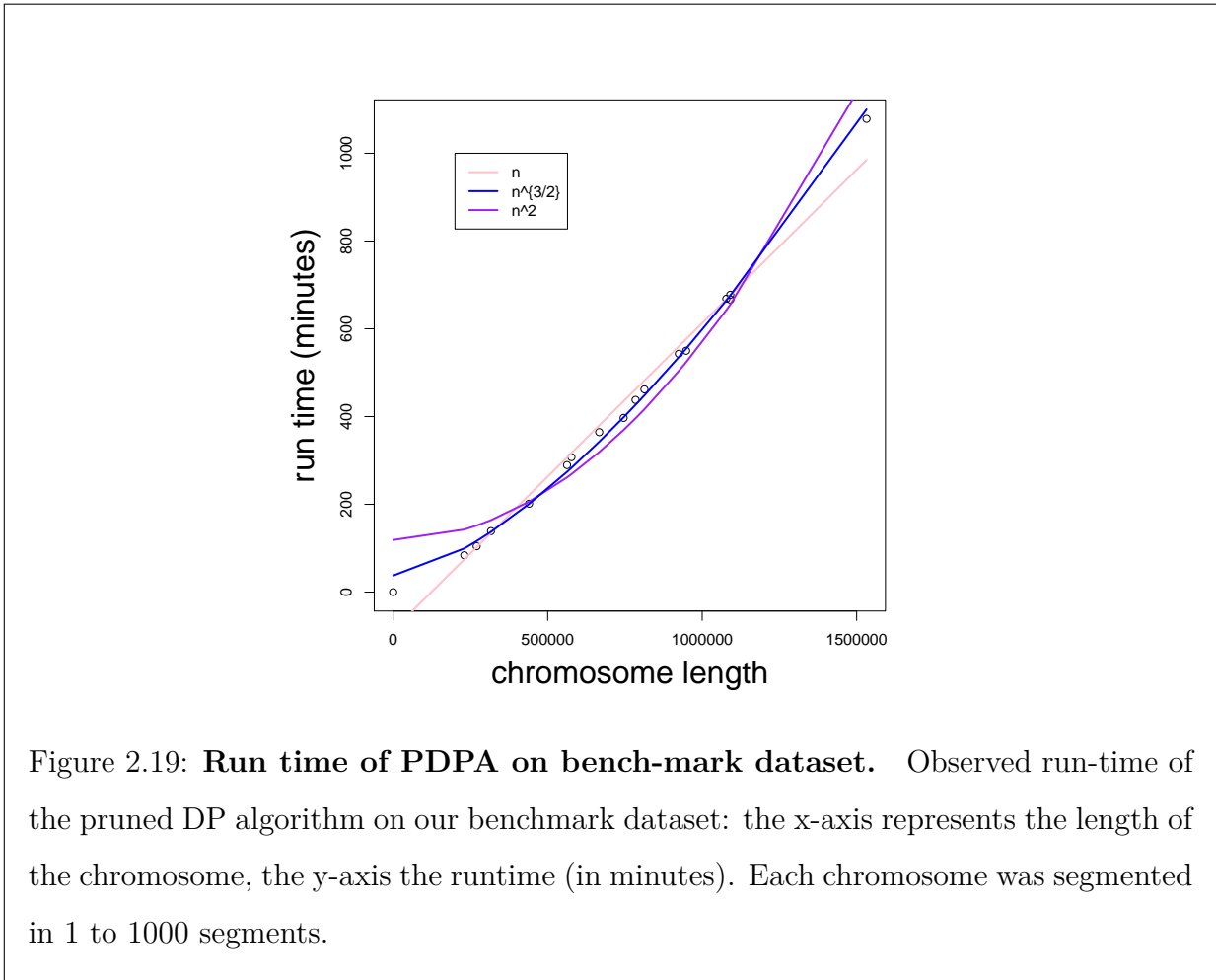
3 prime and 5 prime UTR lengths. Using this information, we defined the maximum acceptable distance between a change-point and the annotation as 250 bases for the 3' UTR and 500 bases for the 5' UTR. Figure 2.18 illustrates the distribution obtained using the PDPA algorithm, and the median lengths associated with it.

We were excited to find that the medians are very similar, as well as the shape of the distributions, apart from a higher peak around zero for the 3' UTR. As for the intron boundaries, we found a median of 3-base difference between the SGD annotation and our change-points corresponding to expressed genes, suggesting that the annotation of transcribed regions might be quite precise.

We conclude this illustration by analyzing the runtime required by the PDPA algorithm on a standard laptop (indicated in the rt columns of Table 2.3) on these datasets of various length, each processed using the same K_{max} value. Since the positive and negative strands



of the chromosome have the same length, we computed the average run-time of both to plot Figure 2.19. This confirmed that the empirical time complexity of the PDPA is almost linear in the size of the profiles.



One question that has not be addressed and which would deserve attention in some future work is that of the coverage needed to perform such analysis. The first difficulty comes from the definition of coverage in the context of RNA-Seq experiment. Indeed, in the case of DNA-Seq, the coverage is simply the average number of times each position is covered by a read. This is equivalent to defining the coverage as

$$C_v = \frac{\sum_{re} L_{re}}{n_c}$$

where n_c is the size of the genome, re are reads of the experiment and L_{re} their length.

However, by definition, RNA-Seq experiments aim at measuring the transcriptome activity, which is expected to differ between genes. It is precisely the difference in coverage observed in each gene that tells us about their expression. It has been proposed to replace the length n_c of the genome by the length of the transcriptome (subset of the genome which corresponds to coding regions) in the definition of the coverage. Though a little more precise, this does not really tell us about coverage since no distinction is made between expressed and unexpressed genes. One quantity which appears more appropriate to us is the number of transcripts detected with at least an X coverage (and in this case the coverage of a transcript is defined by the previous formula with n replaced by the transcript's size).

With this definition, our question becomes 'what should X value for the algorithm to detect a transcript?' One first natural approach is to compute the power of the Wald test associated with the equality of successive probability parameters p_1 and p_2 . More specifically, we test $\phi \frac{1-p_1}{p_1} = \phi \frac{1-p_2}{p_2}$ with the statistic

$$T_W = \frac{\bar{y}_1 - \bar{y}_2 - \left(\phi \frac{1-p_1}{p_1} - \phi \frac{1-p_2}{p_2} \right)}{\sqrt{Var(\bar{y}_1 - \bar{y}_2)}}$$

so that the power of the test, assuming $n_1 = n_2 = n$ depends on a term in

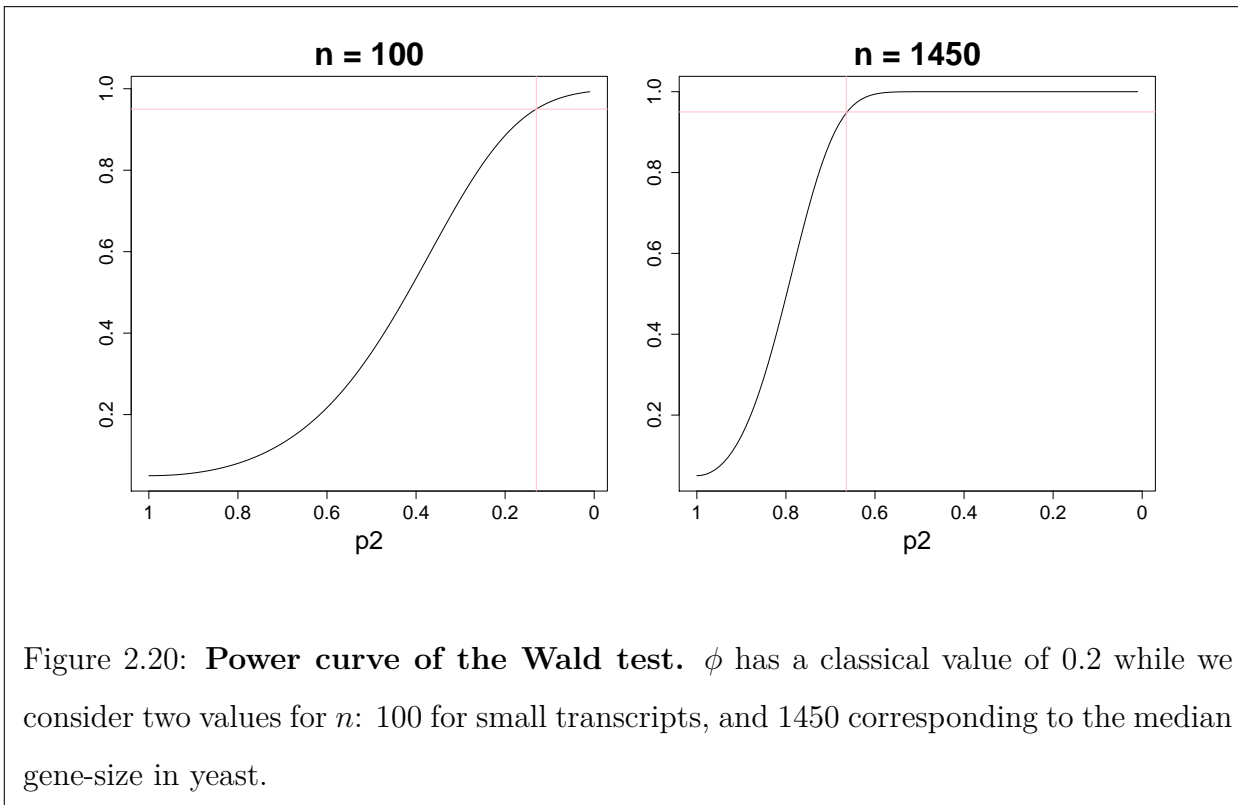
$$\sqrt{n\phi} \frac{|p_2 - p_1|}{\sqrt{p_2^2 + p_1^2 + p_1p_2^2 + p_2p_1^2}}.$$

Now because in non-coding regions the signal is expected to be null, we can approximate this expression by taking $p_1 = 1$, and the previous formulae becomes

$$\sqrt{n\phi} \frac{1 - p_2}{\sqrt{2p_2^2 + p_2 + 1}}.$$

Figure 2.20 illustrates the power of the test for a typical value of $\phi = 0.2$ and two values of n : 100 for small transcripts, and 1450 which corresponds to the median size of yeast genes.

In the case of our experiment, we investigated the effect of the total number of reads by multiplying the data by a constant λ and rounding the value to conserve their count



characteristic. For each value of λ in $(0.1, 0.2, 0.5, 1, 1.5, 2, 3, 5, 7, 10)$, we estimated K and computed the value of the parameter of the least expressed detected segment.

Figure 2.21 shows that the estimated number of segments increases smoothly with λ , and that the parameter of the least expressed detected segment appears to remain stable. However, we observed that the increased number of segments did not result from identifying more transcribed regions, but from fragmenting some regions into pieces separated by segments of length smaller than 10 with no signal. Because the reads have a length of 36 bases, we can ascertain that those non-expressed regions should be classified as false negatives. In fact, multiplying the signal by a constant λ only affects positions with non-zero counts, and it is not surprising that even small zero-regions should stand out between highly expressed regions.

Our conclusion is the following. We can hope that increasing the volume of RNA introduced in the NGS process will not be equivalent to multiplying the signal by a constant but will have a more homogeneous impact. In this case, it is likely that our algorithm will

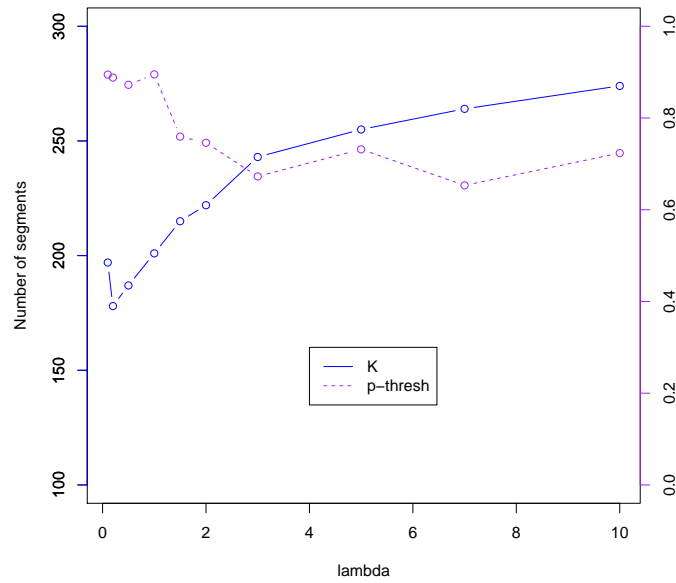


Figure 2.21: **pruned DP algorithm and coverage.** Estimated number of segments and parameter value of least expressed detected region when multiplying true signal by a constant lambda.

be able to identify more transcribed regions if any were missed. If however the technology biases should not allow some homogeneous increase, we would not recommend too large a library size.

Segmentation methods for gene annotation using RNA-Seq data

3.1	Method comparison	163
3.1.1	Introduction	165
3.1.2	Methods	168
3.1.3	Results and discussion	176
3.1.4	Conclusion	187
3.1.5	Choice of the priors	204
3.2	Profile comparison	206
3.2.1	Introduction	208
3.2.2	Model for one series	211
3.2.3	Posterior distribution of the shift	213
3.2.4	Comparison of change point locations	215
3.2.5	Simulation study	217
3.2.6	Comparison of transcribed regions in yeast	221
3.2.7	Conclusion	223
3.3	EBS: R package for Exact Bayesian Segmentation	230
3.3.1	EBS class and associated methods	230
3.3.2	EBSProfiles class and change-point comparison	233
3.4	Results on the yeast dataset	237

In the previous chapters we have proposed methods to segment long subsets of the genome. In those cases, whether the goal is to determine which genes are expressed or to identify new transcripts, the major questions to address are the selection of the number of segments in the profile, and the identification of their locations. While it is important to measure the uncertainty associated with the number of segments K , interpreting any information on the uncertainty in the location of change-points would be very challenging, especially in contexts where K is in the order of 10^2 or even larger.

In the framework we consider now, the annotation of individual genes for which the ultimate goal is the comparison of transcription boundaries between different growth environments, we find ourselves in the opposite situation. Indeed, in most cases it will be very reasonable to assume that the number of segments is known, as it is strongly related to the number of exons of the gene, while on the contrary, as change-points are associated to transcript boundaries, it will be crucial to measure the uncertainty associated with their location.

This section is dedicated to the segmentation issues in the analysis of smaller datasets for which we request more precise information.

3.1 Method comparison

In this section we propose a comparison of segmentation methods which can be applied to RNA-Seq data corresponding to a region containing a gene. Their requirements are two-fold: the ability to model count datasets, and their non-dependence on reference profiles (it may be recalled that we are not, as was the case for microarrays, comparing a mutant profile to a wild-type profile). The paper presented here compares the 5 algorithms presented in Section 1.2.3, and is joint work with Sandrine Dudoit and Stéphane Robin.

Comparing segmentation methods for genome annotation based on RNA-Seq data

Alice Cleynen, Sandrine Dudoit and Stéphane Robin

abstract

Transcriptome sequencing (RNA-Seq) yields massive datasets, containing a wealth of information on the expression of a genome. While numerous methods have been developed for the analysis of differential gene expression, little has been attempted for the localization of transcribed regions, that is, segments of DNA that are transcribed and processed to result in a mature messenger RNA. Our understanding of genomes, mostly annotated from biological experiments or computational gene prediction methods, could benefit greatly from re-annotation using the high precision of RNA-Seq.

We consider five classes of genome segmentation methods to delineate transcribed regions based on RNA-Seq data. The methods provide different functionality and include both exact and heuristic approaches, using diverse models, such as hidden Markov or Bayesian models, and diverse algorithms, such as dynamic programming or the forward-backward algorithm. We evaluate the methods in a simulation study where RNA-Seq read counts are generated from parametric models as well as by resampling of actual yeast RNA-Seq data. The methods are compared in terms of criteria that include global and local fit to a reference segmentation, Receiver Operator Characteristic (ROC) curves, and coverage of credibility intervals based on posterior change-point distributions. All compared algorithms are implemented in packages available on the Comprehensive R Archive Network (CRAN, <http://cran.r-project.org>). The dataset used in the simulation study is publicly available from the Sequence Read Archive (SRA, <http://www.ncbi.nlm.nih.gov/sra>).

While the different methods each have pros and cons, our results suggest that the EBS Bayesian approach of RIGAILL *et al.* (2012) performs well in a re-annotation context, as illustrated in the simulation study and in the application to actual yeast RNA-Seq data.

Keywords

Change-point detection, Confidence intervals, Count data, Genome annotation, Negative binomial distribution, RNA-Seq, Segmentation.

3.1.1 Introduction

Many genomes have been annotated, using approaches ranging from *in vitro* biological experiments to *in silico* gene prediction. Today, with the low cost and high precision of high-throughput sequencing, the question of re-annotation arises. In this context, an interesting problem is the following: given base-level read counts from transcriptome sequencing (RNA-Seq) and approximate knowledge of a gene's location from prior heuristic annotation, is it possible to precisely localize a transcribed region, that is, the set of nucleotides leading to a mature messenger RNA (mRNA), i.e., a mature transcript. This involves identifying the set of nucleotides defining the 5' and 3' untranslated regions (UTR), i.e., the start and end of transcription, as well as the boundaries between exons and introns. In this paper, we use as lax and inclusive definition for a gene, the set of all genomic regions that are transcribed to eventually form a mature transcript (including all exons and introns and the 5' and 3' UTRs) and that can be represented as a discrete interval. Additional motivation for genome re-annotation based on RNA-Seq data is the ability to localize UTRs: while available annotation typically only provides the location of *translated* regions (corresponding to a protein), we consider the annotation of *transcribed* regions, which are usually larger than and include translated regions.

A segmentation of a discrete interval $\{1, \dots, n\}$ of size n (e.g., set of n consecutive nucleotides) is a partition of this interval into disjoint intervals, or segments, whose union is the original interval. The segmentation is usually summarized by a set of change-points, i.e., boundaries between segments. In a statistical inference context, a segmentation is based on random variables indexed by the elements of the interval to be segmented (e.g., RNA-Seq

base-level read counts). The random variables have segment-specific distributions and the change-points correspond to changes in distribution (e.g., in mean parameter). Segmentation methods are particularly adapted to transcript localization using RNA-Seq data: the exons expressed in a given transcript are separated by intronic and non-expressed exonic regions expected to have low read counts (reflecting low transcriptional activity), thus allowing the variation in read counts to be exploited to define the transcript. Because of the discrete nature of RNA-Seq data (number of sequenced reads beginning at each position of the genome), segmentation is based on discrete distributions, such as the Poisson or negative binomial distributions. Figure 3.8 of the Supplementary Materials displays RNA-Seq base-level read counts for a few representative genes in *Saccharomyces cerevisiae*.

This paper is dedicated to the comparison of segmentation methods for the annotation of genomes based on RNA-Seq data. Each segmentation method involves a combination of three choices: (i) a model for the segment-specific read count distributions (e.g., Poisson, negative binomial); (ii) criteria for inferring parameters of the segment-specific distributions (e.g., log-likelihood) and for selecting the number of segments (e.g., penalized log-likelihood); (iii) optimization methods for the criteria in (ii). For ease of implementation, we have limited our comparison to segmentation methods available in R or Bioconductor packages.

We distinguish between two main classes of segmentation methods: those that return a segmentation into a fixed number of segments and those that return a probability for the existence of a change-point at each location. Note that, in some cases, the number of segments might be known (e.g., in the context of re-annotation) and in others it might be part of the statistical inference problem (e.g., in the context of *de novo* annotation or transcript discovery).

The first class includes algorithms that are usually fast enough to deal with long sequences (10^5 to 10^9 base-pairs) and that can be applied to recover an entire set of expressed transcripts or to localize novel transcripts. The *Dynamic Programming Algorithm* (DPA)

is an exact algorithm that returns the optimal segmentation into K segments, according to the log-likelihood criterion, for each K ranging from 1 up to a user-supplied K_{max} (GUTHERY, 1974). Its fast (but still exact) version, the *Pruned Dynamic Programming Algorithm* (PDPA), is implemented in the R package `Segmentor3IsBack` for the negative binomial and Poisson distributions (CLEYNEN *et al.*, under review). Segmentation by binary classification with *CART* (SCOTT and KNOTT, 1974; BREIMAN *et al.*, 1984) is an efficient and extremely fast heuristic algorithm that returns a non-optimal segmentation into K segments, for each K ranging from 1 up to a user-supplied K_{max} , by drastically reducing the number of segmentations explored, but still yielding good results when the signal is not too noisy. When the number of segments is unknown, these algorithms have to be combined with a model selection strategy. Finally, *Pruned Exact Linear Time* (PELT) is an exact algorithm that returns the optimal segmentation according to a penalized log-likelihood criterion and where the number of segments is estimated within the algorithm. These last two algorithms are implemented for the Poisson distribution in the R package `changepoint` (KILLICK and ECKLEY, 2011).

The second class of segmentation approaches includes algorithms with a longer runtime, but that provide credibility intervals (a.k.a., Bayesian confidence intervals) for the location of change-points. They usually deal with shorter sequences (10^3 to 10^4 base-pairs), but can be applied for precise re-annotation of the genome with high confidence. The constrained hidden Markov model (HMM) approach implemented in the package `postCP` (LUONG *et al.*, 2013) uses the PDPA for its parameter initialization. The exact Bayesian approach proposed by RIGAILL *et al.* (2012) is implemented in the R package `EBS` (which is available on the CRAN). Both methods are applicable to the Poisson and negative binomial distributions.

Note that all segmentation methods mentioned thus far are also available for the Gaussian distribution, which is widely-used, for instance, for the identification of copy-number variation based on Comparative Genomic Hybridization (CGH) microarray data.

Numerous other segmentation approaches exist, such as, to only mention a few, least squares regression (BAI and PERRON, 2003), Bayesian inference based on product partition

models and Markov sampling (BARRY and HARTIGAN, 1993), adaptive weights smoothing (HUPÉ *et al.*, 2004), or wavelets (HSU *et al.*, 2005). Since they are not adapted to count data, we do not consider them in our comparison study. Though FREEC (BOEVA *et al.*, 2011) was developed for discrete sequencing data, it applies a Gaussian segmentation method to transformed read counts and is thus not considered here.

The paper is organized as follows. In the next section, we describe our segmentation framework, the methods to be evaluated, and our simulation study design. Then, we present results of the comparison of segmentation methods for different types of biological questions and examine the effect of a classical log-transformation of the data. Finally, we discuss the results and consider extensions to other problems such as copy-number variation.

3.1.2 Methods

Segmentation framework

In the context of re-annotation, the segmentation framework can be formulated as follows. Suppose we have RNA-Seq base-level read counts for a region of the genome represented by nucleotide positions $t \in \{1, \dots, n\}$ and which contains, for simplicity, only one transcript (i.e., we do not consider alternative splicing). For a transcript with K_e exons, the segmentation for the sequence has $K = 2 \times K_e + 1$ segments, where each even (odd) segment corresponds to an exon (intron). Let τ_k , $k = 0, \dots, K$, denote the k^{th} change-point, with the convention that $\tau_0 = 1$ and $\tau_K = n + 1$. Then, the k^{th} segment is defined as the interval $[[\tau_{k-1}, \tau_k[[$ and the corresponding segmentation can be summarized by $\tau = \{\tau_k : k = 0, \dots, K\}$. Finally, the set of all possible segmentations into K segments is denoted by \mathcal{M}_K .

Let Y_t and y_t denote, respectively, the random variable and its realization for the number of aligned reads with first base at position t and let $Y = \{Y_t : t = 1, \dots, n\}$ and $y = \{y_t : t = 1, \dots, n\}$ denote the signal over the entire region to be segmented. Note that strand-specific

reads are mapped and counted separately for each strand and that distinct segmentations are performed on each strand. When comparing segmentation results for actual RNA-Seq data to existing annotation, read length is taken into account by extending the change-point locations τ_k accordingly (this is unnecessary for simulated datasets). We assume that the Y_t are independent random variables with distributions affected by $K - 1$ abrupt changes in their parameters at each of the change-points τ_k . Specifically, the model can be written as

$$Y_t \sim \mathcal{G}(\theta_k, \phi), \quad \forall t \in \llbracket \tau_{k-1}; \tau_k \llbracket, \quad k = 1, \dots, K,$$

where \mathcal{G} is a parametric distribution (e.g., Poisson or negative binomial), θ_k are segment-specific parameters (such as, but not limited to the mean μ_k), and ϕ is a global parameter (e.g., dispersion).

Three statistical inference questions are therefore pertinent in the context of segmentation: (i) the estimation of the number of segments K ; (ii) the estimation of the parameters $\theta = \{\theta_k : k = 1, \dots, K\}$ and ϕ of the distribution \mathcal{G} ; (iii) the estimation of the location $\tau = \{\tau_k : k = 0, \dots, K\}$ of the change-points. Our main concern is the localization of exon/intron boundaries and hence the estimation of $\{\tau_k\}$. While it can be hard in general to estimate K , this parameter is often known in the context of re-annotation. Additionally, although the parameters $\{\theta_k\}$ are typically not of interest, they can often be estimated trivially by maximum likelihood given estimates of K and $\{\tau_k\}$.

Because of the discrete nature of RNA-Seq data, we consider methods that model read counts using a Poisson (\mathcal{P}) or negative binomial (\mathcal{NB}) distribution, that is, assume that

$$\begin{aligned} \mathcal{P} : \quad & \mathcal{G}(\theta_k, \phi) = \mathcal{P}(\theta_k) \\ \mathcal{NB} : \quad & \mathcal{G}(\theta_k, \phi) = \mathcal{NB}(\theta_k, \phi). \end{aligned}$$

Note that, for the Poisson distribution, θ_k coincides with the mean parameter μ_k . For the negative binomial distribution, θ_k denotes the probability parameter ($0 \leq \theta_k \leq 1$) and $\phi > 0$ the dispersion parameter, so that the mean signal on the k^{th} segment is $\mu_k = \phi(1 - \theta_k)/\theta_k$ and the variance $\mu_k(1 + \mu_k/\phi) \geq \mu_k$. Because RNA-Seq read counts typically exhibit over-dispersion, the negative binomial model is most appropriate (ROBINSON *et al.*, 2010;

RISSO *et al.*, 2011). When $\phi \rightarrow +\infty$, with θ_k such that the ratio $\mu_k = \phi(1 - \theta_k)/\theta_k$ remains constant, one recovers the Poisson distribution with parameter μ_k . Our model requires that the dispersion parameter be constant over all segments.

Since the numbers of reads Y_t are assumed to be independent at each position, the log-likelihood can be decomposed into the sum of the log-likelihoods for each segment, i.e.,

$$\log p(y|K, \tau, \theta, \phi) = \sum_{k=0}^{K-1} \sum_{t=\tau_k}^{\tau_{k+1}-1} \log(g(y_t; \theta_k, \phi)),$$

where $g(\cdot; \theta_k, \phi)$ is the probability density function (PDF) of distribution \mathcal{G} . In order to work in a Bayesian framework, one further needs to specify prior distributions $p(K)$, $p(\tau|K)$, and $p(\theta|K, \tau)$. Their choice is discussed in RIGAILL *et al.* (2012).

