The Cross-Entropy Method and Multiple Change-Point Detection in Genomic Sequences

by

Madawa Priyadarshana, W. J. R.

This thesis is submitted in fulfilment for the degree of Doctor of Philosophy in the

Department of Statistics

MACQUARIE University
SYDNEY·AUSTRALIA

FACULTY OF
SCIENCE AND ENGINEERING

April 2015

Declaration

I declare that the PhD thesis titled, 'The Cross-Entropy Method and Multiple Change-Point Detection in Genomic Sequences' and the work presented in it are my own. I confirm that the work described in this thesis was performed between December 2011 and November 2014 at the Department of Statistics, Faculty of Science and Engineering, Macquarie University.

The thesis contains one peer-reviewed book chapter, one-peer reviewed journal article and two peer-reviewed full conference papers. Except where otherwise indicated, this thesis is entirely my own work. Any help and assistance that I have received in my research work and the preparation of the thesis itself have been appropriately acknowledged. Furthermore, this thesis contains no material that has been submitted previously, in whole or in part, for the award of any other academic degree, fellowship or other recognition.

Madawa Priyadarshana, W. J. R.

Date: ________________________________

“An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.”

John Tukey

Abstract

Heterogeneity in observed data is a common feature that statisticians have to deal with when analyzing data. Estimating these changes in an observed process not only helps to better model the underlying phenomena, but also facilitates the process of making more informed decisions. In health informatics, when analyzing patients' genomes with complex diseases, it is a pivotal step in finding disease-causing genes or active regions of the genome that has functional importance when characterizing these diseases. Change-point analysis methods are among the best approaches that can be used to address this problem of locating important genomic variations in genomes. Detection of these variations helps researchers and practitioners to assess disease progression, prognosis and efficacy of treatments. Thus, at patient level it helps to provide more improved personalized medicine to alleviate a disease.

The overall research aim of this thesis is to introduce the Cross-Entropy (CE) method, a model-based stochastic optimization procedure that nests under the branch of evolutionary computing techniques, to estimate both the number of change-points and their locations in biological sequences. Particularly we focused on analyzing array comparative genomic hybridization (aCGH) data and DNA read count data obtained through next generation sequencing (NGS) methods. Several variants of the CE method are proposed in this work to detect change-point locations in both continuous and discrete (count) data. Different model selection criteria are used in the CE method to estimate the optimal number of change-points. It is known that evolutionary computing methods consume more computational resources due to the nature of their implementation. In this thesis we propose two alternative solutions to ameliorate this efficiency issue of the general CE algorithm. At first, we develop a multi-core parallel implementation of the CE algorithm in the R statistical computing environment. Later, for the first time in the literature, we combine two powerful sequential detection techniques with the CE method to further increase its efficiency. We further explore the feasibility of incorporating auxiliary information to the process of change-point detection in the CE method with the use of generalized additive model for location, scale and shape (GAMLSS). A series of extensive simulations were performed in multiple publications to establish the procedures and to ascertain their efficacy. We apply the proposed variants of the CE method to both aCGH and DNA read count data obtained through NGS methods to detect copy number variations. The methods discussed in this thesis are freely available as an R package named "breakpoint" at the website http://cran.r-project.org/web/packages/breakpoint/index.html.

This thesis contains four peer-reviewed publications, which include a book chapter, a journal article and two conference papers. It further includes details of an R package developed to detect multiple change-points in continuous and count data based on the methods developed in this thesis.

Publications, Presentations and Awards

Peer-reviewed book chapter

- **Priyadarshana, W. J. R. M.**, Polushina, T., and Sofronov, G. (2015). Hybrid algorithms for multiple change-point detection in biological sequences. In Sun, C., Bednarz, T., Pham, T., Vallotton, P., and Wang, D., (Eds.), Signal and Image Analysis for Biomedical and Life Sciences, volume 823 of Advances in Experimental Medicine and Biology. Springer International Publishing. http://link.springer.com/chapter/10.1007/978-3-319-10984-8_3.

Peer-reviewed journal publication

- **Priyadarshana, W. J. R. M.**, Sofronov G. (2014). Multiple Break-Points Detection in array CGH Data via the Cross-Entropy Method, IEEE/ACM Transactions on Computational Biology and Bioinformatics, no. 1, pp. 1, PrePrints, doi:10.1109/TCBB.2014.2361639, ISSN: 1545-5963.

Peer-reviewed full conference publications

- **Priyadarshana, W. J. R. M.**, Polushina, T. and Sofronov, G. (2013). A hybrid algorithm for multiple change-point detection in continuous measurements, In Proc. of the International Symposium on Computational Models for Life Sciences, AIP Conference Proceedings, vol. 1559, pp. 108–117, http://dx.doi.org/10.1063/1.4825002.

- **Priyadarshana, W. J. R. M.** and Sofronov, G. (2013). GAMLSS and Extended Cross-Entropy Method to Detect Multiple Change-Points in DNA Read Count Data, In: Muggeo VMR, Capursi V, Boscaino G, Lovison G (Eds.), Proceedings of the 28th International Workshop on Statistical Modelling, vol.1, 453-457, ISBN 978-88-96251-47-8.

- **Priyadarshana W. J. R. M.**, Sofronov G. (2012). A Modified Cross-Entropy Method for Detecting Change-Points in the Sri-Lankan Stock Market. In: B. M. Chen, M. T. Khan, K-K. Tan (Eds.) IASTED International Conference on Engineering and Applied Science, Colombo, Sri Lanka, DOI: 10.2316/P.2012.785-041.

- **Priyadarshana, W. J. R. M.** and Sofronov, G. (2012). The Cross-Entropy Method and Multiple Change-Points Detection in Zero-Inflated DNA read count data. In: Y. T. Gu, S. C. Saha (Eds.) The 4th International Conference on Computational Methods (ICCM2012), 1-8, ISBN 978-1-921897-54-2.

- Sofronov, G., Polushina, T. and **Priyadarshana W. J. R. M.** (2012). Sequential Change-Point Detection via the Cross-Entropy Method. In Proc. of the 11th Symposium on Neural Network Applications in Electrical Engineering (NEUREL), 185-188, DOI: 10.1109/NEUREL.2012.6420004.

- **Priyadarshana, W. J. R. M.** and Sofronov, G. (2012). A Modified Cross-Entropy Method for Detecting Multiple Change-Points in DNA Count Data. In Proc. of the IEEE Conference on Evolutionary Computation (CEC), 1020-1027, DOI: 10.1109/CEC.2012.6256470.

Software

- **Priyadarshana, W. J. R. M.** and Sofronov, G. (2014). breakpoint: An R Package for Multiple Break-Points Detection via the Cross-Entropy Method. R package version 1.1.

Presentations

- Multiple Break-Points Detection in Biological Sequences via the Cross-entropy Method, Australian Statistical Conference in conjunction with the IMS Annual meeting (ASC-IMS), Sydney, Australia, 2014.

- A hybrid algorithm for multiple change-point detection in continuous measurements, International Symposium on Computational Models for Life Sciences, Sydney, Australia, 2013.

- GAMLSS and Extended Cross-Entropy Method to Detect Multiple Change-Points in DNA Read Count Data, 28th International Workshop on Statistical Modelling, Palermo, Italy, 2013.

- An Extended Cross-Entropy Method for Detection of Copy Number Variations in Biological Sequences. 26th European Conference on Operational Research, Rome, Italy, 2013.

- Detection of Copy Number Variation in Next Generation Sequencing Data via the Cross-Entropy Method. Young Statisticians Conference, Melbourne, Australia, 2013.

- A Modified Cross-Entropy Method for Detecting Change-Points in the Sri-Lankan Stock Market. The IASTED International Conference on Engineering and Applied Science, Colombo, Sri-Lanka, 2012.

- The Cross-Entropy Method and Multiple Change-Points Detection in Zero-Inflated DNA read count data, The 4th International Conference on Computational Methods, Gold Coast, Australia, 2012.

- A Modified Cross-Entropy Method for Detecting Multiple Change-Points in DNA Count Data, WCCI 2012 IEEE World Congress on Computational Intelligence, IEEE CEC, Brisbane, Australia, 2012.

- Research congress, Department of Statistics, Macquarie University, Sydney, Australia, 2012 - 2014.
  - A Modified Cross Entropy Method for Detecting Multiple Change Points in DNA Count Data, 2012.
  - GAMLSS and Extended Cross-Entropy Method to Detect Multiple Change-Points in DNA Read Count Data, 2013.
  - The Cross-Entropy method and multiple change-point detection in biological sequences of continuous measurements, 2014.

Awards and Grants

- J.B.Douglas Postgraduate Awards for excellence in postgraduate research in Statistics or Econometrics, The Statistical Society of Australia Inc., 2012.

- Travel grant to attend the 28th International Workshop on Statistical Modelling, Palermo, Italy from the Statistical Modelling Society, 2013.

- Macquarie University Post Graduate Research Fund (PGRF), 2013.

- Travel grant to attend Young Statisticians Conference, Melbourne from The Statistical Society of Australia Inc. - New South Wales Branch, 2013.

- Travel grant to attend the Winter School in Mathematical & Computational Biology, Brisbane from the Bioplatforms Australia and EMBL Australia, 2012.

- Travel grant to attend BioInfoSummer, Melbourne from the Australian Mathematical Sciences Institute, 2011.

Conferences and workshops attended

- The 59th World Statistics Congress (WSC), Hong-Kong, 2013.
- Workshop on introduction to UNIX & HPC, Macquarie University, Sydney, Australia, 2013.
- 10th Annual Australian Mathematical Sciences Institute (AMSI) Summer School, University of New South Wales, Sydney, Australia, 2012.
- Winter School in Mathematical & Computational Biology, The University of Queensland, Brisbane, Australia, 2012.
- BioInfoSummer, The Walter and Eliza Hall Institute (WEHI), Melbourne, Australia, 2011.

Acknowledgements

I wish to express my sincere and profound gratitude to the following persons who have generously contributed to the successful completion of this research study. Their support that encouraged and guided me to perfection was priceless and immeasurable.

In the first place, I would sincerely like to extend my special gratitude to my principal supervisor, Dr. Georgy Sofronov for taking me on as a PhD student. I greatly appreciate his continuous advice, guidance, support, inspiration and mentorship throughout my PhD life. I feel extremely grateful to work under his supervision and I am indebted to him more than he knows.

I sincerely thank my associate supervisor, Dr. David Bulger, for his support, encouragement over the years and especially for proof reading this thesis. I also owe a huge gratitude to Associate Professor Gillian Heller for introducing me to Georgy at the very early stages, when I was searching for a potential supervisor. I am very much glad that I found the right person to work with. I also gratefully acknowledge the support given by Professor Ian Marschner, Head of the Department of Statistics, Macquarie University, Professor Barry Quinn and all the academic/non-academic staff for extending their fullest assistance during my PhD. I would also like to acknowledge Macquarie University for providing me necessary financial support mainly through the International Macquarie University Research Excellence Scholarship (iMQRES) and through other generous research grants. I would like to thank Dr. Tatiana Polushina for her support in our collaborative work and Dr. Vincent Plagnol for providing the DNA read count data of celiac disease patients.

I would also like to thank my fellow PhD candidates, colleagues and students at Macquarie University for making it an enjoyable few years, especially my dear friend Kasun Rathnayake for all the support he has given to me over the years. These past few years have been less stressful and more enjoyable because of the companionship of my friends including Darshana, Anuradhi and the UWS crew.

My family has been a constant source of love, patience and strength all these years. I thank all my family members including my sister, brother and their families for their continuous support. My heartfelt gratitude goes out to my parents who were the guiding stars throughout my life; my mother who took pride and shed tears on every little achievement in my life, and my father who has been the unwavering pillar of strength throughout the years. Without their endless support and unconditional love I would not have achieved this much.

I have a wonderful extended family who has supported and encouraged me throughout this journey. I would like to express a special gratitude to my sister-in-law and her husband for all the support given during these three years and my mother-in-law and father-in-law for the continuous words of encouragement.

Last but not the least, it is my darling wife who has provided me an overwhelming love and caring with all the support and encouragement to finish this monumental work. These past several years have not been an easy ride in many ways. Thank you for being there for me in all seasons. I am truly lucky to have you in my life.

Contents

List of Figures

List of Tables

Abbreviations

aCGH	**A**rray - **C**omparative **G**enomic **H**ybridization
ARL	**A**verage **R**un **L**ength
ASP	**A**ssociated **S**tochastic **P**roblem
BIC	**B**ayesian **I**nformation **C**riterion
CE	**C**ross - **E**ntropy
CNV	**C**opy **N**umber **V**ariation
CUSUM	**Cu**mulative **Sum**
DNA	**D**eoxyribo **N**ucleic **A**cid
FWER	**F**amily **W**ise **E**rror **R**ate
GAMLSS	**G**eneralised **A**dditive **M**odels for **L**ocation **S**cale and **S**hape
GLM	**G**eneralised **L**inear **M**odels
K-L	**K**ullback - **L**eibler
MAD	**M**edian **A**bsolute **D**eviation
mBIC	**M**odified **B**ayesian **I**nformation **C**riterion
MCPP	**M**ultiple **C**hange **P**oint **P**roblem
NB	**N**egative **B**inomial
NGS	**N**ext **G**eneration **S**equencing
RMSE	**R**oot **M**ean **S**quared **E**rror
SC	**S**topping **C**riterion
SD	**S**tandard **D**eviation
SNR	**S**ignal - to - **N**oise **R**atio
SR	**S**hiryaev - **R**oberts
TSP	**T**raveling **S**alesman **P**roblem
ZINB	**Z**ero **I**nflated **N**egative **B**inomial

I dedicate this thesis to
my darling wife Kaushala and my parents...

Chapter 1

Introduction and Thesis Outline

1.1 Change-point problem

The change-point problem (or disorder problem) is commonly described as the problem of identifying changes of the parameter(s) of interest at unknown times and estimating their corresponding locations in stochastic processes. We utilize change-point models to detect these changes in order to provide an improved, detailed interpretation of the underlying properties of the process. The standard practice in statistical modelling is to consider homogeneity of the parameter(s) of interest across the entire data sequence [29]. However, in real life this assumption is frequently violated in many scenarios. This heterogeneity may be due to a change in mean, variance, skewness, distributional family or any other characteristic of the observed data. Thus, there exist change-points (or break-points, disorder points, structural breaks etc.) that divide an entire process into piece-wise homogeneous sections with respect to the parameter(s) of interest. These sections are commonly referred to as segments or regimes. The ultimate goal of change-point analysis is to "*correctly answer the question of whether the data obtained are generated by one or by many probabilistic mechanisms*" [29]. From the statistical point of view, a change-point is a location or time point such that the observations follow a particular distribution up to that point and follow a different distribution after that point. Therefore, one should carefully consider these change-points as well as their corresponding locations before developing a statistical model to carryout further analysis [29]. In the last few decades the change-point problem has received increasing attention for these reasons and has attracted a wide range of applications in many scientific streams. These models are employed in biological sequences analysis, financial and economic data analysis, EEG analysis, signal processing, surveillance and security system analysis, oceanographic studies, industrial quality control, etc.

In the literature, there exists a number of methodologies that attempt to solve this issue addressing different aspects or dimensions of the change-point problem. Broadly, there are two main classes of change-point problems. They are "retrospective (off-line or a posteriori)" and "sequential (on-line or prospective)". In the retrospective change-point problem the entire data set is considered to obtain inferences about the change-points. In the retrospective case, the process of obtaining inferences about the change-points can be considered as a static process. However, in the sequential change-point problem data get updated continually and the future observations are not known. Therefore, it follows a dynamic process in obtaining inferences about the change-points. The retrospective change-point problem tests for a single or multiple change-points in an observed data sequence, while the sequential change-point problem looks for the first time a change is detected. In general, both of these statistical diagnosis problems are concerned with the detection of change-points in stochastic processes and the estimation of their corresponding locations. Furthermore, they can be formulated as a model selection problem. In the literature, there are number of methodologies available to solve both of these change-point problems using frequentist as well as Bayesian approaches.

In this thesis, we are primarily interested in addressing the retrospective multiple change-point problem with a special interest in detecting abrupt changes in genomic sequences. The majority of the current detection methods are deterministic and use dynamic programming or different smoothing techniques to obtain the estimates of change-points. These approaches limits the search space of the problem due to different assumptions made in the methods and do not represent the true nature of the uncertainty associated with the unknown change-points in genomic sequences. The ultimate objective of this work is to develop and introduce more improved methodologies to accurately address the uncertainty associated with the change-points and their locations.

In genomic sequences, the change-point locations (loci) are the core estimates that facilitate to find disease-causing genes in many complex health problems. Thus, it helps to develop better diagnostic methods as well as to alleviate the disease, which is the ultimate goal. Therefore, the accuracy of change-point locations is considered as one of the most important factors when analyzing the effectiveness of a proposed procedure. In this thesis, we propose to utilize an evolutionary computing technique: the Cross-Entropy (CE) method, which is a model-based stochastic optimization procedure to estimate number of change-points and locations in genomic sequences more precisely.

Later, for the first time in the literature of change-point problems, we combine powerful sequential change-point detection techniques, such as the cumulative sum method and Shiryaev-Roberts procedure with the CE method to obtain estimates of both the number as well as the locations of change-points in biological sequences. Moreover, a multi-core parallel implementation is introduced to improve computational efficiency of the proposed procedures. The R package "breakpoint" is created based on some of the proposed methodologies in this thesis and it is freely available on R CRAN (`http://cran.r-project.org/web/packages/breakpoint/index.html`).

1.2 Review of segmentation methods

In the literature there exists a number of methodologies that can be utilized to detect change-points in observed data sequences. Broadly these methodologies can be applied to solve both the retrospective (off-line) and prospective (online) classes of change-point problems. First, we shall review some of the main techniques available in the literature and in sections 1.3 and 1.4 we provide a review of the two broader classes of change-point problems in general.

Recursive segmentation methods

In the recursive segmentation methods, the observed data sequence is segmented recursively into domains with a homogeneous composition [86]. The process of segmentation is repeated until no statistically significant improvement is achieved. Binary segmentation [136, 138], circular binary segmentation [100, 161] and wild binary segmentation [56] are some of the popular recursive segmentation methods in the literature. In these methods, based on a statistical test (non-parametric or model-based) the most significant change-point is identified at first and that location is kept in the memory. Then, the original sequence is partitioned into two independent segments based on the identified change-point and searched for any significant change-points. This process of splitting the sequence is carried out until no statistically significant change-point is found. These recursive segmentation approaches have obtained significant attention in the change-point analysis literature due to the simplicity in the implementation and the efficient computation time. However, numerous studies have raised questions about the behavior of these recursive methods when there are slight departures from the model assumptions [98]. In bioinformatics, the circular binary segmentation method [161] is used as one of the gold standards in detecting change-points in biological sequences of continuous measurements, such as micro-array data.

Sliding window analysis

Sliding window is a naive and a popular approach in determining the variability along data sequences [83, 99, 121, 155]. In terms of bioinformatics applications, it is being used to perform sequence alignment in biological sequences [155], to detect isochores [99], to investigate evidences for recombination [25] and to determine variation within a gene [121]. However, technically it lacks multiple attributes to be considered as an optimal segmentation approach. This is mainly due to the fact that, it averages out the observation with respect to a pre-determined size of sliding window [162]. The choice of window size is a crucial step in forming the analysis and the averaging effect is also questionable in some circumstances. For example if the window size is five, we obtain the first point by taking the average of the 1-5 observations, second point by taking the average of 2-6 observations , and so on until the end of the data sequence [155] . If the length of the observed data sequence is "n" and the window size is "W", then the length of the average points sequence is $n - W + 1$. However, the choice of an optimal window size has been an open research question for more than two decades [155].

Dynamic programming methods

Dynamic programming (DP) is a general optimization method which provides a framework to solve a complex problem with the use of multistage optimization approach. Simply, it transforms or splits the complex problem into a pool of simpler problems to obtain an optimal solution. The mathematical formulation of the DP was first introduced in [19] as a technical report. In bioinformatics there are many variants of the DP methods developed to perform protein folding, sequence alignment and comparison [1], gene recognition, copy number variation detection and structural predictions of the biological sequences. We refer to [59] for a gentle review of DP in bioinformatics. Many variants of DP have been proposed to detect multiple change-points in a given time series or sequence [10–12, 75, 76, 92] while improving many computational drawbacks in the standard DP implementation.

Hidden Markov models

Hidden Markov Models (HMM) are one of the popular techniques in temporal pattern recognition. The general idea was initially conveyed in [154] as a solution to a non-linear filtering problem. Later, with the work of [18] to establish its mathematical properties, it attracted a wide audience in different scientific streams to apply the HMM in practice and to further improve its capabilities. For instance, HMMs have been applied in speech recognition [123], fault detection [149] and computational biology [55, 80, 107]. In change-point analysis, HMMs are commonly used to obtain inferences about the change-points in an observed data sequence. In a change-point problem when forming the HMM the data are referred to as the observations and the unknown segments are defined as the hidden states. The hidden states affect the transition probabilities within each segment, whereas at each segment boundary a transition between states occurs. In bioinformatics, HMM was developed to segment biological sequences [107] with respect to the composition of the basic nucleotides. Later, [55] first proposed HMMs to detect copy number variations in biological sequences, followed by further extensions proposed in [14, 164, 165]. HMMs are popular mainly due to the fact that they produce statistically sound post inferences about the change-points. However, at the initial stages the ambiguity in the process of selecting prior probability distributions for the unknown states is considered as one of the drawbacks of the HMMs.