Segmentation methods

Most segmentation methods comprise two steps. The first combines inference questions (ii) and (iii) by estimating, for a given number of segments K , the location of the change-points $\{\tau_k\}$ and the parameters $\{\theta_k\}$ and ϕ using, for example, maximum likelihood. The second step is then to estimate the number of segments K , resolving inference question (i). Note that some methods such as PELT combine the two steps into one, estimating the parameters of \mathcal{G} , the change-point locations, and the number of segments directly, using, for example, a penalized version of the likelihood.

Estimating K can be viewed as a model selection problem, for which natural approaches include cross-validation and penalized likelihood criteria. Although cross-validation methods have been proposed in the context of segmentation (ARLOT and CELISSE, 2011), the interpretation of cross-validation is problematic due to the spatial structure and hence dependence of the data. Furthermore, the approach is time-consuming and no software implementation is currently available. We therefore focus on likelihood-based goodness-of-fit criteria, where the estimator \hat{K} of the number of segments K maximizes some function $\text{crit}(K; y)$ of the data y with respect to K (for simplicity, we adopt the shorter notation

$c(K)$). We consider specifically the following three criteria, corresponding to three different penalties for the likelihood.

- The natural approach in a Bayesian framework is to maximize the posterior probability of K given the data, i.e., select the \hat{K} maximizing $c(K) = \log p(K|y)$, which in the case of EBS can be computed exactly. A crude approximation leads to a penalized version of the likelihood, $c(K) = \log p(y|K, \tau, \theta, \phi) - K \log(n)$. In the sequel, we refer to both criteria as the Bayesian Information Criterion (BIC), $BIC(K)$.
- The penalized likelihood criterion of CLEYNEN and LEBARBIER (2013), proposed in a non-asymptotic framework that takes into account the complexity of the visited segmentation, is defined as $PL(K) = \log p(y|K, \tau, \theta, \phi) - \beta K \left(1 + 4\sqrt{1.1 + \log\left(\frac{n}{K}\right)}\right)^2$, where β is a constant tuned according to the data.
- The Integrated Completed Likelihood (ICL) criterion is defined in a Bayesian framework as $ICL(K) = \log p(K|y) + \mathcal{H}(K)$, where the left term $\mathcal{H}(K) = -\sum_{m \in \mathcal{M}_K} p(\tau|y, K) \log p(\tau|y, K)$ is the posterior entropy. Indeed, the segmentation τ can be viewed as an unobserved variable, in the sense that the segment labels of each data point y_t are unknown. RIGAILL *et al.* (2012) introduced this criterion in the context of change-point detection and showed that it performs better than other criteria such as the BIC or DIC (Deviance Information Criterion, i.e., the expected deviance of the model). In a frequentist framework, the ICL criterion can be approximated by $ICL(K) = BIC(K) - \sum_{m \in \mathcal{M}_K} p(\tau|y, K, \theta, \phi) \log p(\tau|y, K, \theta, \phi)$.

If one is not concerned with obtaining an estimate for the number of segments K or if one does not trust the estimation of K , the following two Bayesian approaches are available. The first applies the BIC directly to the segmentation, so that $BIC(\tau) = \log p(\tau|y)$, and chooses the $\hat{\tau}$ that maximizes this criterion. In our study, results using $BIC(\tau)$ and $ICL(K)$ were very similar and only the later will be discussed. The second approach is to integrate posterior probabilities of interest (as those mentioned next) over the possible values of K rather than choose an optimal one (e.g., method EBS-a discussed below). Such model averaging presupposes the ability to compute posterior distributions for K and τ .

Finally, to allow precise and confident re-annotation, it is useful to obtain credibility intervals. The posterior distribution of the j^{th} change-point of a segmentation into k segments is

$$p_{\tau_j, y, k}(t) = \mathbf{P}\{\tau_j = t | Y = y, K = k\}, \quad \forall t \in \llbracket 1, n \rrbracket, \quad (3.1)$$

from which we can derive, by model averaging, the probability of, say, the first change-point occurring at position t ,

$$\mathbf{P}\{\tau_1 = t | Y = y\} = \sum_k p_{\tau_1, y, k}(t) \mathbf{P}\{K = k | Y = y\}, \quad (3.2)$$

where $\mathbf{P}\{A\}$ is the probability of event A . One can then define 95% credibility intervals by, for instance, selecting values of highest posterior probability until 95% coverage.

In our comparison of segmentation algorithms, we are therefore interested in the following functionality: (i) the ability to model RNA-Seq read counts using a discrete distribution, such as the Poisson or negative binomial; (ii) the ability to estimate the number of segments K according to criteria such as those mentioned above; (iii) the possibility to obtain credibility intervals. The left part of Table 3.1 summarizes the available functionality for the algorithms introduced in Section 3.1.1.

Simulation study design

Datasets

The simulation study was conceived to mimic typical RNA-Seq data. We used as benchmark strand-specific, poly(A)-selected *S. cerevisiae* RNA-Seq data from the Sherlock Laboratory at Stanford University (RISSE *et al.*, 2011) and publicly available from the NCBI's Sequence Read Archive (SRA, <http://www.ncbi.nlm.nih.gov/sra>, accession number SRA048710). Reads were mapped to the reference genome using Bowtie (LANGMEAD *et al.*, 2008) and strand-specificity and read-length information was taken into account in our analysis. We selected a set of five genes (YAL038W, YAL035W, YAL030W, YAL019W, and YAR008W) that were previously annotated by NAGALAKSHMI *et al.* (2008)

Algorithm	Functionality							Comparison criteria				
	Distribution		Model selection				Model averaging	Global	Local	ROC	Credibility	Run-
	\mathcal{P}	\mathcal{NB}	$BIC(K)$	$BIC(\tau)$	$PL(K)$	ICL		fit	fit	curves	intervals	time
CART	⊗		⊗					×	×			×
PELT	⊗		⊗					×	×			×
PDPA	×	⊗	×		⊗			×	×			×
postCP	×	⊗	×			⊗		×	×	×	×	×
EBS	×	⊗	×	×		⊗	⊗	×	×	×	×	×

Table 3.1: **Properties of segmentation algorithms.** Left: \times indicate available functionality in terms of distribution and model selection; \otimes indicate methods retained for the simulation study (see RESULTS AND DISCUSSION section). Right: \times indicate comparison criteria that can be computed for each method.

and that span representative scenarios for yeast RNA-Seq data, in terms of gene length, number of exons, and read counts. Figure 3.8 of the Supplementary Materials shows the original unnormalized base-level read counts for these five genes.

In order to choose realistic values for the parameters of the distributions used to simulate read counts, we considered the true, known segmentations of the five genes. For each gene, we first fit a negative binomial distribution $\mathcal{NB}(\theta_k, \phi)$ to each segment k and estimated θ_k using the method of moments and ϕ using a modified version of the Johnson and Kotz’s estimator (JOHNSON *et al.*, 2005). Specifically, for each sliding window of size h equal to twice the size of the longest zero band, we computed the method of moments estimator of ϕ , using the formula $\phi = \mathbf{E}^2(X)/(\mathbf{V}(X) - \mathbf{E}(X))$, and retained the median over all windows. We also computed the maximal value of the read counts over the entire region. The results are given in Table 3.2.

Following LAI *et al.* (2005), we created an artificial four-exon gene, with $K = 9$ segments defined by $m = (1, 101, 121, 221, 271, 371, 471, 571, 1071, 1171)$. An odd segment corresponds to an intronic region (average size 100 bases), while an even segment corresponds to an exon (length varying from 20 to 500 bases).

Gene	Length	Dispersion, $\hat{\phi}$	Empirical mean, $\hat{\mu}_k$	$\max y_t$
YAL038W	2003	0.3121	(0.2928, 325, 0.2738)	6216
YAL035W	3509	0.2523	(0.2174, 7.81, 0.1232)	690
YAL030W	967	0.2966	(0.0044, 1.55, 0.08, 3.62, 0.103)	167
YAL019W	3896	0.2721	(0, 1.196, 0.0266)	25
YAR008W	1328	0.2758	(0, 1.325, 0.0466)	34

Table 3.2: **Estimates of model parameters for each of the five yeast genes.** For each segment, parameters correspond to a negative binomial distribution with mean $\mu_k = \phi(1 - \theta_k)/\theta_k$ and dispersion ϕ .

We considered three simulation scenarios, corresponding to two parametric distributions and one resampling-based distribution.

- For the *Negative Binomial* (NB) scenario, with $\mathcal{G}(\theta_k, \phi) = \mathcal{NB}(\theta_k, \phi)$, we used the artificial four-exon gene segmentation and set the dispersion parameter ϕ to 0.27 for all segments. For odd segments (i.e., introns), we chose $\theta_{2k+1} = 0.9$, and for even segments (i.e., exons), we allowed θ_{2k} to vary smoothly between 0.2 and 0.001. For each value of θ_{2k} , we simulated 100 datasets.
- For the *Mixture of discrete Uniforms* (MU) scenario, with $\mathcal{G}(\theta_k, \phi) = \frac{1}{2}\mathcal{U}(\llbracket 0, \theta_k/2 \rrbracket) + \frac{1}{2}\mathcal{U}(\llbracket 0, \theta_k \rrbracket)$, we again used the artificial four-exon gene segmentation and set $\theta_{2k+1} = 4$ and allowed θ_{2k} to vary smoothly between 24 and 6,250. For each value of θ_{2k} , we simulated 100 datasets.
- For the *Resampling* (RS) framework, we considered the true segmentation of each of the five yeast genes (NAGALAKSHMI *et al.*, 2008) and resampled the counts of each segment at random, with replacement, i.e., $\mathcal{G}(\theta_k, \phi) = \text{sample}(\llbracket y_{\tau_k}; y_{\tau_{k+1}} \rrbracket)$. For each gene, we repeated this procedure 100 times.

In the remainder of the paper, we let μ represent the mean signal intensity over even segments (i.e., exons), so that μ_{2k} is equal to $\phi(1 - \theta_{2k})/\theta_{2k}$ in the NB simulations and $3\theta_{2k}/8$ in the MU simulations and refers to the qualitative level of expression of the genes in the RS simulations. With parameters chosen as above in the NB and MU simulations, the different θ_{2k} yield comparable signal intensities μ_{2k} . Note that we have associated μ with the level of expression of a gene, but that it can also relate to the sequencing coverage of an experiment. While we will only refer to the former in the manuscript, low-expressed genes from experiments with higher coverage might present the same characteristics as highly-expressed genes from experiments with lower coverage.

Comparison criteria

In the simulation study, the segmentation methods are compared according to the following criteria.

- The *global fit* index gf assesses the global quality of a proposed segmentation, in the sense that it reflects the agreement between the true segmentation τ and the estimated segmentation $\hat{\tau}$ over all pairs of bases in the region. Specifically, let C_t be the true index of the segment to which base t belongs and let \hat{C}_t be the index estimated by the method, then

$$gf = \frac{2}{(n-1)(n-2)} \sum_{s=1}^n \sum_{t=s+1}^n \left[\mathbf{1}_{C_t=C_s} \mathbf{1}_{\hat{C}_t=\hat{C}_s} + \mathbf{1}_{C_t \neq C_s} \mathbf{1}_{\hat{C}_t \neq \hat{C}_s} \right].$$

- The *local fit* index lf assesses the ability to recover a particular change-point c and is defined by

$$lf(c) = \delta_c / P_{k(\hat{\tau})},$$

where δ_c is equal to 1 if the method finds a change-point at most three bases away from c and 0 otherwise, $k(\hat{\tau})$ is the number of segments of the segmentation $\hat{\tau}$, and P_k is the probability that a segmentation into k segments has a change-point at c , i.e., $P_k = \frac{k-1}{n-1}$. Note that while the choice of a three-base tolerance threshold is somewhat subjective and allows change-points to be detected more easily, the ranking of the methods is robust to the value of the threshold (results not shown).

- *Receiver Operator Characteristic* (ROC) curves for methods yielding change-point

probabilities as in Equations (3.1) and (3.2).

- For the Resampling Simulation scenario and methods yielding change-point probabilities, the percentage of true change-points covered by 95% credibility intervals defined by starting from the mode of the distributions in Equation (3.1) or (3.2) and adding the next most probable location until 95% (or slightly more because of the discrete nature of the distribution) of the mass has been reached. This leads to intervals that may not be contiguous, but have the smallest possible length.
- The average run-time.

The right part of Table 3.1 indicates which criteria are applicable for each of the algorithms to be evaluated.

3.1.3 Results and discussion

Preliminary remarks

We first compared all algorithms with every available distribution. Our results show (see Figure 3.9 of Supplementary Materials) that when an algorithm was implemented for both the Poisson and negative binomial distributions (PDPA, postCP, and EBS), the latter always performed better. This is to be expected, as read counts typically exhibit over-dispersion. For this reason, as well as to simplify the reporting of results and figures, we only retained PDPA, postCP, and EBS with the negative binomial distribution and CART and PELT with the Poisson distribution, as indicated by \otimes symbols in the left part of Table 3.1. Although this may appear to bias the results in favor of PDPA, postCP, and EBS, the comparison is still fair, as the restriction of CART and PELT to the Poisson distribution and their inability to accommodate over-dispersion is a clear limitation of these methods. Furthermore, the Poisson distribution is included as special case of the negative binomial implementation of PDPA, postCP, and EBS.

Method postCP failed to return a segmentation for a number of simulations (221 times for the mixture of uniforms scenario, 111 times for the negative binomial scenario and, for

the RS scenario, 21 times for gene YAL030W, 20 times for gene YAL035W, and 5 times for gene YAR008W). The results presented in this section exclude these cases for postCP.

Figure 3.1 displays the segmentations obtained with each of five methods for gene YAL030W. In this particular example, postCP, EBS, and PDPA recover the true segmentation, while PELT largely over-estimates the number of exons and CART misses the 3' boundary of the first exon and erroneously splits the second exon into three.

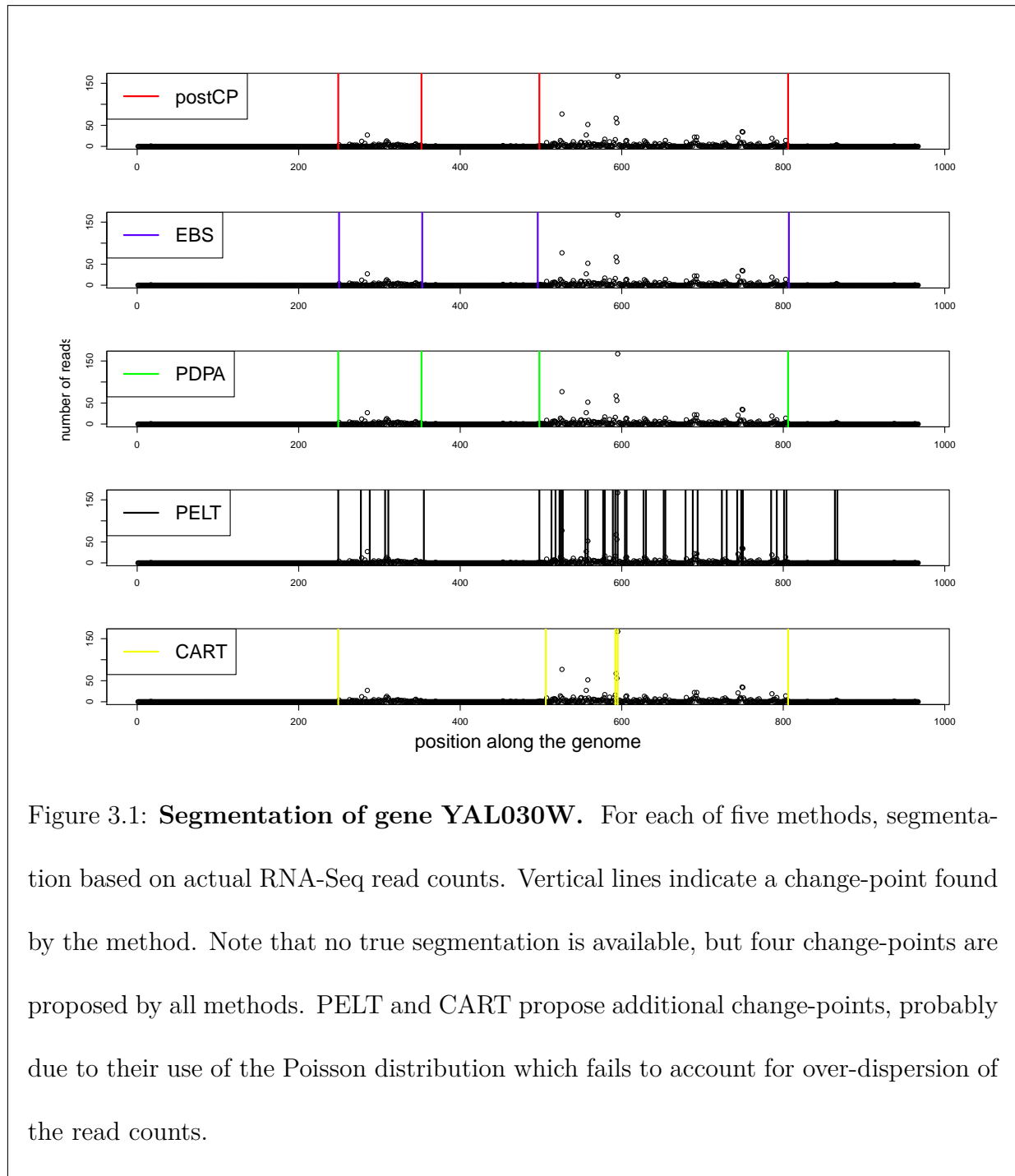
Quality, precision, and confidence in the change-point localization

For both the negative binomial and mixture of uniforms simulation scenarios, methods implemented with the negative binomial distribution performed better than others according to the global fit criterion. As expected, we observed a general trend of slight improvement as the signal intensity μ increased. The left side of Figure 3.2 illustrates the performance of each method according to the global fit index for a particular value of μ corresponding to a moderate level of expression.

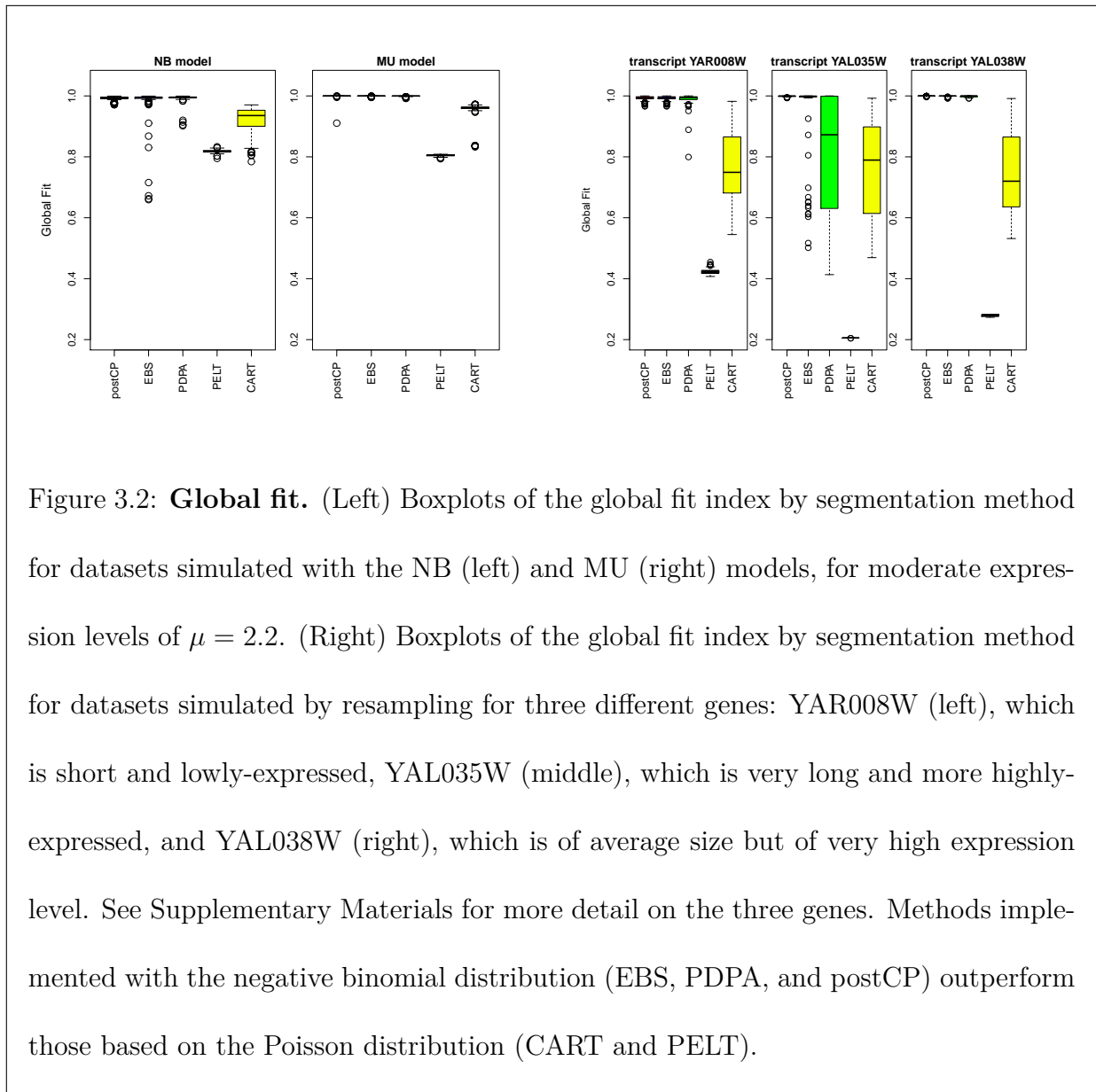
On the datasets simulated by resampling, however, we noticed that the effect of length was significant (right side of Figure 3.2). For instance, for the long gene YAL035W, methods PDPA and PELT drastically worsened. EBS and postCP consistently showed satisfying results and CART and PELT remained the least accurate methods.

Figure 3.3 displays the performance of each method in terms of local fit averaged over the first and last change-points, which are of particular interest in context of UTR annotation. We observe that local fit improves for all methods as the expression level μ increases, although the methods tend to overestimate the number of segments when μ is high (see Figure 3.14 of Supplementary Materials).

Methods EBS and postCP yield posterior change-point probabilities for any given genomic location (see Equation (3.1)). An example is given in Figure 3.10 of the Supplemen-



tary Materials. This can be used to evaluate false positive and false negative rates for, say, the first change-point τ_1 . Specifically, for a given simulation and threshold s , a position t is declared as first change-point if $\mathbf{P}\{\tau_1 = t|y, K\} \geq s$. Averaging the resulting proportions of false positives and false negatives over simulations and varying s leads to the ROC-like



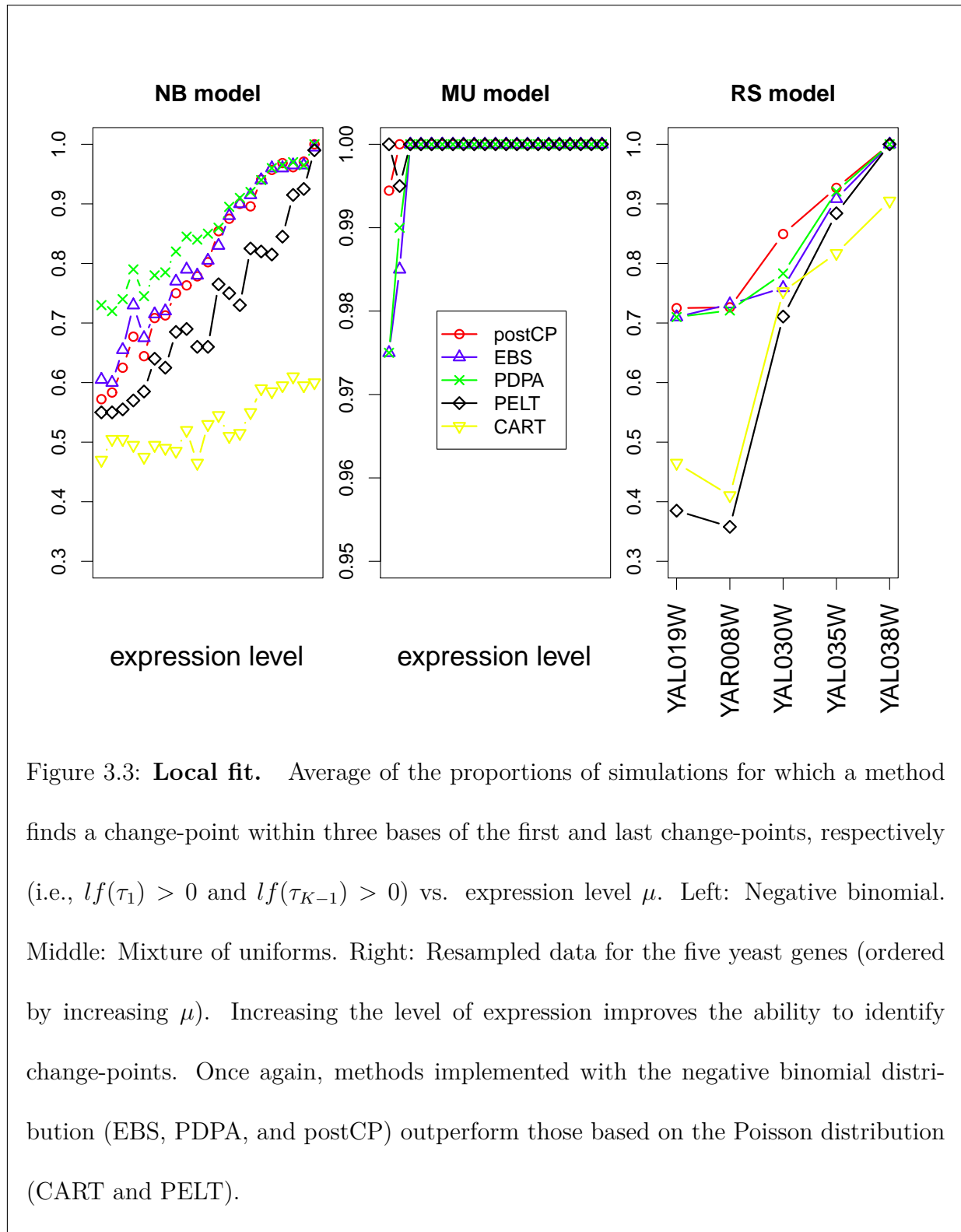
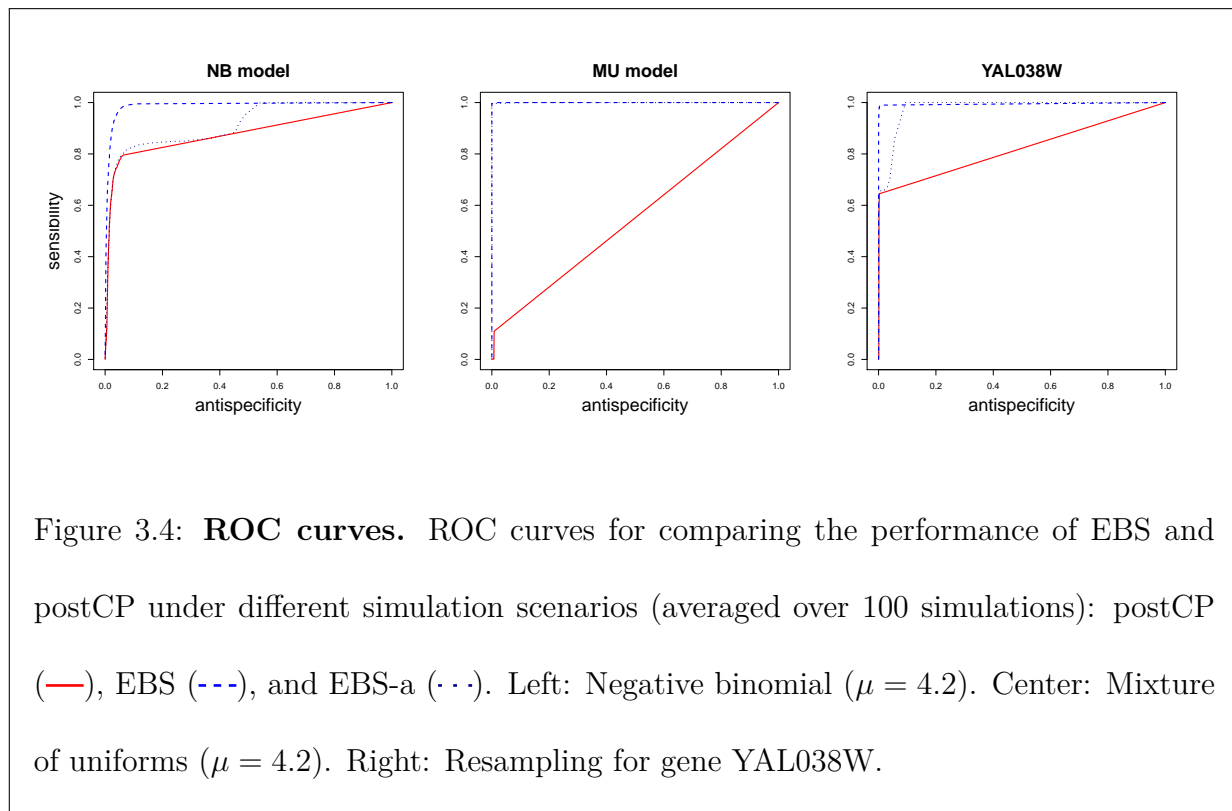


Figure 3.3: **Local fit.** Average of the proportions of simulations for which a method finds a change-point within three bases of the first and last change-points, respectively (i.e., $lf(\tau_1) > 0$ and $lf(\tau_{K-1}) > 0$) vs. expression level μ . Left: Negative binomial. Middle: Mixture of uniforms. Right: Resampled data for the five yeast genes (ordered by increasing μ). Increasing the level of expression improves the ability to identify change-points. Once again, methods implemented with the negative binomial distribution (EBS, PDPA, and postCP) outperform those based on the Poisson distribution (CART and PELT).

curves of Figure 3.4.



postCP's performance is acceptable when the data are simulated according to its negative binomial model, but very poor for the mixture of uniforms. Furthermore, performance deteriorates with increasing expression level μ (results not shown). A possible explanation is the very sharp aspect of the posterior distribution for the change-point location, which leads to false positives as soon as the mode is not equal to the true change-point. For the Resampling scenario, postCP's performance is good, but again worsens as the level of expression increases (see Figure 3.11 of the Supplementary Materials). Method EBS has good overall performance, with nearly perfect ROC for each model. Averaging over the number of segments K (EBS-a), as in Equation (3.2), doesn't seem to improve the results.

Both postCP and EBS also provide posterior credibility intervals. Table 3.3 presents the average width and the percentage of nominal 95% credibility intervals covering the true change-point τ_1 (over 100 simulations). We display the results for the first change-point τ_1 for ease of comparison with the Bayesian aggregation method EBS-a (indeed, studying the k^{th} change-point would require a segmentation into at least $k + 1$ segments and thus

Gene	Interval length			Coverage		
	postCP	EBS	EBS-a	postCP	EBS	EBS-a
YAL019W	18	20	393	0.35	0.95	1
YAR008W	16	17	354	0.2	0.98	1
YAL030W	10	37	398	0.12	0.99	1
YAL035W	8	10	322	0.1	0.97	1
YAL038W	4	7	198	0.1	0.99	1

Table 3.3: **Credibility intervals.** Median length of the 95% credibility intervals and percentage of simulations for which the intervals covered the true change-point (out of 100).

the modification of the prior used for K), but results are similar for other change-points. The empirical coverage of EBS is close to the nominal credibility of 95%, with reasonably narrow intervals. The empirical coverage of EBS-a is 100%, at the price of huge credibility intervals, precluding its use in practice. This observation and the ROC curves of Figure 3.4 suggest that EBS-a yields a well-located posterior mode, but too large a posterior variance. postCP showed very poor coverage (fewer than 40% of the simulations had a 95% credibility interval covering the true location), due to its small credibility intervals that do not account for uncertainty in the estimation of the parameters θ_k and ϕ . Results of the comparison are similar across expression levels μ .

Number of change-points

All results presented up to now are based on methods that involve estimating the number of segments K . The accuracy of the resulting segmentation could therefore be affected by a poor choice of K . Figures 3.12 to 3.16 of the Supplementary Materials show the distribution

of estimates of K . In the context of genome re-annotation, where it is reasonable to assume that the number of segments is known (for instance, a gene with K_e exons will have $K = 2 \times K_e + 1$ segments), it is of interest to compare segmentations based on the true and estimated K . Note that PELT cannot take advantage of the knowledge of K , as the estimation of K is embedded in the algorithm.

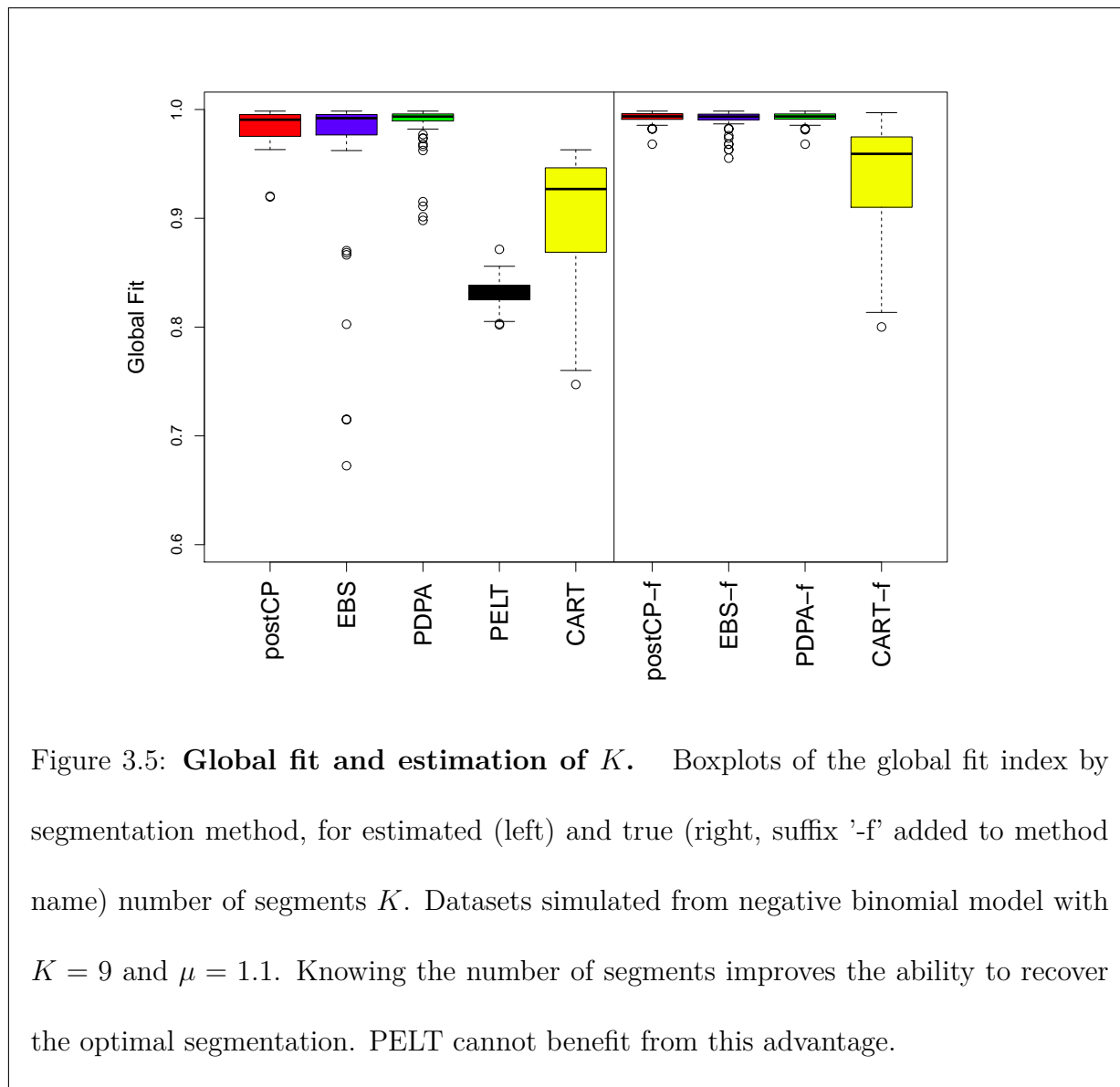
The boxplots in Figure 3.5 illustrate the advantage of providing the true K to a segmentation method: for methods postCP, EBS, and CART, the global fit index gf is less variable and higher with the true K (right) than with estimated K (left). This trend is mostly observed on datasets simulated with the negative binomial model. As expected, as the expression level μ increases and the segmentation becomes more obvious, the impact of the choice of K for methods postCP and EBS lessens, as the ICL criterion becomes more accurate. In the case of PDPA, the model selection criterion already provides the true value of K in more than 90% of the simulations, thus the knowledge of K does not yield a noticeable gain.

The ROC curves in Figure 3.6 illustrate the impact of the estimation of K for methods EBS and postCP in terms of false positive and false negative rates. The gain from using the true K lessens as the level of expression increases, regardless of the performance of the methods; while performance improves for method EBS, postCP worsens with expression level.

Extension to more complex organisms

A natural question is how the methods would compare for an organism with a more complex gene structure than *S. cerevisiae*. We have therefore considered the artificial scenario in which two of the isoforms of the *Drosophila melanogaster* gene Inr-a, Inr-a-RB (six exons) and Inr-a-RC (two exons), are expressed at different levels (Figure 3.17 of the Supplementary Materials illustrates the gene and its isoforms).

In our simulation, we used the annotation of the Inr-a gene from FlyBase (<http://>



www.flybase.org) to define the true segmentation. To simulate read counts, we pooled the observed counts from several yeast genes to create three classes of expression: intronic, low, and medium. Then, on each segment, the read counts were obtained by re-sampling at random, with replacement from the three groups. Specifically, we used the intronic class for segments corresponding to intronic regions of the two isoforms, the medium expression class for exons of Inr-a-RB, and the low expression class for exons of Inr-a-RC. Thus, for exons shared by the two isoforms, read counts are sums of counts from the low and medium classes. This created a synthetic signal for a gene of length 5,000 nucleotides with 14

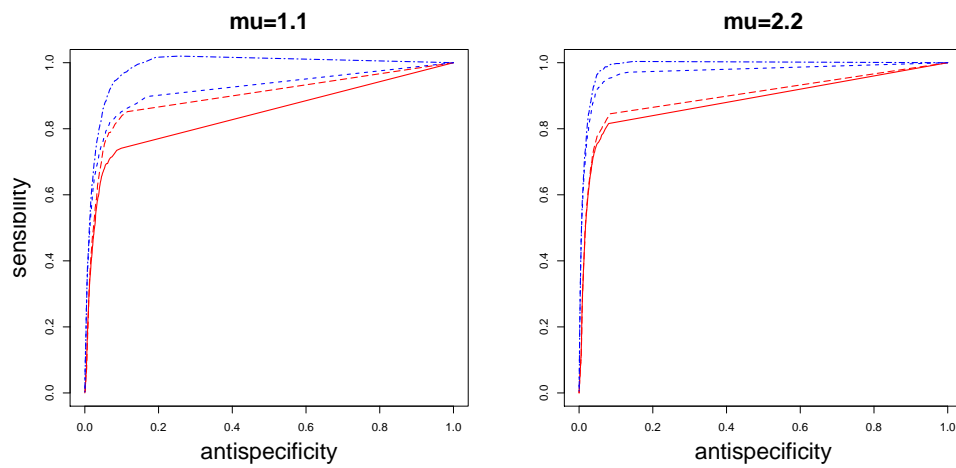


Figure 3.6: **ROC curves and estimation of K .** ROC curves for comparing the performance of EBS and postCP with known and estimated number of segments K : postCP with estimation of K (—), postCP with known K (---), EBS with estimation of K (-.-), and EBS with known K (·-·). Data simulated from negative binomial model with $\mu = 1.1$ and $\mu = 2.2$. Once again, the knowledge of K improves the performance of the algorithms.

segments.

The right side of Figure 3.18 of the Supplementary Materials shows that all methods but PDPA fail to recover the right number of segments. This leads to poor results in terms of local and global fit. However, the prior information on K allows method EBS to yield almost perfect ROC curves (see the left side of Figure 3.18 of Supplementary Materials), while methods PELT and CART still fail to retrieve an acceptable number of true change-points.

The segmentation methods considered in this article can be applied to organisms of a large range of complexity (in the number of exons, isoforms, etc.), provided that the exons of different genes of interest do not overlap, in which case it would not be possible to assign a segment to a specific gene. Fast methods such as CART, PELT or PDPA can be applied regardless of gene length. However, the EBS algorithm is restricted to sequences no longer than 10^4 bases.

Transformation of the data

One might be interested in transforming the discrete RNA-Seq read counts to allow the use of a wider range of methods, for instance, continuous data segmentation methods developed for microarrays. Because of encouraging results with EBS, we compared its performance on the resampled datasets, with the negative binomial distribution, as above, and with the Gaussian distribution applied to log-transformed counts ($\tilde{y}_t = \log(y_t + 1)$). We also applied the variance-stabilization transformation corresponding to the negative binomial distribution (which involves the *arsinh* function), but the results were similar to the widely-used and dispersion-independent log-transformation and thus are not presented here.

We observed that on the RS simulations, the two approaches yield very similar results. However, as illustrated by the ROC curves in Figure 3.19 of the Supplementary Materials, the negative binomial distribution is better for more complex scenarios, where some segments can be very small, as is the case with *D. melanogaster* introns.

3.1.4 Conclusion

This simulation study showed that each method is adapted to a different type of problem. CART and PELT perform worse in all situations for distinct reasons: CART is a heuristic that is most appropriate when the signal is long and segments are well-delimited (for example, large changes in the mean), while PELT fails because of its inability to choose an appropriate number of segments. Indeed, PELT was designed to segment profiles in which the number of segments increases with the length of the signal, which is not the case in our framework.

PDPA showed excellent results in proposing a segmentation close to the true one, especially when the signal was not both very long and high. The criterion used for the choice of the number of segments yielded good performance even when other methods failed. Its use is promising in a range of biological settings, such as transcript discovery or assessment of which genes are expressed.

Finally, postCP and EBS demonstrated the ability to both propose a segmentation that is very close to the true one and return distributions for the location of change-points, thereby allowing precise and confident re-annotation. Both methods showed equivalent results for their optimal segmentation, but EBS had better results in terms of ROC curves on the true datasets and showed a clear improvement when the number of segments was known. Figure 3.7 illustrates the results of method EBS on actual RNA-Seq data for the five yeast genes of interest.

Segmentation methods are of interest in related contexts such as whole-genome (re-)annotation or copy-number estimation using DNA-Seq data. However, in practice, few methods are fast enough to be applied in those frameworks. Indeed, in our simulation study, performed on a standard computer (Intel-Core2 Duo CPU P8400, 2.26 GHz x 2 with 3 Gio of RAM), the average run-times were very different among the methods. PELT and CART were almost instantaneous, while PDPA (respectively postCP) needed a few seconds per simulation (about 4s (resp. 10s) on model-simulated datasets, up to 20s (resp. 50s) on the longer genes). EBS was by far the slowest, needing about 15 seconds for model-

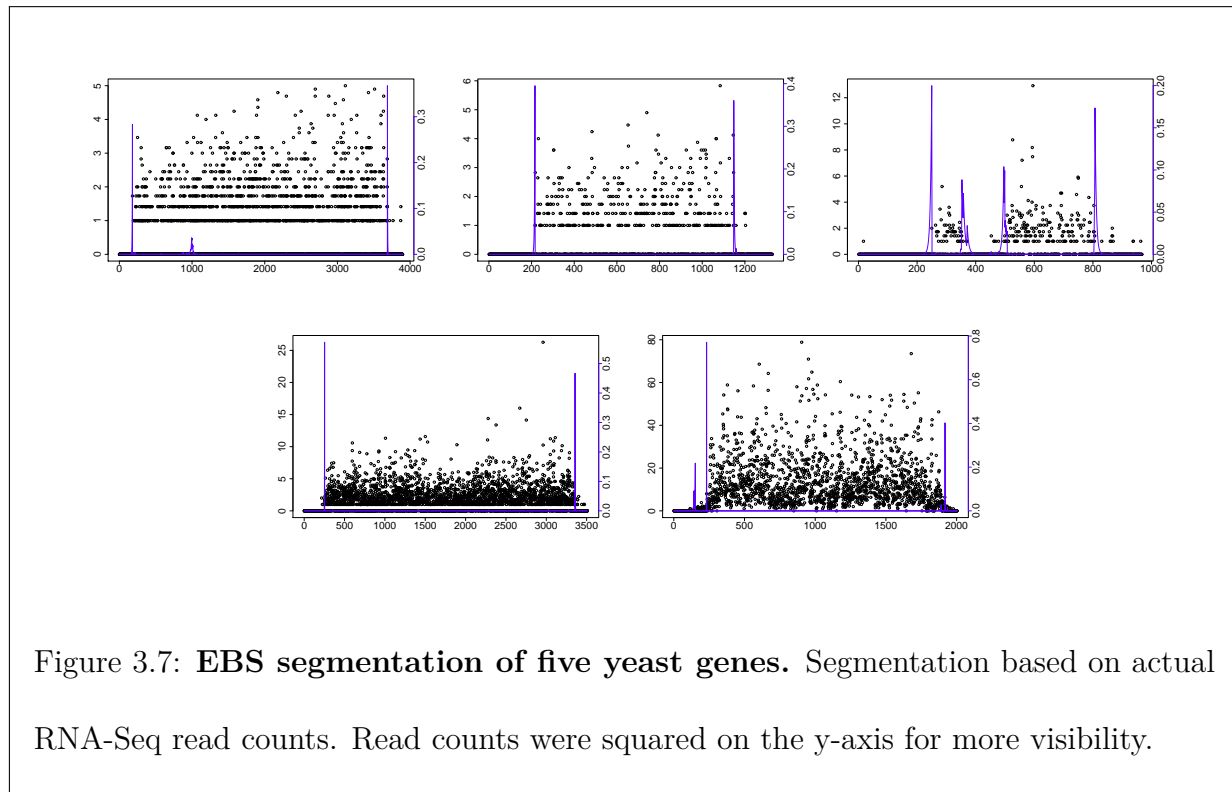


Figure 3.7: **EBS segmentation of five yeast genes.** Segmentation based on actual RNA-Seq read counts. Read counts were squared on the y-axis for more visibility.

simulated datasets and up to 6.5 minutes on the longer genes. While we would recommend using method EBS (with prior information on K when available) for targeted transcript re-annotation, its run-time prohibits its use for larger segmentation problems. A possible strategy would consist in first applying PDPA to large regions in order to delimit smaller regions of interest and then using EBS to obtain confidence intervals on the change-point locations within the smaller regions.

Availability of supporting data The dataset supporting the results of this article is available in the Sequence Read Archive repository, <http://www.ncbi.nlm.nih.gov/sra>, with the accession number SRA048710.

Additional Files *Additional file 1 — Supplementary figures Additional figures referred to in the main article as Figures in Supplementary Materials.

Supplementary Materials

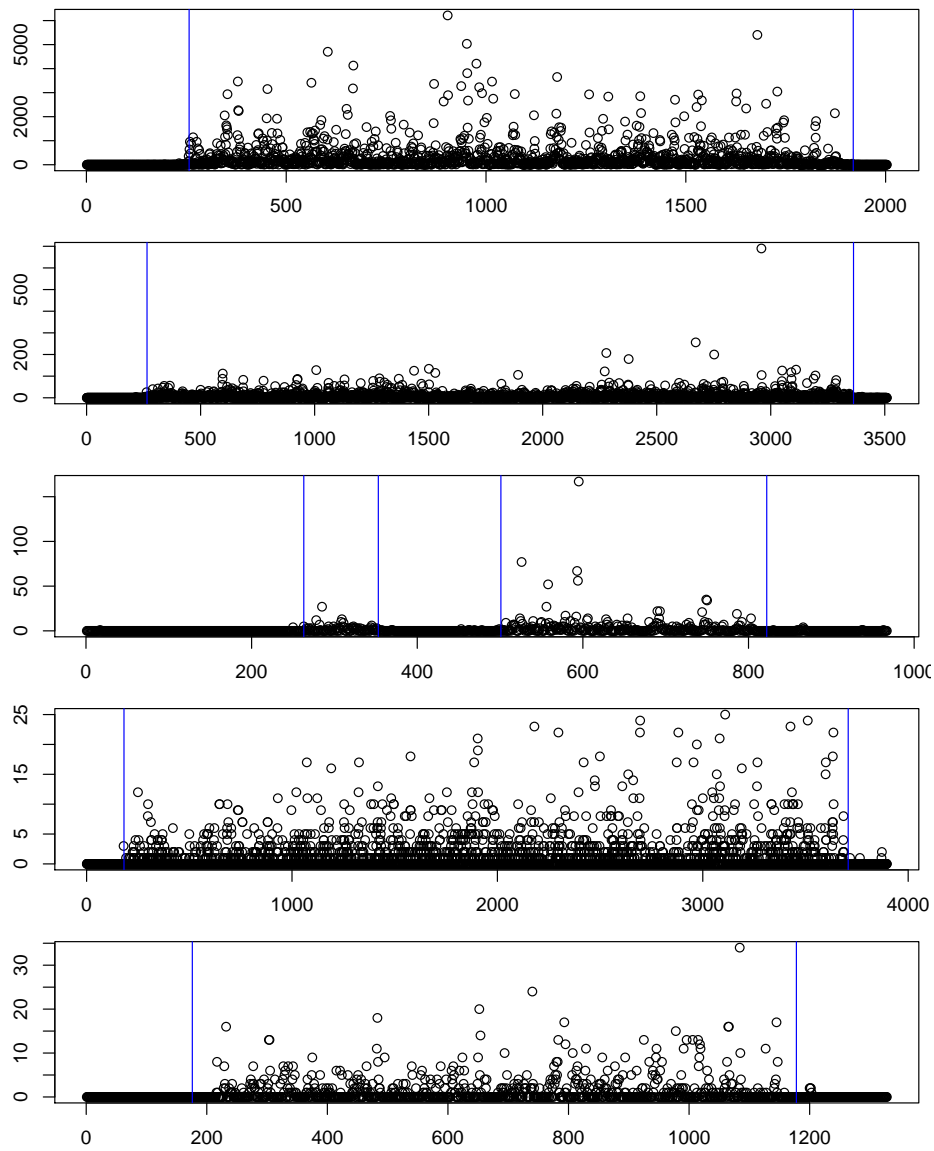
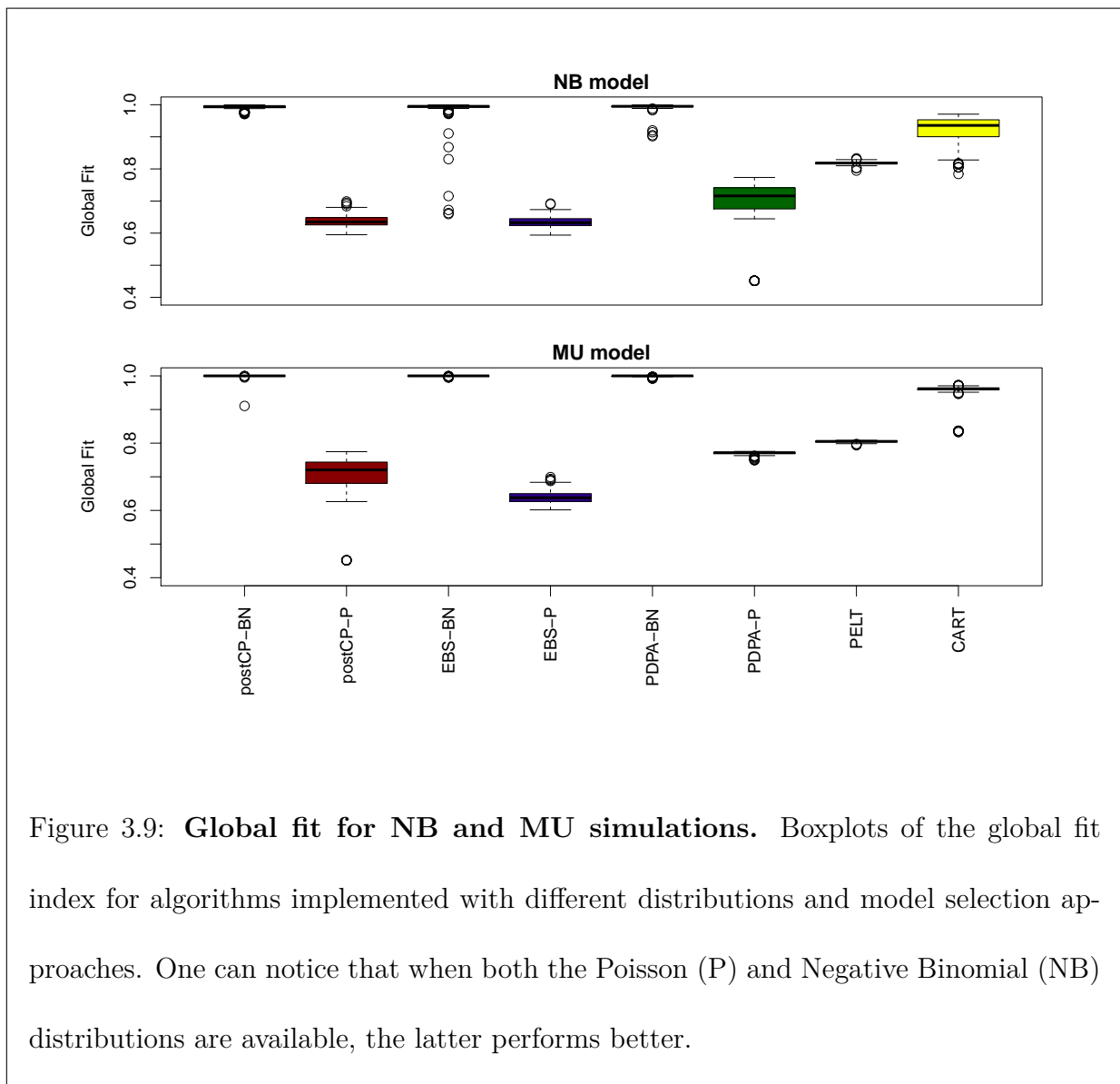
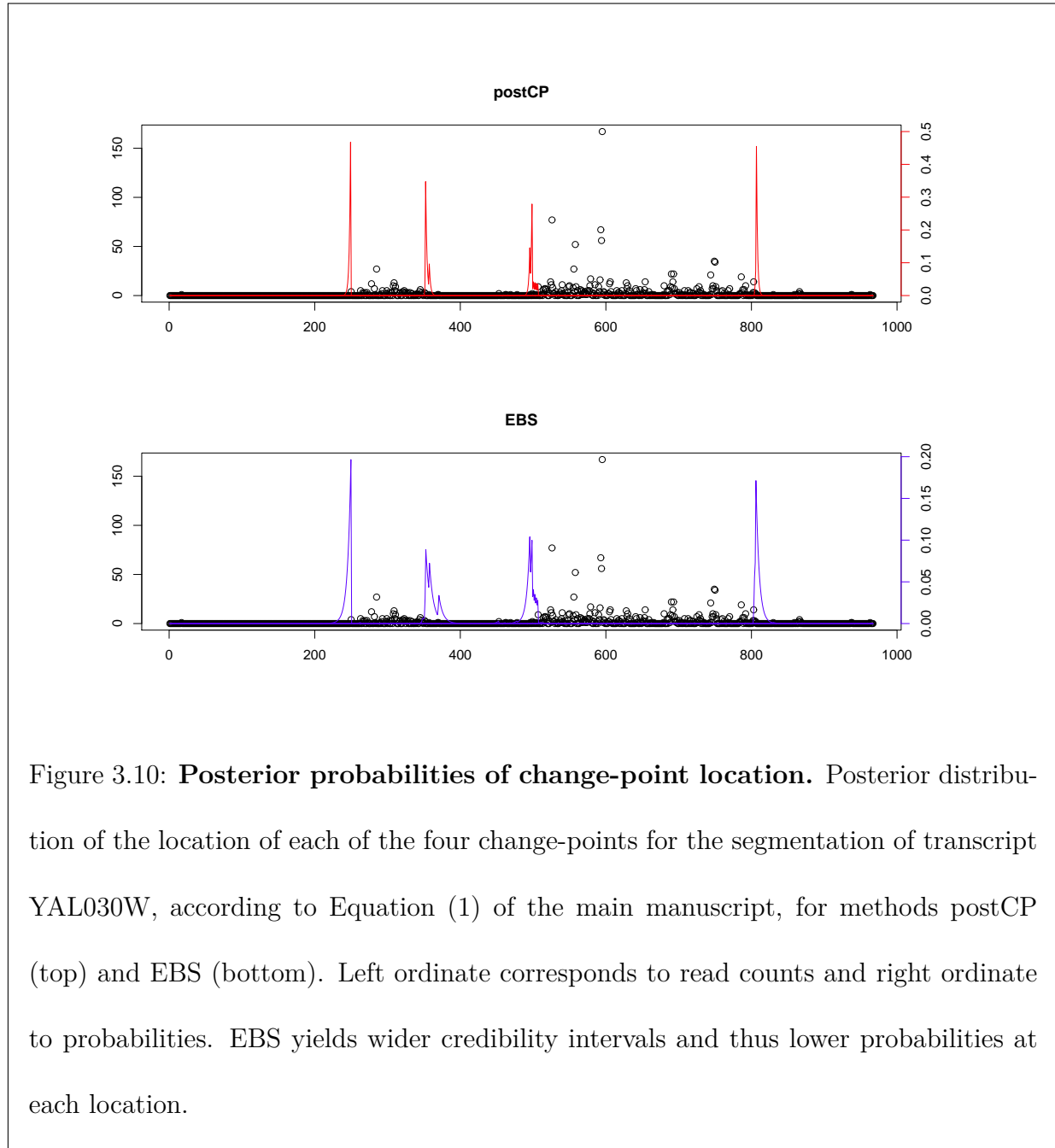