Penalized likelihood methods

The derivation of the likelihood concept by Fisher [53, 54] opened up a wide spectrum of applications in statistical modelling. In change-point analysis, the use of likelihood approaches is an essential and inevitable step in majority of the methods proposed in the literature. This is mainly due to the fact that change-point analysis can be considered as a model selection problem. The first application of the use of penalized likelihood methods in change-point analysis was proposed in the seminal paper by Yao [172]. In [172], the Bayesian information criterion (BIC; [135]) was used to estimate the number of change-points in an observed data sequence, which is independent and normally distributed. It was found in [172] that the BIC provides weak consistency when estimating the true number of change-points. Later, a quasi-likelihood based method with a modified Schwarz criterion was proposed in [26] that utilizes a data-driven approach to select the penalty parameter of the original BIC. Recently, multiple change-point detection methods developed on the least absolute shrinkage and selection operator (Lasso; [157, 158]) were proposed in [159]. However, it was found that the Lasso-type approaches in change-point analysis generally tend to over-estimate the true number of change-points [139]. Thus, to overcome this issue, an adaptive Lasso method was proposed recently with a post change-point analysis by using multivariate t simultaneous confidence intervals in [139].

Evolutionary algorithms

Evolutionary algorithms (EAs) are population-based metaheuristic optimization algorithms, which nests under the wider scope of evolutionary computation techniques [13]. The attention on EAs have significantly increased over the last decades mainly due to the exponential growth in technological advancements. The computational time of a process in a computer has significantly reduced over time due to these improvements. There are a wide variety of EAs currently being developed in many scientific streams. These EAs are mainly utilized to solve optimization-related search problems. Change-point analysis can also be considered as a mixture of estimation and optimization problems. On one hand the estimates of the change-points have to be obtained, and on the other hand, the solution has to be optimized over the range of the feasible solution set. Thus, it naturally contains the characteristics that can be solved by utilizing EAs. In the literature, there are multiple publications on applying different variants of the genetic algorithm[34, 42], which is a sub-class of EAs to solve the change-point estimation problem [70, 85, 113]. We refer to [103] for a detailed review on some of the evolutionary computation techniques utilized in bioinformatics. Later, the Cross-Entropy (CE) method [49] was considered to estimate change-point locations in binary data. However, their method was not developed to estimate the number of change-points.

In this thesis, we contribute to the existing literature on EAs by considering further extensions, modifications and applications of the the cross-entropy method, which is one of the EAs to estimate the unknown change-point locations in different biological sequences [115–117, 119, 120, 167].

1.3 Retrospective class of change-point problems

In the retrospective (or off-line, a posteriori) change-point problem, we observe the entire sample at once to obtain the estimates for both the number of change-points as well as their locations. The retrospective change-point problem was first introduced by Page

[101]. In his paper, some attributes of the single change-point problem were discussed. Over the years there have been significant methodological advancements in the branch of retrospective change-point problem. We refer [20, 36, 41, 65, 102, 144, 146] for more information about different methodologies as well as for detailed reviews of them.

We shall define the single change-point and multiple change-point problem in the domain of retrospective class of change-point problems. This is mainly due to the scope of this thesis, where in the context of analyzing biological sequences, we observed the entire data sequence before hand. Thus, the change-point analysis on the biological sequences can be primarily attributed to the retrospective class of change-point problems.

1.3.1 Single change-point problem

In the single change-point problem, we detect only one abrupt change in the parameter(s) of interest. The single change-point problem has been addressed by various authors using both frequentist and Bayesian approaches. The pioneering work of Shewart [140], Page [101, 102], Roberts [129], Shiryaev [141–143], Hinkley [65] and Lorden [87] initiated the discussion about different approaches that can be utilized to solve the single change-point problem with the use of sequential detection techniques. In the Bayesian context, Smith's [146] contributions are imperative for the evolution of advanced Bayesian techniques for the change-point problem.

The general formulation of the single change-point problem is as follows. Consider a sequence of observations $\mathbf{y} = (y_1, y_2, \ldots, y_L)$ of length L. Let y_i's, $i = 1, \ldots, L$ are independent random variables with probability distribution functions $G_1, G_2, \ldots, G_L$. Let c be defined as the unknown location of the change-point $(1 < c < L)$ in the observed data sequence. In general, the single change-point problem is to test the following null hypothesis,

$$H_0 : G_1 = G_2 = \ldots = G_L,$$

versus the alternative,

$$H_1 : G_1 = G_2 = \ldots = G_{c-1} \neq G_c = \ldots = G_L.$$

If the distributions $G_1, \ldots, G_L$ belong to a common parametric distribution family $G(\theta)$, then the change-point problem can be illustrated as a hypothesis test on population parameters $\theta_i, i = 1, \ldots, L$:

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_L = \theta \text{ (unknown)},$$

versus the alternative,

$$H_1 : \theta_1 = \ldots = \theta_{c-1} \neq \theta_c = \ldots = \theta_L.$$

Figure 1.1 is an example of a single change-point process. It illustrates a change in the mean of normally distributed random variables.

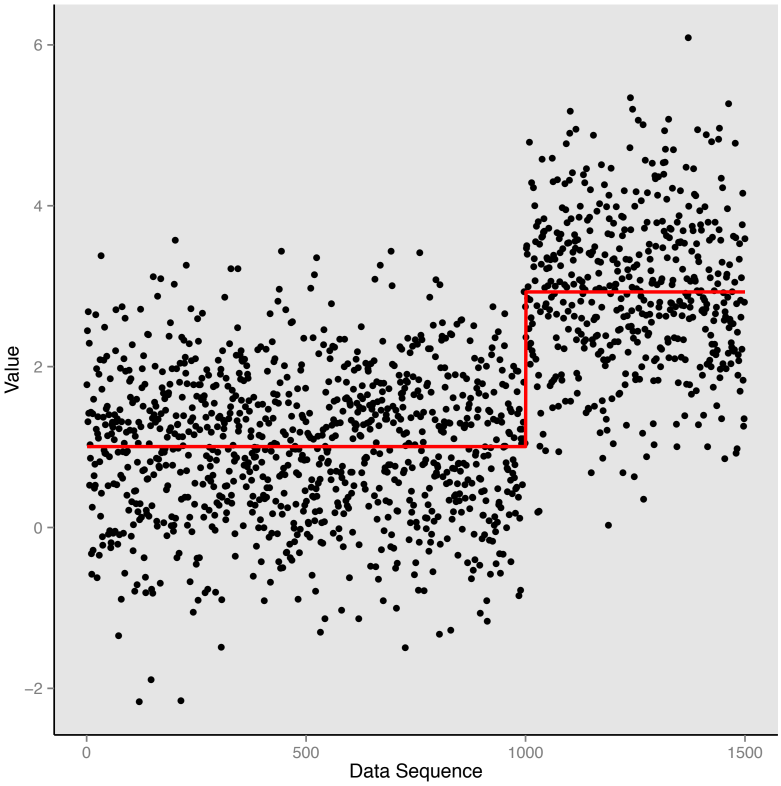


FIGURE 1.1: Single change-point problem.

1.3.2 Multiple change-point problem

In the multiple change-point problem (MCCP), we encounter more than one change-point in the process. It is the natural extension of the single change-point problem. The general formulation of the problem is as follows.

Let $\mathbf{y} = (y_1, y_2, \ldots, y_L)$ be a sequence of independent random variables with probability distribution functions $G_1, G_2, \ldots, G_L$, respectively. Let $c_1, c_2, \ldots, c_N$ be the unknown locations of N, the number of change-points, where $c_1 < c_2 < \ldots < c_N$. The sequence of observations are divided into $N+1$ segments based on the N change-points. In general, the change-point problem is to test the following hypothesis,

$$H_0 : G_1 = G_2 = \ldots = G_L,$$

versus the alternative,

$$H_1 : G_1 = \ldots = G_{c_1-1} \neq G_{c_1} = \ldots = G_{c_2-1} \neq G_{c_2} = \ldots = G_{c_N-1} \neq G_{c_N} \ldots = G_L.$$

Figure 1.2 is an example of a multiple change-point process. It illustrates multiple changes in the mean of normally distributed random variables.

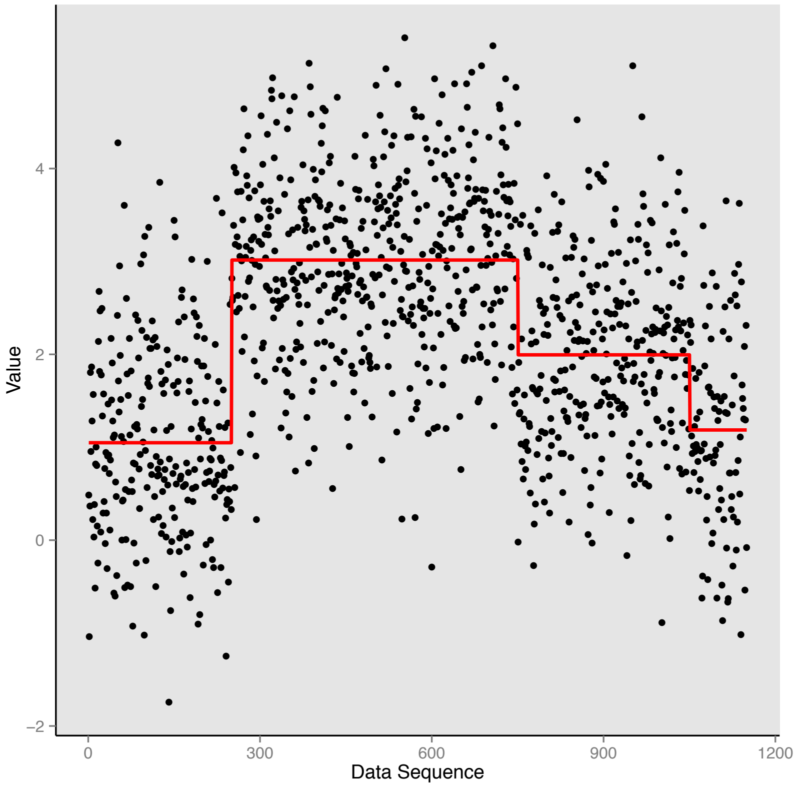


Figure 1.2: Multiple change-point problem.

1.4 Sequential class of change-point problems

In the sequential (or quickest, prospective) change-point problem, a sequence of random variables is observed on-line, that is, future observations are not known. The goal of these methods is to identify a change as soon as possible and to avoid false alarms. In general, the sequential change-point problem is considered as the origin of the change-point analysis, where much of the earlier work was initiated and carried out to solve quality control issues in industrial processes [65, 101, 102, 129, 140, 141]. To date there are two main procedures that have been extensively studied in this area: the Cumulative Sum (CUSUM) procedure [101] and the Shiryaev-Roberts (SR) procedure [129, 141–143]. We refer to [82] for a detailed review of sequential change-point methods.

The sequential problem can be described in mathematical terms as follows. Let $\{Y_n\}_{n\geq 1}$ be independent random variables which are observed sequentially, one by one. Suppose that initially the sequence is in so-called "controlled" state for $n = 1, 2, \ldots, \tau - 1$, that is, the random variables are distributed with $f_0(y)$, a common probability density function (pdf) with parameter values of $\boldsymbol{\theta}_0$. At some unknown moment "τ" a breakage occurs and the observed sequence runs "out of control", which means that after the breakage (change-point) the probabilistic characteristics of the sequence have changed. From moment "τ" we observe random variables with $f_1(y)$, $f_1(y) \neq f_0(y)$, another probability density function with a different set of parameter values $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_1$. Our objective is to detect the change-point as soon as it occurs with a minimum number of false alarms. In other words, in the sequential change-point problem, we would like to detect the moment "τ" as quickly as possible after it has occurred and, at the same time, we would like to keep the false alarm rate at a low pre-defined level.

1.5 Biological background

One of the major breakthroughs of modern science has been the elucidation of the chemical nature of the gene. The transmission of traits from parents to offspring depends on the transfer of a specific giant molecule that carries a coded blueprint in its molecular structure. This complex molecule, the basic component of the chromosome, is *deoxyribonucleic acid*, or in abbreviated form, DNA [166]. It is the heredity material or the information carrier in humans and in all living organisms. DNA consists of two long polymers of nucleotides. The information in DNA is stored as a code made up of four chemical bases known as Adenine (A), Guanine (G), Cytosine (C) and Thymine (T) (see

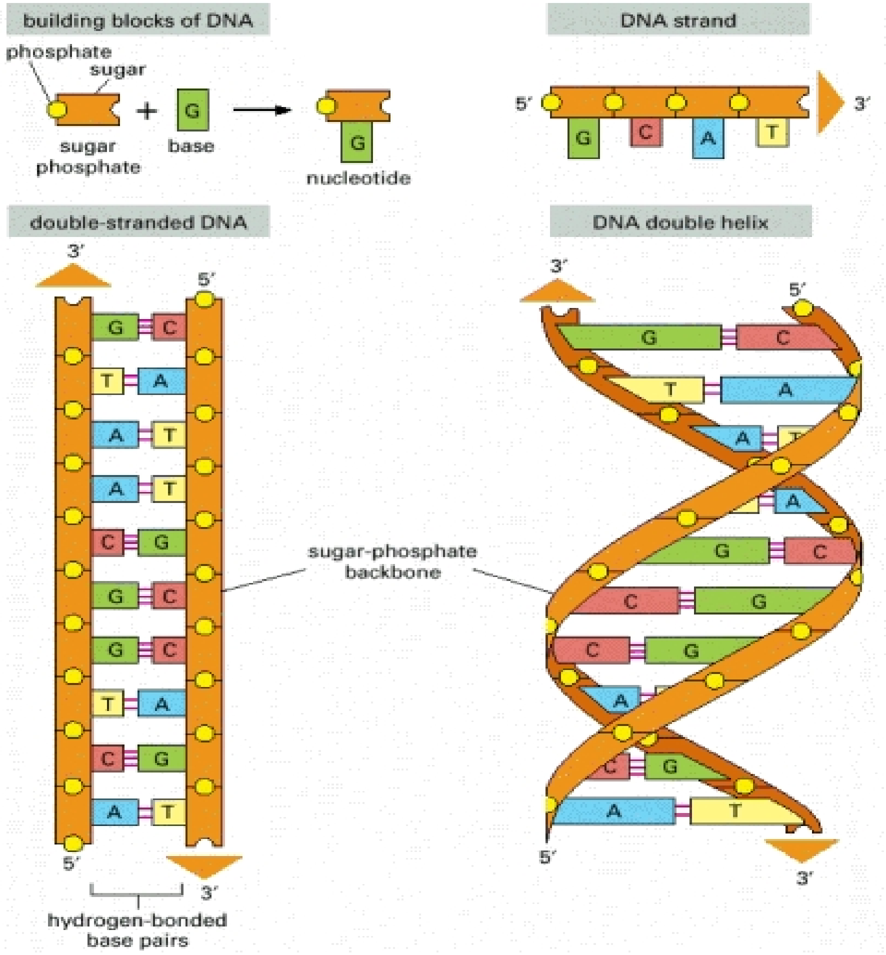


FIGURE 1.3: The structure of DNA.
Source:http://www.ncbi.nlm.nih.gov/books/NBK26821/

Figure 1.3). The order or the sequence of these chemical bases determines the information available for building and maintaining a living organism. The DNA molecules are packed into thread-like structures called chromosomes in the nucleus of each cell (see Figure 1.4). In almost every cell of a human being, there are 23 pairs of chromosomes, which makes 46 chromosomes in total. The sex chromosomes determine the gender (male if XY or female if XX) and all other chromosomes are known as autosomes. Figure 1.5 is a photograph of a person's chromosomes and it is called a karyotype.

The information carried in DNA molecule can be divided into a number of separable units, which we identify as the genes. More specifically, a gene is a particular segment of a DNA that codes for a specific protein [58, 106]. Bases in the gene determines the order in which amino acids are put together to make the protein. In this modern era a gene is defined as "a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions" [104, 106]. Thus, genes are the basic unit of inheritance that ferry a characteristic from parent to offspring. It is estimated that there are around 20,000 to 25,000 protein-coding genes in the human genome. These genes account for only about two percent of our DNA.

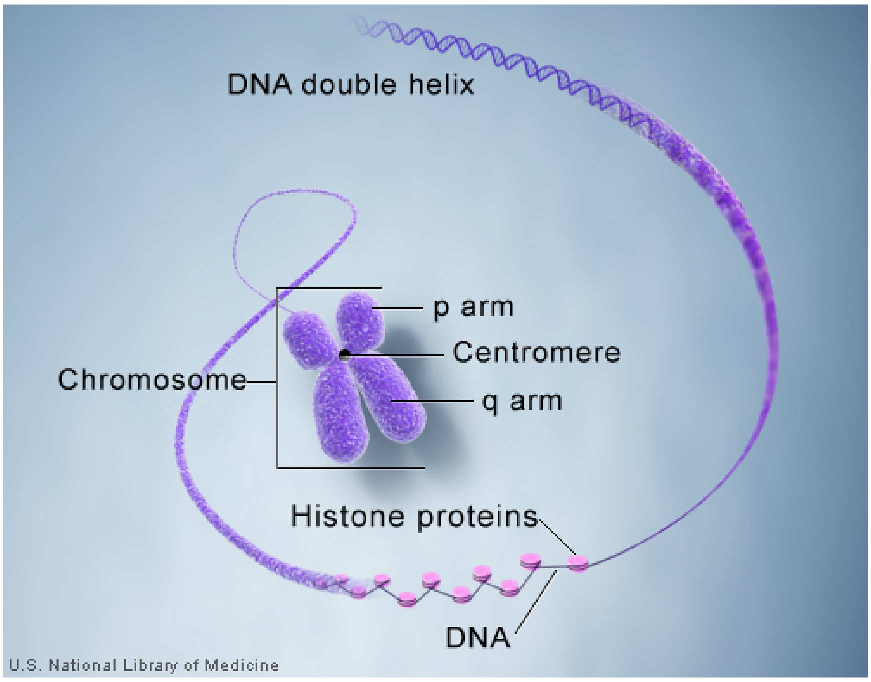


FIGURE 1.4: The composition of a chromosome.
Source:http://ghr.nlm.nih.gov/handbook/basics/chromosome

In health informatics, detection and characterization of genomic structural variations are essential in identifying disease-causing genes that have functional importance in causing genome-wide complex diseases, such as cancer, autism, immune disorders, etc. These complex genetic alterations in the human genome are key driving forces behind tumour development and progression. They also affect the degree of response to drugs, vaccines, pathogens and other forms of cure to genomic diseases. In the last few decades, there has been a significant improvement in sequencing technologies as well as data analyzing methodologies, which aim to detect these important variations. These structural alterations can be inherited through germline or be somatically acquired. Copy number variation (CNV) is a form of structural variation in the genome. CNV is defined as *"a segment of DNA that is one kilobase (kb) or larger and is present at a variable copy number in comparison with a reference genome"* [51]. Recent studies have shown that around twelve percent of the human genome varies in copy number [125] which includes important genetic information [32]. Furthermore, it is confirmed through multiple studies that CNV plays an important role in genetic susceptibility to common diseases [37] such as cancer [30, 40, 68, 108], autism [137], HIV [61], immune disorders [50], intellectual disabilities [78], etc. In the literature, there are multiple platforms and developed procedures to detect CNV [33, 69, 72, 88, 170] with the use of advanced technologies. In this thesis, I shall briefly discuss the array comparative genomic hybridization and some of the next generation sequencing techniques that are used in detecting CNVs.