Figure 3.8: **Yeast dataset.** Original unnormalized base-level read counts for five *Saccharomyces cerevisiae* transcripts (RISSE *et al.*, 2011). Vertical blue lines correspond to the annotation given by NAGALAKSHMI *et al.* (2008). Note the different scales for the abscissa and ordinate: transcripts differ in length, number of exons, and expression level.





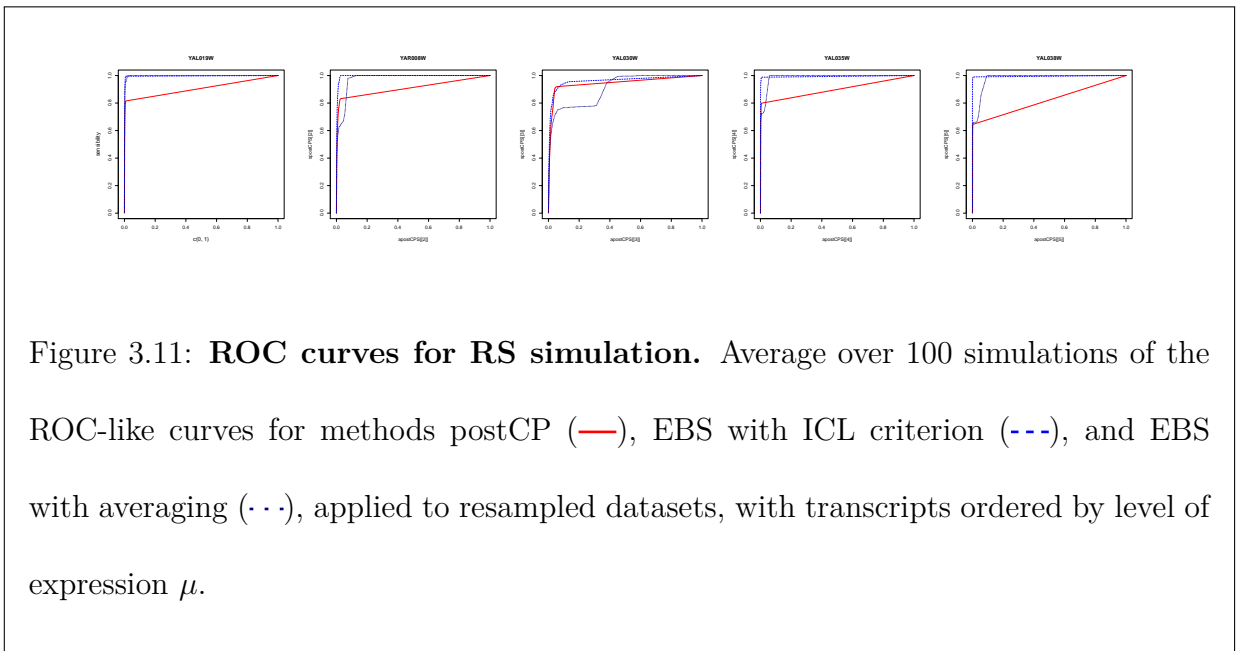
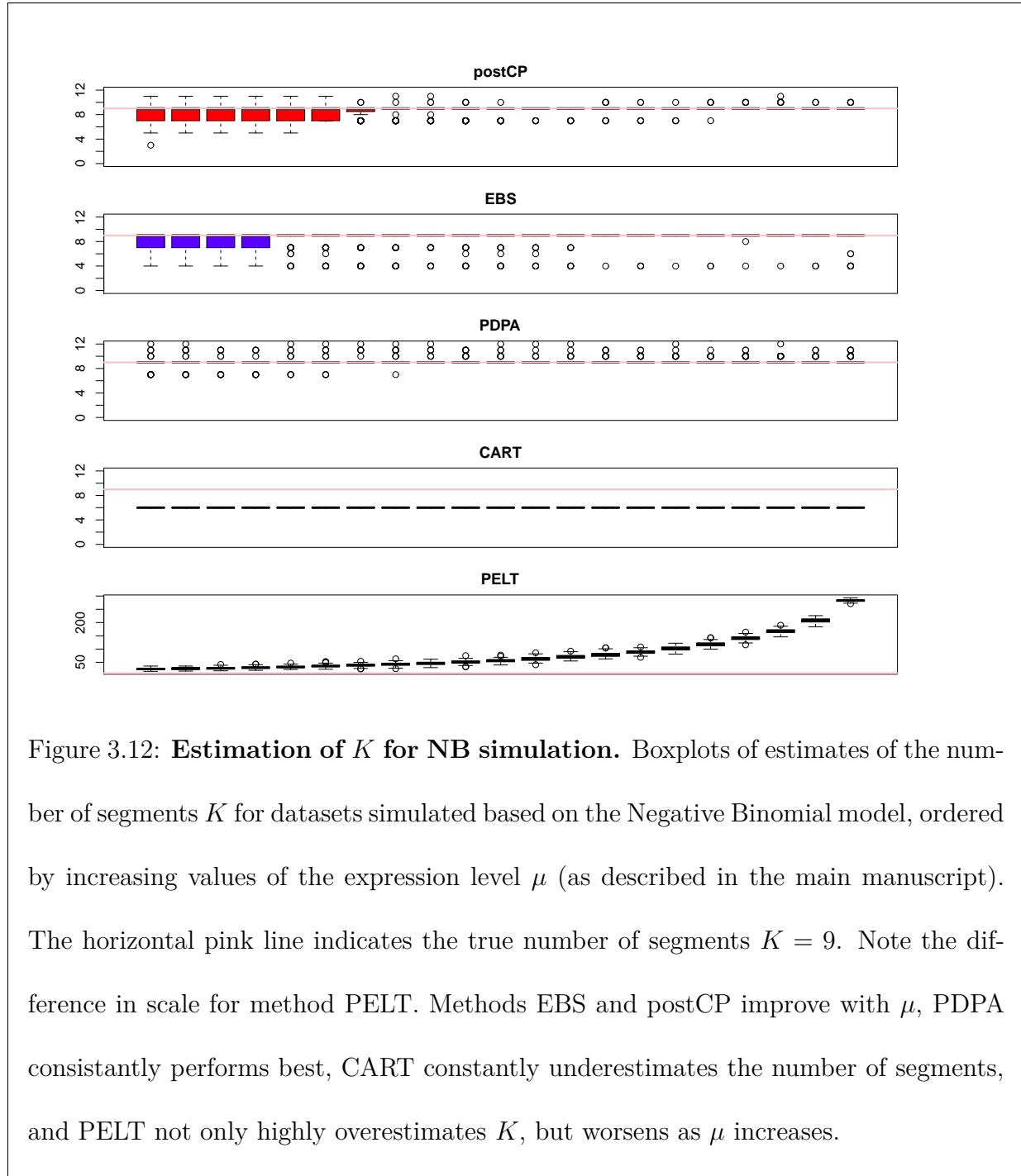


Figure 3.11: **ROC curves for RS simulation.** Average over 100 simulations of the ROC-like curves for methods postCP (—), EBS with ICL criterion (---), and EBS with averaging (···), applied to resampled datasets, with transcripts ordered by level of expression μ .



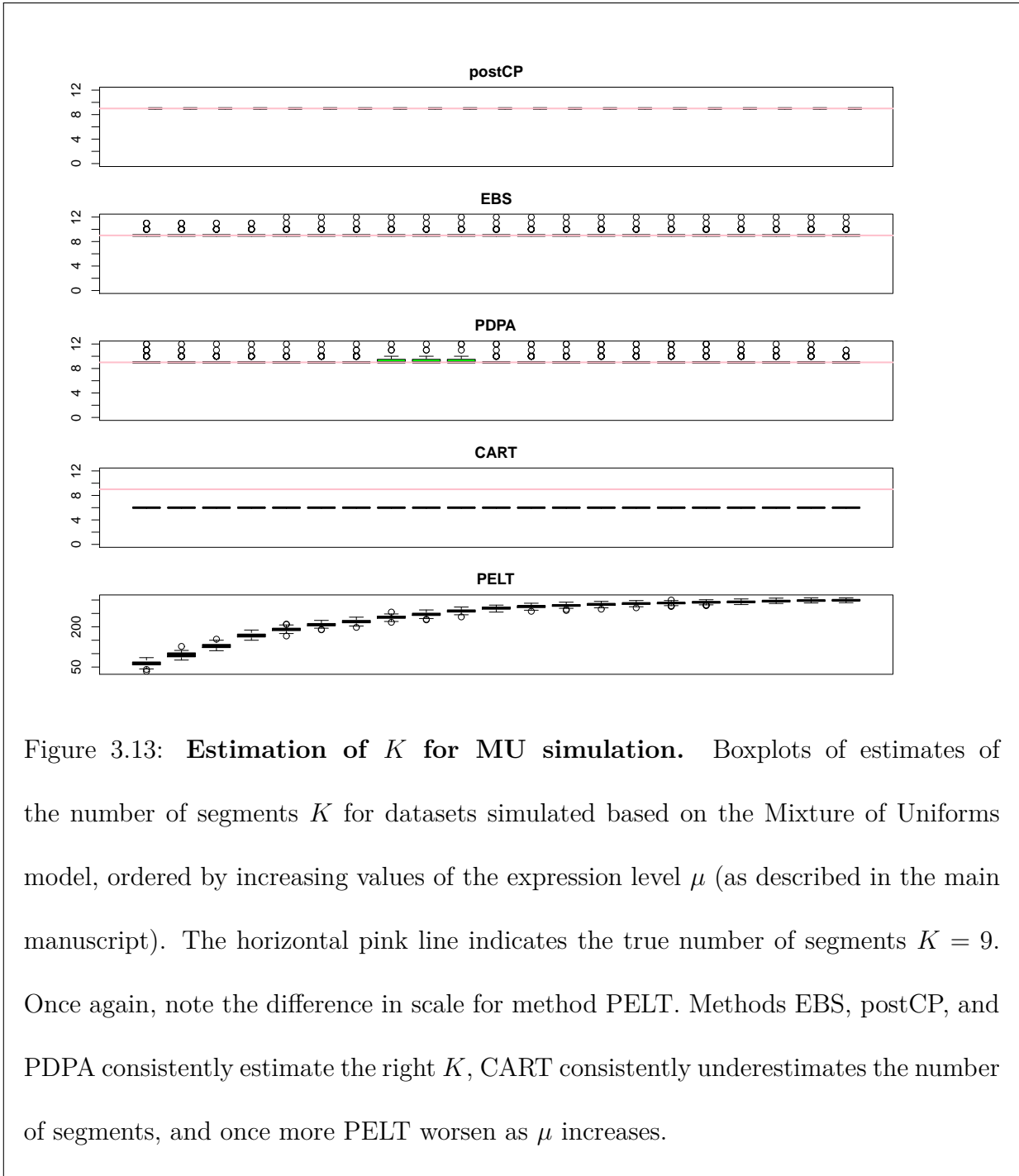
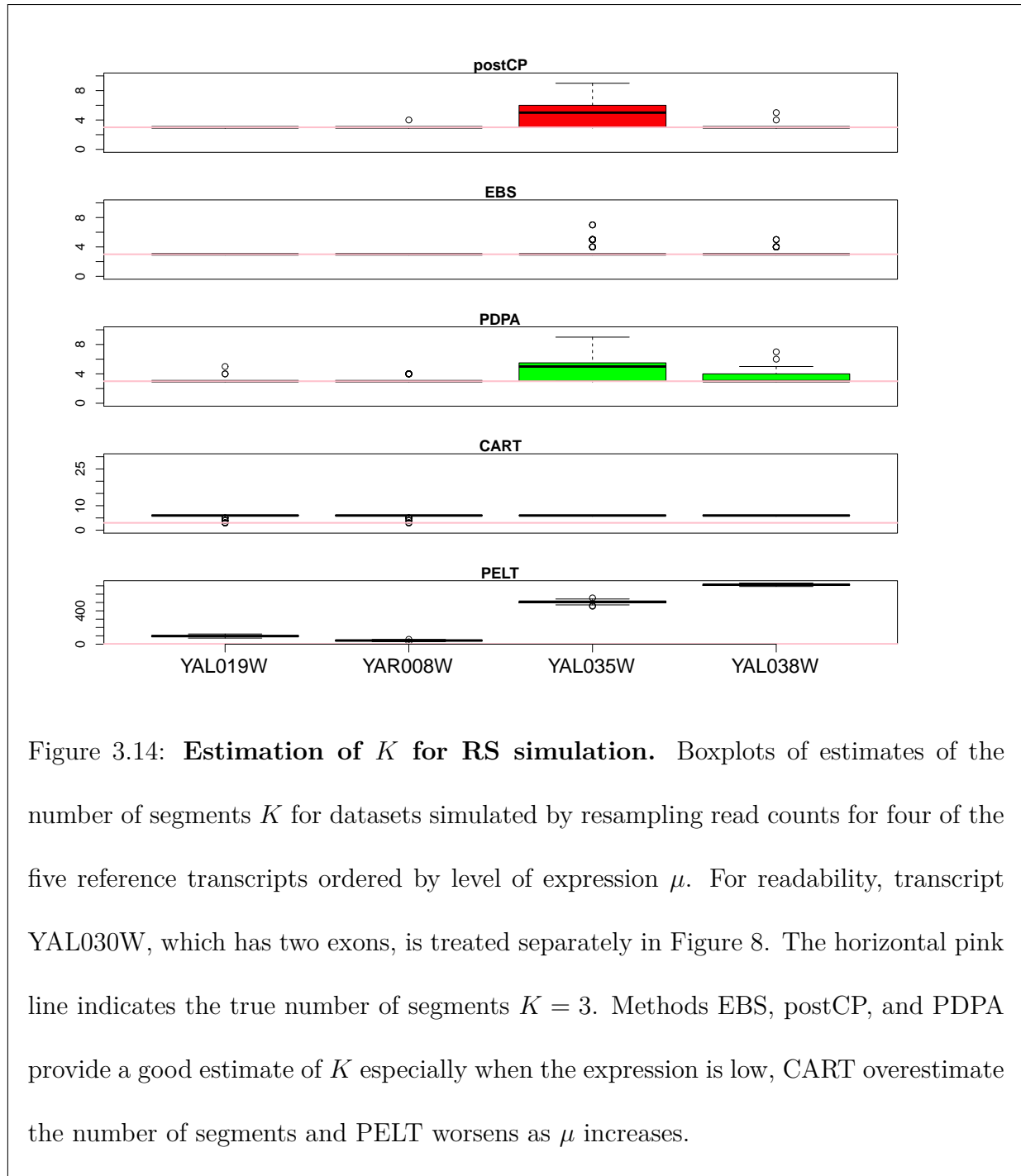
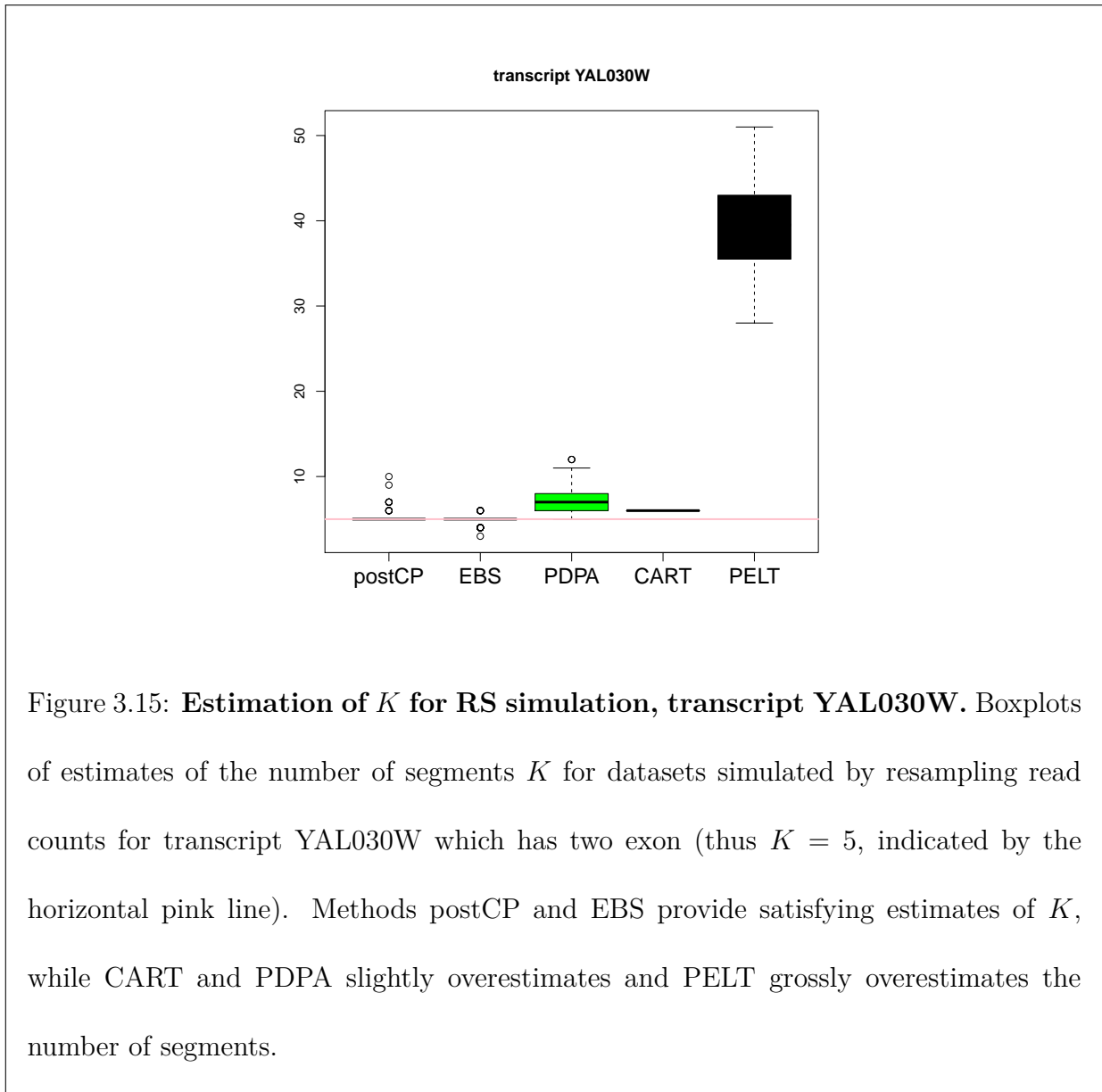


Figure 3.13: **Estimation of K for MU simulation.** Boxplots of estimates of the number of segments K for datasets simulated based on the Mixture of Uniforms model, ordered by increasing values of the expression level μ (as described in the main manuscript). The horizontal pink line indicates the true number of segments $K = 9$. Once again, note the difference in scale for method PELT. Methods EBS, postCP, and PDPA consistently estimate the right K , CART consistently underestimates the number of segments, and once more PELT worsen as μ increases.





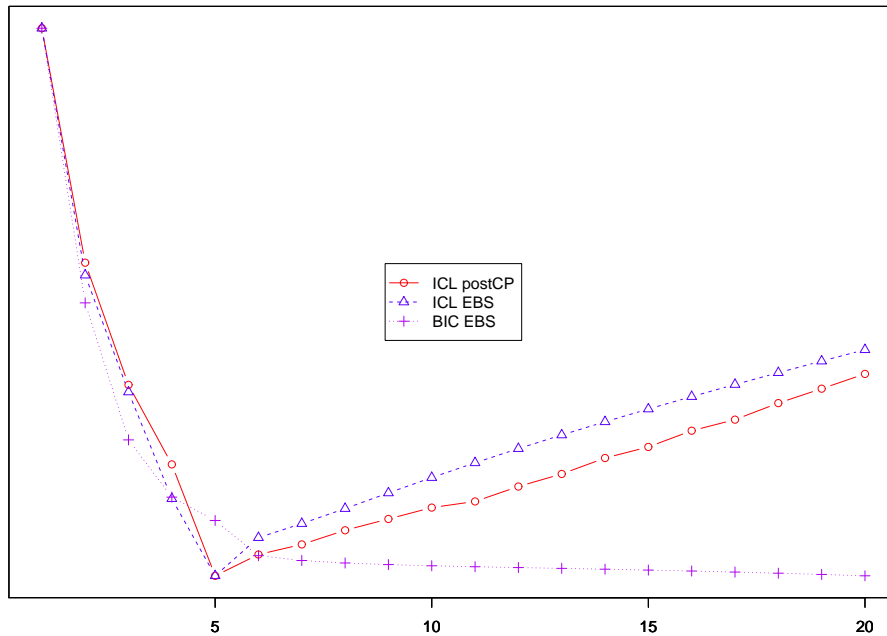


Figure 3.16: **BIC and ICL criterion for postCP and EBS.** Posterior ICL criterion and BIC, averaged over the resampling simulations for transcript YAL030W, for methods EBS and postCP. The estimate of K is the value which minimizes each criterion; the true value of K is 5. Note that here the ordinate is irrelevant and only the shape of the curve matters. Indeed, the different criteria are on different scales, as likelihood functions are computed up to an unknown constant.