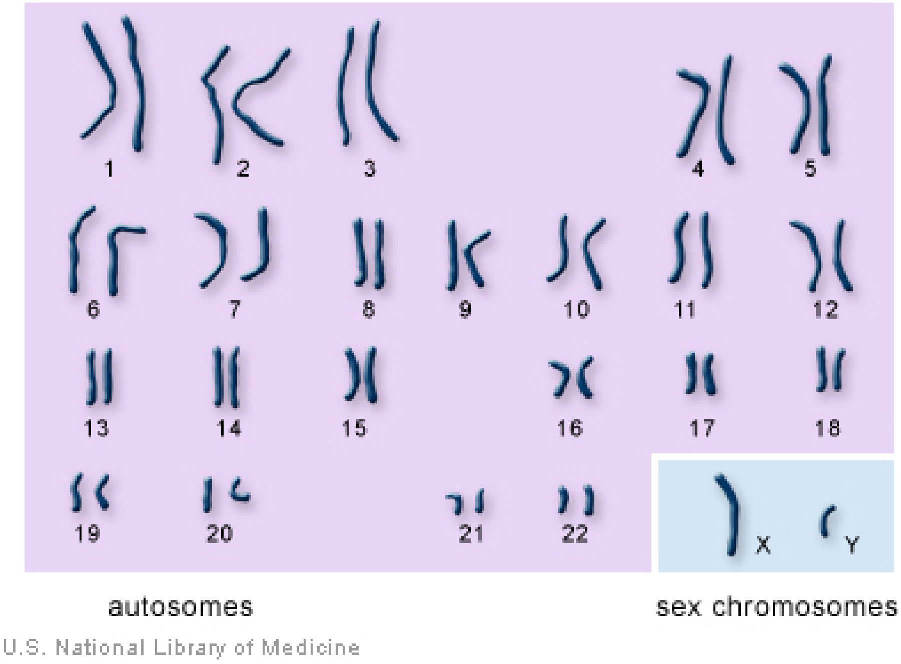


FIGURE 1.5: A karyotype of the the human chromosomes.
Source:http://ghr.nlm.nih.gov/handbook/basics/howmanychromosomes

1.5.1 Array-based comparative genomic hybridization

Array comparative genomic hybridization (aCGH) is one of the most popular techniques that is commonly used to detect and map copy number variation in DNA sequences. It is developed on the principles of the conventional comparative genomic hybridization (CGH) technique [72], which produces a map of DNA sequence copy number with respect to chromosomal locations. The CGH method was developed to detect small-scale chromosomal aberrations that are not visible through the microscope. It bridges the gap between techniques that look for large chromosomal variations (e.g. karyotyping) and those that concentrate on a specific region or section of DNA (targeted sequencing). In CGH experiments, the differentially-labelled test and control genomes are hybridized to metaphase chromosomes. The metaphase stage of a chromosome is defined as "*the process that separates duplicated genetic material carried in the nucleus of a parent cell*

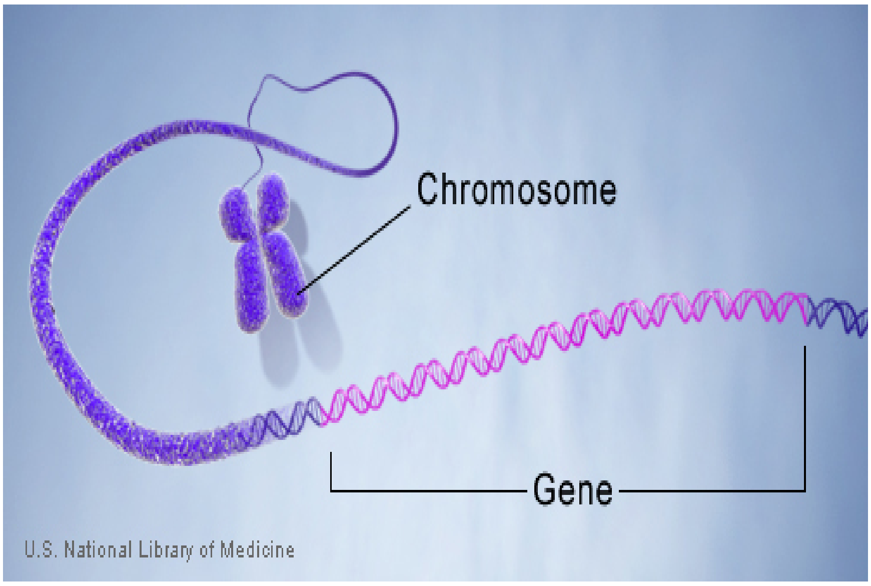


Figure 1.6: The composition of a gene.
Source:http://ghr.nlm.nih.gov/handbook/basics/gene

into two identical daughter cells". A chromosome is most condensed at this stage and it facilitates to identify the structure for further studies. Metaphase chromosomes are used in traditional and spectral karyotyping to identify large-scale genomic abnormalities. In the traditional karyotyping, researchers are able to view the full set of chromosomes in black and white. However, in the spectral karyotyping a multi-colored picture of chromosomes is obtained. In a standard CGH experiment, a tumor sample labeled red (Cy5) is hybridized to a reference normal sample labeled green (Cy3). The fluorescent signal intensity of the tumor DNA relative to the reference DNA along the chromosome is linearly plotted to identify CNVs [33, 72].

The aCGH technique uses slides arrayed with small segments of DNA as the targets for analysis [88] in contrast to the use of metaphase chromosomes in CGH. Figure 1.7 shows the complete process of aCGH analysis. The aCGH technique offers high resolution for CNV detection. Moreover, the ability to detect different types of alterations in a single process is one of the advantages of the CGH technique [156]. It has been also proven that aCGH is a powerful tool to detect submicroscopic chromosomal abnormalities in individuals with idiopathic mental retardation and various birth defects. Refer [33, 156] for a detailed review of DNA micro-arrays and CNV detection.

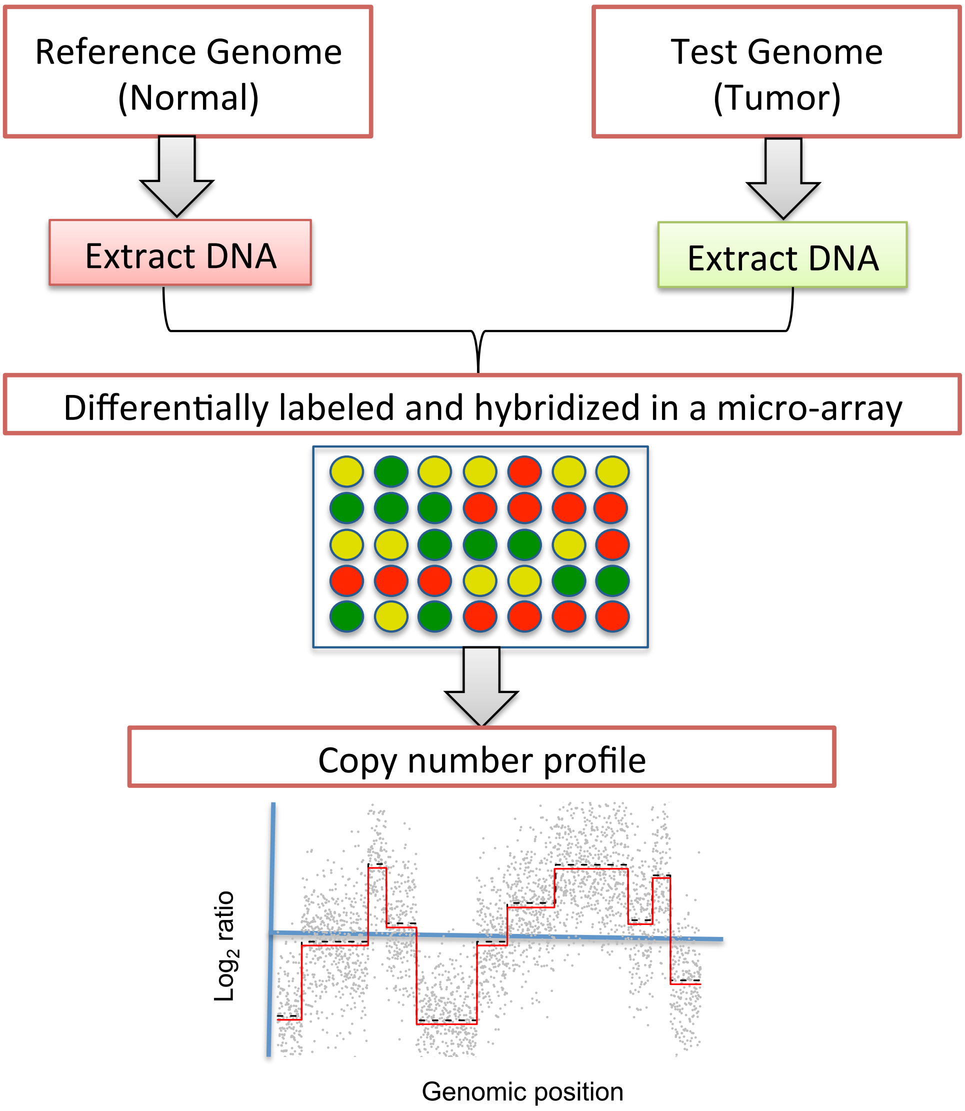


FIGURE 1.7: Diagram of the aCGH process

1.5.2 Next generation sequencing

The discovery of the use of dideoxy nucleotides for chain termination by Sanger et al. [133] marked a milestone in the history of DNA sequencing. This concept provided a basis for the development of automated Sanger sequencing [9, 148] which has been the method of choice for DNA sequencing for almost two decades.

In the last decade a revolution occurred in the field of DNA sequencing. This was due to the development of new high-throughput sequencing technologies, known as Next-Generation Sequencing (NGS) technologies, that are now widely adopted [62, 93]. These technologies are considered as the successors to the first-generation sequencing methods (e.g. aCGH). They are also known as ultra-high-throughput or massively parallel genomic sequencing technologies. These technologies have fundamentally changed the way in which we think about genetic and genomic research, opening new frontiers and new pathways that were not even thinkable or achievable with Sanger sequencing. They have provided the opportunity for a global investigation of multiple genomes and transcriptomes across the genome at single base resolution, in an extremely efficient and timely manner at much lower costs compared to the previous (e.g. Sanger sequencing) sequencing methods. One of the main advantages offered by NGS technologies is their ability to produce an incredible volume of data, cheaply, that in some cases exceeds one billion of short reads per instrument run. Applications that have already benefited from these technologies include: disease-targeted sequencing [126], polymorphism discovery [147], non-coding RNA discovery [94], large-scale chromatin immunoprecipitation [71], gene-expression profiling [15], clinical studies [171], mutation mapping and whole transcriptome analysis [63, 91]. Thus, the advent of the next generation sequencing (NGS) technologies have greatly increased the availability of data with a high level of sensitivity. As a result it has also opened up avenues to develop novel computational and statistical methodologies to effectively analyze these data.

1.6 Thesis outline

The remainder of this thesis is organized as follows. Chapter 2 provides an overview of the theory and methodologies utilized in this thesis. It contains detailed information about the Cross-Entropy (CE) method and its convergence properties, the multiple change-point problem and the sequential change-point problem. Furthermore, it gives an overview of the parallel computing techniques that can be carried-out in the R statistical computing environment both in the UNIX and WINDOWS operating systems. Chapters 3 and 4 contain two peer-reviewed publications that discuss the implementation of the CE method to detect multiple change-points in biological sequences of continuous measurements. Chapter 3 includes a peer-reviewed journal article that introduces the CE method to detect multiple change-points in aCGH data. Chapter 4 contains a peer-reviewed book chapter that introduces two hybrid algorithms combining powerful sequential detection techniques with the CE method to detect multiple change-points in continuous measurements. The work in chapter 4 can be considered as a significant improvement to the methodology proposed in chapter 3. Chapters 5 and 6 focus on the implementation of the CE method to detect multiple change-points in NGS data. Chapter 5 contains a peer-reviewed full conference paper that introduces the CE method to detect copy number variation in NGS read count data. Chapter 6 contains a peer-reviewed conference paper that utilizes Generalized additive models for location scale and shape (GAMLSS [127]) to detect copy number variation in NGS data with the use of auxiliary information. Chapter 7 describes the 'breakpoint' R package [114] that is developed on some of the methodologies discussed in the thesis to detect multiple change-points in continuous and count data. Finally, chapter 8 concludes the thesis providing a summary and a discussion of the overall findings together with future research directions.

Chapter 2

Methods

In this section of the thesis, we provide detailed information on the methodologies utilized in the subsequent chapters corresponding to multiple publications. Those methods discussed briefly in the papers, are elaborated and generalized to obtain an overview of the scope of the thesis.

2.1 The General Cross-Entropy Method

The Cross-Entropy (CE) method [44, 130–132] is a general model-based stochastic optimization technique developed on one of the fundamental principles of modern information theory called Kullback-Leibler [81] information (or cross-entropy). The CE method was first introduced by Reuven Y. Rubinstein in 1997 [132] as an adaptive importance sampler for estimating probabilities of rare events. Later, with further developments [130, 131], it was extended to solve complex combinatorial, continuous and multi-extremal optimization problems [79, 130]. The CE method has been successfully applied in solving a wide range of traditionally hard test problems including the maximal-cut problem, the traveling salesman problem (TSP) and the quadratic assignment problem [131]. Additionally, the CE method has been applied to buffer allocation problems [4], queuing models for telecommunication systems [43], traffic assignment problems [89], probability density estimation [22], medical image segmentation [169], prototype-based learning [24], DNA sequence alignment [73], binary data segmentation [49], vehicle routing [35] etc.

In general, the CE method is summarized by a three-step iterative procedure:

- **Step 1:** Simulate candidate solutions for the problem based on a statistical distribution.
- **Step 2:** Calculate the performance function score relative to the problem that needs to be solved.
- **Step 3:** Update the parameters of the statistical distribution (in step 1) in order to obtain an improved solution set in subsequent iterations by minimizing the cross-entropy.

The CE method carries out all its computations in the optimization steps based on a statistical distribution (parametrized probability distribution) and it is similar to the estimation of distribution algorithms [57]. The probabilistic approach in the CE algorithm transfers information from the current best solutions to the next iteration in order to increase the probability that similar solutions appear in subsequent iterations. This model-based approach is a key feature of the CE method as compared to the other competing stochastic optimization methodologies, such as simulated annealing [77], tabu search [60] and genetic algorithms [67], which are operating directly on a population of candidate solutions [38].

2.1.1 The Kullback-Leibler Information

The Kullback-Leibler (K-L) information [81] is a fundamental concept in information theory. It measures the dissimilarity or directed distance between two probability distributions. It is also known as K-L discrepancy, distance, CE divergence or K-L number. Many statistical procedures including the likelihood and penalized likelihood methods directly or indirectly utilize the K-L information to make inferences. The CE method is developed on the principles of the K-L information.

Let F and G be two probability distributions with probability distribution functions f and g that are defined on the same sample space $\mathbf{Y}$ of length L. Then the expectation

$$I_{KL}(\mathrm{F},\mathrm{G}) = \mathbb{E}_f \ln\left(\frac{f(Y)}{g(Y)}\right)$$

is the K-L information of G with regard to F, $I_{KL}(\mathrm{F},\mathrm{G})$ is the information lost when G is used to approximate F. We can obtain the formulas for discrete and continuous probability distributions as follows.

$$I_{KL}(\mathrm{F},\mathrm{G}) = \begin{cases} \sum_{i=1}^{L} f(y_i)\ln\left(\dfrac{f(y_i)}{g(y_i)}\right) & \text{if } F \text{ and } G \text{ are discrete distributions, and} \\ \displaystyle\int_{-\infty}^{\infty} f(y)\ln\left(\dfrac{f(y)}{g(y)}\right)dy & \text{if } F \text{ and } G \text{ are continuous distributions.} \end{cases}$$

Properties of the K-L information:

1. $I_{KL}(\mathrm{F},\mathrm{G}) \geq 0$ (i.e. it is always non-negative);
2. $I_{KL}(\mathrm{F},\mathrm{G}) = 0$ only if $f(y) = g(y)$ (i.e. the two distributions are identical);
3. $I_{KL}(\mathrm{F},\mathrm{G}) \neq I_{KL}(\mathrm{G},\mathrm{F})$ in general (i.e. it is not symmetric);
4. I_{KL} is invariant under transformations of the sample space.

2.1.2 The General CE algorithm for optimization

Consider the global optimization of a real-valued deterministic performance function $F(\mathbf{Y})$, which is evaluated by a candidate solution vector of $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_M)$ defined on the feasible solution space of $\mathcal{Y}$. The objective of this optimization problem is to find the maximum of F over $\mathcal{Y}$ and the corresponding maximizer $\mathbf{Y}^*$. We can define this problem as

$$F(\mathbf{Y}^*) = \gamma^* = \underset{\mathbf{Y}\,\in\,\mathcal{Y}}{\arg\max}\ F(\mathbf{Y}),$$

where the γ^* is the global maximum of $F(\mathbf{Y})$ and $\mathbf{Y}^*$ is the corresponding optimal solution. In the CE method, the deterministic optimization problem is translated into an associated stochastic problem (ASP) and then rare event simulation techniques discussed in [131, 132] are used to solve the original problem. Let the ASP be defined [131] as follows:

$$\mathbb{P}_{\boldsymbol{\nu}}(F(\mathbf{Y}) \geq \gamma) = \mathbb{E}_{\boldsymbol{\nu}} I_{\{F(\mathbf{Y})\geq\gamma\}}, \tag{2.1}$$

where $\mathbf{Y}$ is drawn from a parametric distribution with probability distribution function $f(\cdot;\boldsymbol{\nu})$ and $\boldsymbol{\nu}$ is the parameter vector of its that controls the sampling of the candidate solutions. $\mathbb{P}_{\boldsymbol{\nu}}$ is the probability measure associated with $f(\cdot;\boldsymbol{\nu})$. The left-hand-side of equation 2.1 is the probability under $f(\cdot;\boldsymbol{\nu})$ that $\{F(\mathbf{Y}) \geq \gamma\}$. $\mathbb{E}_{\boldsymbol{\nu}}$ is the expectation operator and $I_{\{\cdot\}}$ is an indicator function. Equation 2.1 can be estimated using a log-likelihood estimator with parameter $\boldsymbol{\nu}$,

$$\hat{\boldsymbol{\nu}}^* = \arg\max_{\boldsymbol{\nu}} \frac{1}{M_{elite}} \sum_{i=1}^{M_{elite}} I_{\{F(\mathbf{Y}_i)\geq\gamma\}} \ln f(\mathbf{Y}_i;\boldsymbol{\nu}),$$

where $M_{elite} \leq M$ is the size of the candidate solutions simulated from the statistical distribution $f(\cdot;\boldsymbol{\nu})$ where $F(\mathbf{Y}_i) \geq \gamma$.

The general CE algorithm can be described as follows.

Algorithm 1 The general CE algorithm.

1. Parameter initialization: Set initial parameters of the statistical distribution as $\hat{\boldsymbol{\nu}}^{(0)}$. Set iteration counter $t = 1$.

2. Update $\hat{\gamma}^{(t)}$: Generate M samples $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_M$ from $f(\cdot;\hat{\boldsymbol{\nu}}_{t-1})$ and calculate the performances $F(\mathbf{Y}_1), F(\mathbf{Y}_2), \ldots, F(\mathbf{Y}_M)$ for the candidate solutions. These values are then sorted in increasing order, $F_{(1)} \leq \ldots \leq F_{(M)}$. Let $\gamma^{(t)}$ be the $(1-\rho)$ quantile of $F(\mathbf{Y})$ satisfying

$$\mathbb{P}_{\boldsymbol{\nu}^{(t-1)}}(F(\mathbf{Y}) \geq \gamma^{(t)}) \geq \rho,$$
$$\mathbb{P}_{\boldsymbol{\nu}^{(t-1)}}(F(\mathbf{Y}) \leq \gamma^{(t)}) \geq 1-\rho.$$

   The estimate of $\hat{\gamma}^{(t)}$ is computed as

$$\hat{\gamma}^{(t)} = F_{(\lceil(1-\rho)M\rceil)}.$$

3. Update $\hat{\boldsymbol{\nu}}^{(t)}$: Given $\hat{\boldsymbol{\nu}}^{(t-1)}$ we obtain estimates for the $\hat{\boldsymbol{\nu}}^{(t)}$ by utilizing the CE (or KL divergence) technique. The optimization problem can be obtained as

$$\hat{\boldsymbol{\nu}} = \arg\max_{\boldsymbol{\nu}} \mathbb{E}_{\hat{\boldsymbol{\nu}}^{(t-1)}} I_{\{F(\mathbf{Y}_i)\geq\hat{\gamma}^t\}} \ln f(\mathbf{Y}_i;\boldsymbol{\nu}).$$

   We obtain the following optimization problem if we utilize the best performing fraction of the samples (M_{elite}) to update the parameters of the statistical distribution.

$$\hat{\boldsymbol{\nu}} = \arg\max_{\boldsymbol{\nu}} \frac{1}{M_{elite}} \sum_{i=1}^{M_{elite}} I_{\{F(\mathbf{Y}_i)\geq\gamma\}} \ln f(\mathbf{Y}_i;\boldsymbol{\nu}). \tag{2.2}$$

4. Smooth update of $\hat{\boldsymbol{\nu}}^{(t)}$ (Optional):

   The smooth updating of the parameters [131] can be utilized to decrease the probability of the CE procedure converging too quickly to a sub-optimal solution. Thus, it will prevent the CE algorithm from being trapped at local optimums. Let the constant smoothing coefficient be defined as a $(0 \leq a \leq 1)$. The smoothed update of $\hat{\boldsymbol{\nu}}^{(t)}$ can be obtained as:

   $$\hat{\boldsymbol{\nu}}^{(t)} = a\tilde{\boldsymbol{\nu}}^{(t)} + (1-a)\hat{\boldsymbol{\nu}}^{(t-1)}.$$

   where $\tilde{\boldsymbol{\nu}}^{(t)}$ is the parameter estimates obtain in the current iteration by utilizing 2.2. If $a = 1$ the update will not be smoothed.

5. If the stopping criterion (SC, see 2.1.7) is met, then stop the process and identify the solution that maximizes the performance function. Otherwise, set $t = t + 1$ and repeat steps 2 to 4 until the SC is satisfied.

Detailed information about the convergence properties of the CE algorithm discussed in this thesis can be found in [90] and [38]. Convergence properties of two versions of the CE method was discussed in [90]. They have named the first version Graphed-Based CE Algorithm/Conservative Modification (GBCE/CM), which uses a smoothed parameter update scheme. The other version was named Graphed-Based CE Algorithm/-Conservative Modification with Lower Bound (GBCE/CMLB), which has a normalized probability updating rule. The procedures followed in [64] to prove the convergence of two ant colony optimization algorithms were directly adopted in [90] to form theoretical convergence results of the CE methods. Later, in [38], convergence properties of the CE method for discrete optimization was discussed in detail. They have proved several convergence properties of the CE method that uses an "elite sample" to obtain parameter estimates with smoothed updating schemes. The work presented in this thesis falls under the branch of discrete optimization. Thus, the convergence properties of the CE method discussed in [38] is directly related.

2.1.3 The CE method for retrospective MCPP

The change-point problem can be described as a combinatorial optimization problem. The CE method is one of the best evolutionary computing techniques that utilizes a stochastic framework to solve both estimation and optimization problems. Thus, it makes an ideal methodology to address the change-point problem which has a greater amount of uncertainty in the parameters of interest. In real world applications, the number of change-points and their corresponding locations are not known in advance. Our goal in this thesis is to propose effective and efficient procedures to estimate both the number and the locations of change-points in genomic data sequences by using the CE method. Identifying critical change-points in genomic sequences is a crucial step in understanding the genetic mechanism of complex diseases. The discovery of these change-points in genomic sequences helps researchers and practitioners to further improve the treatment procedures, drug development and assess disease progression.

In the context of MCPP, the CE method is used to solve the complex discrete optimization problem in this thesis. Similar combinatorial optimization problems including the maximal cut problem, optimal buffer allocation, DNA sequence alignment and the travelling salesman problem have been successfully solved by utilizing the CE method [4, 49, 73, 131, 169]. The general structure of the algorithm is discussed in this section. More specific implementations of the CE method along with relevant applications are discussed in chapters 3, 4, 5 and 6.

Recall the change-point problem from Chapter 1, with the number of change-points denoted by N, the length of the observed data sequence by L, change-point location vector by $\mathbf{C}$ and the elite sample fraction by ρ.

The CE algorithm for estimating the locations of change-points (given the value of N):

Algorithm 2 The CE algorithm for estimating the locations of change-points.

```
1. Given the value of $N$, set the initial parameters of the statistical distribution $f(\cdot;\boldsymbol{\nu}^0)$ that simulates the change-point locations. The parameter vectors are $N$ dimensional. Set the iteration counter $t = 1$.

2. Generate $M$ random samples $\mathbf{C}^{(1)}, \mathbf{C}^{(2)}, \ldots, \mathbf{C}^{(M)}$ from the statistical distribution with parameters $\boldsymbol{\nu}^{t-1}$, where $\mathbf{C}^{(i)} = (c_1^{(i)}, c_2^{(i)}, \ldots, c_N^{(i)})$, $i = 1, 2, \ldots, M$.

3. For each $i = 1, 2, \ldots, M$ sort the simulated change-point locations in ascending order and calculate the performance function score $F$ for each candidate solution. Sort the $F$ score in descending order if the model selection criterion is defined as finding a maximum, otherwise sort it in ascending order.
```

4. Let the size of the best performing fraction of the samples be defined as $M_{elite} = \rho \times M$. Based on the size of the M_{elite} sample, we obtain the upper quantile of the candidate solutions with respect to the performance function score. This elite sample is used to calculate and update the parameters ($\boldsymbol{\nu}^t$) of the statistical distribution considered in the step 1. The parameters of the statistical distribution can be updated either using the smoothing technique discussed in Section 2.1.2 or without any smoothing.
5. If the process has met a stopping criterion, we stop the algorithm and obtain the current solution as estimates for the locations of the change-points.

The above algorithm 2 defines the general work-flow of the CE method in estimating the locations of the change-points with respect to the number of change-points (N). The flow chart in Figure 2.1 further illustrates the procedure. In order to estimate the number of change-points, we propose to utilize one or more model selection criteria. In this thesis, we have explored the feasibility of applying the Bayesian information criterion [135] and the modified Bayesian information criterion [174] to estimate the number of change-points.

An overview of the overall algorithm to find both the number of change-points and their corresponding locations is as follows:

Algorithm 3 Overall algorithm to estimate both the number and their locations of change-points.

1. Define the search space for the number of change-points, i.e., define the minimum (N_{min}) (default is set at 0) and the maximum number of change-points (N_{max}).
2. For each value of N from N_{min} to N_{max}, carry-out the CE algorithm 2.1. Obtain the optimal solution of change-point locations and its performance function score (i.e., Information Criterion (IC) score).
3. Plot the IC score vs. N and find the value of N that minimizes (or maximizes) the IC. We denote this N value as $\tilde{N}$ and it is used as the estimate for the number of change-points.
4. Finally, the value of $\tilde{N}$ and the corresponding locations are given as the optimal estimates for the number of change-points and their locations.

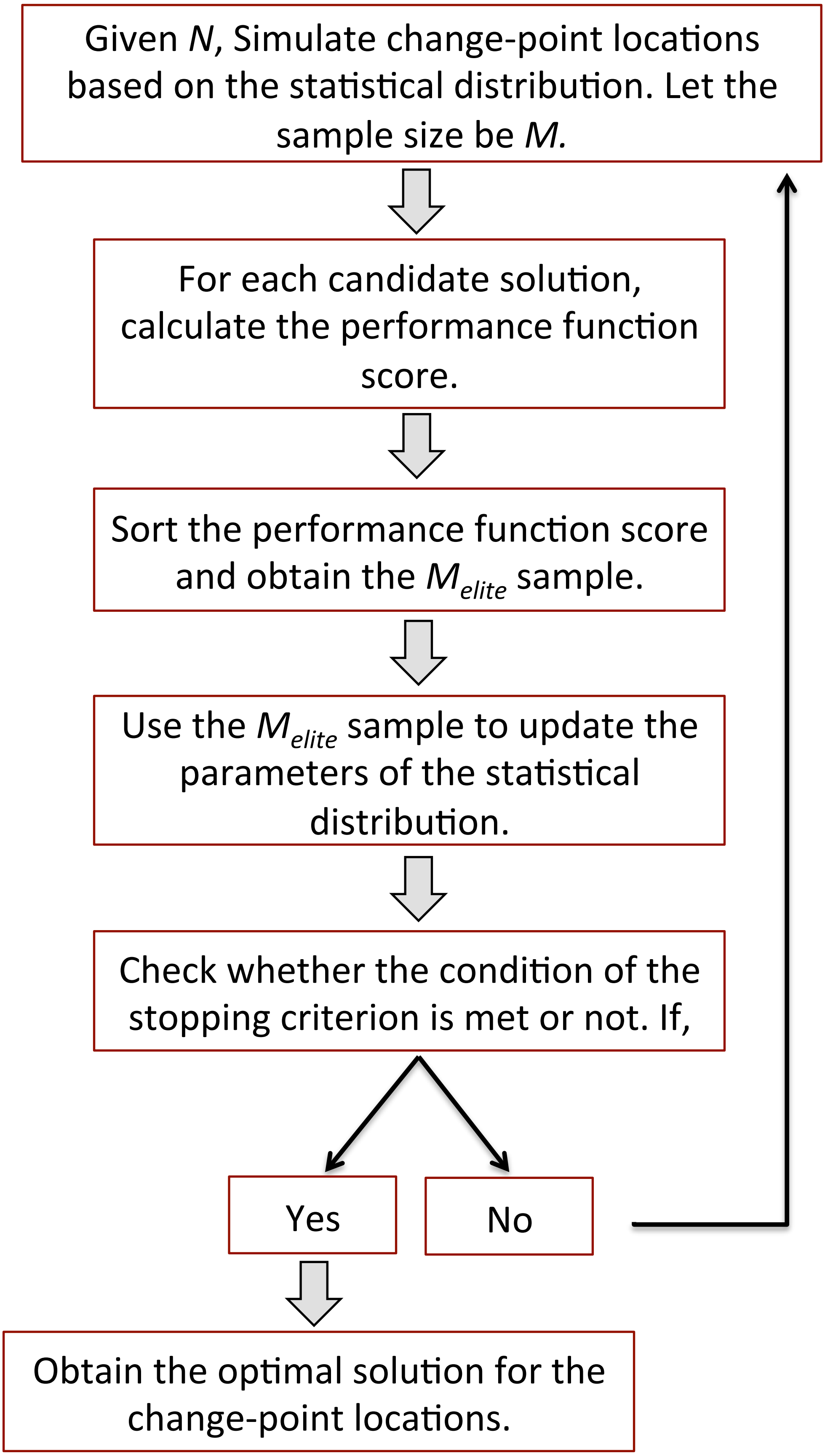

Given N, Simulate change-point locations based on the statistical distribution. Let the sample size be M.
For each candidate solution, calculate the performance function score.
Sort the performance function score and obtain the M_{elite} sample.
Use the M_{elite} sample to update the parameters of the statistical distribution.
Check whether the condition of the stopping criterion is met or not. If,
Yes
No
Obtain the optimal solution for the change-point locations.

FIGURE 2.1: Flow chart of the CE algorithm in MCPP.

2.1.4 External and Internal parameters of the CE algorithm

We classify the parameters of the CE algorithm into two broader categories based on their impact on the overall performance. The classification of the parameters not only simplify the overall nature of the algorithm, it also provides a better overview to the end-user to appropriately alter the parameter values with a proper understanding of its overall impact to the algorithm.

2.1.4.1 External parameters

We define the "external" parameters as the non-core parameters that have marginal impact to the overall performance of the CE algorithm. They are the minimum and maximum number of change-points (N_{min} and N_{max}) and the segment width (h). Segment width (aberration width) h is defined as the minimum allowed distance between two adjacent change-points.

2.1.4.2 Internal parameters

The "internal" parameters are defined as the core (or crucial) parameters that directly affect the performance of the CE algorithm. They are the simulated sample size (M), elite sample fraction (ρ), cut-off value for the stopping criterion (ε, see page 49) and smoothing coefficient (a).

2.1.5 Statistical distributions to simulate change-point locations

One of the major steps in the CE algorithm is to select an appropriate statistical distribution to simulate change-point locations. In this thesis, we have proposed two statistical distributions to simulate change-point locations in the CE algorithm. They are the four-parameter beta distribution and the truncated Gaussian distribution. They both can be utilized in analyzing aCGH data as well as the NGS data.

2.1.5.1 Four-parameter beta distribution

The four-parameter beta distribution is the generalization of the standard beta distribution which is defined on the interval $[0, 1]$. It is also known as the generalized beta distribution [2]. In the change-point problem, locations of the change-point estimates may vary along the length of the observed data sequence. Thus, the support of the standard beta distribution needs to be expanded to equal the length of the sequence. The beta distribution is characterized by two positive shape parameters α and β. The probability density function of the four-parameter beta distribution defined on the interval $[L_L, L_U]$ with shape parameters α and β is given by

$$f(y \mid \alpha, \beta, L_L, L_U) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \dfrac{(y-L_L)^{\alpha-1}(L_U-y)^{\beta-1}}{(L_U-L_L)^{\alpha+\beta-1}} & \text{if } L_L \le y \le L_U, \\ 0 & \text{otherwise,} \end{cases}$$

where L_L is the lower limit and L_U is the upper limit satisfying $L_L < L_U < \infty$, two shape parameters $\alpha, \beta > 0$ and the gamma function $\Gamma(\cdot)$ is defined as

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \, dt \quad \text{for all } z > 0.$$

The cumulative distribution function (cdf) of the four-parameter beta distribution is

$$F(y \mid \alpha, \beta, L_L, L_U) = \begin{cases} 0 & \text{if } y < L_L, \\ \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \displaystyle\int_{L_L}^{y} \dfrac{(t-L_L)^{\alpha-1}(L_U-t)^{\beta-1}}{(L_U-L_L)^{\alpha+\beta-1}} \, dt & \text{if } L_L \le y \le L_U, \\ 1 & y > L_U. \end{cases}$$

In this thesis, we have considered utilizing the method of moments (MoM) technique [39, 105], rather than the maximum likelihood estimation [31, 52] to obtain parameter estimates for the four-parameter beta distribution [2]. This is mainly because there are no closed form solutions for the parameters of the four-parameter beta distribution under the maximum likelihood approach, thus it requires more computational resources to generate estimates. In contrast to that, the MoM approach is an efficient estimation procedure that gives reasonably accurate estimates for the parameters of interest. We can obtain mean (μ) and variance (σ^2) of the four parameter beta distribution as below [2].

$$\mu = \mathbb{E}(Y) = L_L + (L_U - L_L)\frac{\alpha}{\alpha+\beta}, \tag{2.3}$$

$$\sigma^2 = \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 = (L_U - L_L)^2 \cdot \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{2.4}$$

Method of Moments

The method of moments (MoM) is a population parameter estimation technique that depends on the law of large numbers [39, 105]. The MoM technique is a three-step procedure.

1. Derive equations of the theoretical moments with respect to the parameters of interest.
2. Compute corresponding sample moments based on the observed data.
3. Equate the sample moments to the theoretical moments to obtain a system of equations and solve it to obtain the corresponding MoM estimates for the population parameters.

In the MCPP, the lower and upper limits of the four-parameter beta distribution is known. They are the starting and end location points of the observed data sequence. We can obtain the MoM estimates for the two-shape parameters by equating the first two theoretical moments (2.3 and 2.4) to the corresponding sample moments [2] as:

$$\bar{y} = \frac{1}{L}\sum_{i=1}^{L} y_i = L_L + (L_U - L_L)\frac{\hat{\alpha}_{MoM}}{\hat{\alpha}_{MoM} + \hat{\beta}_{MoM}},$$

$$s^2 = \frac{1}{L-1}\sum_{i=1}^{L}(y_i - \bar{y})^2 = (L_U - L_L)^2 \cdot \frac{\hat{\alpha}_{MoM}\hat{\beta}_{MoM}}{(\hat{\alpha}_{MoM} + \hat{\beta}_{MoM})^2(\hat{\alpha}_{MoM} + \hat{\beta}_{MoM} + 1)},$$

where $\bar{y}$ and s^2 are the unbiased estimates of the sample mean and sample variance of the observed data sequence. We obtain the MoM estimates for α and β as:

$$\hat{\alpha}_{MoM} = \left(\frac{\bar{y} - L_L}{L_U - L_L}\right)\left[\frac{(\bar{y} - L_L)(L_U - \bar{y})}{s^2} - 1\right],$$

$$\hat{\beta}_{Mom} = \left(\frac{L_U - \bar{y}}{L_U - L_L}\right)\left[\frac{(\bar{y} - L_L)(L_U - \bar{y})}{s^2} - 1\right].$$

2.1.5.2 Truncated Gaussian distribution

The truncated Gaussian (or normal) distribution [128] is a variant of the general Gaussian distribution, which is defined on the range of $(-\infty, +\infty)$. Generally, there are three versions of the truncated normal distributions that are defined on specific boundary conditions. The three versions are defined with a support of $[LB, +\infty)$, $(-\infty, UB]$ and $[LB, UB]$ respectively. The third version of the truncated Gaussian distribution with bounded support in both directions (a.k.a. two-sided truncated normal distribution [128]) is the most appropriate distribution to simulate locations of change-points in the CE method under the MCPP. The two-sided truncated normal distribution was considered in [49] to simulate change-point locations in the context of segmenting binary data.

Let us define the two-sided truncated normal distribution as simply the normalized restriction of normal distribution $N(\mu, \sigma^2)$ on a bounded interval $[L_L, L_U]$, $-\infty < L_L < L_U < +\infty$. The probability density function of the two-sided truncated normal distribution is given by:

$$f(y \mid \mu, \sigma^2, L_L, L_U) = \begin{cases} \dfrac{\dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(\dfrac{-(y-\mu)^2}{2\sigma^2}\Big)}{\Phi\Big(\dfrac{L_U - \mu}{\sigma}\Big) - \Phi\Big(\dfrac{L_L - \mu}{\sigma}\Big)} & \text{if } L_L \leq y \leq L_U, \\ 0 & \text{otherwise,} \end{cases}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. The maximum likelihood method is used to obtain parameter estimates for μ and σ^2.

2.1.6 Model selection

In the MCPP, the goal is to obtain estimates for both the number of change-points and their locations. A performance function (F) is used in the CE method to evaluate the fitness of each candidate solution. In [49] the log-likelihood function is considered as the F to assess the performance of solutions in the CE procedure. The aim of their paper was only to obtain estimates of the change-point locations and they did not consider estimating the number of change-points. To fill this gap, in this thesis, we propose using different model selection criteria not only to obtain estimates of the change-point locations, but also to estimate the number of change-points.

In the literature, there are a few popular model selection criteria that have been proposed in different contexts. The Akaike Information Criterion (AIC) [3] and the Bayesian Information Criterion (BIC) [135] are the most widely known and applied model selection procedures. All these model selection methods can be categorized as penalized likelihood methods. They choose a model that maximizes a criterion of the form

$$ll_{model}(\hat{\theta}) - p(\hat{\theta}),$$

where ll_{model} is the model log-likelihood, $\hat{\theta}$ is the maximum likelihood parameter estimates and p is the penalization (penalty) function. We can further classify these penalized likelihood approaches into two broader categories based on the way in which the penalty function is defined. The first category can be defined as methods in which the penalty function is precisely formulated [3, 135, 174]. The other category contains methods where the values of the penalty function parameters are obtained by direct simulation [26], data driven methods [84], or cross-validation [21, 157, 158]. In this thesis, we only consider to use several methods under the first category which has an explicit formulation of the penalty function. Particularly, we use the general BIC [135] (see 5.6) and the modified Bayesian Information Criterion [174] (see 3.1) for model selection when analyzing biological sequences with the CE method.

2.1.7 Stopping Criteria

In the CE algorithm one has to define a stopping criterion to terminate the iterative process. Several methodologies are discussed in [49, 73, 79, 131]. We will elaborate the three methods discussed in [131] in the context of change-point problem. They can be described as follows.

1. Selecting a maximum number of iterations. Let us define that number as T. Thus, at time $t = T$, the process is stopped and the corresponding solution is considered the optimal solution for the problem.

2. Require a minimum performance improvement. Compare the performance function score for upto "t^*" and setup a cut-off value for the differences. In other words, if the difference of scores from F_{t-t^*} to F_t is less than a cut-off value (ε), then stop the process and report the best solution found [119].

3. Setting up a cut-off value for the dispersion of the solutions in the elite sample [49, 116, 119, 167]. In [49, 119] a cut-off value for the variance of the elite sample solutions was considered. In this thesis, we propose to use a cut-off value to the median absolute deviation [66] of the solutions in the elite sample [116, 167].

The second and the third stopping criteria are discussed in this thesis.

2.2 Sequential change-point problem

The sequential problem of statistical diagnosis considers identifying a change in a stochastic process as quickly as possible with a pre-defined (controlled) false alarm rate. Therefore, sequential methods of statistical diagnosis are also known as quickest detection techniques. In the sequential detection problem, we observe data on-line (i.e., data are updated sequentially, one-by-one) and raise an alarm once a change is detected. The quality of a sequential detection technique is mainly characterized by the "average detection delay" and "average time between false alarms". In the literature many methodologies have been developed as on-line approaches to detect change-points. Some of the most popular methods are Shewhart method [140], Page's cumulative sum (CUSUM) method [101], Shiryaev-Roberts (SR) procedure [129, 141–143] and exponentially-weighted moving average method [129]. In this thesis, we will discuss the CUSUM and SR procedures. Both CUSUM and SR procedures are based on Wald's theory for sequential hypothesis testing [163] with certain optimality conditions.

The general sequential change-point problem can be described as follows.

Let $\{\mathbf{y}_n\}_{n\geq 1}$ be independent random variables which are observed sequentially, one by one. Here, n is the current sample size. Each y value follows a probability density function (pdf) $y_j \sim f(y_j;\boldsymbol{\theta})$, depending on some deterministic parameter(s) $\boldsymbol{\theta}$. Suppose at time τ a change in the process has occurred. This change is modelled by an instantaneous modification of the value of $\boldsymbol{\theta}$ at time τ. We shall define the parameter values of the probability distribution as $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ before the change-point at time τ, and $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ after the change-point.

Hypothesis to test :

$$\begin{aligned} &H_0 : \text{There is no change in the sequence (in-control)},\\ &H_1 : \text{A change occurred at time } \tau = k \text{ for } 0 \leq k < \infty \text{ (out-of-control)}. \end{aligned}$$

Under H_0, the pdf becomes:

$$f_0(\mathbf{y}_n \mid H_0) = \prod_{j=1}^{n} f(y_j; \boldsymbol{\theta_0}),$$

and under H_1, the pdf can be written as:

$$f_1(\mathbf{y}_n \mid H_1) = \prod_{j=1}^{k-1} f(y_j; \boldsymbol{\theta_0}) \prod_{j=k}^{n} f(y_j; \boldsymbol{\theta_1}), k \leq n.$$

In order to decide between H_0 and H_1, we can perform a likelihood ratio test. The likelihood ratio is as follows:

$$LR_k = \frac{f_1(\mathbf{y}_n \mid H_1)}{f_0(\mathbf{y}_n \mid H_0)} = \prod_{j=k}^{n} \frac{f(y_j; \hat{\boldsymbol{\theta}}_{\mathbf{1}})}{f(y_j; \hat{\boldsymbol{\theta}}_{\mathbf{0}})}, \tag{2.5}$$

where $\hat{\boldsymbol{\theta}}_{\mathbf{0}}$ and $\hat{\boldsymbol{\theta}}_{\mathbf{1}}$ are the maximum likelihood estimates of the pre-change and post-change pdf parameter values. In the simplest setting, let us assume we know the parameter values of the pre and post change pdf of the observed sequence. In this case, we can develop an algorithm based on the on-line or sequential detection techniques to estimate the change time. In the on-line approach, sample after sample, sequentially we can test the two hypotheses. The general form of a sequential change-point detection algorithm can be formulated below.