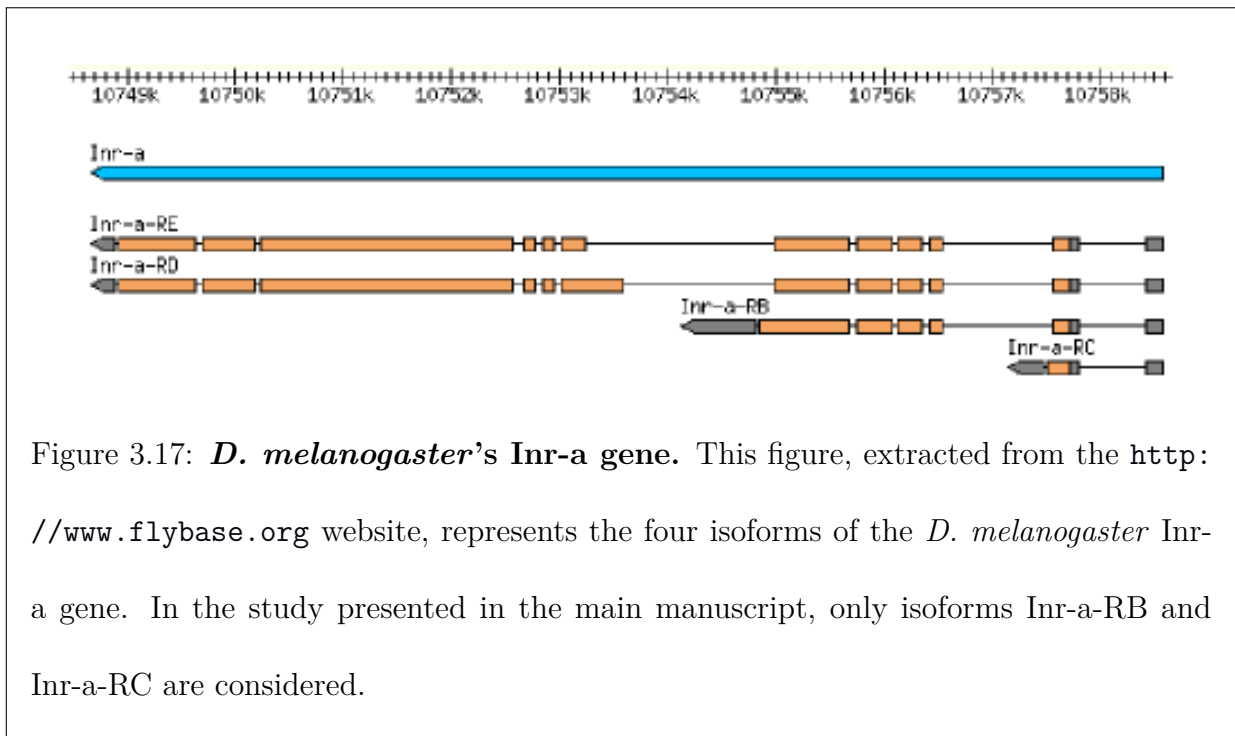


Figure 3.17: *D. melanogaster*'s *Inr-a* gene. This figure, extracted from the <http://www.flybase.org> website, represents the four isoforms of the *D. melanogaster* *Inr-a* gene. In the study presented in the main manuscript, only isoforms *Inr-a-RB* and *Inr-a-RC* are considered.

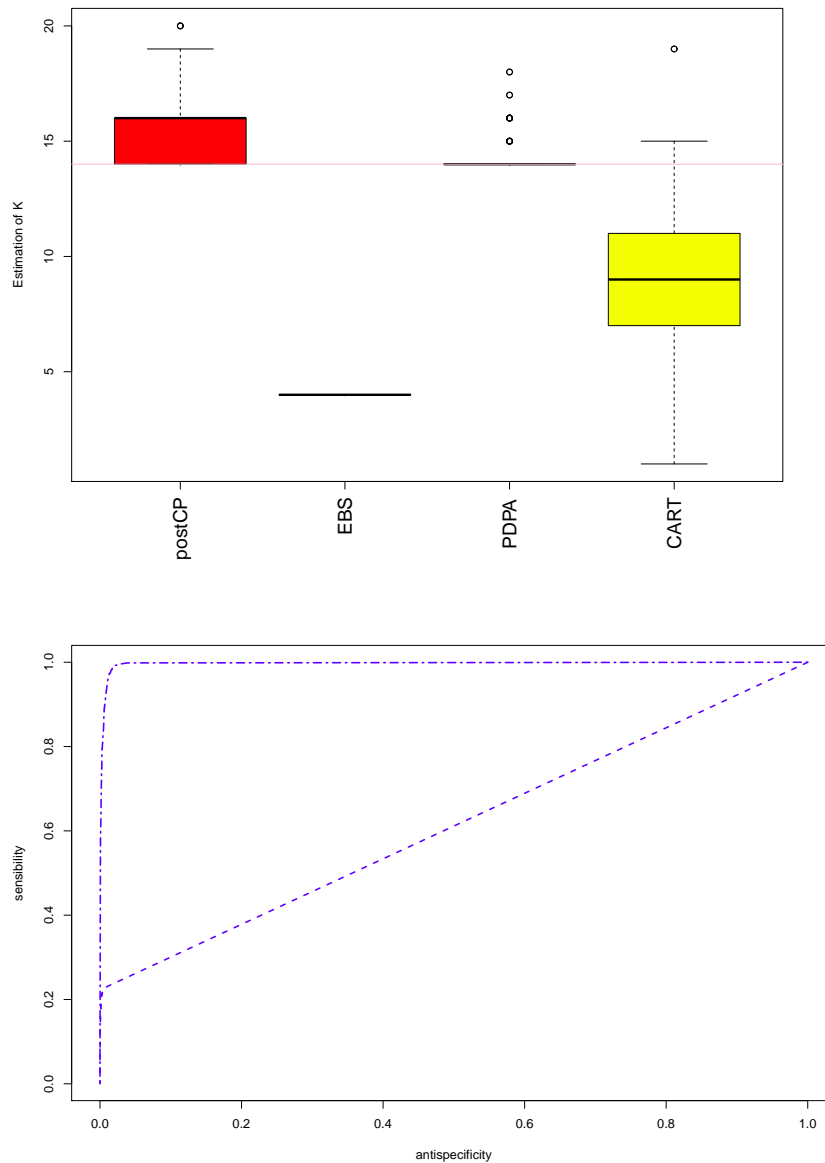


Figure 3.18: **Estimation of K and ROC curves for *Drosophila*-like simulations.**

Top: Boxplots of estimates of K for methods postCP, EBS, PDPA, and CART, for one hundred simulations based on the Inr-a *D. melanogaster* gene. Bottom: ROC-like curves for method EBS with the Negative Binomial distribution, with estimation of K (---) and with known K (-.-)..

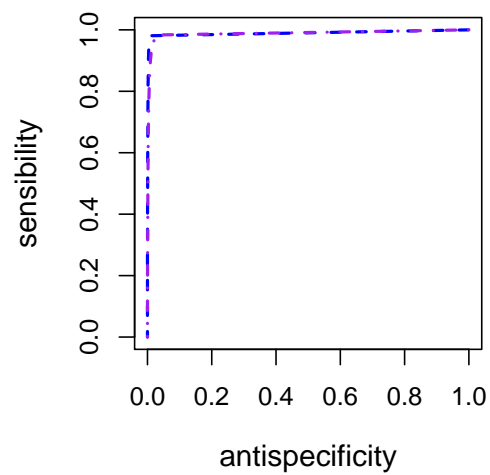


Figure 3.19: **ROC curves for EBS with Negative Binomial and Gaussian distributions.** ROC curves for method EBS applied with the Gaussian distribution on log-transformed counts (·-·) and the Negative Binomial distribution on untransformed counts (- - -). Datasets simulated by resampling pooled read counts from yeast genes for segments defined by the *D. melanogaster* Inr-a gene annotation.

References

- Sylvain Arlot and Alain Celisse. Segmentation of the mean of heteroscedastic data via cross-validation. *Statistics and Computing*, 21(4):613–632, 2011.
- Jushan Bai and Pierre Perron. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22, 2003.
- Daniel Barry and John A Hartigan. A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319, 1993.
- Valentina Boeva, Andrei Zinovyev, Kevin Bleakley, Jean-Philippe Vert, Isabelle Janoueix-Lerosey, Olivier Delattre, and Emmanuel Barillot. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics*, 27 (2):268–269, 2011.
- Breiman, Friedman, Olshen, and Stone. Classification and regression trees. Wadsworth and Brooks, 1984.
- Alice Cleynen and Emilie Lebarbier. Segmentation of the Poisson and negative binomial rate models: a penalized estimator. *arXiv preprint arXiv:1301.2534*, 2013.
- Alice Cleynen, Michel Koskas, and Guillem Rigau. A generic implementation of the pruned dynamic programming algorithm. *Arxiv preprint arXiv:1204.5564*, under review.
- Scott B Guthery. Partition regression. *Journal of the American Statistical Association*, 69 (348):945–947, 1974.
- LI Hsu, Steven G Self, Douglas Grove, Tim Randolph, Kai Wang, Jeffrey J Delrow, Lenora Loo, and Peggy Porter. Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, 6(2):211–226, 2005.
- 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.
- N. Johnson, A.W. Kemp, and S. Kotz. Univariate discrete distributions. *John Wiley & Sons, Inc.*, 2005.
- Rebecca Killick and Idris A Eckley. Changepoint: an R package for changepoint analysis. *Lancaster University*, 2011.
- Weil R Lai, Mark D Johnson, Raju Kucherlapati, and Peter J Park. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, 21(19):3763–3770, 2005.
- B Langmead, C Trapnell, M Pop, and SL Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10, 2008.
- T.M. Luong, Y. Rozenholc, and G. Nuel. Fast estimation of posterior probabilities in change-point models through a constrained hidden Markov model. *Arxiv preprint arXiv:1203.4394*, 2012.

Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349, 2008.

G Rigaiil, E Lebarbier, and S Robin. Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and Computing*, 22(4):917–929, 2012.

Davide Risso, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, 12(1):480, 2011.

Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26 (1):139–140, 2010.

A.J. Scott and M. Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30:507–512, 1974.

3.1.5 Choice of the priors

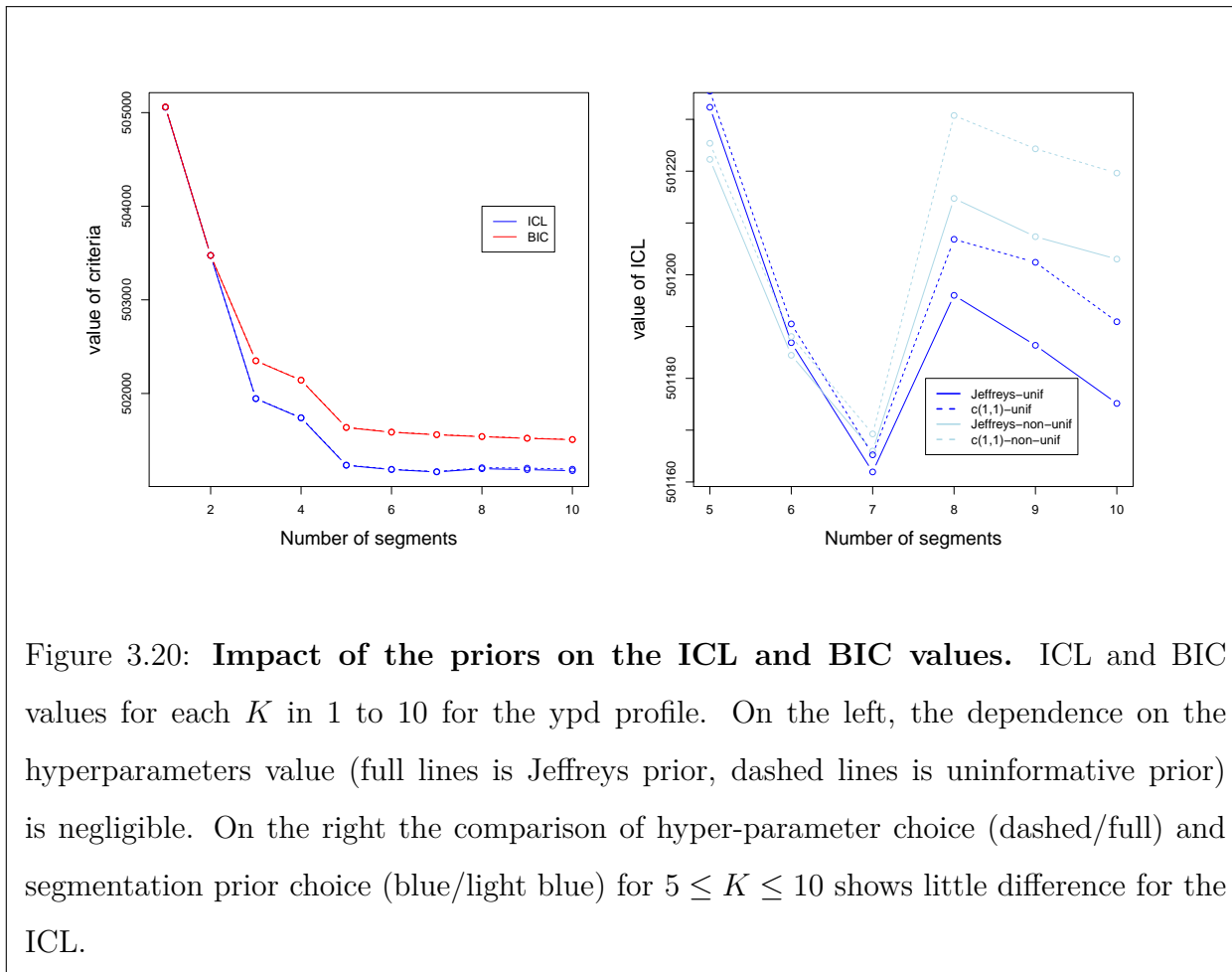
In the paper we have not discussed the choice of the priors even though as in any Bayesian analysis this question should deserve further study. In the case of EBS, we can distinguish two types of priors: that used on the segmentation m , and the values of the hyperparameters from the conjugate priors.

Concerning the segmentation, the possibilities for the priors are restricted by the requirement that the model verifies the factorability assumption. Then the difference between two choices of priors will be observed in the constant C as well as in the factors a_J .

As for the hyperparameters, their influence is directly related to the terms in matrix A . More precisely, in the decomposition of $[A]_{i,j}$ as in Table 1.1, we can see that term 2 depends only on their value, and the K^{th} power of this term is needed in the computation of $P(Y, K)$. While they also appear in the first term, typically when the data are informative enough, its dependency on their value is very small. It is therefore mostly term 2 which drives the dependency, and indicates that if not chosen carefully, the hyperparameters can constitute a penalization on the number of segments.

Note that the prior $\mathcal{B}eta(1/2, 1/2)$ which was used in the paper for the negative binomial distribution corresponds to Jeffreys' prior (JEFFREYS, 1946) which main properties is the non-dependency to the parametrization of the distribution. However, these hyperparameters lead to a penalization in the number of segments K with the order of $\mathcal{B}(1/2, 1/2)^{-K}$. Hyperparameters 1, 1 would thus be recommended if no penalization was wanted.

We illustrate some of these remarks in the particular case where the question is the choice of the number of segments, since as explained above it is in this context that the choice of the hyperparameters should be most crucial. We thus computed the ICL and BIC criteria on the segmentation of a profile corresponding to the ypd data of a given yeast gene. On the right of Figure 3.20 are displayed the values obtained for Jeffrey's prior $\mathcal{B}eta(1/2, 1/2)$ (in full lines) and the non-informative prior $\mathcal{B}eta(1, 1)$ (in dashed lines) when the prior on the segmentation m is the uniform conditional on K . On the left are displayed the values of the ICL criterion for $5 \leq K \leq 10$ for both priors, but comparing the uniform on



m to the prior which favors the segmentation with segments of equal length (*i.e.* $a_J = n_J^{-1}$). While there is a difference between ICL, which presents a minimum in $K = 7$ and BIC which is non-increasing, the impact of the choice of hyperparameters is negligible.

3.2 Profile comparison

We have shown in the previous chapter that the Bayesian model of RIGAILL *et al.* (2012) has almost optimal performance in the context of gene re-annotation, providing tools to measure the uncertainty associated with the change-point localization. We have subsequently retained this model to perform our further analysis on the comparison of change-point location between independent profiles.

Our motivation is the following: in frameworks where we have data corresponding to the same genome subset but for biologically different (and independent) subjects, can we identify phenomena such as differential splicing which would result in different change-point location? Situations abound in which this question is relevant. For instance, chromosomal aberrations are common in patients with cancer, and some of them are recurrently observed in subclasses of a particular disease. Identifying these aberrations and their location is crucial since they might result in fusion genes responsible for some aspect of the disease. Thus the comparison of their locations, typically assessed by segmentation methods, between different patients can tell us a lot on the importance and the role of those fusion genes.

Another example is inspired by our benchmark dataset. It may be recalled that it measures the transcriptional activity of a yeast species that has been grown into three different conditions, one of which results in cells respiring instead of fermenting. One biological intuition to explain this difference is differential splicing: some transcript boundaries are expected to differ in those conditions. As an example, Figure 3.21 corresponds to a subset of our benchmark dataset limited to a region surrounding gene EFB1. Its two exons can clearly be identified, but are their boundaries the same in the three growth conditions?

The paper presented here, submitted in collaboration with Stéphane Robin and available at <http://arxiv.org/abs/1307.3146>, proposes two methods to address this question which are based on the Bayesian model described before. They are implemented in the R package **EBS** as detailed in the next section.

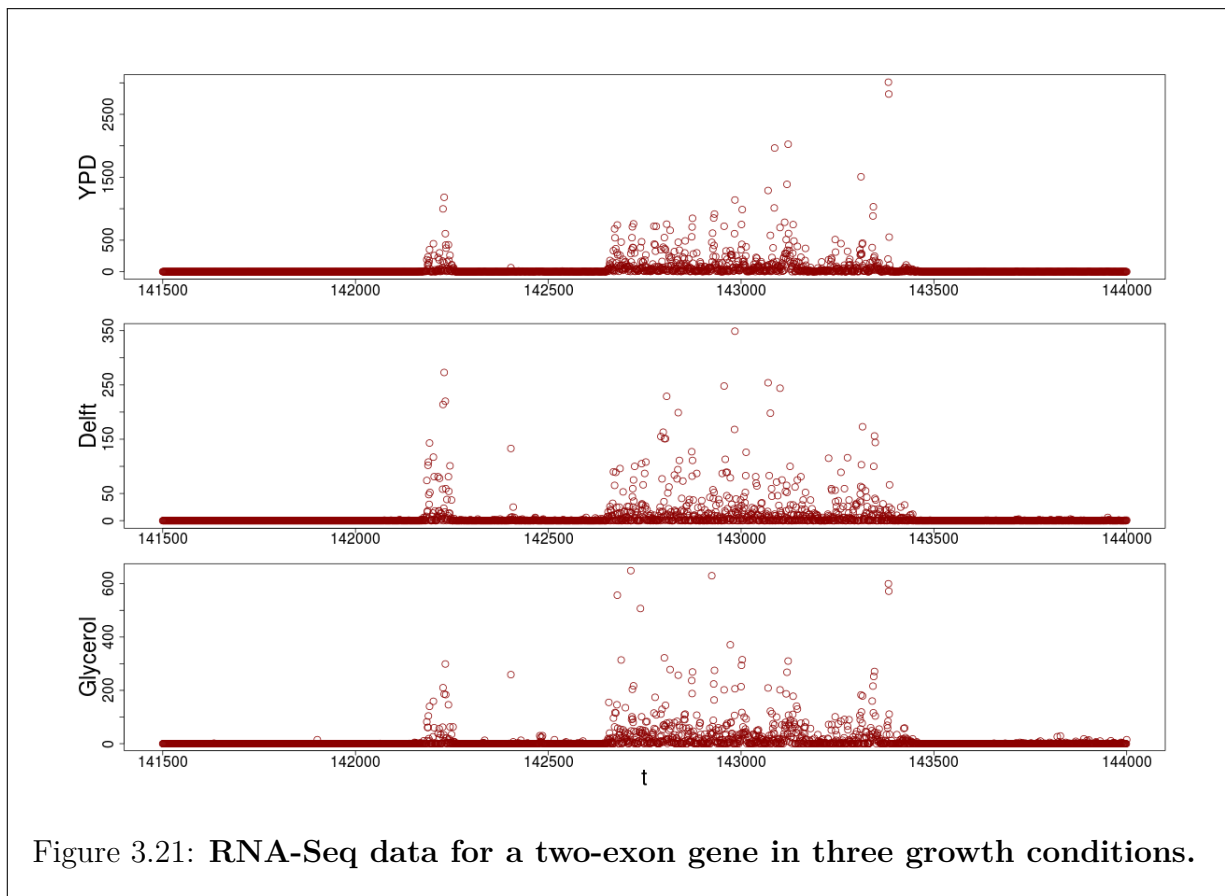


Figure 3.21: RNA-Seq data for a two-exon gene in three growth conditions.

Comparing change-point locations of independent profiles with application to gene annotation

Alice Cleynen and Stéphane Robin

abstract

We are interested in the comparison of transcript boundaries from cells which originated in different environments. The goal is to assess whether this phenomenon, called alternative splicing, is used to modify the transcription of the genome in response to stress factors. We address this question by comparing the change-points locations in the individual segmentation of each profile, which correspond to the RNA-Seq data for a gene in one growth condition. This requires the ability to evaluate the uncertainty of the change-point positions, and the work of RIGAILL *et al.* (2012) provides an appropriate framework in such case. Building on their approach, we propose two methods for the comparison of change-points, and illustrate our results on a dataset from the yeast specie. We show that the UTR boundaries are subject to alternative splicing, while the intron boundaries are conserved in all profiles. Our approach is implemented in an R package called EBS which is available on the CRAN.

Keywords

segmentation; change-point comparison; Bayesian inference; negative binomial; differential splicing

3.2.1 Introduction

Segmentation problems arise in a large range of domains such as economics, biology or meteorology, to name a few. Many methods have been developed and proposed in the literature in the last decades to detect change-points in the distribution of the signal along

one single series. Yet, more and more applications require the analysis of several series at a time to better understand a complex underlying phenomenon. Such situations refer for example to the analysis of the genomic profiles of a cohort of patients (PICARD *et al.*, 2011), of meteorological series observed in different locations (EHSANZADEH *et al.*, 2011) or of sets astronomical series of photons abundance (DOBIGEON *et al.*, 2007).

When dealing with multiple series, two approaches can be typically considered. The first consists in the *simultaneous* segmentation of all series, looking for changes that are common to all of them. This approach amounts to the segmentation of one single multivariate series but might permit the detection of change-points in series with too low a signal to allow their analysis independently. The second approach consists in the *joint* segmentation of all the series, each having its specific number and location of changes. This allows to account for dependence between the series without imposing that the changes occur simultaneously.

We are interested here in a third kind of statistical problem, which is the comparison of change-point locations in several series that have been segmented separately. To our knowledge, this problem has not yet been fully addressed.

Indeed, comparing change-point is connected to the evaluation of the uncertainty of the change-point positions. An important point is that the standard likelihood-based inference is very intricate, since the required regularity conditions for the change-point parameters are not satisfied (FEDER, 1975). Most methods to obtain change-point confidence intervals are based on their limit distribution estimators (FEDER, 1975; BAI and PERRON, 2003) or the asymptotic use of a likelihood-ratio statistic (MUGGEO, 2003). Bootstrap techniques have also been proposed (see HUŠKOVÁ and KIRCH (2008) and references therein). Comparison studies of some of these methods can be found in REEVES *et al.* (2007) for climate applications or in TOMS and LESPERANCE (2003) for ecology. Recently, RIGAILL *et al.* (2012) proposed a Bayesian framework to derive the posterior distributions of various quantities of interest – including change-point locations – in the context of exponential family distributions with conjugate prior.

As for the comparison of change-points, the most common approaches rely on classifi-

cation comparison techniques such as the Rand Index (RAND, 1971); and aim at assessing the performances of segmentation methods on single datasets, by comparing their outputs between themselves or using the truth as reference. The notion of change-point location difference as a quantity of interest has, to our knowledge, never been considered.

Our work is a generalization of RIGAILL *et al.* (2012) to the comparison of change point location. It is motivated by a biological problem detailed in the next paragraph.

Differential splicing in yeast

Differential splicing is one of the mechanism that living cells use to modify the transcription of their genome in response to some change in their environment, such as a stress. More precisely, differential splicing refers to the ability for the cell to choose between versions (called isoforms) of a given gene by changing the boundaries of the regions to be transcribed.

New sequencing technologies, including RNA-Seq experiments, give access to a measure of the transcription at the nucleotide resolution. The signal provided by RNA-Seq consists in a count (corresponding to a number of reads) associated to each nucleotide along the genome. This count is proportional to the transcription level of the nucleotide. This technology therefore allows to locate precisely the boundaries of the transcribed regions, to possibly revise the known annotation of the genomes and to study the variation of these boundaries across conditions.

We are interested here in an RNA-Seq experiment made on a given specie, yeast, grown under several conditions. The biological question to be addressed is 'Does yeast use differential splicing of a given gene as a response to a change in its environment?'

Contribution

In this paper we develop a Bayesian approach to compare the change-point location of independent series corresponding to the same gene under several conditions. We suppose

that we have information on the structure of this gene (such as the number of introns) so that the number of segments of each segmentation is assumed to be known. In Section 3.2.2, we recall the Bayesian segmentation model introduced in RIGAILL *et al.* (2012) and its adaptation to our framework. In Section 3.2.3 we derive the posterior distribution of the shift between the change-point locations in two independent profiles, while in Section 3.2.4 we introduce the calculation of the posterior probability for change-points to share the same location in different series. The performances are assessed in Section 3.2.5 via a simulation study designed to mimic real RNA-Seq data. We finally apply the proposed methodology to study the existence of differential splicing in yeast in Section 3.2.6. Our approach is implemented in an R package **EBS** which is available on the CRAN repository.

All the results we provide are given conditional on the number of segments in each profiles. Indeed comparing the location of, say, the second change-points in each series implicitly refers to a total number of change-points in each of them. Yet, most of the results we provide can be marginalized over the number of segments.

3.2.2 Model for one series

In this section we introduce the general Bayesian framework for the segmentation of one series and recall preceding results on the posterior distribution of change-points.

Bayesian framework for one series

The general segmentation problem consists in partitioning a signal of n data-points $\{y_t\}_{t \in [1, n]}$ into K segments. The model is defined as follows: the observed data $\{y_t\}_{t=1, \dots, n}$ are supposed to be a realization of an independent random process $Y = \{Y_t\}_{t=1, \dots, n}$. This process is drawn from a probability distribution \mathcal{G} which depends on a set of parameters among which one parameter θ is assumed to be affected by $K - 1$ abrupt changes, called change-points and denoted τ_k ($1 \leq k \leq K - 1$). A partition m is defined as a set of change-points: $m = (\tau_0, \tau_1, \dots, \tau_K)$ with conventions $\tau_0 = 1$ and $\tau_K = n + 1$ and a segment J is said to belong to m if $J = \llbracket \tau_{k-1}; \tau_k \rrbracket$ for some k .

The Bayesian model is fully specified with the following distributions:

- the prior distribution of the number of segments $P(K)$;
- the conditional distribution of partition m given K : $P(m|K)$;
- the parameters θ_J for each segment J are supposed to be independent with same distribution $P(\theta_J)$;
- the observed data $Y = (Y_t)$ data are independent conditional on m and (θ_J) with distribution depending on the segment:

$$(Y_t|m, J \in m, \theta_J, t \in J) \sim \mathcal{G}(\theta_J, \phi)$$

where ϕ is some parameter that is constant across the segments that will be supposed to be known.