Algorithm 4 General form of a sequential change-point detection algorithm

```
Initialization
while the algorithm is not stopped do
    obtain the current sample
    test for the two hypotheses and decide between $H_0$ and $H_1$
    if $H_1$ decided then
        store the detection time and estimate the change-point location
        stop or reset the algorithm
    end
end
```

2.2.1 Motivating example

Let us consider a sequence of observations $y_j, j = 1, 2, \ldots,$ that are normally distributed with a common variance σ^2. Our interest is to estimate a process change in the mean levels. The pre-change parameters of the distribution can be denoted as $\boldsymbol{\theta}_0 = \{\mu_0, \sigma^2\}$ and the post-change parameters are denoted as $\boldsymbol{\theta}_1 = \{\mu_1, \sigma^2\}$. For simplicity and without loss of generality, we assume that $\mu_0 = 0$ and $\sigma^2 = 1$, then

$$f_0(y_j \mid \mu_0, \sigma^2) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y_j^2}{2}\right\},$$

$$f_1(y_j \mid \mu_1, \sigma^2) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(y_j - \mu_1)^2}{2}\right\}.$$

We can calculate the LR_k by utilizing (2.5):

$$LR_k = \prod_{j=k}^{n} \left\{ \frac{\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(y_j - \hat{\mu}_1)^2}{2}\right\}}{\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y_j^2}{2}\right\}} \right\},$$

$$= \prod_{j=k}^{n} \exp\left\{-\frac{1}{2}(y_j - \hat{\mu}_1)^2 + \frac{1}{2}y_j^2\right\},$$

$$= \exp\left\{\sum_{j=k}^{n} y_j\hat{\mu_1} - \frac{1}{2}\sum_{j=k}^{n} \hat{\mu_1}^2\right\}.$$

Let $\hat{\mu}_1$ be the maximum likelihood estimate of the post-change data sequence mean

$$\hat{\mu}_1 = \frac{\sum_{j=k}^{n} y_j}{n-(k-1)}.$$

Then

$$LR_k = \exp\left\{ \frac{\left(\sum_{j=k}^{n} y_j\right)^2}{2(n-k+1)} \right\}. \tag{2.6}$$

2.2.2 Cumulative sum procedure

The cumulative sum (CUSUM) algorithm was first introduced in [101] as an improvement to the sensitivity aspects of the Shewart charts. There are different derivations of the CUSUM algorithm in the literature [17]. The formal online approach is based upon a repeated use of the sequential probability ratio test [17]. In the context of observing a current sample and testing for multiple hypotheses regarding a change has occurred, this is another form of the CUSUM procedure, which we discuss in this thesis. Detailed information about the exact or near optimality conditions of the CUSUM procedure can be found in [87, 95, 109]. In the CUSUM approach, the stopping time is defined as

$$T_{A_C} = \inf\{n \geq 1 : W_n \geq A_C\}. \tag{2.7}$$

Here, W_n in 2.7 is known as the CUSUM statistic and $A_C > 0$ is an unknown positive threshold that controls the false alarm rate in the CUSUM procedure. We will define W_n for the case described in section 2.2.1 as

$$W_n = \max_{1 \leq k \leq n} LR_k = \max_{1 \leq k \leq n} \prod_{j=k}^{n} \frac{f(y_j; \hat{\boldsymbol{\theta}}_{\mathbf{1}})}{f(y_j; \hat{\boldsymbol{\theta}}_{\mathbf{0}})} = \max_{1 \leq k \leq n} \exp\left\{ \frac{\left(\sum_{j=k}^{n} y_j\right)^2}{2(n-k+1)} \right\}, \quad k = 1, 2, \ldots.$$

2.2.3 Shiryaev-Roberts procedure

The CUSUM procedure uses a frequentist approach (maximum likelihood) to obtain an estimate for the change-point. A Bayesian formulation was proposed in [141, 143] assuming a geometric prior distribution for the change-point. Thus, it has formulated the optimal sequential change-point problem as an optimal stopping problem. Later [129] extended the work of [141] to a more general context, where it made no assumption on the prior distribution of the change-point. Different optimality conditions of the SR procedure were discussed extensively in [109, 111, 112].

The stopping time of the SR procedure is defined as

$$T_{A_{SR}} = \inf\{n \geq 1 : R_n \geq A_{SR}\}, \tag{2.8}$$

where R_n is known as the SR statistic and $A_{SR} > 0$ is an unknown positive threshold that controls the false alarm rate. We will define R_n in the context of the example considered in section 2.2.1.

$$R_L = \sum_{k=1}^{n} LR_k = \sum_{k=1}^{n} \prod_{j=k}^{n} \frac{f(y_j; \hat{\boldsymbol{\theta}}_{\mathbf{1}})}{f(y_j; \hat{\boldsymbol{\theta}}_{\mathbf{0}})} = \sum_{k=1}^{n} \exp\left\{ \frac{\left(\sum_{j=k}^{n} y_j\right)^2}{2(n-k+1)} \right\}, \quad k = 1, 2, \ldots.$$

There are currently few studies that discuss the performance differences of the CUSUM and SR procedures in detail [96, 110, 153]. A comprehensive asymptotic study was performed in [110] to compare the effectiveness of the two procedures in detecting a change in the drift of the Brownian motion. They have found that the CUSUM performs better than the SR procedure for changes that occur at the beginning and the SR performs better than the CUSUM with respect to the conditional average detection delay [96, 110]. However, it is noted in [96] that the performance differences of the two procedures are significant for small changes, visible for moderate changes and not significant at all for large changes.

2.3 Parallel programming in R

Being an evolutionary computing method the CE method naturally requires a higher order of computer resources to carry out its computations. This factor is considered as one of the bottlenecks of the CE method [119] especially in the context of detecting change-points in biological sequences. To mitigate this problem, we have proposed an implementation of the CE algorithm by utilizing multi-core parallel computing techniques [114–117, 120, 167] in the R statistical computing environment [122]. The proposed parallel computing procedures were applied in both Microsoft Windows and UNIX operating systems [167]. We refer to [45] for a detailed review of parallel computation techniques in general and refer to [134] to obtain detailed information on parallel computing in theR statistical software.

R is a powerful and extensive open source statistical environment, but the default behaviour of R is to execute arguments serially or in a non-parallel way. Thus, a step-by-step (or line-by-line) approach is taken in command execution. However, with the use of other R packages it is possible to carry out parallel calculations. In R, there are three main streams of code (program) parallelization [134]. They are:

- computer cluster-based parallel computation,
- grid computing,
- multi-core systems-related parallel computation.

In this thesis, we only consider multi-core parallel computation techniques in R to efficiently perform calculations of the CE method both in Microsoft Windows and UNIX operating systems [167]. In muti-core parallel computation, we distribute a common set of R commands to run on the available cores in the computer. Even though the overall process is controlled by the common set of R commands that is initially specified, the assignment of those instructions to the cores are controlled by the operating system.

In Microsoft Windows, we have developed a SNOW (Simple Network Of Workstations) parallel process by utilizing the snow [160], doSNOW [6] and foreach [7] packages. The doSNOW R package works as the foreach parallel adapter for the snow package. In UNIX, we used doMC [5], parallel [122] and foreach [7] to form a multi-core parallel computing environment.

Chapter 3

Multiple Break-Points Detection in array CGH Data via the Cross-Entropy Method

This chapter constitutes a peer-reviewed journal article published in the IEEE Transactions on Computational Biology and Bioinformatics. The content of the paper including notation, plot sizes and wording has been slightly changed to improve the flow and the overall coherence of the thesis. The supplementary material originally published with the paper is included as an appendix (see Appendix A). The citation for the article is:

Priyadarshana, W. J. R. M., Sofronov, G. (2014), "Multiple Break-Points Detection in array CGH Data via the Cross-Entropy Method," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol.PP, no.99, pp.1,doi: 10.1109/ TCBB.2014.2361639
Writing: 100% , analysis: 100% , conception and design: 95%

Specific contribution of joint authors
Georgy Sofronov: Overall supervision of data analysis and interpretation. Provided continual feedback and suggestions on both analysis and writing.

Reprinted with permission from Priyadarshana, W. J. R. M. and Sofronov, G. (2014) IEEE Transactions on Computational Biology and Bioinformatics, Copyright © 2014 IEEE.

3.1 Summary

This chapter discusses a study in which we present the Cross-Entropy (CE) Method, as described in Chapter 2, to detect multiple change-points (break-points) in biological sequences of continuous measurements. Particularly, we applied the CE method to detect copy number variations (CNVs) in array comparative genomic hybridization (aCGH) data.

There exists a number of methods that have been developed to detect CNVs using high-throughput sequencing techniques. The aCGH data falls under the branch of first generation high-throughput sequencing technologies. Even though there exists a wide spectrum of methodologies, the majority of currently available detection methods are sub-optimal in finding change-points. These methods do not represent the underlying uncertainty of the number as well as the locations of the change-points precisely. Furthermore, there is currently no attempt to utilize model-based evolutionary stochastic methods to estimate the number and their corresponding locations of change-points in aCGH data. In order to fill this gap in the current literature, we have proposed the CE method, which is a model-based stochastic evolutionary computing technique, to detect multiple change-points in aCGH data.

The proposed CE algorithm contains several improvements over its erstwhile implementations [49, 119]. We have proposed to incorporate the modified Bayesian information criterion [174], as described in Chapter 2, to effectively estimate the number of change-points. A detailed simulation study was carried out to obtain the optimal set of parameters from the CE algorithm. Furthermore, we compared the performance of the CE algorithm with four other well-known methods: the circular binary segmentation method (R package name - *DNACopy* [161]) proposed in [100, 161]; an efficient segmentation method proposed in [98] using the least angle regression [46] and models selection methods (R package name - *cumSeg* [97]); the Pruned Exact Linear Time proposed in [76] (R package name - *changepoint* [74, 124]); and a Bayesian change-point detection method discussed in [47, 48] (R package name - *bcp* [47, 48]) based on the product partition model [16]. All four methods are freely available as R packages in the comprehensive R archive network [122]. A comprehensive simulation study was carried out using artificially generated data with different aberration widths and signal strengths. We applied the proposed CE method to three publicly available aCGH data sets to identify CNVs. The first data set on fibroblast cell lines, was introduced in [150] and further discussed in [98, 161]. The second data set contains cDNA (complementary DNA) aCGH data from glial brain tumours, and was introduced in [28]. The last real data example is on breast cancer cell lines data, originally published in [108]. In general, it was revealed in both the analysis of artificially generated and real data that the proposed CE method performs evenly or better than competing techniques in terms of detecting the number of change-points and their locations. In the simulation study, it was further observed that the CE method is more effective in estimating the change-point locations than the change-point number. However, in terms of processing time, being an evolutionary computing algorithm, the CE method consumes more computational resources than other methods. In order to mitigate this computational drawback of the general CE method, we have implemented a parallel computation approach to increase its efficiency both in WINDOWS and UNIX operating systems. The proposed methodology is freely available as an R package (package name: breakpoint) at the Comprehensive R Archive Network (`http://cran.r-project.org/web/packages/breakpoint/index.html`).

Multiple Break-Points Detection in array CGH Data via the Cross-Entropy Method

W. J. R. M. Priyadarshana[1,*], Georgy Sofronov[1]

[1] Department of Statistics, Faculty of Science, Macquarie University, Sydney NSW 2109, Australia.

*E-mail: madawa.weerasinghe@mq.edu.au

Abstract

Array comparative genome hybridization (aCGH) is a widely used methodology to detect copy number variations of a genome in high resolution. Knowing the number of break-points and their corresponding locations in genomic sequences serves different biological needs. Primarily, it helps to identify disease-causing genes that have functional importance in characterizing genome wide diseases. For human autosomes the normal copy number is two, whereas at the sites of oncogenes it increases (gain of DNA) and at the tumour suppressor genes it decreases (loss of DNA). Majority of the current detection methods are deterministic in their set-up and use dynamic programming or different smoothing techniques to obtain the estimates of copy number variations. These approaches limit the search space of the problem due to different assumptions considered in the methods and do not represent the true nature of the uncertainty associated with the unknown break-points in genomic sequences. We propose the Cross-Entropy method, which is a model-based stochastic optimization technique as an exact search method, to estimate both the number and locations of the break-points in aCGH data. We model the continuous scale log-ratio data obtained by the aCGH technique as a multiple break-point problem. The proposed methodology is compared with well established publicly available methods using both artificially generated data and real data. Results show that the proposed procedure is an effective way of estimating number and especially the locations of break-points with high level of precision.

Availability: The methods described in this article are implemented in the new R package breakpoint and it is available from the Comprehensive R Archive Network at http://CRAN.R-project.org/package=breakpoint.

Introduction

Complex genetic alterations in the human genome is one of the key driving forces behind tumour development and progression. It also affects the degree of response to drugs, vaccines, pathogens and other forms of cure to genomic diseases. In the last decade, there has been a significant improvement in the sequencing technologies as well as the data analyzing methodologies, which aim to detect these important variations. These structural alterations can be inherited through germline as well as be somatically acquired. Copy number variations (CNVs) is a form of structural variation in the genome. CNV is defined as a DNA segment that is 1 kb or larger and present at variable copy number in comparison with a reference genome [15]. Recent studies have shown that around 12% of the human genome vary in copy number [32] which includes important genetic information. Furthermore, it has been identified that CNV plays an important role in genetic susceptibility to common diseases [37] such as cancer [1, 6, 19, 26], autism [38], HIV [16], immune disorders [13], intellectual disabilities [21], etc.

Comparative genome hybridization to DNA micro-arrays (aCGH) is one of the most popular techniques that can be utilized to detect and map copy number variation in DNA sequences. It measures copy number variations of an individual's DNA with respect to a reference or pool of reference DNAs at a fixed set of genomic locations. Thus, CNV detection in aCGH data is essentially a multiple break-point (or change-point) problem. Refer [7] for a detailed review on CNV using DNA microarrays.

Reviewing the literature on break-point modelling in aCGH data, [4] discussed some of the methodologies that were used to segment DNA sequences. They have proposed a local segmentation method called split polynomial fitting. Furthermore, [3] proposed a quasi-likelihood approach to DNA sequence segmentation with the use of a simulation based modified version of the Schwarz criterion [34]. A non-parametric approach based on the circular binary segmentation (CBS) procedure was proposed in [52] to detect abrupt changes in the mean levels of aCGH data. A test statistic based on mean differences to identify the number of break-points is utilized in the CBS. Later, [40] applied fused lasso method to the hot spot detection in aCGH data. [42] proposed a scan statistic based on summing a chi-squared statistic for each individual sample in order to detect simultaneous change-points in multiple sequences. [26] proposed *cumSeg* procedure which uses *lars* algorithm and a generalized version of the BIC to estimate break-points in aCGH data. They have provided a detailed comparison study with other competing methodologies to signify the advantages of their proposed approach. A fast Bayesian change-point detection method *bcp* was proposed in [6] with the use of Product

Partition Models (PPM), which is originally proposed by [2]. Recently, a Pruned Exact Linear Time (PELT) method is introduced in [17] to detect multiple break-points in a sequence of observations with different specifications. Refer [19] for a detailed review of some of the methodologies available. Furthermore, various approaches to detect multiple break-points in DNA sequences were discussed in [29] and [39] with respect to different measurement scales.

In general, the problem of finding break-points in a sequence of observations is a two-fold problem. First, an estimate of the number of break-points is obtained. Then, based on the estimated number, corresponding locations are identified. In this study, we propose the Cross-Entropy (CE) method [33] to successfully solve both of these optimization problems simultaneously. Even though both the problems of estimating the number as well as the locations are important, the locations (loci) are the core estimates that we are ultimately interested in, since it facilitates the process of finding disease-causing genes in genomic sequences.

This paper introduces the CE method proposed in [2, 12] to detect number of break-points as well as their corresponding genome locations in aCGH data. In this study, our method assumes that within each segment the aCGH data are approximately distributed as normal. A simulation study was carried out to obtain the best possible values for the parameters of the CE method for the case of continuous measurements, in which they are pivotal in the performance of the methodology. The work in this paper is significantly different from [2, 12] in multiple aspects. Firstly, this paper considers continuous scale measurements (aCGH data) to estimate break-points and their locations as opposed to the count data used in [2] and binary data used in [12]. Thus, it outreaches to wider audience to utilize the CE method in detecting break-points in genomic data. Secondly, we implement a multi-core architecture based parallel implementation of the CE method as compared to the non-parallel versions discussed in [2, 12]. Moreover, in [12] the CE method is only used to obtain the locations of the break-points and not as an optimization methodology to estimate the number of break-points. However, this study as well as [2] utilizes the CE method as a stochastic optimization technique to obtain both the number and the corresponding locations of break-points. Furthermore, in [12] the model log-likelihood is considered as the performance function in the CE method, whereas in this paper we introduce an information criterion as the performance function to facilitate the process of estimating the number of break-points. In terms of the parameter estimation procedure, the approach followed in [12] is completely different from the approach considered in this paper. In the proposed CE method, there is no supposition that break-points are uniformly distributed along the sequence contrary to the equally spaces assumption considered in [12]. Furthermore, in the proposed CE procedure users can specify the minimum aberration width (segment width) to control the number of break-points based on the user requirements. Thus, it adds more versatility to the CE method to control false alarms and to meet user specifications.

In this study, we propose an improved version of the CE algorithm compared to that used in [2]. We developed a parallel implementation of the CE algorithm with the use of a different objective function as well as a more robust stopping criterion, to significantly improve the efficiency levels of the CE algorithm. In [2] the general Bayesian Information Criterion (BIC) was used as the objective function in the CE algorithm. In this study we propose to utilize the modified Bayesian Information Criterion discussed in [43], which is theoretically justified for the break-point problem. The proposed methodology is freely available as an R package ("*breakpoint*") from the Comprehensive R Archive Network (http://CRAN.R-project.org/package=breakpoint).

The paper is structured as follows. Section 2 introduces the multiple break-point problem in mathematical terms and describes in detail the CE method and the work-flow associated with the algorithm in segmenting aCGH data. It also discusses the modified BIC used in this study. Section 3 presents a detailed simulation study that is carried out to obtain the best set of parameter values for the CE method in detecting multiple break-points in aCGH data. Section 4 discusses the advantages of the parallel implementation of the CE method over the non-parallel standard implementation. Section 5 presents the results of numerical experiments both for artificially generated and real data. Comparison studies were carried out to compare the proposed methodology with some of the best performing techniques available in the literature to estimate the number of break-points as well as their corresponding locations. We conclude the paper with a general discussion and future research directions.

Methods

The Multiple Break-Point Problem

Let us formulate the multiple break-point problem in mathematical terms. Consider a sequence of observations $\mathbf{y} = (y_1, y_2, \ldots, y_L)$ of length L, in which y_i's are independently distributed Gaussian random variables.

A segmentation of the sequence is specified by the number of break-points N and the positions of the break-points $\mathbf{C} = (c_1, c_2, \ldots, c_N)$, where $1 = c_0 < c_1 < \cdots < c_N < c_{N+1} = L + 1$. In this context, a break-point is a boundary between two adjacent segments. The value of c_i is the sequence position of the rightmost character of the segment to the left of the i-th break-point. Segments are numbered from 0 to N as there will be one or more segments than the number of break-points. The model assumes that within each segment the observations are distributed as normal with mean μ and common variance σ^2. Both the piecewise constant means and the common variance are not known in advance.