Exact calculation of posterior distributions

RIGAILL *et al.* (2012) show that if distribution \mathcal{G} possesses conjugate priors for θ_J , and if the model satisfies the factorability assumption, that is, if

$$\begin{aligned} P(Y, m) &= C \prod_{J \in m} a_J P(Y_J|J), \\ \text{where } P(Y_J|J) &= \int P(Y_J|\theta_J) P(\theta_J) d\theta_J, \end{aligned} \quad (3.1)$$

quantities such that $P(Y, K)$, posterior change-point location distributions or the posterior entropy can be computed exactly and in a quadratic time. Examples of satisfying distributions are

- the Gaussian heteroscedastic:

$$\mathcal{G}(\theta_J, \phi) = \mathcal{N}(\mu_J, \sigma_J^2) \text{ with } \theta_J = (\mu_J, \sigma_J^2), \phi = \emptyset,$$

- the Gaussian homoscedastic with known variance σ^2 :

$$\mathcal{G}(\theta_J, \phi) = \mathcal{N}(\mu_J, \sigma^2) \text{ with } \theta_J = \mu_J, \phi = \sigma^2,$$

- the Poisson:

$$\mathcal{G}(\theta_J, \phi) = \mathcal{P}(\lambda_J) \text{ with } \theta_J = \lambda_J, \phi = \emptyset,$$

- or the negative binomial homoscedastic with known dispersion ϕ :

$$\mathcal{G}(\theta_J, \phi) = \mathcal{NB}(p_J, \phi) \text{ with } \theta_J = p_J, \phi = \phi.$$

Note that the Gaussian homoscedastic does not satisfy the factorability assumption if σ is unknown, and that the negative binomial heteroscedastic does not belong to the exponential family and does not have a conjugate prior on ϕ .

The factorability assumption (3.1) also induces some constraint on the distribution of the segmentation $P(m|K)$. In this paper, we will limit ourselves to the uniform prior:

$$P(m|K) = \mathcal{U}(\mathcal{M}_K^{1,n+1})$$

where $\mathcal{M}_K^{1,n+1}$ stands for the set of all possible partitions of $\llbracket 1, n+1 \rrbracket$ into K non-empty segments.

3.2.3 Posterior distribution of the shift

The framework described above allows to compute a set of quantities of interest in an exact manner. In this paper, we are mostly interested in the location of change-points. We first remind how posterior distributions can be computed and then propose a first exact comparison strategy.

Posterior distribution of the change-points

The key ingredient for most of the calculations is the $(n+1) \times (n+1)$ matrix A that contains the probabilities of all segments:

$$[A]_{i,j} = \begin{cases} P(Y_{\llbracket i,j \rrbracket} | \llbracket i,j \rrbracket) & \forall 1 \leq i < j \leq n+1 \\ 0 & \textit{else} \end{cases} \quad (3.2)$$

where $P(Y_J|J)$ is given in (3.1).

The posterior distribution of change-points can be deduced from this matrix in a quadratic time with the following proposition:

Proposition 1. Denoting $p_k(t; Y; K) = P(\tau_k = t | Y, K)$ the posterior distribution of the k th change-point, we have

$$p_k(t; Y; K) = \frac{[(A)^k]_{1,t} [(A)^{K-k}]_{t,n+1}}{[(A)^K]_{1,n+1}}.$$

Proof. We have

$$p_k(t; Y; K) = \frac{\sum_{m \in \mathcal{B}_{K,k}(t)} p(Y|m) p(m|K)}{P(Y|K)}$$

where $\mathcal{B}_{K,k}(t)$ is the set of partitions of $\{1, \dots, n\}$ in K segments with k th change-point at location t . Note that $\mathcal{B}_{K,k}(t) = \mathcal{M}_k^{1,t} \otimes \mathcal{M}_{K-k}^{t,n+1}$ (i.e. all $m \in \mathcal{B}_{K,k}(t)$ can be decomposed uniquely as $m = m_1 \cup m_2$ with $m_1 \in \mathcal{M}_k^{1,t}$ and $m_2 \in \mathcal{M}_{K-k}^{t,n+1}$ and reciprocally). Then using the factorability assumption, we can write

$$p_k(t; Y; K) = \frac{\sum_{m_1 \in \mathcal{M}_k^{1,t}} p(Y|m_1) \sum_{m_2 \in \mathcal{M}_{K-k}^{t,n+1}} p(Y|m_2) p(m|K)}{\sum_{m \in \mathcal{M}_K^{1,n+1}} p(Y|m) p(m|K)}$$

□

Comparison of two series

We now propose a first procedure to compare the location of two change-points in two independent series. Consider two independent series Y^1 and Y^2 with same length n and respective number of segments K^1 and K^2 . The aim is to compare the locations of the k_1 th change-point from series Y^1 (denoted $\tau_{k_1}^1$) with the k_2 th change-point of series Y^2 (denoted $\tau_{k_2}^2$). The posterior distribution of the difference between the location of the two change-points can be derived with the following Proposition.

Proposition 2. Denoting $\delta_{k_1,k_2}(d; K^1, K^2) = P(\Delta = d | Y^1, Y^2, K^1, K^2)$ the posterior distribution of the difference $\Delta = \tau_{k_1}^1 - \tau_{k_2}^2$, we have

$$\delta_{k_1,k_2}(d; K^1, K^2) = \sum_t p_{k_1}(t; Y^1; K^1) p_{k_2}(t-d; Y^2; K^2).$$

Proof. This simply results from the convolution between the two posterior distributions p_{k_1} and p_{k_2} . □

The posterior distribution of the shift can therefore be computed exactly and in a quadratic time. The non-difference between the two change-point locations $\tau_{k_1}^1$ and $\tau_{k_2}^2$ can then be assessed, looking at the position of 0 with respect to the posterior distribution δ .

3.2.4 Comparison of change point locations

We now consider the comparison of change-point locations between more than 2 series. In this case, the convolution methods described above does not apply anymore so we propose a comparison based on the exact computation of the posterior probability for the change-points under study to have the same location.

Model for I series

We now consider I independent series Y^ℓ (with $1 \leq \ell \leq I$) with same length n . We denote m^ℓ , their respective partitions and K^ℓ their respective number of segments. We further denote τ_k^ℓ the k th change-point in Y^ℓ so $m^\ell = (\tau_0^\ell, \tau_1^\ell, \dots, \tau_{K^\ell}^\ell)$. Similarly, θ_J^ℓ denotes the parameter for the series ℓ within segment J provided that $J \in m^\ell$ and ϕ^ℓ the constant parameter of series ℓ . In the following, the set of profiles will be referred to as \mathbf{Y} and respectively for the vector of segment numbers (\mathbf{K}), the set of all partitions (\mathbf{m}) and the set of all parameters ($\boldsymbol{\theta}$).

In the perspective of change-point comparison, where one is interested in the k_l th change-point of series l , for $1 \leq l \leq I$, we introduce the following event:

$$E_0 = \{\tau_{k_1}^1 = \dots = \tau_{k_I}^I\}.$$

We further denote E_1 its complement and define the binary random variable

$$\mathbf{E} = \mathbb{I}\{E_1\} = 1 - \mathbb{I}\{E_0\}.$$

The complete hierarchical model is displayed in Figure 3.22 and is defined as follows:

- The random variable \mathbf{E} is drawn conditionally on \mathbf{K} as a Bernoulli $\mathcal{B}(1 - p_0(\mathbf{K}))$ where $p_0(\mathbf{K}) = P(E_0|\mathbf{K})$;

- The parameters $\boldsymbol{\theta}$ are drawn independently according to $P(\boldsymbol{\theta}|\mathbf{K})$;
- The partitions are drawn conditionally on \mathbf{E} according to $P(\mathbf{m}|\mathbf{K}, \mathbf{E})$;
- The observations are generated according to the conditional distribution $P(\mathbf{Y}|\mathbf{m}, \boldsymbol{\theta})$.

More specifically, denoting $\mathcal{M}_{\mathbf{K}}^{1,n+1} = \otimes_{\ell} \mathcal{M}_{K_{\ell}}^{1,n+1}$, the partitions are assumed to be uniformly distributed, conditional on E , that is

$$P(\mathbf{m}|\mathbf{K}, E_0) = \mathcal{U}(\mathcal{M}_{\mathbf{K}}^{1,n+1} \cap E_0), \quad P(\mathbf{m}|\mathbf{K}, E_1) = \mathcal{U}(\mathcal{M}_{\mathbf{K}}^{1,n+1} \cap E_1).$$

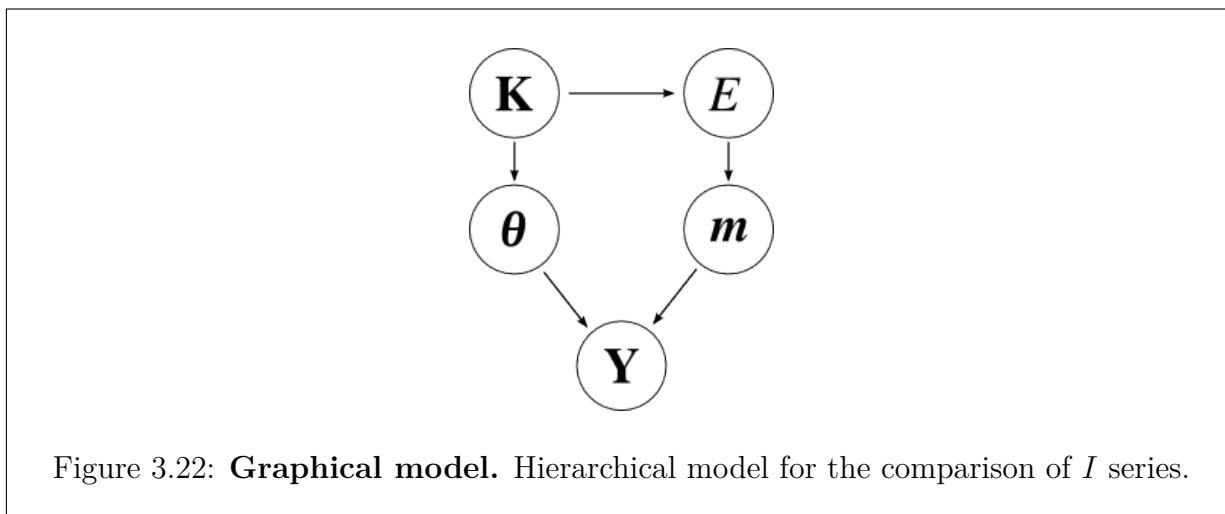


Figure 3.22: **Graphical model.** Hierarchical model for the comparison of I series.

Posterior probability for the existence of a common change-point

We propose to assess the existence of a common change-point location between the I profiles based on the posterior probability of this event, namely $P(E_0|\mathbf{Y}, \mathbf{K})$.

Proposition 3. *The posterior probability of E_0 can be computed in $O(Kn^2)$ as*

$$P(E_0|\mathbf{Y}, \mathbf{K}) = \frac{p_0(\mathbf{K})}{q_0(\mathbf{K})} Q(\mathbf{Y}, E_0|\mathbf{K}) .$$

$$\left[\frac{1 - p_0(\mathbf{K})}{1 - q_0(\mathbf{K})} Q(\mathbf{Y}|\mathbf{K}) + \frac{p_0(\mathbf{K}) - q_0(\mathbf{K})}{q_0(\mathbf{K})[1 - q_0(\mathbf{K})]} Q(\mathbf{Y}, E_0|\mathbf{K}) \right]^{-1}$$

where

$$Q(\mathbf{Y}|\mathbf{K}) = \prod_{\ell} [(A_{\ell})^{K_{\ell}}]_{1,n+1},$$

$$Q(\mathbf{Y}, E_0|\mathbf{K}) = \sum_t \prod_{\ell} [(A_{\ell})^{k_{\ell}}]_{1,t} [(A_{\ell})^{K_{\ell}-k_{\ell}}]_{t+1,n+1},$$

and $q_0(\mathbf{K}) = Q(E_0|\mathbf{K}) = \sum_t \prod_{\ell} \binom{t-2}{k_{\ell}-1} \binom{n-t}{K_{\ell}-k_{\ell}-1} / \binom{n-1}{K_{\ell}-1}.$

and A_ℓ stands for the matrix A as defined in (3.2), corresponding to series ℓ .

Proof. We consider the surrogate model where the partition \mathbf{m} is drawn uniformly and independently from \mathbf{E} , namely $Q(\mathbf{m}|\mathbf{K}) = \mathcal{U}(\mathcal{M}_{\mathbf{K}}^{1,n+1})$ (note that this corresponds to choosing $p_0(\mathbf{K}) = q_0(\mathbf{K})$). All probability distributions under this model are denoted by Q along the proof. The formulas for probabilities $Q(\mathbf{Y}|\mathbf{K})$ and $Q(\mathbf{Y}, E_0|\mathbf{K})$ derive from RIGAILL *et al.* (2012). It then suffices to apply the probability change as

$$P(\mathbf{Y}, E_0|\mathbf{K}) = \frac{p_0(\mathbf{K})}{q_0(\mathbf{K})}Q(\mathbf{Y}, E_0|\mathbf{K}), \quad P(\mathbf{Y}, E_1|\mathbf{K}) = \frac{1 - p_0(\mathbf{K})}{1 - q_0(\mathbf{K})}Q(\mathbf{Y}, E_1|\mathbf{K}).$$

The result then follows from the decomposition of $P(\mathbf{Y}|\mathbf{K})$ as $P(\mathbf{Y}, E_0|\mathbf{K}) + P(\mathbf{Y}, E_1|\mathbf{K})$ and the same for $Q(\mathbf{Y}|\mathbf{K})$. \square

The Bayes factor is sometimes preferred for model comparison; it can be computed exactly in a similar way:

Corollary 4. *The Bayes factor can be computed in $O(Kn^2)$ as*

$$\frac{P(\mathbf{Y}|E_0, \mathbf{K})}{P(\mathbf{Y}|E_1, \mathbf{K})} = \frac{1 - q_0(\mathbf{K})}{q_0(\mathbf{K})} \frac{Q(\mathbf{Y}, E_0|\mathbf{K})}{Q(\mathbf{Y}|\mathbf{K}) - Q(\mathbf{Y}, E_0|\mathbf{K})}$$

using the same notations as in Proposition 3.

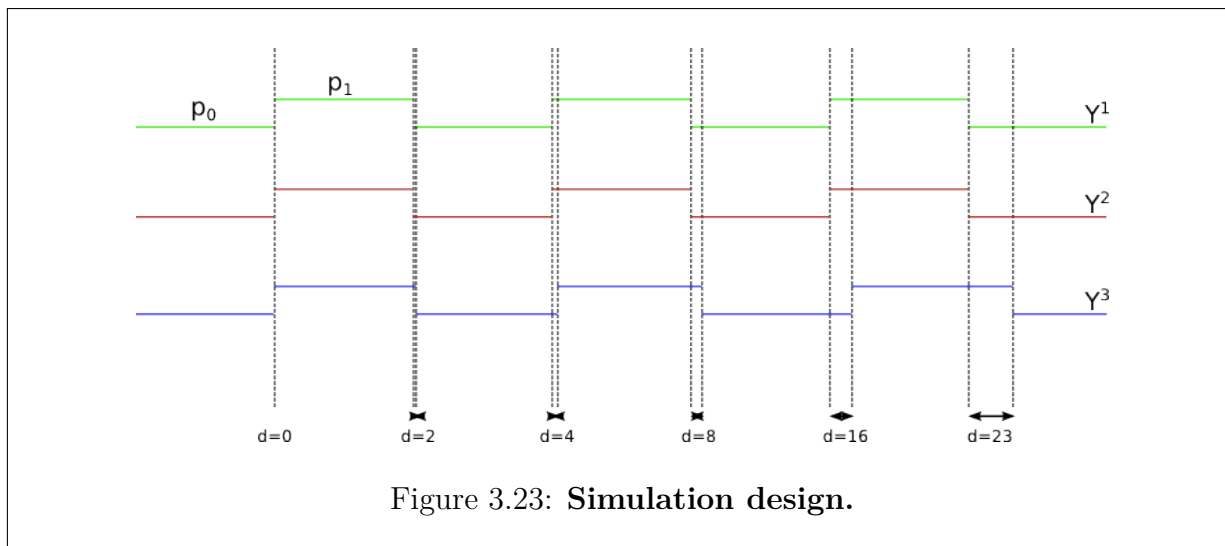
Proof. The proof follows that of Proposition 3. \square

3.2.5 Simulation study

Simulation design

We designed a simulation study to identify the influence of various parameters on the performances of our approach. The design is illustrated in Figure 3.23: we compared 3 independent profiles with 7 segments, with all odd (respectively even) segments sharing the same distribution. The first two profiles have identical segmentation m given by

$m = (1, 101, 201, 301, 401, 501, 601, 701)$ and the change-point locations of the third one are progressively shifted apart as $\tau_k^3 = \tau_k^1 + 2^{k-1}$, for each $1 \leq k \leq 6$. We shall denote $d_k = \tau_k^3 - \tau_k^1$ and drop the index k when there is no ambiguity on it.



Our purpose is to mimic data obtained by RNA-Seq experiments, so that the parameters for the negative binomial distribution were chosen to fit typical real-data. Considering the model where odd segments are sampled with distribution $\mathcal{NB}(p_0, \phi)$, and even with $\mathcal{NB}(p_1, \phi)$, we chose two different values of p_0 , 0.8 and 0.5, and for each of them, we made p_1 vary so that the odd-ratio $s := p_1/(1-p_1)/[p_0/(1-p_0)]$ is 4, 8 and 16. Finally, we used different values of ϕ as detailed in Table 3.4 in order to explore a wide range of possible dispersions while keeping a signal/noise ratio not too high. Note that the higher ϕ , the less overdispersed the signal. From our experience, the configuration of parameter combinations with $p_0 = 0.5$ is the more typical of observed values for RNA-Seq data.

Provided that the ratio $\lambda = \phi(1-p)/p$ remains constant, the negative binomial distribution with dispersion parameter ϕ going to infinity converges to the Poisson distribution $\mathcal{P}(\lambda)$. We propose an identical simulation study based on the Poisson distribution for the comparison with non-dispersed datasets. Specifically, we used for λ_0 the values 1.25 and

$p_0 = 0.8$		$p_0 = 0.5$	
p_1	ϕ	p_1	ϕ
0.5	5	0.2	$0.08^{1/8}$
0.33	$\sqrt{5}$	0.1	$0.08^{1/4}$
0.2	0.8	0.05	$0.08^{1/2}$
	0.64		0.08

Table 3.4: Values of parameters used in the simulation study

0.73 so that the odd-ratios $s = 4; 8; 16$ corresponded to the respective values $\lambda_1 = 5; 10; 20$ and 2.92; 5.83; 11.7

In practice there is little chance that the overdispersion is known. We propose to estimate this parameter from the data and use the obtained value in the analysis. The results presented here used the estimator inspired from JOHNSON *et al.* (2005): starting from sliding window of size 15, we compute the method of moments estimator of ϕ , using the formula $\phi = \mathbf{E}^2(X)/(V(X) - \mathbf{E}(X))$, and retain the median over all windows. When this median is negative (which is likely to happen in datasets with many zeros), we double the size of the window. In practice however, results are very similar when using maximum likelihood or quasi-maximum likelihood estimators on sliding windows.

Results

We compute the posterior probability $P(E_0|\mathbf{Y}, \mathbf{K})$ for each simulation and each value of d . Figures 3.27 to 3.29 in Appendix 3.2.7 represent the boxplots of this probability for each configuration. For sake of visibility, the outliers were not drawn in those figures. Note that in each figure, the first boxplot corresponds to $d = 0$ and thus to model E_0 , while $d \neq 0$ for left boxplots so that the true model is E_1 . These plots can be understood as abacus for the detection power of the proposed approach. For example, the perfect scenario corresponds

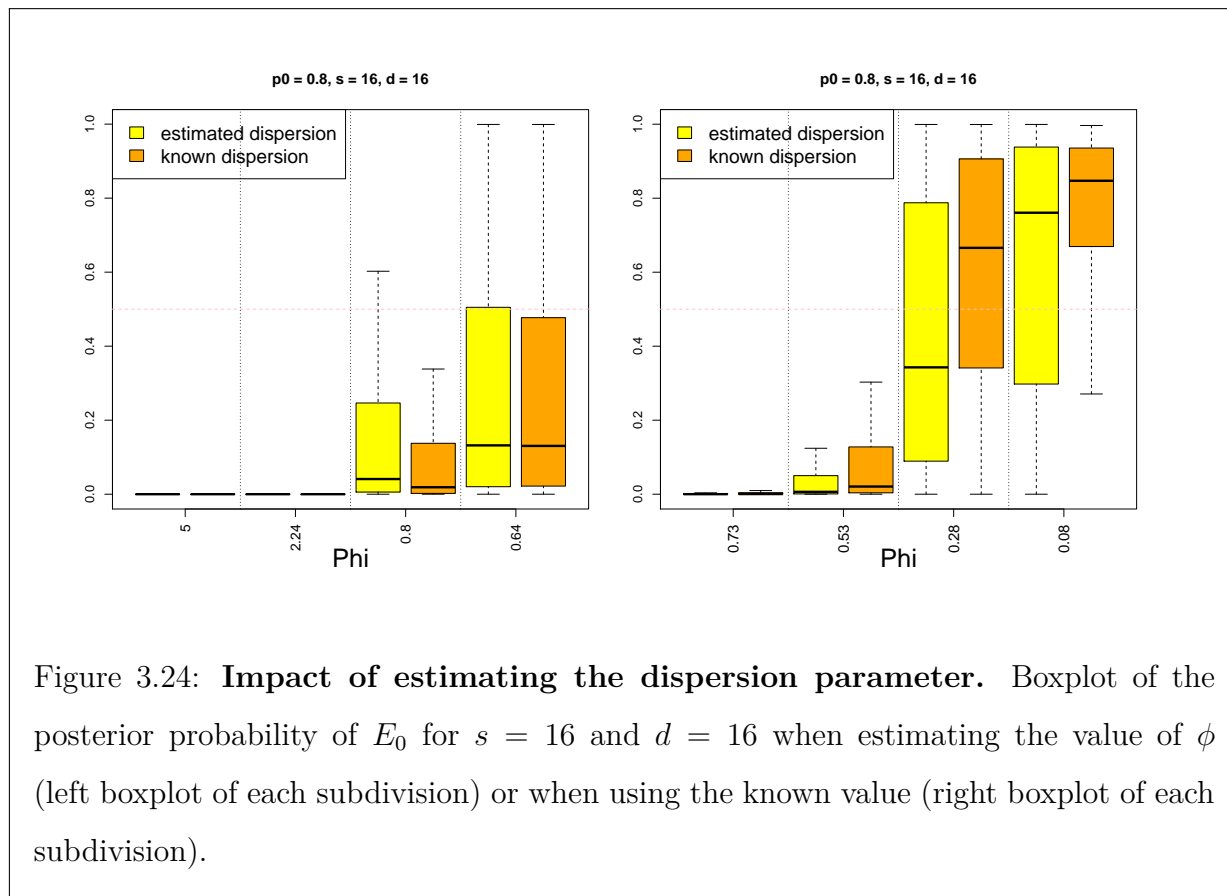


Figure 3.24: **Impact of estimating the dispersion parameter.** Boxplot of the posterior probability of E_0 for $s = 16$ and $d = 16$ when estimating the value of ϕ (left boxplot of each subdivision) or when using the known value (right boxplot of each subdivision).

to $s = 16$ in the Poisson case of Figure 3.27.

As expected, these results show that the lower the value of ϕ (the Poisson distribution is interpreted here as $\phi = +\infty$), the most difficult the decision becomes. The trend is identical for decreasing values of the odd-ratio s and decreasing values of d . In the most difficult scenario of very high dispersion compared to signal value, the method fails to provide satisfying decisions whatever the level of odd-ratio or distance between change-points. However, in most configurations, the method is adequate as soon as $d \geq 16$.

An important question is the impact of the estimation of the dispersion parameter. Interestingly, in the simulation study with $p_0 = 0.8$, our estimator tended to under-estimate ϕ (and thus over-estimate the dispersion) while it was the contrary in the simulation study

with $p_0 = 0.5$. This affects the performance of the decision rule, which behaves better when ϕ is higher. For instance, Figure 3.24 shows, for $s = 16$ and $d = 16$, that knowing the true value of ϕ improves the results when $p_0 = 0.8$ but worsens them when $p_0 = 0.5$.

3.2.6 Comparison of transcribed regions in yeast

Experimental design.

We now go back to our first motivation and consider a study from the Sherlock lab in Stanford (RISSO *et al.*, 2011). In their experiment, they grew a yeast strain, *Saccharomyce Cerevisiae*, in three different environments: ypd, which is the traditional (rich) media for yeast, delft, a similar but poorer media, and glycerol. In the last decade many studies (see for instance PROUDFOOT *et al.*, 2002; TIAN *et al.*, 2005) have showed that a large proportion of genes have more than one polyadenylation sites, thus can express multiple transcripts with different 3' UTR sizes. Similarly, the 5' capping process is dependent on environment conditions (MANDAL *et al.*, 2004), and the 5' UTR size may vary according to stress factors. We may therefore expect that the yeast cells grown in different conditions (they ferment in the first two media, while they respire in glycerol) will produce transcripts of unequal sizes. On the contrary, the intron-exon boundaries are not expected to differ between conditions.

Change-point location.

We applied our procedure to gene YAL013W which has two exons. The RNA-Seq series were segmented into 5 segments to allow one segment per transcribed region separated by segments of non-coding regions. Figure 3.25 illustrates the posterior distribution of each change-point in each profile.

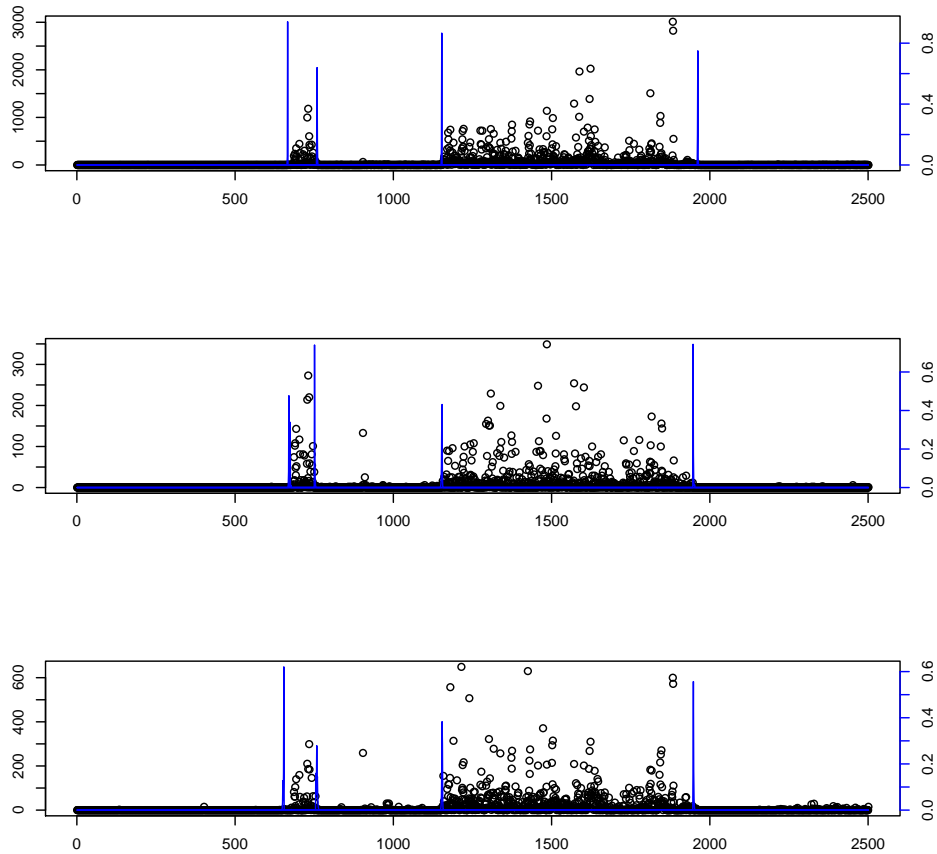


Figure 3.25: **Posterior distribution of change-point location.** Segmentation in 5 segments of gene YAL013W in three different media: ypd (top), delft (middle) and glycerol (bottom). Black dots represent the number of reads starting at each position of the genome (left scale) while blue curves are the posterior distribution of the change-point location (right scale).

Credibility intervals on the shift.

For each of the first to the fourth change-point, we computed the posterior distribution of the difference between change-point locations for each pairs of conditions. For the biological reasons stated above, we expect to observe more differences for the first and last change-points than for the other two, which can be used as a verification of the decision rule.

Figure 3.26 provides the posterior distribution of these differences, as well as the 95% credibility intervals.

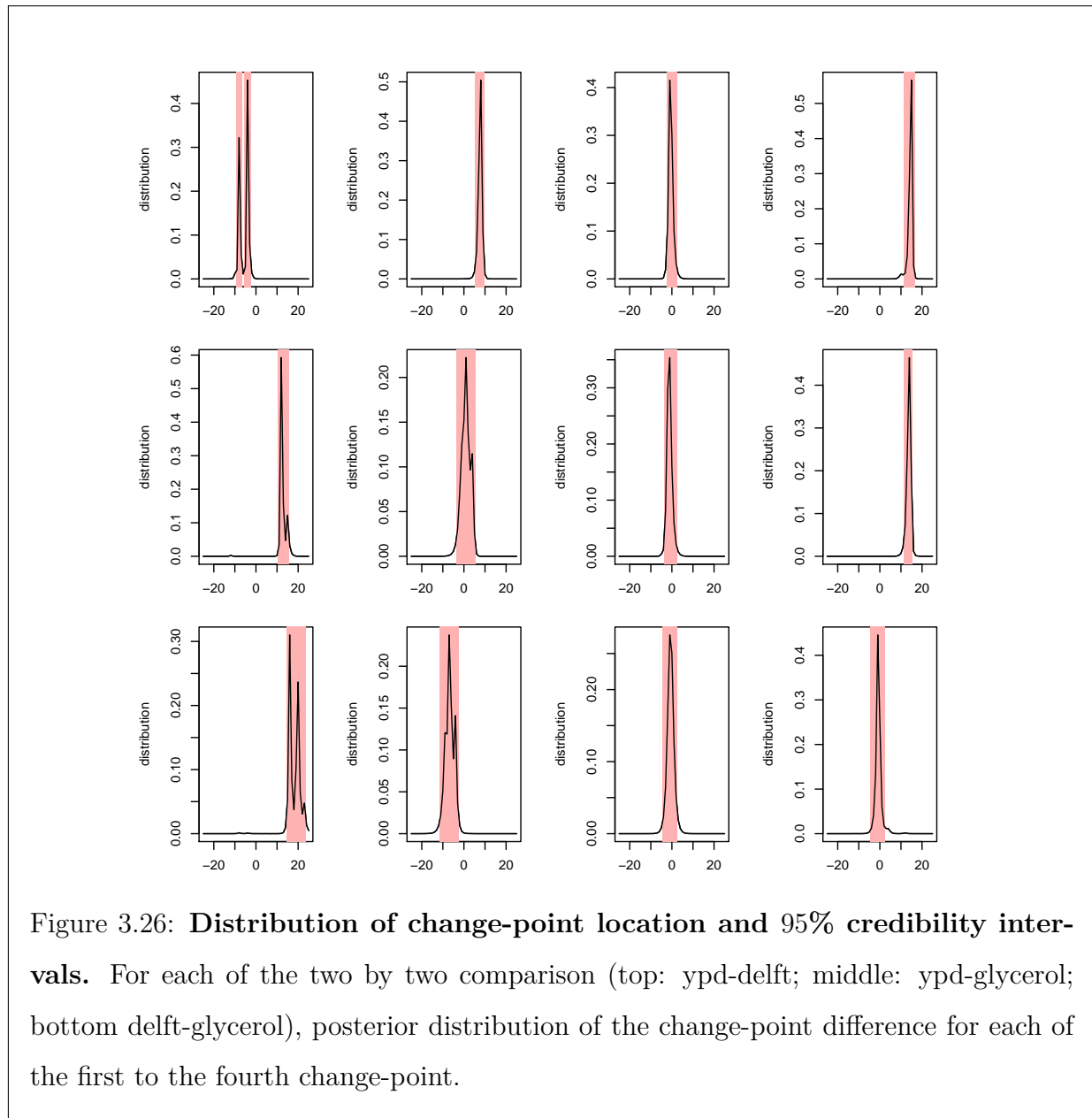
Posterior probability of common change-point.

We then computed the probability that the change-point is the same across several series, taking $p_0 = 1/2$. Table 3.5 provides, for the simultaneous comparison of the three conditions and for each pair of conditions, the value of the posterior probability of E_0 at each change-point (τ_1^ℓ is associated with the 5' UTR, τ_2^ℓ to the 5' intron boundary, τ_3^ℓ to the 3' intron boundary and τ_4^ℓ to the 3' UTR). Reassuringly, in most cases the change-point location is identical when corresponding to intron boundaries. On the contrary, UTR boundaries seem to differ from one condition to another.

3.2.7 Conclusion

We have proposed two exact approaches for the comparison of change-point location. The first is based on the posterior distribution of the shift in two profiles, while the second is adapted to the comparison of multiple profiles and studies the posterior probability of having a common change-point. These procedures, when applied to RNA-Seq datasets, confirm the expectation that transcription starting and ending sites may vary between growth conditions while the localization of introns remains the same.

While we have illustrated these procedures with count datasets, they can be adapted to all distributions from the exponential family verifying the factorability assumption as described in Section 3.2.2. They are in fact implemented in an R package **EBS** for the negative binomial, Poisson, Gaussian heteroscedastic and Gaussian homoscedastic with known variance parameter. This package is available on the CRAN repository at <http://cran.r-project.org/web/packages/EBS/index.html>.



Acknowledgments The authors deeply thank Sandrine Dudoit, Marie-Pierre Etienne, Emilie Lebarbier, Eric Parent and Gavin Sherlock for helpful conversations and comments on this works.

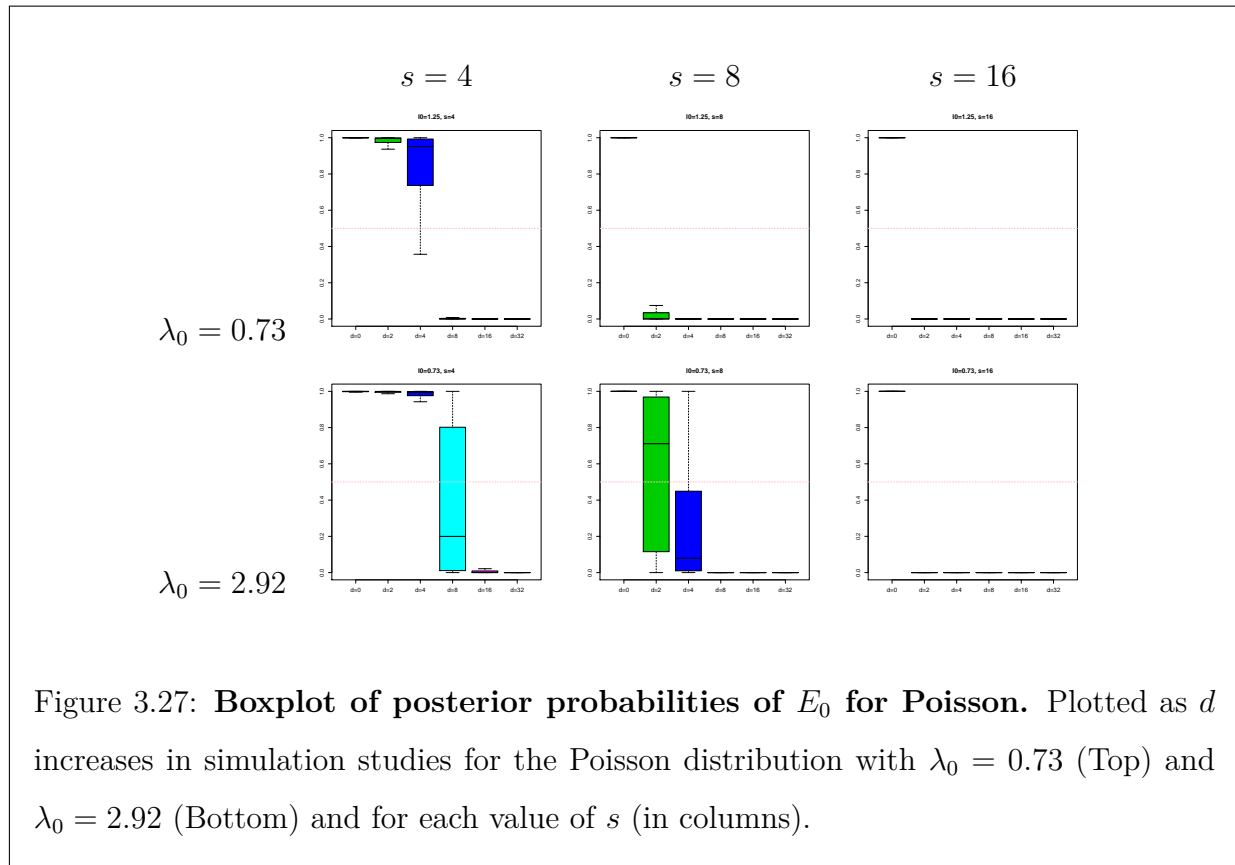
comparison	change-point			
	τ_1	τ_2	τ_3	τ_4
all media	10^{-3}	0.99	0.99	$6 \cdot 10^{-3}$
ypd-delft	0.32	0.30	0.99	10^{-5}
ypd-glycerol	$4 \cdot 10^{-4}$	0.99	0.99	$6 \cdot 10^{-3}$
delft-glycerol	$5 \cdot 10^{-2}$	0.60	0.99	0.99

Table 3.5: Posterior probability of a common change point across conditions for gene YAL013W

Appendix section

References

- Jushan Bai and Pierre Perron. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22, 2003.
- Nicolas Dobigeon, Jean-Yves Tournet, and Jeffrey D Scargle. Joint segmentation of multivariate astronomical time series: Bayesian sampling with a hierarchical model. *Signal Processing, IEEE Transactions on*, 55(2):414–423, 2007.
- Eghbal Ehsanzadeh, Taha BMJ Ouarda, and Hadiza M Saley. A simultaneous analysis of gradual and abrupt changes in Canadian low streamflows. *Hydrological Processes*, 25(5):727–739, 2011.
- Paul I Feder. The log likelihood ratio in segmented regression. *The Annals of Statistics*, 3 (1):84–97, 1975.
- Marie Hušková and Claudia Kirch. Bootstrapping confidence intervals for the change-point of time series. *Journal of Time Series Analysis*, 29(6):947–972, 2008. ISSN 1467-9892. doi: 10.1111/j.1467-9892.2008.00589.x. URL <http://dx.doi.org/10.1111/j.1467-9892.2008.00589.x>.
- N. Johnson, A.W. Kemp, and S. Kotz. Univariate discrete distributions. *John Wiley & Sons, Inc.*, 2005.
- Subhrangsu S Mandal, Chun Chu, Tadashi Wada, Hiroshi Handa, Aaron J Shatkin, and Danny Reinberg. Functional interactions of RNA-capping enzyme with factors that positively and negatively regulate promoter escape by RNA polymerase II. *Proc Natl Acad Sci U S A*, 101(20):7572–7, 2004.
- V. M. Muggeo. Estimating regression models with unknown break-points. *Stat Med*,



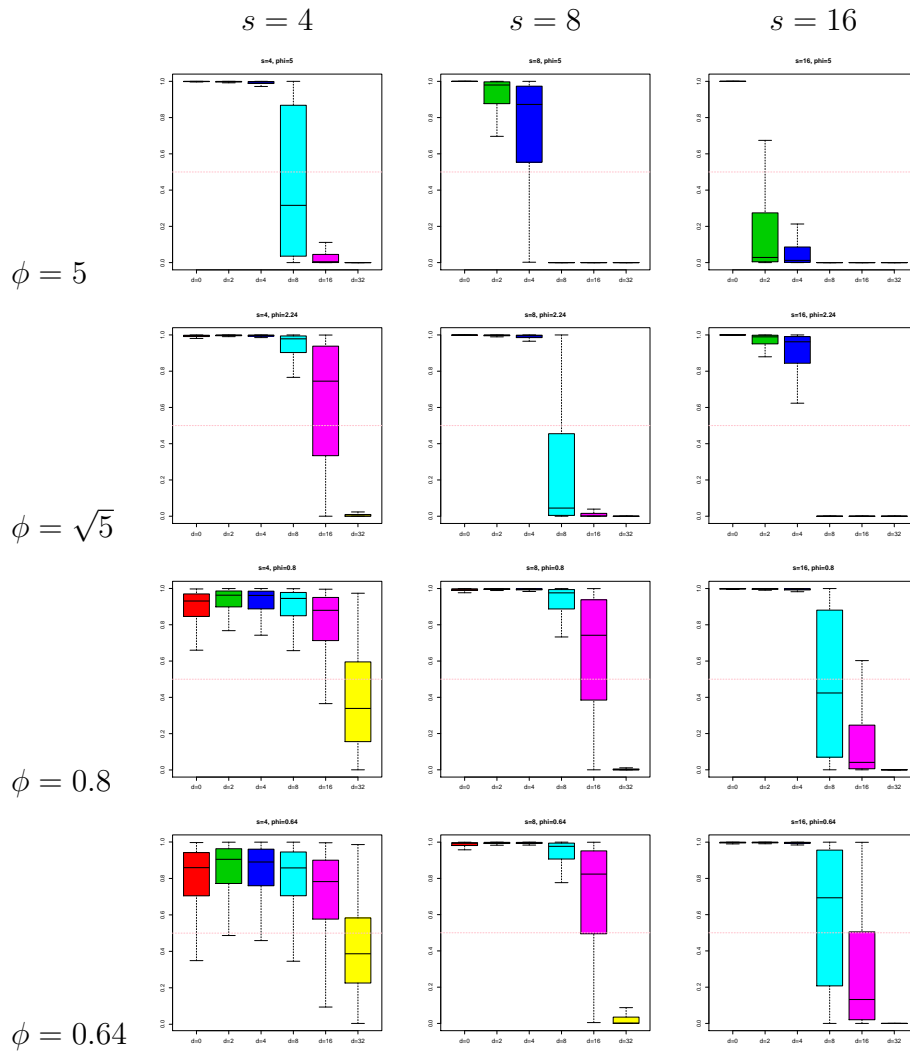


Figure 3.28: **Boxplot of posterior probabilities of E_0 for negative Binomial, with $p_0 = 0.8$.** Plotted as d increases in simulation studies for the negative binomial distribution with $p_0 = 0.8$ and for each value of s (in columns) and each value of ϕ (in rows) as detailed in the left side of Table 3.4. The overdispersion is estimated as detailed in Section 3.2.5.

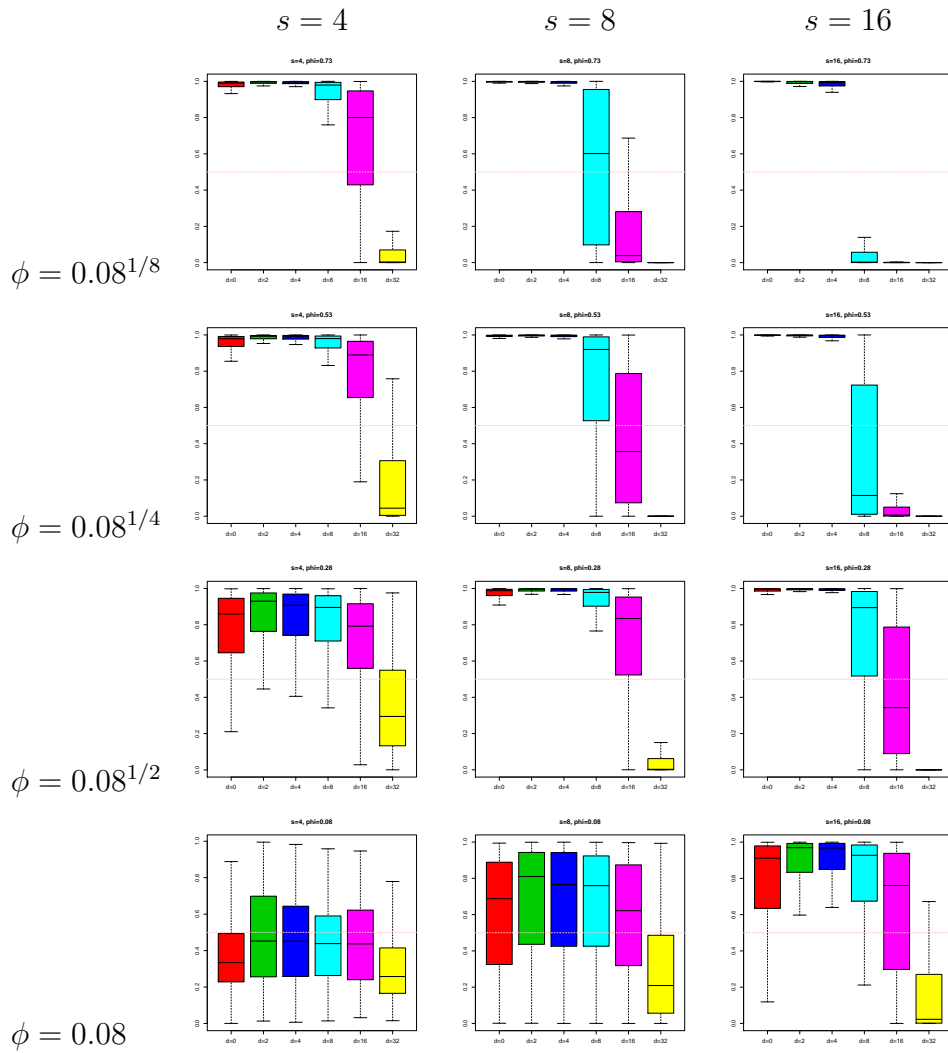


Figure 3.29: **Boxplot of posterior probabilities of E_0 for negative Binomial, with $p_0 = 0.5$.** Plotted as d increases in simulation studies for the negative binomial distribution with $p_0 = 0.5$ and for each value of s (in columns) and each value of ϕ (in rows) as detailed in the right side of Table 3.4. The overdispersion is estimated as detailed in Section 3.2.5.

22: 3055–3071, 2003.

Franck Picard, Emilie Lebarbier, Mark Hoebeke, Guillem Rigaiil, Baba Thiam, and Stéphane Robin. Joint segmentation, calling, and normalization of multiple CGH profiles. *Biostatistics*, 12(3):413–428, 2011.

Nick Proudfoot, Andre Furger, and Michael Dye. Integrating mRNA processing with transcription. *Cell*, 108:501–512, 2002.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

J. Reeves, J. Chen, X. L. Wang, R. Lund, and L. QiQi. A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46(6):900–915, 2007.

G Rigaiil, E Lebarbier, and S Robin. Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and Computing*, 22(4): 917–929, 2012.

Davide Risso, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, 12(1):480, 2011.

Bin Tian, Jun Hu, Haibo Zhang, and Carol Lutz. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research*, 33:201–212, 2005.

J.D. Toms and M.L. Lesperance. Piecewise regression: A tool for identifying ecological thresholds. *Ecology*, 84(8):2034–41, 2003.

3.3 EBS: R package for Exact Bayesian Segmentation

The framework described in RIGAILL *et al.* (2012) and all further developments presented in the previous sections are implemented in the R package EBS. This package relies on the construction and manipulation of two S4 classes of objects: an `EBS` class for the segmentation and analysis of a single profile, and an `EBSProfiles` class, built on the former one, for the analysis of independent profiles. To illustrate the implementation of the package, we will use the EFB1 gene of the yeast genome as a continuing example. We will suppose that the data for this gene is contained in the three vectors `yypd`, `ydel` and `ygly` for the respective conditions `ypd`, `delft` and `glycerol`.

3.3.1 EBS class and associated methods

An object of the `EBS` class is constructed using the method `EBSegmentation`, and is used to store the fundamental probability matrix A as well as the first row (generically noted L_1) and last column (generically noted C_{n+1}) of its k^{th} powers. This method takes as input the dataset as well as various options such as the model to use among 4 possibilities, the maximum number of segments to be considered, the hyperparameters to use for the model, the value to use for an eventual global parameters and the prior on the segmentation to use. Specifically, the method is used in the following way:

```
> out<- EBSegmentation(y, model, Kmax, hyper, theta, var, unif)
```

- `y` is the vector of data;
- `model` takes as input an integer between 1 and 4: 1 corresponds to the Poisson model, 2 to the Gaussian model with global and known variance parameter, 3 to the negative binomial model with global and known dispersion parameter, and 4 to the Gaussian heteroscedastic model;
- `Kmax` is an integer for the maximum number of segments to consider;
- `hyper` are the hyperparameters to use for the conjugate distributions of the models.

The conjugate distributions for models 1 to 4 are given in Section 1.2.3.

Default values can be used: (1, 1) for model 1, (0, 1) for model 2, (1/2, 1/2) for model 3 and $(0, 2\beta, \alpha, \beta)$ for model 4 where α and β are obtained by fitting an inverse gamma distribution on the MAD (HAMPEL, 1974; DONOHO, 1995) estimation for the variance;

- **theta** is the value of the dispersion parameter ϕ used in the negative binomial model. Default value is the estimator proposed in Section 1.2.1;
- **var** is the value of the variance σ^2 used in the Gaussian homoscedastic model. Default value is the estimator proposed by HALL *et al.* (1990), and finally
- **unif** is a boolean stating whether the prior on m is a uniform conditional on K (**unif=TRUE**) or a prior which favors segments of equal length. Default value is **TRUE**.

With those options specified, we can compute the probability matrix A . As stated in Table 1.1, the generic element (up to factor a_J) of this matrix is made of three components which have different roles.

- Term 1 depends only on the data and will appear in each possible segmentation, whichever its number of segments. This means that this component needs not be computed and stored in matrix A as it will not intervene in the comparisons of interest to be made, be it at the segmentation (\mathcal{M}) or at the number of segments (\mathcal{K}) levels.
- As stated in Section 3.1.5, term 2 depends only on the value of the hyperparameters, and will intervene multiplicatively as many times as the number of segments. It thus needs not be computed in matrix A but stored to perform comparisons at level (\mathcal{K}).
- Finally, term 3 is altogether data, segment and hyperparameter dependent. It therefore constitutes the true value to be computed and stored in A .

Note that in practice, this does not require to compute the powers of the whole matrix. Indeed, L_1^k and C_{n+1}^k can be obtained by recursion as respectively $L_1^{k-1} \times A$ and $A \times C_{n+1}^{k-1}$. Moreover, to avoid numerical issues, the logarithm of the generic A terms are stored, and the technique $\log(e^{l_1} + e^{l_2}) = \max(l_1, l_2) + \log(\exp(\min(l_1, l_2) - \max(l_1, l_2)))$ is used for the computation of the product of matrices to ensure that the terms do not go to infinity.

Therefore, the command


```
> Eypd<- EBSegmentation(yypd, 3, 10)
```

computes the A matrix for the y_{pd} profile modeled with the negative binomial distribution, as well as the first line and last column of its 10 first powers.

From EBS class objects, the quantities of interest presented in the previous sections can easily be computed. For instance, the commands

```
> ICL <- EBSICL(out, prior)
> BIC <- EBSBIC(out, prior)
```

compute respectively the ICL and BIC criterion of the data, given some prior on the number of segments. By default, the uniform prior on $1, \dots, K_{max}$ is used, with the K_{max} value used when creating the `out` object. Both those commands return the value of the criterion for each $1 \leq k \leq K_{max}$ and their optimal K . Note that the BIC is computed up to a constant, so that it is proportional to the posterior probability of the number of segments. Its true value can be obtained with the method `EBSPostK(out, prior)` if they are needed, for instance for model averaging.

Finally, information on the posterior change-point distributions can be obtained with the commands

```
> EBSDistrib(out, k, Kk)
> EBSPlotProba(out, K, data)
```

The former computes the posterior distribution of the k^{th} change-point of a segmentation in Kk segments (returned under the form of a vector of length $n + 1$), while the second plots the posterior distribution of all $K - 1$ change-points of a segmentation in K segments, on top of the data if `data` is true (value by default is false). Figure 3.30 is the output of the command

```
> EBSPlotProba(Eypd, EBSICL(Eypd)$NbICL, data=TRUE)
```

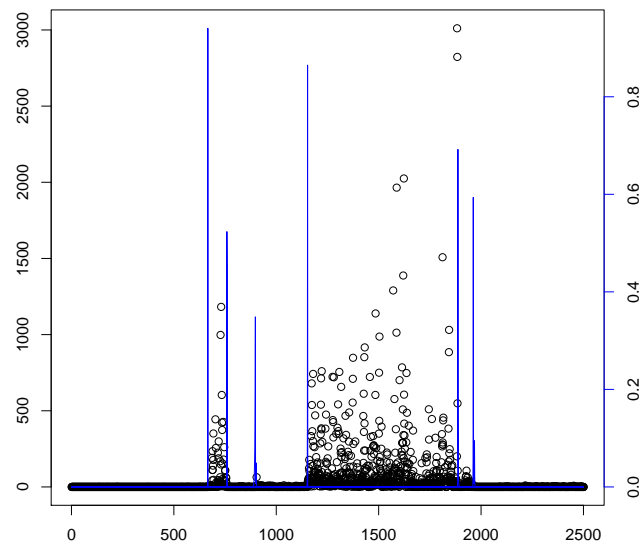


Figure 3.30: **Posterior distribution of change-points location.** Output of the command `EBSPlotProba` for the `ypd` profile with K chosen with the ICL criterion.