Modified Bayesian Information Criterion (mBIC)

We utilized the modified BIC [43] as the performance function to be used in the CE algorithm. In [43], authors have mentioned that the classic BIC proposed in [34] is not theoretically justified for break-point problem due to the fact that the likelihood function does not satisfy the standard regularity conditions (see page 1 in [43]). A detailed simulation study has been carried out in [43] to signify the performance differences of the mBIC over the classic BIC. We refer the reader to [43] for more details. In this study, we considered the mBIC (Theorem 2 in [43]) developed for the independent normally distributed observations with unknown constant variance and piecewise constant means:

$$\begin{aligned}
&\log\frac{P(M_N|\mathbf{y})}{P(M_0|\mathbf{y})} = \left(\frac{L-N+1}{2}\right)\log\left[1+\frac{SS_{bg}(\hat{\mathbf{c}})}{SS_{wg}(\hat{\mathbf{c}})}\right] \\
&+\log\left[\frac{\Gamma\left(\frac{L-N+1}{2}\right)}{\Gamma\left(\frac{L+1}{2}\right)}\right]+\frac{N}{2}\log(SS_{all}) \\
&-\frac{1}{2}\sum_{i=1}^{N+1}\log n_i(\hat{\mathbf{c}})+\left(\frac{1}{2}-N\right)\log(L)+O_p(1),
\end{aligned} \tag{3.1}$$

with

$$
\begin{aligned}
&SS_{bg} = \sum_{i=1}^{N+1} n_i(\hat{\mathbf{c}}) \left[\bar{y}_i - \bar{y}\right]^2, \quad SS_{all} = \sum_{j=1}^{L} (y_j - \bar{y})^2, \\
&SS_{wg} = SS_{all} - SS_{bg}, \quad \hat{\mathbf{c}} = \underset{\mathbf{c} \in D_N}{\operatorname{argmax}} \frac{SS_{bg}(\mathbf{c})}{SS_{wg}(\mathbf{c})}, \\
&n_i(\hat{\mathbf{c}}) = \hat{c}_{i-1} - \hat{c}_i, \quad \bar{y}_i(\hat{\mathbf{c}}) = \frac{1}{n_i(\hat{\mathbf{c}})} \sum_{j=\hat{c}_{i-1}}^{\hat{c}_i - 1} y_j, \\
&\bar{y} = \frac{1}{L} \sum_{j=1}^{L} y_j,
\end{aligned}
$$

where M_N is defined as the Gaussian model with N number of break-points and M_0 is the simplest Gaussian model with no break-points.

The Cross-Entropy method

The Cross-Entropy (CE) method [33] is an evolutionary computing technique originally developed to estimate probabilities of rare events. Later, the introduction of CE minimization has lead path not only to estimate probabilities of rare events but also for solving complex combinatorial and multi-extremal optimization problems. The CE method is developed on the Kullback-Leibler divergence [18], which is one of the fundamental concepts of modern information theory. Generally the CE method is an iterative procedure and each iteration consists of two main phases:

- simulate a random solution set (vectors, trajectories, etc.) based on a specified random mechanism (e.g. statistical distribution),
- score each of the solution set based on a performance function and update the parameters of the random mechanism to produce an improved solution set in the next iteration.

Thus, the CE method is an iterative optimization procedure. It starts with a parametrized sampling distribution from which a random sample is generated as possible solutions for the problem. Then, each solution or the combinatorial arrangement is given a score based on a performance function. A fixed number of best performing combinatorial arrangements are retained and it is denoted as the elite sample. This elite sample is subsequently used to update the parameters of the sampling distribution in the next iterations. Thus, adaptive parameters are utilized in each iteration. The sampling distribution eventually converges to a degenerate distribution about a locally optimal solution, which ideally will be globally optimal [25, 38].

In this study, the process of multiple break-point detection is considered as a combinatorial optimization problem. We utilized the variant of the CE method discussed in [2] with further modifications and improvements to analyze biological sequences of continuous measurements. The proposed methodology gives more flexibility to the user in terms of specifying the minimum aberration width or the segment width (h). A stopping criterion (SC) based on Median Absolute Deviation (MAD) was used as opposed to the variance based SC proposed in [2]. Median is a robust measure of location, that is not affected by extreme values contrary to the standard deviation, which is sensitive to outliers. Empirically we found out that the performance of the algorithm with MAD was better than the standard deviation (SD), especially in terms of the processing time without compromising its accuracy. We refer the readers to Section 2 of the supplementary material on the comparative analysis of the CE method with the use of MAD and SD. Furthermore, a multi-core architecture based parallel implementation of the algorithm was introduced in order to carry out the calculations more efficiently.

The CE method for break-point problem is a model-based iterative stochastic optimization procedure that starts with a parametrized sampling distribution, from which a random sample of size M is generated with respect to the number of break-points (N). We used a four-parameter beta distribution as the sampling distribution to simulate break-points. Each combinatorial arrangement was scored for its performance based on the performance function F, where, the mBIC was used as the F to score each of the arrangements. Then we obtained a best performing fraction of the samples based on the performance function score. We defined this sample as elite sample and $M_{elite} = \rho \times M$ as its size, where rho (ρ) was defined as the elite sample fraction. This elite sample was used to update the parameters of the sampling distribution until a SC is met. The MAD, which is a robust estimator for dispersion [1] was used as the SC.

In the CE method, there are few parameters to be specified prior to the initialization. We categorized these parameters into two groups based on their impact to the performance of the CE algorithm. We defined them as "internal" and "external" sets of parameters. The internal parameters are the core parameters that directly affect the performance of the CE algorithm. They are the sample size (M), elite sample fraction (ρ) and the cut-off value for the SC (ε). The external parameters; segment width (h) and the maximum number of break-points (N_{max}); are the non-core parameters for the performance of the CE algorithm. These parameters only affect the overall processing time but not the internal performance of the CE algorithm. The external parameters can freely be changed by the user based on their requirements. However, changes in the internal parameters have to be done with caution.

Based on the user-defined external set of parameters (N_{max}, h), the CE algorithm can be outlined as follows:

1. Choose initial values for $\boldsymbol{a}^{\mathbf{0}} = (1, 1, \dots, 1)$ and $\boldsymbol{b}^{\mathbf{0}} = (1, 1, \dots, 1)$. In this case we have set both parameters equal to one and both parameter vectors are N dimensional. Set $t = 1$.
2. Generate a random sample $\mathbf{C}^{(1)}, \mathbf{C}^{(2)}, \dots, \mathbf{C}^{(M)}$ from the $\mathsf{Beta}(\boldsymbol{a}^{t-1}, \boldsymbol{b}^{t-1})$ distribution, where $\mathbf{C}^{(i)} = (c_1^{(i)}, c_2^{(i)}, \dots, c_N^{(i)})$, $i = 1, 2, \dots, M$.
3. For each $i = 1, 2, \dots, M$ order $c_1^{(i)}, \dots, c_N^{(i)}$ from smallest to largest and set $\mathbf{C}^{(i)} = (c_1^{(i)}, c_2^{(i)}, \dots, c_N^{(i)})$, where $\mathbf{C}^{(i)}$ is the break-point vector defined earlier.
4. Evaluate the performance score F of each $\mathbf{C}^{(1)}, \mathbf{C}^{(2)}, \dots, \mathbf{C}^{(M)}$. Obtain the elite sample (M_{elite}), which is the best performing combinations of the break-point locations.
5. For all $j = 1, 2, \dots, N$ estimate the two shape parameters $\boldsymbol{a}^t = (a_1^t, a_2^t, \dots, a_N^t)$, $\boldsymbol{b}^t = (b_1^t, b_2^t, \dots, b_N^t)$ using the elite sample and update the current parameter set.
6. If the SC is met, then stop the process and identify the combination of the locations of break-points $\mathbf{C}^{(i)}$ that optimizes the performance function F. Otherwise set $t = t + 1$ and iterate from step 2.

The SC considered in the CE procedure is:

$$\text{SC : Stop the process if } \max_j \text{MAD}_j < \varepsilon,$$

where,

$$\text{MAD}_j = \underset{i=1,2,\dots,M}{\text{Median}} \left| c_j^{(i)} - \text{Median}\left(c_j^{(1)}, c_j^{(2)}, \dots, c_j^{(M)}\right) \right|$$

for all $j = 1, 2, \dots, N$.

The final solution is a single vector of break-point locations ($\mathbf{C}$).

Selection of best parameter values for the CE algorithm

We carried out two simulation studies to obtain the best set of internal parameters for the CE algorithm. In the first study, we fixed the cut-off value for the SC (ε) as 0.01. An artificially spiked-in data set having four break-points with aberration widths (w) of 80, 70, 100, 75 and 125 was considered for the first simulation study. The length of the sequence is 450. The corresponding signal-to-noise ratio (SNR) values that we used to generate data for each of the segments were 0, 2, 1, 3.5 and 4. The SNR is defined as the ratio of the mean of the aberration width divided by the standard deviation of the super imposed Gaussian noise [19]. The standard deviation of the Gaussian noise was set as 0.25 in SNR calculations. The choice of standard deviation value is based on the real aCGH examples [19]. We compared the results obtained from our methodology with four other well-known methods: *DNAcopy* ([52], a.k.a CBS), *cumSeg* [26], *changepoint* [17] and *bcp* [6]. In *DNAcopy* the default configuration was considered. In *cumSeg*, generalized BIC with the penalty function of "log(log(n))" was utilized. In *changepoint* "cpt.mean" with "PELT" specifications was utilized. Finally, for *bcp* a conservative threshold value of 0.25 was considered in posterior probabilities to obtain the break-points, since it does not give the break-points explicitly as an output.

The external parameters (h, N_{max}) were set as (5, 10) and we carried out the analysis varying both the sample size (M) and elite sample fraction (ρ) value, which is used to obtain the elite sample of size M_{elite}. The set of parameter values considered for M and ρ are: $M = \{100, 200, 300, 400, 500\}$ and $\rho = \{0.02, 0.03, 0.04, \ldots, 0.1\}$. In each of the combinations of M and ρ, we generated 100 random sequences with respect to the aforementioned aberration widths and SNR values.

Figure 5.2 shows the average RMSE of the results with respect to different M and ρ values. The error profile for $M = 100$ is comparatively higher than the other M values. In general, a logarithmic decline in the error rates can be seen in all the profiles with the increase of ρ values. Thus, the average RMSE values are affected by the changes in ρ values. However, the changes in the average RMSE values are not in larger scale after the ρ value of 0.06 irrespective of the M value. It can also be observed that there is a gradual reduction in the RMSE values with the increase of M values. However, after $M = 300$ advantages of increasing the M value is not significant at all. Table 3.1 shows the average RMSE values for the CE method as well as the other four methods. In terms of the accuracy of estimates the proposed CE method outperforms competing methods

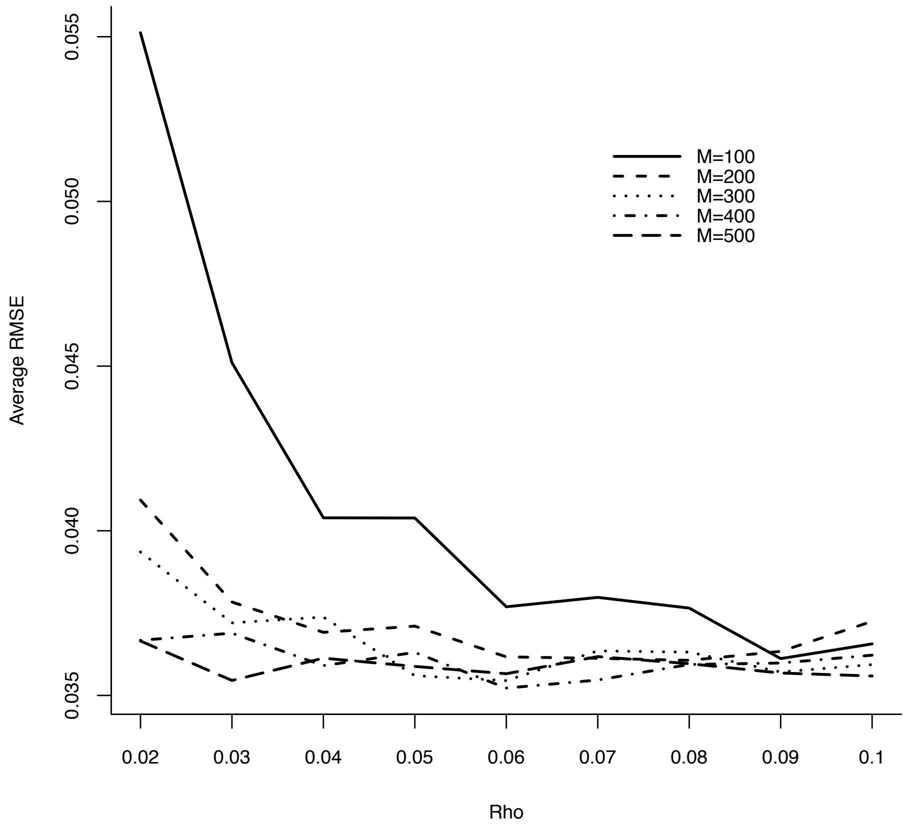


Figure 3.1: Average RMSE values of the CE method with respect to different M and ρ values

Table 3.1: Average RMSE values of the CE method for different M and $rho(\rho)$ values

rho (ρ)	M				
	100	200	300	400	500
0.02	0.0551	0.0409	0.0394	0.0367	0.0367
0.03	0.0451	0.0378	0.0372	0.0369	0.0355
0.04	0.0404	0.0369	0.0374	0.0359	0.0361
0.05	0.0404	0.0371	0.0356	0.0363	0.0359
0.06	0.0377	0.0362	0.0355	0.0352	0.0357
0.07	0.0380	0.0361	0.0364	0.0355	0.0362
0.08	0.0377	0.0361	0.0363	0.0359	0.0360
0.09	0.0361	0.0363	0.0357	0.0360	0.0357
0.1	0.0366	0.0372	0.0359	0.0362	0.0356

in all the situations even at lower ρ values. It is observed that there is a significant reduction in the average RMSE especially at lower ρ values with the increase of M values. However, the reduction in average RMSE at higher ρ values is not significant

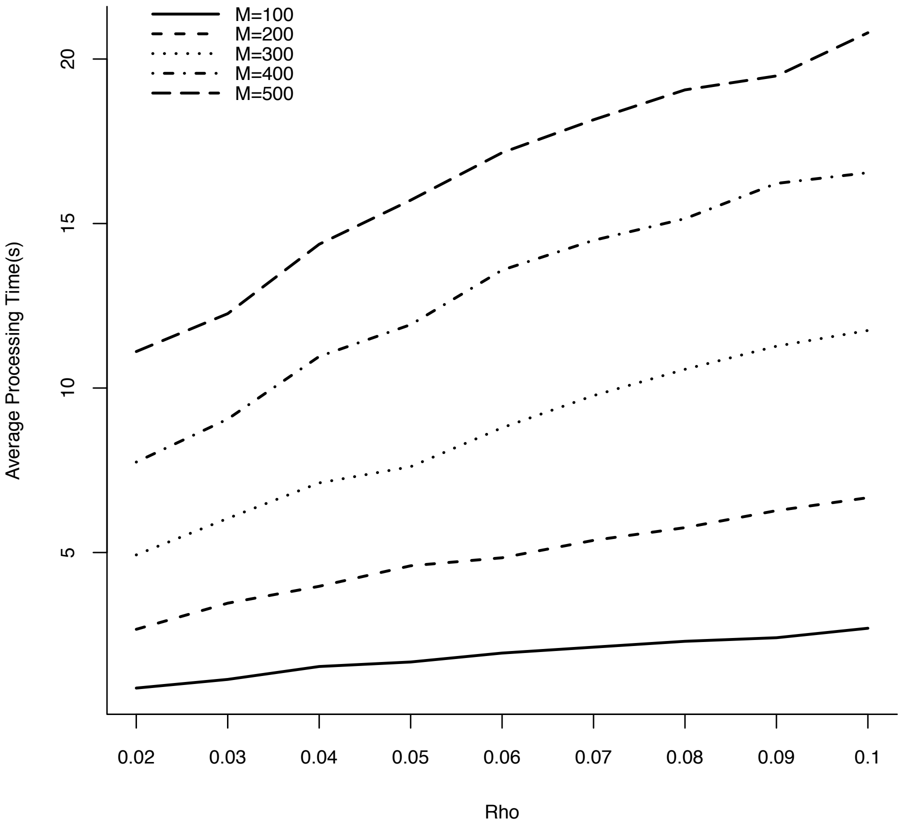


Figure 3.2: Average processing time (s) of the CE method with respect to different M and ρ values

Table 3.2: Average RMSE values of the other four methods

Method	Average RMSE
DNAcopy	0.0572
cumSeg	0.0639
changepoint	0.1272
bcp	0.0776

at all. This agrees with the discussion in [38], that the effect of elite sample to the performance of CE algorithm is not significant at higher M values, even though there is a substantial effect of ρ values at lower M values.

Figure 5.3 exhibits the processing time with respect to different M and ρ values. It is observed that the overall processing time significantly increases with the increase of sample size M (Note: computation times are relative to a 2.3GHz Intel Core i7 processor with 8GB physical RAM in Mac OS X Version 10.6.8). It further reveals that the processing time has an approximate linear type relation with the sample size

M. Furthermore, at all the M values the processing time is linearly increasing with the increase of ρ values. Thus, it shows that there is an impact of the choice of ρ values and the M values to the efficiency level of the CE algorithm. Therefore, we need to consider an optimal choice of M and ρ values that balance the trade-off between the precision and processing time. We propose to consider M of 200 and ρ value of 0.06 as the default values for the CE algorithm based on the simulation results.

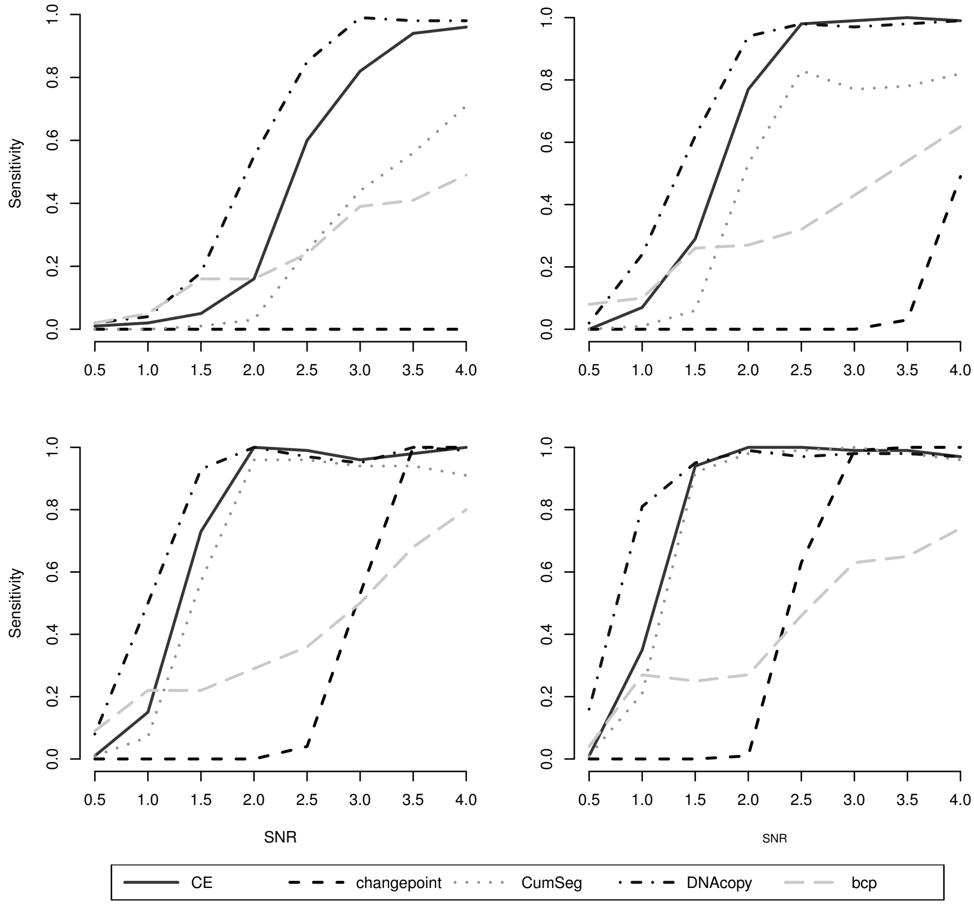


Figure 3.3: Sensitivity of estimating the number of break-points for aberration widths of 5, 10, 20 & 40 (From top left to bottom right)

The second simulation study was carried out to determine the best parameter value for the cut-off value in the SC (ε). The M and ρ values were set at 200 and 0.06 as identified in the first simulation study. A range of values from 0.001 to 0.1 were considered for the ε. We carried out 100 simulations for each ε value with four break-points with a similar framework considered in the first simulation study. It is observed that there is no significant impact of the ε value on both the precision and processing time based on the range of values considered for the ε. Therefore, we set the default value for the ε as 0.01 in the CE algorithm. More details are in supplementary material: Section 1.

Comparison of parallel and non parallel implementation of the CE method

The proposed CE methodology utilized a multi-core architecture based parallel implementation in both WINDOWS and UNIX type operating systems with the use of R statistical computing environment [35]. In [2], it has been identified that the processing time as the major drawback of the CE method. Being an evolutionary optimization technique the CE algorithm naturally yields higher computing time. Non-parallel implementation also adds an extra burden on the overall processing time. In this paper, we solve this critical bottleneck of the CE method to a greater extent with the use of parallel computing techniques.

In this study the parallel implementation was carried out in multiple cores by utilising the parallel [35], doMC [11] and doSNOW [10] R packages. The doSNOW R package registers the SNOW [37] parallel back-end with the use of foreach R package [14] in WINDOWS operating systems. In UNIX like operating systems parallel computation was carried out with the help of doMC, parallel and foreach R packages.

We carried out a similar simulation study as described in Section 3 for parameter selection for the CE algorithm. Where, four break-points were incorporated with the same SNR values with ε value of 0.01 as the cut-off value for the SC and ρ value of 0.06 was considered to obtain the elite sample. We repeated the analysis with different sample sizes (100, 1000 and 10000) to illustrate the significance of the proposed parallel implementation of the CE algorithm over the non-parallel version. In each of the sample sizes 100 random sequences were simulated to obtain the average results. Table 3.3 shows the processing time of the proposed methodology with the parallel and non-parallel implementation for different sample sizes in MAC OS X and WINDOWS OS. The parallel implementation of the methodology has significantly improved the efficiency level of the proposed methodology when compared with the non-parallel implementation.

On average, in MAC OS X there is more than 100% improvement (1-fold) in parallel implementation with regard to the three sample sizes considered. In Windows OS the proposed parallel implementation enjoys a significant improvement over the non-parallel implementation with more than 200% (2-fold) improvement. Thus, it is evident that the proposed parallel implementation of the CE algorithm significantly reduces the processing time in both MAC OS X and Windows OS.

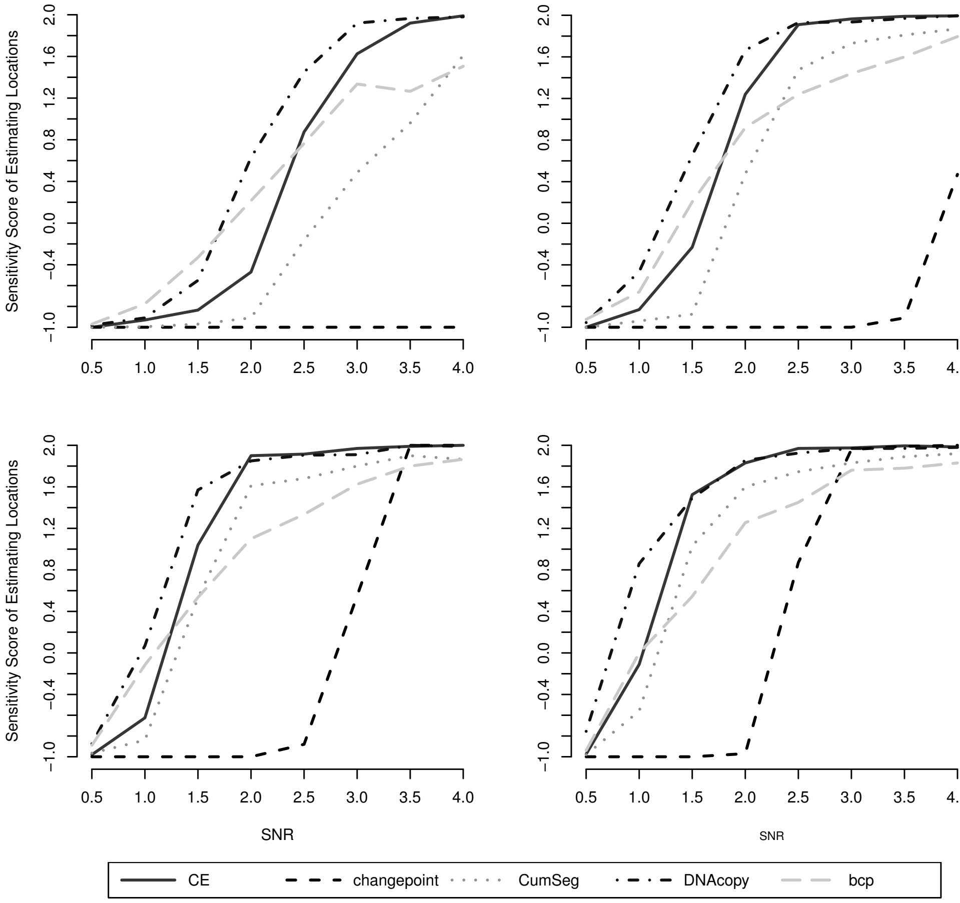


FIGURE 3.4: Sensitivity score of estimating the true locations of break-points for aberration widths of 5, 10, 20 & 40 (From top left to bottom right)

TABLE 3.3: Average processing time (in seconds) for non-parallel and parallel implementations for different sample sizes

	Sample size (L)		
	100	1000	10000
MAC OS X[1]			
non-parallel	5.52	15.87	38.78
Parallel	2.38	7.14	18.58
Improvement (%)	131.93	122.27	108.72
WINDOWS OS[2]			
non-parallel	8.92	24.14	46.86
Parallel	2.84	7.02	14.95
Improvement (%)	214.08	243.87	213.44

[1] Relative to Mac OS X Version 10.6.8, 2.3GHz Intel Core i7 processor with 8GB physical RAM.

[2] Relative to Windows 7, 2.0GHz Intel Core i7 processor with 8GB physical RAM

Numerical Results

In this section, we include results of numerical experiments to assess the performance of the proposed methodology in identifying the correct number of break-points (N) as well as their locations ($\mathbf{C}$). We considered an artificially generated data set having two break-points, with different aberration widths as well as different signal-to-noise (SNR) ratios. Then, we applied our methodology and four other competing approaches (*DNAcopy* , *cumSeg*, *changepoint* and *bcp*) on three well-known publicly available real aCGH data to further signify the importance of the proposed methodology.

Results on artificially generated data

An artificially generated data set having two break-points with aberration widths of 5, 10, 20 and 40 probes and SNR of 0.5 to 4 with an increment of 0.25 was considered in this simulation study. We set up the standard deviation of the Gaussian noise as 0.5 and generated 100 random sequences in each of the SNR values to assess the effectiveness of the methodology. The length of the whole sequence was 100 probes. We introduced a square wave of aberration to the middle of the sequence.

A sensitivity analysis was carried out both on the estimation of the number of break-points as well as their locations. If the number of break-points estimated is equal to the actual number, then it is considered as a valid result (true positive result). The ratio of number of times that correctly identified the true number of break-points out of the total runs is plotted against different SNR values to assess the sensitivity of identifying the true number of break-points. Figure 5.6 shows the results. It is observed that none of the methods have demonstrated superior performance in detecting the number of break-points at low SNR values and small aberration widths. However, *DNAcopy* method has performed better on average in low SNR values as well as in small aberration widths followed by the proposed CE method. Note that, at small aberration widths *changepoint* has significantly under-performed even at higher SNR values, while all other methods have performed reasonably better.

In assessing the sensitivity of estimating the break-point locations, we utilized the procedure discussed in [18] and [26]. We defined the true number of break-points as N_0 and the estimated number from the algorithm as $\hat{N}$. Let $c_j^0; j = 1, \dots, N_0$, be the true locations of the break-points. We gave a score (S_1) of '1' if the estimated break-point location $\hat{c}_j \in [c_j^0 - 2, c_j^0 + 2]$ and '0' otherwise. To account for the number of break-points

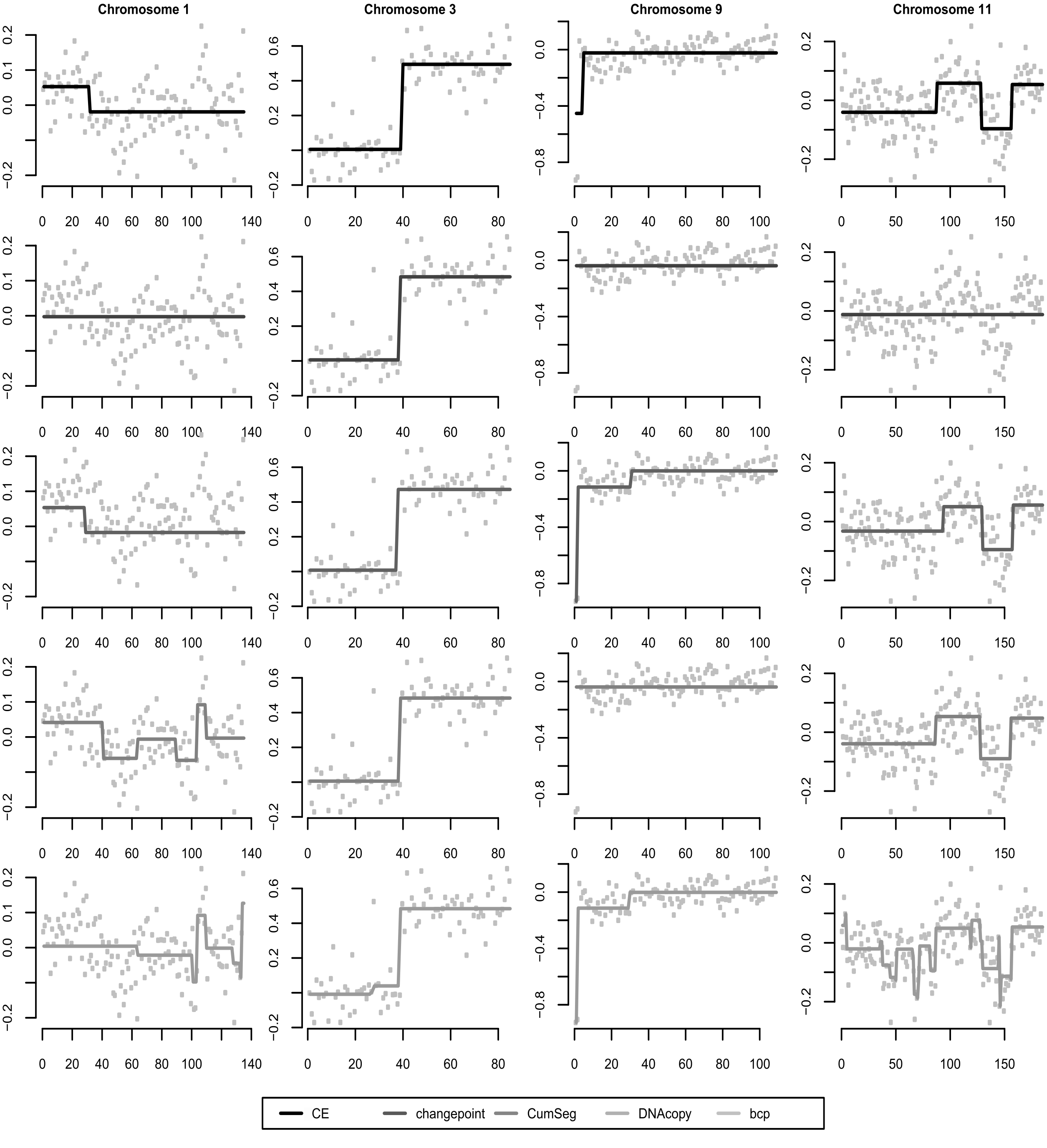


FIGURE 3.5: Array-CGH profiles of chromosome 1, 3, 9 and 11 for fibroblast cell line GM03563 data. Each column represents a chromosome and each row refers to a different estimating methodology (i.e. row 1 - CE, row 2 - changepoint, row 3 - cumSeg, row 4 - DNAcopy and row 5 - bcp, respectively)

estimated ($\hat{N}$) and its accuracy, we considered the same methodology used in [26], i.e. $S_2 = -0.5|N_0 - \hat{N}|$. Finally, we considered the grand score of $S = S_1 + S_2$ as the sensitivity score for the break-point location estimation. If the true number of break-points (N_0) and its locations were correctly estimated the grand score (S) equals to N_0. Figure 3.4 shows the results. In terms of the accuracy of the estimated break-point locations, it can be observed that the proposed methodology performs equally well or better than the competing methods as the SNR value increases. However, at lower aberration widths with lower SNR values, it is observed that the *DNAcopy* and *bcp* performs marginally better than the other methods. A notable performance difference can be observed in the *changepoint* method. It has significantly underperformed at lower SNR values irrespective of the aberration widths. Furthermore, it can be observed that at lower aberration widths the *changepoint* method has underperformed even at higher SNR values. It is also visible that *cumSeg* and *bcp* behave in a similar pattern except at lower aberration widths. In lower aberration widths and at lower SNR values *bcp* estimates the locations of the break-points more precisely than the *cumSeg* even though it has a lower sensitivity in estimating the break-point number. The performance of the *DNAcopy* method closely follows the proposed CE method, except at the lower SNR values and lower abberation widths the *DNAcopy* method performs marginally better. In general, *DNAcopy* shows a relatively high level of sensitivity in estimating the true number of break-points followed by the CE method at lower SNR values as compared to the other methods.

Validation on real aCGH data

Example 1: Fibroblast cell lines (GM03563) data

We use the fibroblast cell lines data set, which has already been discussed by several authors in [26, 52]. These data has been introduced in [45] and freely available to download at `http://www.nature.com/ng/journal/v29/n3/suppinfo/ng754_S1.html`. This data set has been referred as “Coriell data set” in [52] and it consists of single experiments on 15 fibroblast cell lines. We analyzed the data in the fibroblast cell line GM03563 with respect to the chromosomes $1, 3, 9, 11$. In this data set break-point locations were already known and verified by the spectral karyotyping method as well. The real alterations were only found in chromosome 3 and 9 with a single break-point ($N = 1$). Refer [26] for a detailed analysis on this data and a comparison study with few other well known methods. Figure 3.5 shows the results.

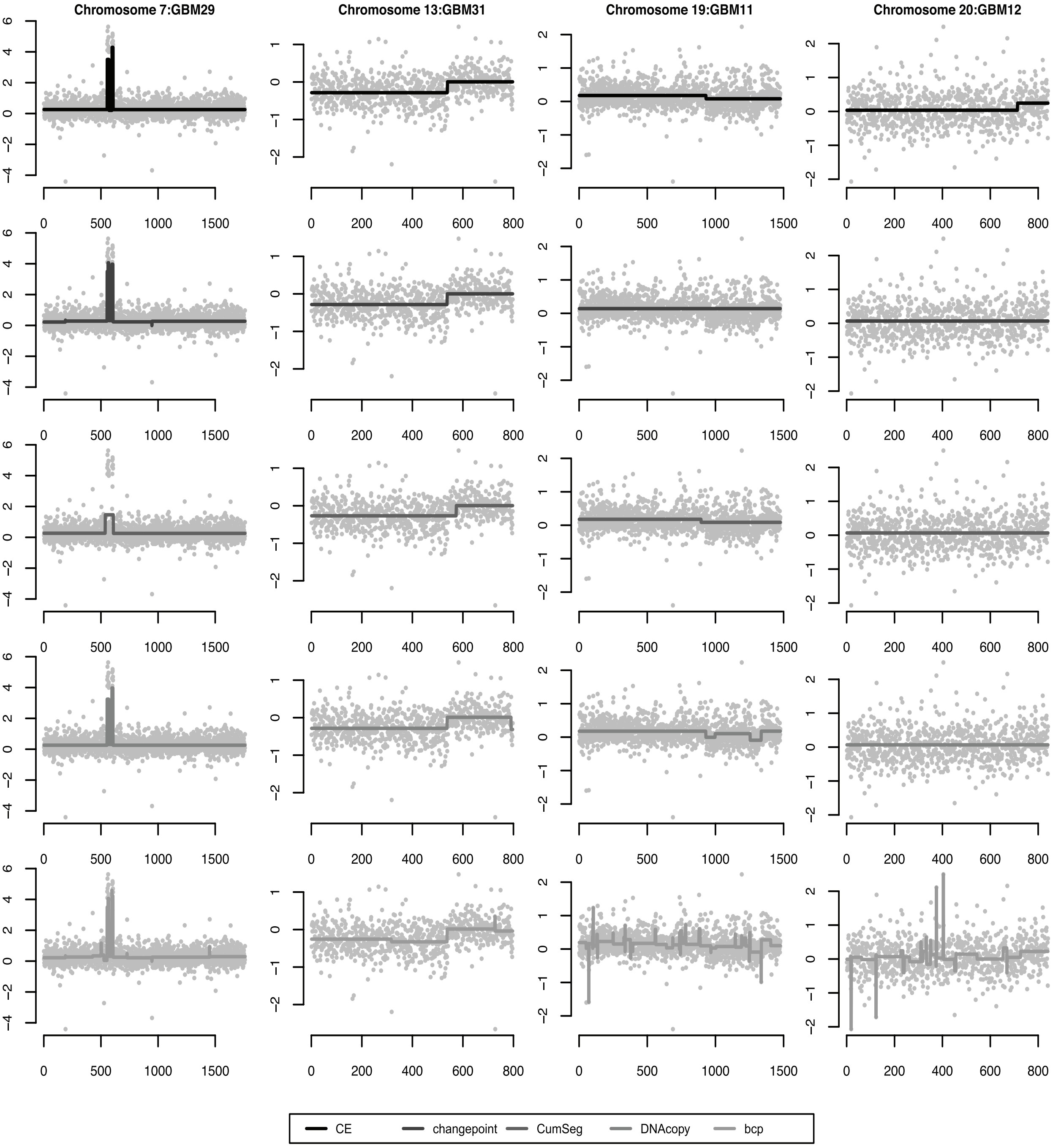


FIGURE 3.6: Array-CGH profiles of chromosome 7 (GBM29), 13 (GBM31), 19 (GBM11) and 20 (GBM12). Each column represents a chromosome and each row refers to a different estimating methodology (i.e. row 1 - CE, row 2 - changepoint, row 3 - cumSeg, row 4 - DNAcopy and row 5 - bcp, respectively)

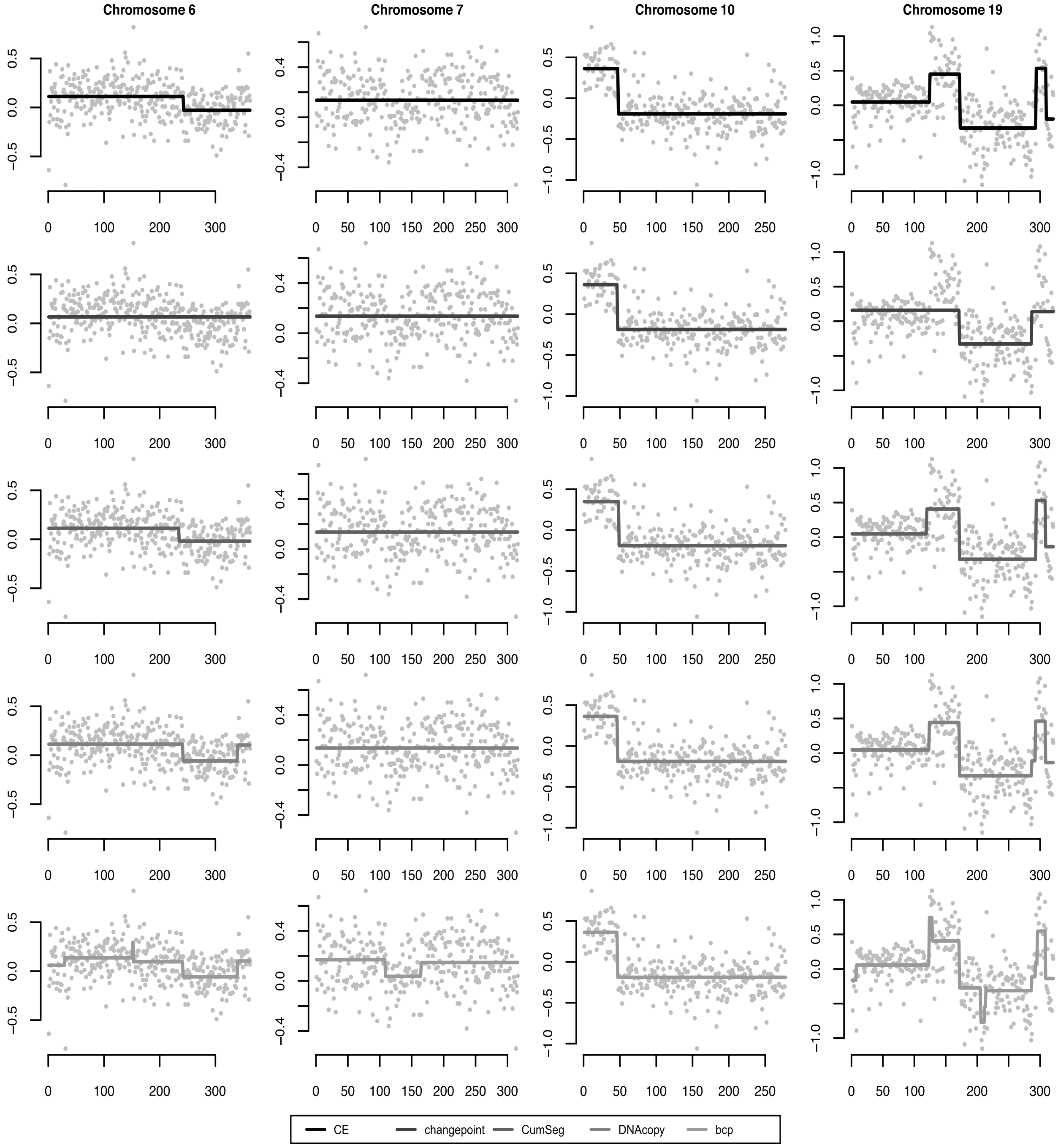


FIGURE 3.7: Array-CGH profiles for chromosome 1, 3, 9 and 11 of breast cancer cell lines (MDA157) data. Each column represents a chromosome and each row refers to a different estimating methodology (i.e. row 1 - CE, row 2 - changepoint, row 3 - cumSeg, row 4 - DNAcopy and row 5 - bcp, respectively)

In chromosome 9 only the proposed methodology CE has correctly identified the true number of break-points as 1. Where as *changepoint* and *DNAcopy* have estimated zero break-points leading to an under-estimation. The *cumSeg* and bcp methods have estimated two break-points for chromosome 9 resulting in an over-estimation. Furthermore, in [26], it has been identified that all other competing methodologies considered have also failed to identify the correct number of break-points in chromosome 9. However, in chromosome 3 all of the methods have correctly estimated N as 1 except for *bcp* which over-estimates the true number. In chromosome 1 and 11 except for *changepoint* all other methods have over-estimated the true number of break-points. For chromosome 1 CE and *cumSeg* have over-estimated the true number by 1 whereas both *DNAcopy* and *bcp* have over-estimated the true number significantly. Processing time of the proposed CE method and the other methods are detailed in the supplementary material: Section 3.1.