3.3.2 EBSProfiles class and change-point comparison

The class `EBSProfiles` is dedicated to the analysis of multiple and independent profiles, and is based on the structure of objects of class `EBS`. The construction of such an object is performed with the command

```
> EBSProfiles(data, model, K, hyper, theta, var, homoscedastic, unif)
```

- `data` is the $\ell \times n$ matrix of data, (with ℓ the number of profiles and n the length of each profile);
- `model` takes as input an integer between 1 and 4 as for `EBS` class objects,
- `K` is a vector of integer for the maximum number of segments to consider for each profile (thus not requiring all profiles to have the same K_{max});
- `hyper` are the hyperparameters to use for the conjugate distributions of the models. This entry requires a vector of size $2 \times \ell$ for models 1, 2 and 3, and $4 \times \ell$ for model 4.

- `theta` is the vector of values of the dispersion parameter ϕ used in the negative binomial model;
- `var` is the vector value of the variance σ^2 used in the Gaussian homoscedastic model,
- `homoscedastic` is a boolean stating whether or not the global parameters should be shared by each profile, if appropriate (default value being set to false), and finally,
- `unif` is a boolean stating whether the prior on m is a uniform conditional on K .

Note that the same default values are included as in the EBS objects if needed.

While stored slightly differently, an object of class `EBSProfiles` is the set of ℓ objects of class `EBS`, which can be extracted through the command

```
> GetCondition(x, Condition)
```

where `x` is the `EBSProfiles` object and `Condition` is the index of the profile to be extracted. It is then possible to apply all methods of Section 3.3.1 to the output of `GetCondition`, even though some analysis can be performed in parallel for all profiles. It is for instance the case for the computation of the ICL criterion, with the method `EBSICLProfiles`, or for the plotting of the posterior distribution of the change-point location of each profile, with the method `EBSPlotProbaProfiles`. Figure 3.31 illustrate the output of the commands

```
> Eall <- EBSProfiles(rbind(yypd, ydel, ygly), 3, 10)
> K<-c(5,5,5)
> EBSPlotProbaProfiles(Eall, K, data=TRUE)
```

The two main tools for the comparison of profiles described in Section 3.2 are then implemented in the methods:

```
> CompCredibility(x, Conditions, Tau, K)
> EBSStatistic(x, Conditions, Tau, K, p0)
```

Both take as input an object of class `EBSProfiles`, the index of the conditions to compare, the index of the change-points of interest, and the number of segments of each

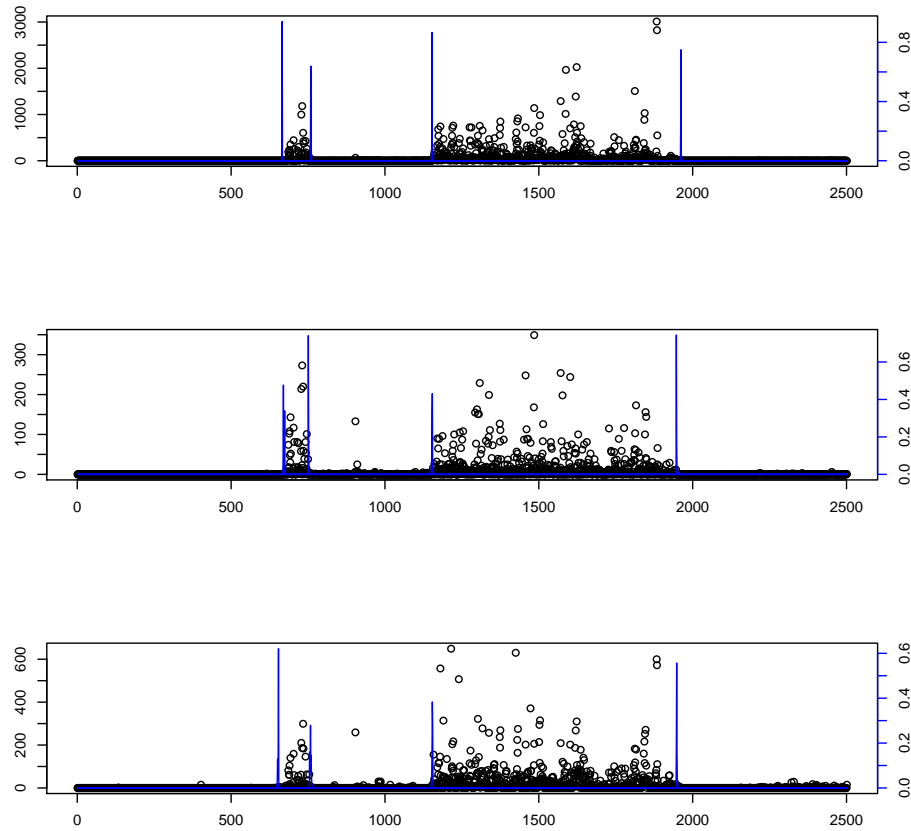


Figure 3.31: **Posterior distribution of change-points location of three profiles.**

Output of the command `EBSPlotProba` for all profiles segmented into 5 segments.

condition. The former is limited to the comparison of 2 profiles and thus `Conditions` must be a vector of integer of length 2. It returns the posterior distribution of the change-point difference, as well as the level of both the smallest credibility interval containing zero and before reaching zero. A plot of this distribution and associated α credibility interval can be obtained with method `plot.Credibility(x,level)` where `x` is an output of the `CompCredibility` method and α is given by `level`, as illustrated in Figure 3.32 which compares the first change-point of profiles corresponding to the `ypd` and `delft` growth conditions segmented into 5 segments. The latter, `EBSStatistic` is dedicated to the comparison of as many profiles as wanted. It computes, given its prior probability `p0`, the posterior prob-

ability of event E_0 introduced in Section 3.2 . A default value of $1/2$ is used in case the user does not have an informative prior. For instance, the posterior probability that the second change-point of each profile segmented into 5 segments (cf Figure 3.31) are located at the same position is computed with command `EBSStatistic(Eall, 1:3, c(2,2,2), c(5,5,5), 1/2)` and is equal to 0.9923283.

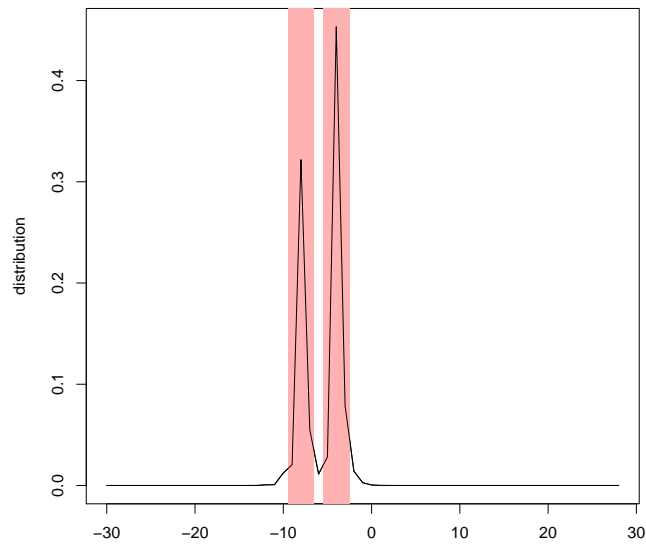


Figure 3.32: **Credibility interval of difference in change-point location.** Posterior distribution of the first change-point location difference between profiles ypd and delft segmented into 5 segments, and associated 0.95 credibility interval.

The examples provided in this section were obtained in June 2013 with version 2.7 of the EBS package.

3.4 Results on the yeast dataset

We have already illustrated throughout the presentation of the R package (Chapter 3.3), and the development of our comparison methods (Chapter 3.2) how this Bayesian framework applies to the analysis of a profile corresponding to a gene.

We have applied those methods to a set of 50 genes from the yeast genome which all have 2 exons, and which were expressed at the time of the experiment. Figure 3.33 illustrates the distribution of the posterior probability of E_0 (event that all change-points have same location) for each of the four change-points of those genes, when using $p_0 = 1/2$ as a prior value. As expected, the change-point corresponding to the intron boundaries have very high posterior probability, while almost half of the other have a posterior probability lower than $1/2$.

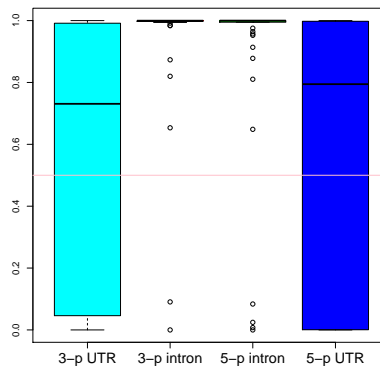
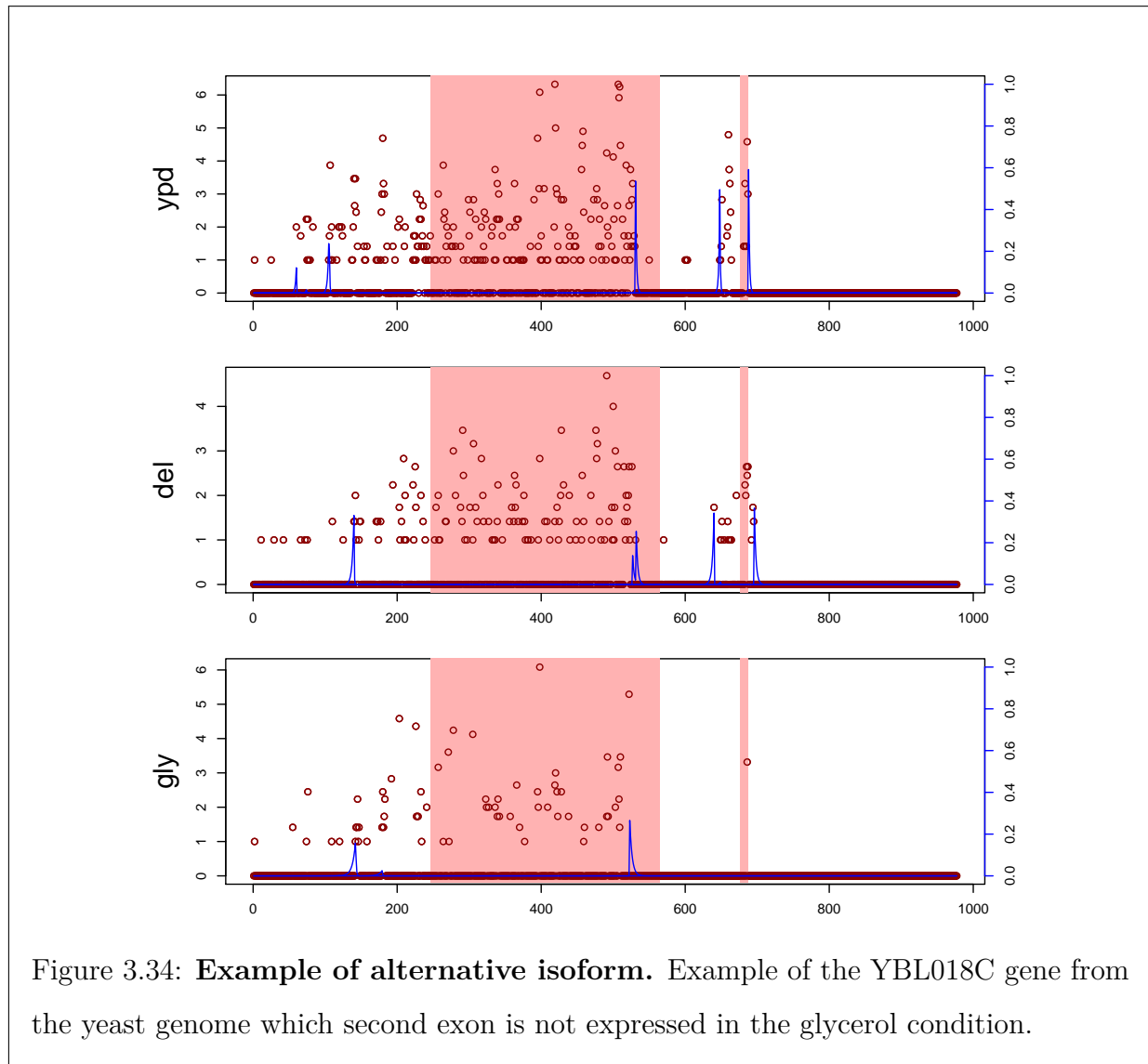


Figure 3.33: **Posterior probability of E_0 .** Distribution of the $P(E_0|\mathbf{Y}, \mathbf{K})$ obtained for each change-point when using $p_0 = 1/2$ for each of the 50 genes.

In fact, when looking at the 5 genes for which the posterior probability of E_0 suggests that at least one of the intron boundaries should be classified as differing between conditions, we found that all of them had one of their two exons which was not expressed in the Glycerol medium (see for instance Figure 3.34).



While methods dedicated to the identification of isoforms and their respective expression proportions would probably have identified these cases, one might note that all traditional differential expression approaches would not have noticed this. Indeed, the first step of such methods consists in summing over all nucleotides of a given gene their associated read counts in order to obtain one number per gene per condition, independent of the number of expressed exons. While it is probable that such analysis would have resulted in classifying these 5 genes as under-expressed in glycerol, the alternative splicing would have gone unnoticed.

Moreover, some further discussion with Dr Sherlock suggested that about 10% of the genes should be liable to differential splicing. We therefore performed the analysis over again removing the 5 previously identified outliers and setting $p_0 = 0.9$ for the change-points corresponding to UTR boundaries and $p_0 = 0.99$ for the intron boundaries. These new results are illustrated in Figure 3.35. For these new prior values, we observe that 9 genes have a 3' UTR length which varies, and 16 for the 5' UTR.

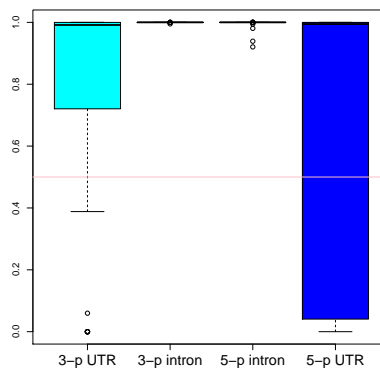


Figure 3.35: **Posterior probability of E_0 with informative priors.** Distribution of the $P(E_0|\mathbf{Y}, \mathbf{K})$ obtained for each change-point when using $p_0 = 0.9$ for UTR boundaries and $p_0 = 0.99$ for intron boundaries.

We can conclude from this analysis that yeast cells have differential UTR lengths when grown in different media and that this phenomenon might be responsible for differences observed in the phenotype, such as switching between respiration and fermentation. But this differential splicing process is not limited to UTRs, it appears that it actually extends to whole exons, despite the small amount of yeast genes which possess more than one. Moreover, as expected, intron boundaries are conserved between conditions.

Bibliography

- AKAIKE, H., 1973 Information theory and extension of the maximum likelihood principle. Second international symposium on information theory : 267–281.
- AKAKPO, N., 2011 Estimating a discrete distribution via histogram selection. *ESAIM: Probability and Statistics* **15**: 1–29.
- ANDRIEU, C., P. DJURIĆ, and A. DOUCET, 2001 Model selection by MCMC computation. *Signal Processing* **81**: 19–37.
- ARLOT, S., and A. CELISSE, 2011 Segmentation of the mean of heteroscedastic data via cross-validation. *Statistics and Computing* **21**: 613–632.
- ARLOT, S., and P. MASSART, 2009 Data-driven calibration of penalties for least-squares regression. *The Journal of Machine Learning Research* **10**: 245–279.
- BAI, J., and P. PERRON, 2003 Computation and analysis of multiple structural change models. *Journal of Applied Econometrics* **18**: 1–22.
- BARAUD, Y., and L. BIRGÉ, 2009 Estimating the intensity of a random measure by histogram type estimators. *Probability Theory Related Fields* **143**: 239–284.
- BARRON, A., L. BIRGÉ, and P. MASSART, 1999 Risk bounds for model selection via penalization. *Probability Theory Related Fields* **113**: 301–413.
- BARRY, D., and J. A. HARTIGAN, 1993 A Bayesian analysis for change point problems. *Journal of the American Statistical Association* **88**: 309–319.
- BAUM, L. E., T. PETRIE, G. SOULES, and N. WEISS, 1970 A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics* **41**: 164–171.
- BIERNACKI, C., G. CELEUX, and G. GOVAERT, 2000 Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**: 719–725.

- BIRGÉ, L., 2007 Model selection for Poisson processes. In *Asymptotics: particles, processes and inverse problems*, volume 55 of *IMS Lecture Notes Monogr. Ser.*. Inst. Math. Statist., Beachwood, OH, 32–64.
- BIRGÉ, L., and P. MASSART, 1997 From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*. Springer, New York, 55–87.
- BIRGÉ, L., and P. MASSART, 2001 Gaussian model selection. *Journal of the European Mathematical Society* **3**: 203–268.
- BIRGÉ, L., and P. MASSART, 2007 Minimal penalties for Gaussian model selection. *Probability Theory Related Fields* **138**: 33–73.
- BOCKHORST, J., and N. JOJIC, 2012 Discovering patterns in biological sequences by optimal segmentation. arXiv preprint arXiv:1206.5256 .
- BOEVA, V., A. ZINOVYEV, K. BLEAKLEY, J.-P. VERT, I. JANOUÉIX-LEROSEY, *et al.*, 2011 Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* **27**: 268–269.
- BRAUN, J. V., R. BRAUN, and H.-G. MÜLLER, 2000 Multiple changepoint fitting via quaslikelihood, with application to dna sequence segmentation. *Biometrika* **87**: 301–314.
- BRAUN, J. V., and H.-G. MULLER, 1998 Statistical methods for DNA sequence segmentation. *Statistical Science* : 142–162.
- BREIMAN, FRIEDMAN, OLSHEN, and STONE, 1984 Classification and regression trees. Wadsworth and Brooks .
- CASTELLAN, G., 1999 Modified Akaike’s criterion for histogram density estimation. preprint **99**.
- CHIANG, D. Y., G. GETZ, D. B. JAFFE, M. J. O’KELLY, X. ZHAO, *et al.*, 2008 High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods* **6**: 99–103.
- CHIB, S., 1998 Estimation and comparison of multiple change-point models. *Journal of Econometrics* **86**: 221–241.
- CLEYNEN, A., M. KOSKAS, and G. RIGAILL, under review A generic implementation of the pruned dynamic programming algorithm. Arxiv preprint arXiv:1204.5564 .
- CLEYNEN, A., and E. LEBARBIER, 2013 Segmentation of the poisson and negative binomial rate models: a penalized estimator. ArXiv preprint arXiv:1301.2534 .
- COMTE, F., and Y. ROZENHOLC, 2004 A new algorithm for fixed design regression and denoising. *Annals of the Institute of Statistical Mathematics* **56**: 449–473.

- DAVIS, R. A., T. C. M. LEE, and G. A. RODRIGUEZ-YAM, 2006 Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association* **101**: 223–239.
- DOBIGEON, N., J.-Y. TOURNERET, and J. D. SCARGLE, 2007 Joint segmentation of multivariate astronomical time series: Bayesian sampling with a hierarchical model. *Signal Processing, IEEE Transactions on* **55**: 414–423.
- DONOHO, D. L., 1995 De-noising by soft-thresholding. *Information Theory, IEEE Transactions on* **41**: 613–627.
- DUROT, C., E. LEBARBIER, and A. TOCQUET, 2009 Estimating the joint distribution of independent categorical variables via model selection. *Bernoulli* **15**: 475–507.
- EHSANZADEH, E., T. B. OUARDA, and H. M. SALEY, 2011 A simultaneous analysis of gradual and abrupt changes in Canadian low streamflows. *Hydrological Processes* **25**: 727–739.
- ERDMAN, C., and J. W. EMERSON, 2008 A fast Bayesian change point analysis for the segmentation of microarray data. *Bioinformatics* **24**: 2143–2148.
- FEARNHEAD, P., 2005 Exact Bayesian curve fitting and signal segmentation. *Signal Processing, IEEE Transactions on* **53**: 2160–2166.
- FEDER, P. I., 1975 The log likelihood ratio in segmented regression. *The Annals of Statistics* **3**: 84–97.
- FRANKE, J., C. KIRCH, and J. T. KAMGAING, 2012 Changepoints in times series of counts. *Journal of Time Series Analysis* **33**: 757–770.
- GREEN, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**: 711–732.
- GUTHERY, S. B., 1974 Partition regression. *Journal of the American Statistical Association* **69**: 945–947.
- HACCOU, P., and E. MEELIS, 1988 Testing for the number of change points in a sequence of exponential random variables. *Journal of Statistical Computation and Simulation* **30**: 285–298.
- HALL, P., J. KAY, and D. TITTERINTON, 1990 Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77**: 521–528.
- HAMPEL, F. R., 1974 The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**: 383–393.
- HARTIGAN, J. A., and M. A. WONG, 1979 A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28**: 100–108.

- HERNANDO, D., V. CRESPI, and G. CYBENKO, 2005 Efficient computation of the hidden Markov model entropy for a given observation sequence. *Information Theory, IEEE Transactions on* **51**: 2681–2685.
- HOCKING, T. D., G. SCHLEIERMACHER, I. JANOUÉIX-LEROSEY, V. BOEVA, J. CAPPO, *et al.*, 2013 Learning smoothing models of copy number profiles using breakpoint annotations. *BMC Bioinformatics* **14**: 164.
- HSU, L., S. G. SELF, D. GROVE, T. RANDOLPH, K. WANG, *et al.*, 2005 Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics* **6**: 211–226.
- HUPÉ, P., N. STRANSKY, J.-P. THIERY, F. RADVANYI, and E. BARILLOT, 2004 Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* **20**: 3413–3422.
- HUŠKOVÁ, M., and C. KIRCH, 2008 Bootstrapping confidence intervals for the change-point of time series. *Journal of Time Series Analysis* **29**: 947–972.
- JEFFREYS, H., 1946 An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **186**: 453–461.
- JOHNSON, N., A. KEMP, and S. KOTZ, 2005 *Univariate discrete distributions*. John Wiley & Sons, Inc. .
- KILLICK, R., and I. A. ECKLEY, 2011 *Changepoint: an R package for changepoint analysis*. Lancaster University .
- KILLICK, R., P. FEARNHEAD, and I. ECKLEY, 2012 Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association* **107**: 1590–1598.
- LAI, W. R., M. D. JOHNSON, R. KUCHERLAPATI, and P. J. PARK, 2005 Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21**: 3763–3770.
- LANGMEAD, B., C. TRAPNELL, M. POP, and S. SALZBERG, 2008 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**.
- LAVIELLE, M., 2005 Using penalized contrasts for the change-point problem. *Signal Processing* **85**: 1501–1510.
- LEBARBIER, E., 2005 Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing* **85**: 717–736.
- LUONG, T. M., Y. ROZENHOLC, and G. NUEL, 2013 Fast estimation of posterior probabilities in change-point analysis through a constrained hidden Markov model. *Computational Statistics & Data Analysis* .

- MANDAL, S. S., C. CHU, T. WADA, H. HANDA, A. J. SHATKIN, *et al.*, 2004 Functional interactions of RNA-capping enzyme with factors that positively and negatively regulate promoter escape by RNA polymerase II. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 7572–7577.
- MARTIN, D. E., and J. A. ASTON, 2012 Distribution of statistics of hidden state sequences through the sum-product algorithm. *Methodology and Computing in Applied Probability* **15**: 1–22.
- MASSART, P., 2007 Concentration inequalities and model selection .
- MUGGEO, V. M., 2003 Estimating regression models with unknown break-points. *Statistics in Medicine* **22**: 3055–3071.
- NAGALAKSHMI, U., Z. WANG, K. WAERN, C. SHOU, D. RAHA, *et al.*, 2008 The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344–1349.
- OLSHEN, A. B., E. VENKATRAMAN, R. LUCITO, and M. WIGLER, 2004 Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**: 557–572.
- PICARD, F., E. LEBARBIER, M. HOEBEKE, G. RIGAILL, B. THIAM, *et al.*, 2011 Joint segmentation, calling, and normalization of multiple CGH profiles. *Biostatistics* **12**: 413–428.
- PICARD, F., S. ROBIN, M. LAVIELLE, C. VAISSE, and J.-J. DAUDIN, 2005 A statistical approach for array CGH data analysis. *BMC Bioinformatics* **6**: 27.
- PICARD, F., S. ROBIN, E. LEBARBIER, and J.-J. DAUDIN, 2007 A segmentation/clustering model for the analysis of array cgh data. *Biometrics* **63**: 758–766.
- PROUDFOOT, N., A. FURGER, and M. DYE, 2002 Integrating mRNA processing with transcription. *Cell* **108**: 501–512.
- RAND, W. M., 1971 Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* **66**: 846–850.
- REEVES, J., J. CHEN, X. L. WANG, R. LUND, and L. QIQI, 2007 A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology* **46**: 900–915.
- REYNAUD-BOURET, P., 2003 Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probability Theory Related Fields* **126**: 103–153.
- RIGAILL, G., 2010 Pruned dynamic programming for optimal multiple change-point detection. Arxiv:1004.0887 .

- RIGAILL, G., E. LEBARBIER, and S. ROBIN, 2012 Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and Computing* **22**: 917–929.
- RISSE, D., K. SCHWARTZ, G. SHERLOCK, and S. DUDOIT, 2011 GC-content normalization for RNA-Seq data. *BMC Bioinformatics* **12**: 480.
- RIVERA, C., and G. WALTHER, 2012 Optimal detection of a jump in the intensity of a Poisson process or in a density with likelihood ratio statistics. *arXiv:1211.2859* .
- ROBINSON, M. D., D. J. MCCARTHY, and G. K. SMYTH, 2010 *edgeR*: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140.
- SCOTT, A., and M. KNOTT, 1974 A cluster analysis method for grouping means in the analysis of variance. *Biometrics* **30**: 507–512.
- SHEN, J. J., and N. R. ZHANG, 2012 Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next-generation DNA sequencing. *The Annals of Applied Statistics* **6**: 476–496.
- TIAN, B., J. HU, H. ZHANG, and C. LUTZ, 2005 A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research* **33**: 201–212.
- TOMS, J., and M. LESPERANCE, 2003 Piecewise regression: A tool for identifying ecological thresholds. *Ecology* **84**: 2034–41.
- VENKATRAMAN, E., and A. B. OLSHEN, 2007 A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics* **23**: 657–663.
- VITERBI, A., 1967 Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on* **13**: 260–269.
- XIE, C., and M. TAMMI, 2009 CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10**: 80.
- YAO, Y.-C., 1988 Estimating the number of change-points via schwarz' criterion. *Statistics & Probability Letters* **6**: 181–189.
- YOON, S., Z. XUAN, V. MAKAROV, K. YE, and J. SEBAT, 2009 Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research* **19**: 1586–1592.
- ZHANG, N. R., and D. O. SIEGMUND, 2007 A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63**: 22–32.