Example 2: Glioblastoma Multiforme (GBM) data

In this example we considered cDNA (complementary DNA) microarray based comparative genomic hybridization data of glial brain tumours, which was introduced in [5]. The original study considered 54 gliomas (GBM) of varying histogenesis and tumour grade. This data set has already been discussed in the literature by several authors (see [19], [26]). GBM is a type of malignant brain tumour. We considered the GBM29, GBM31, GBM11 and GBM12 cell lines data for the analysis. Particularly, we analysed chromosome 7 in GBM29, chromosome 13 in GBM31, chromosome 19 in GBM11 and chromosome 20 in GBM12 data.

It has been found in numerous microarray studies on gliomas that there exists copy number variation in chromosomes 7, 10, 13, 19 and 22 ([5, 22]) as a mix of losses and gains. Especially in GBM31 it has been identified a large region of loss on chromosome 13. Figure 5.8 shows the results. In GBM31 CE, *changepoint* and *cumSeg* show a similar profile, where as the other methods have identified more break-points. However, in GBM29 except for CE, *cumSeg* and *changepoint* all other methods have identified more segments. In GBM11, only CE and *cumSeg* have estimated one break-point. Thus, they both share a similar profile. It is observed that *changepoint* failed to estimate any break-point, where as the the other two methods have estimated two or more break-points. In GBM12 only the CE and *bcp* have estimated break-points. It can be noted that *bcp* has identified several single-probe outliers in GBM29, GBM11 and GBM12. These outliers can be a result of a real aberration, error in the experiment or some type of polymorphism as discussed in [19]. Furthermore, it is observed that in GBM11 and GBM12 *bcp* estimates a large number of break-points as compared to the other methods. Refer Section 3.2 of the supplementary material for processing time information.

Example 3: Breast cancer cell line (MDA157) data

We considered the breast cancer cell line (MDA157) data which has been discussed in [26, 26, 40]. The cDNA microarray CGH was profiled across 6691 mapped human genes in 44 breast tumor samples and 10 breast cancer cell lines. This dataset can be downloaded from `http://www.pnas.org/content/99/20/12963/suppl/DC1`.

We applied the proposed methodology as well as the other methods on four chromosomes (6, 7, 10 and 19) of breast cancer cell line MDA157. Figure 5.4 shows the aCGH profiles for the four chromosomes. We observe that the proposed CE method performs quite similar to the *DNAcopy* and *cumSeg*. In chromosome 7 none of the methods except for *bcp* have identified any break-point. Furthermore, *changepoint* method has not detected any break-point in chromosome 6, where all other methods have estimated at least one break-point and share a similar profile. In general, *bcp* method has over-estimated the number of break-points in most of the cases, especially in chromosome 19 it is more evident. Processing time information is detailed in the supplementary material: Section 3.3.

Discussion and Future Directions

In this study, we proposed an improved parallel implementation of the CE method, which is a model-based stochastic optimization procedure, to detect multiple break-points in array-CGH data with high level of accuracy. The proposed methodology was applied to both artificially generated data and real aCGH experiment data to assess the performance and to signify its effectiveness.

The procedure discussed in this paper concentrates especially on the task of estimating the **true locations** of the break-points with a high level of precision than the estimation of the number of break-points. It was observed in both artificial and real data the CE method performs equally well or better than the competing methods in terms of the accuracy of the estimated locations of break-points. However, in terms of the computational time the proposed CE method is not as efficient as the other methods considered in this study (refer supplementary material for details on the computational time). This is mainly due to the fact that the CE method is an evolutionary computing technique. Therefore, it naturally inherits a higher order of computing resources as compared to the most of the other methods.

Current implementation of the methodology does not detect single copy number changes (i.e. trisomies and monosomies) and it is developed to detect changes of an aberration width (h) of at least two probes. In the CE algorithm the default values of some of the parameters can be altered to obtain the desired level of accuracy. For instance increasing the sample size (M) render better estimates of the break-point locations as illustrated in Figure 5.2. However, as a result of that it will increase the processing time (Figure 5.3).

While the results of this work are encouraging, there are plenty of opportunities available for future research work. Especially on improving the processing time of the algorithm by means of limiting the search space for estimating the number of break-points (N). Present implementation of the method is developed as an exact search method, which considers all possibilities from no break-point to the user specified maximum value for the number of break-points. In the algorithm we call this maximum number of break-points as N_{max}. Even though this approach addresses the problem comprehensively, it makes the process more computationally expensive as well. In order to overcome this computational issue, in this study a parallel implementation of the methodology is carried out in multiple cores utilising the parallel [35], doMC [11] and doSNOW [10] R packages . These parallel computation techniques can be performed in Unix/Linux/MAC OS X and Windows operating systems. All the computations are carried out in the R statistical computing environment [35] and the proposed methodology is freely available as an R package ("breakpoint") from the Comprehensive R Archive Network (http://CRAN.R-project.org/package=breakpoint).

We have initiated further investigations on this problem and propose several other directions that can be utilized to overcome the issue of efficiency in the CE method apart from the parallel implementation. For example, the use of faster sequential techniques [38] to obtain initial estimates for the number of break-points and then use that information as the input for the CE algorithm to estimate the locations of the break-points. This will directly affect the performance of the methodology in estimating the number as well as the locations of the break-points, since it will limit the search space for the number of break-points. Thus, making it more efficient.

Acknowledgment

We would like to thank editors and referees for their constructive comments that improved this manuscript. W. J. R. M. Priyadarshana acknowledges the funding received towards his PhD from the International Macquarie University Research Excellence (iMQRES) scholarship.

Bibliography

[1] D. G. Albertson, C. Collins, F. McCormick and J. W. Gray, "Chromosome aberrations in solid tumors," *Nature Genetics*, vol. 34, pp. 369–376, 2003.

[2] D. Barry and J. A. Hartigan, "Product partition models for change point problems," *The Annals of Statistics*, vol. 20, pp. 260–279, 1992.

[3] J. V. Braun, R. K. Braun and H.-G. Müller, "Multiple change point fitting via quasilikelihood, with application to DNA sequence segmentation," *Biometrika*, vol. 2, pp. 301–314, 2000.

[4] J. V. Braun and H. G. Müller, "Statistical methods for DNA sequence segmentation," *Statistical Science*, vol. 13, pp. 142–162, 1998.

[5] M. Bredel, C. Bredel, D. Juric, G. R. Harsh, H. Vogel, L. D. Recht and B. I. Sikic, "High-Resolution Genome-Wide Mapping of Genetic Alterations in Human Glial Brain Tumors," *Cancer Research*, vol. 65, pp. 4088–4096, 2005.

[6] P. J. Campbell, P. J. Stephens, E. D. Pleasance, S. O'Meara, H. Li, T. Santarius, L. A. Stebbings, C. Leroy, S. Edkins, C. Hardy, J. W. Teague, A. Menzies, I. Goodhead, D. J. Turner, C. M. Clee, M. A. Quail, A. Cox, C. Brown, R. Durbin, M. E. Hurles, P. A. Edwards, G. R. Bignell, M. R. Stratton and P. A. Futreal, "Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing," *Nature Genetics*, vol. 40, pp. 722–729, 2008.

[7] N. Carter, "Methods and strategies for analyzing copy number variation using DNA microarrays," *Nature Genetics*, vol. 39, pp. S16–S21, 2007.

[8] A. Costa, O. D. Jones and D.P. Kroese, "Convergence Properties of the Cross-Entropy Method for Discrete Optimization," *Operations Research Letters*, vol. 35, pp. 573–580, 2007.

[9] C. Erdman and J. W. Emerson," A Fast Bayesian Change Point Analysis for the Segmentation of Microarray Data," *Bioinformatics*, vol. 24, pp. 2143–2148, 2008.

[10] Revolution Analytics,“ doSNOW: Foreach parallel adaptor for the snow package,” R package version 1.0.9, 2013.

[11] Revolution Analytics,“ doMC: Foreach parallel adaptor for the multicore package,” R package version 1.3.2, 2013.

[12] G. E. Evans, G. Y. Sofronov, J. M. Keith and D. P. Kroese, “Identifying change-points in biological sequences via the cross-entropy method,” *Annals of Operation Research*, vol. 189, pp. 155—165, 2011.

[13] M. Fanciulli, P. J. Norsworthy, E. Petretto, R. Dong, L. Harper, L. Kamesh, J. M. Heward, S. C. Gough, A. de Smith, A. I. Blakemore, P. Froguel, C. J. Owen, S. H. Pearce, L. Teixeira, L. Guillevin, D. S. Graham, C. D. Pusey, H. T. Cook, T. J. Vyse and T. J. Aitman, “FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity,” *Nature Genetics*, vol. 39, pp. 721–723, 2007.

[14] Revolution Analytics and S. Weston,“ foreach: Foreach looping construct for R,” R package version 1.4.1, 2013.

[15] L. Feuk, A. R. Carson and S. W. Scherer, “Structural variation in the human genome,” *Nature Genetics*, vol. 7, 85–97, 2006.

[16] E. Gonzalez, H. Kulkarni, H. Bolivar, A. Mangano, R. Sanchez, G. Catano, R. J. Nibbs, B. I. Freedman, M. P. Quinones, M. J. Bamshad, K. K. Murthy, B. H. Rovin, W. Bradley, R. A. Clark, S. A. Anderson, R. J. O’connell, B. K. Agan, S. S. Ahuja, R. Bologna, L. Sen, M. J. Dolan and S. K. Ahuja, “The influence of CCL3L1 gene-containing segmental duplications on HIC-1/AIDS susceptibility,” *Science*, vol. 307, pp. 1434–1440, 2005.

[17] D. C. Hoaglin, F. Mosteller and J. W. Tukey, *Understanding Robust and Exploratory Data Analysis*. John Wiley and Sons Inc., New York, 1983.

[18] T. Huang, B. Wu, P. Lizardi and H. Zhao, “Detection of DNA copy number alterations using penalized least squares regression,” *Bioinformatics*, vol. 21, pp. 3811–3817, 2005.

[19] E. Hyman, P. Kauraniemi, S. Hautaniemi, M. Wolf, S. Mousses, E. Rozenblum, M. Ringnér, G. Sauter, O. Monni, A. Elkahloun, O. P. Kallioniemi and A. Kallioniemi, “Impact of DNA amplification on gene expression patterns in breast cancer,” *Cancer Research*, vol. 62, pp. 6240–6245, 2002.

[20] R. Killick, P. Fearnhead and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association*, vol. 107, pp. 1590–1598, 2012.

[21] S. J. Knight, R. Regan, A. Nicod, S. W. Horsley, L. Kearney, T. Homfray, R. M. Winter, P. Bolton and J. Flint, "Subtle chromosomal rearrangements in children with unexplained mental retardation," *The Lancet*, vol. 354, pp. 1676–1681, 1999.

[22] R. Koschny, T. Koschny, U. G. Froster, W. Krupp and M. A. Zuber, "Comparative genomic hybridization in glioma: a meta-analysis of 509 cases," *Cancer Genet. Cytogenet.*, vol. 135, pp. 147–159, 2002.

[23] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.

[24] W. R. Lai, M. D. Johnson, R. Kucherlapati and P. J. Park, "Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data," *Bioinformatics*, vol. 21, pp. 3763–3770, 2005.

[25] L. Margolin, "On the convergence of the cross-entropy method," *Ann. Oper. Res.*, vol. 134, 201–214, 2004.

[26] V. M. Muggeo and G. Adelfio, "Efficient change point detection for genomic sequences of continuous measurements," *Bioinformatics*, vol. 27, 161–166, 2011.

[27] A. B. Olshen, E. S. Venkatraman, R. Lucito and M. Wigler, "Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostatistics*, vol. 5, pp. 557–572, 2004.

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

[29] T. V. Polushina and G. Y. Sofronov, "A Hybrid Genetic Algorithm for Change-Point Detection in Binary Biomolecular Sequences," In Proc. IASTED International Conference Artificial Intelligence and Applications (AIA 2013), pp. 1–8, 2013.

[30] W. J. R. M. Priyadarshana and G. Sofronov, "A Modified Cross Entropy Method for Detecting Multiple Change Points in DNA Count Data," In Proc. IEEE World Congress on Computational Intelligence (CEC'2012), pp. 1020–1027, 2012.

[31] R Core Team:R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2013. http://www.R-project.org/

[32] R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. González, M. Gratacós, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer and M. E. Hurles, "Global variation in copy number in the human genome," *Nature*, vol. 444, pp. 444–454, 2006.

[33] R. Rubinstein and D. P. Kroese, *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Springer-Verlag, New York, 2004.

[34] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[35] J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y. H. Lee, J. Hicks, S. J. Spence, A. T. Lee, K. Puura, T. Lehtimäki, D. Ledbetter, P. K. Gregersen, J. Bregman, J. S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M. C. King, D. Skuse, D. H. Geschwind, T. C. Gilliam, K. Ye and M. Wigler, "Strong association of De Novo copy number mutations with autism," *Science*, vol. 316, pp. 445–449, 2007.

[36] A. M. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A. K. Hindle, B. Huey, K. Kimura, S. Law, K. Myambo, J. Palmer, B. Ylstra, J. P. Yue, J. W. Gray, A. N. Jain, D. Pinkel and D. G. Albertson, "Assembly of microarrays for genome-wide measurement of DNA copy number," *Nature Genetics*, vol. 29, pp. 263–264, 2001.

[37] L. Tierney, A. J. Rossini, N. Li and H. Sevcikova," snow: Simple Network of Workstations," R package version 0.3-13, 2013.

[38] G. Sofronov, T. Polushina and W. J. R. M. Priyadarshana, "Sequential Change-Point Detection via the Cross-Entropy Method," In: B. Reljin, S. Stankovic (Eds.) The 11th Symposium on Neural Network Applications in Electrical Engineering (NEUREL2012), pp. 185–188, 2012.

[39] G. Sofronov, G. E. Evans, J. M. Keith and D. P. Kroese, "Identifying Change-points in Biological Sequences via Sequential Importance Sampling," *Environmental Modeling & Assessment*, vol. 14, pp. 577–584, 2009.

[40] R. Tibshirani and P. Wang, "Spatial smoothing and hot spot detection for CGH data using the fused lasso," *Biostatistics*, vol. 9, pp. 18–29, 2008.

[41] WTCCC: "Genome-wide association study of copy number variation in 16,000 cases of eight common diseases and 3,000 shared controls," *Nature*, vol. 464, pp. 713–720, 2010.

[42] N. R. Zhang, D. O. Siegmund, H. Ji and J. Z. Li, "Detecting simultaneous change-points in multiple sequences," *Biometrika*, vol. 93, pp. 631–645, 2010.

[43] N. R. Zhang and D. O. Siegmund, "A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data," *Biometrics*, vol. 63, pp. 22–32, 2007.

W. J. R. M. Priyadarshana received his BSc degree in Statistics from the University of Colombo, Sri Lanka, in 2009. He is currently a PhD candidate in Statistics at the Department of Statistics, Macquarie University, Sydney, Australia. His research interests are in the areas of break-point modelling, Cross-Entropy method, generalized linear models, statistical genetics and applied statistics.

Georgy Sofronov obtained his PhD in Probability Theory and Mathematical Statistics from Moscow State University in 2002. He has held teaching and research positions at Mari State University, Russia, the University of Queensland, Australia and the University of Wollongong, Australia. Currently he is a Senior Lecturer in Statistics at Macquarie University, Australia. His research interests include Markov chain Monte Carlo simulation, the Cross-Entropy method, DNA sequencing, small area estimation, change-point problem and optimal stopping rules.

Chapter 4

Hybrid Algorithms for Multiple Change-Point Detection in Biological Sequences

This chapter constitutes a peer-reviewed book chapter published in “Advances in Experimental Medicine and Biology”. The content of the book chapter including notation, plot sizes and wording has been slightly changed to improve the flow and the overall coherence of the thesis. The citation for the book chapter is:

Priyadarshana, W. J. R. M., Polushina, T., and Sofronov, G. (2015). Hybrid algorithms for multiple change-point detection in biological sequences. In Sun, C., Bednarz, T., Pham, T., Vallotton, P., and Wang, D., (Eds.), Signal and Image Analysis for Biomedical and Life Sciences, volume 823 of Advances in Experimental Medicine and Biology. Springer International Publishing. http://link.springer.com/chapter/10.1007/978-3-319-10984-8_3.
Writing: 85% , analysis: 85% , conception and design: 75%

Specific contribution of joint authors
Tatiana Polushina: Contributions to the analysis and writing the manuscript, particularly the section relating to the sequential change-point analysis.
Georgy Sofronov: Overall supervision of data analysis and interpretation. Provided continual feedback and suggestions on both analysis and writing.

Reprinted with permission from Priyadarshana, W. J. R. M., Polushina, T., and Sofronov, G. (2015) Signal and Image Analysis for Biomedical and Life Sciences, volume 823 of

Advances in Experimental Medicine and Biology, Copyright © 2015 Springer Science and Business Media (License number: 3512961325759).

4.1 Summary

This chapter describes a study in which we propose two hybrid algorithms, each combining a popular sequential detection technique that was discussed in Chapter 2 with the Cross-Entropy (CE) method, to detect multiple change-points in biological sequences of continuous measurements. This study can be considered as a major improvement to the work in Chapter 3.

The motivation of this study was to amalgamate powerful sequential detection techniques with the CE method to effectively and efficiently detect change-points in continuous biological sequences. In Chapter 3, the CE method is used as an exact search procedure, where it performs calculations for all possible solutions from no change-point to the maximum value of change-points that a user has provided. Thus, it consumes significant amount of computational resources and the CE method is not essentially optimized near the true change-point location. To alleviate this issue, we proposed to obtain initial estimates for the number and the locations of change-points by using powerful sequential detection methods. Then, we initiate the CE algorithm with these preliminary estimates to carry out the calculations and to obtain more refined final estimates.

The proposed improved implementations of the CE method, which incorporate sequential detection techniques, were applied particularly to detect change-points in aCGH data. In this study two hybrid algorithms were proposed [115, 116]. The first hybrid algorithm, which combines the cumulative sum (CUSUM [101, 102]) procedure and the CE method was named "CUSUM-CE". The second hybrid method was denoted as "SR- CE", combining both the Shiryaev-Roberts (SR) procedure [129, 141, 142] and the CE method. In the CE algorithm, the log-likelihood function was used as the performance function to obtain the final estimates. Furthermore, we conducted multiple hypothesis tests to identify the most significant change-points with an improved Bonferroni correction, which was proposed in [145]. The Bonferroni correction is used to control the family-wise error rate when conducting multiple hypothesis tests. The parallel implementation of the CE method discussed in Chapter 3 was considered in this study when developing the hybrid algorithms. A detailed simulation was carried out to study properties of the proposed procedures and to analyze their segmentation capabilities. Finally, the proposed two hybrid methods were utilized to find CNVs in two aCGH data sets. In order to compare the performance of the two procedures with other methods, the four well-known procedures (*DNACopy, cumSeg, changepoint, bcp*) described in Chapter 3 and the general CE procedure were considered.

The first example is on the fibroblast cell lines data [150] as considered in Chapter 3. We analyzed chromosomes 1, 3 and 7 of the GM03563 cell lines data. The second example is on the breast cancer cell lines data originally discussed in [108]. We have applied the proposed procedures on chromosomes 3, 5, 9 and 13 of MDA157 cell lines data. The proposed two hybrid methods have performed significantly better than the CE method proposed in Chapter 3, predominantly in efficiency, where the proposed hybrid approaches significantly improve computational time. Particularly, the CUSUM-CE method's improvement in processing time was more than two-fold. In general, it was found, both in artificial and real data analysis, that the proposed two approaches have performed evenly or better than the other competing methods.

Hybrid Algorithms for Multiple Change-Point Detection in Biological Sequences

W. J. R. M. Priyadarshana[1,*], Tatiana Polushina[2] and Georgy Sofronov[1]

[1] Department of Statistics, Faculty of Science, Macquarie University, Sydney NSW 2109, Australia.

[2] Department of Clinical Science, Faculty of Medicine and Dentistry, University of Bergen, Postboks 7804, NO-5020 Bergen, Norway. This work was carried out when the author was at the Department of Mathematics, Mari State University, Russia.

[*]E-mail: madawa.weerasinghe@mq.edu.au

Abstract

Array comparative genomic hybridization (aCGH) is one of the techniques that can be used to detect copy number variations in DNA sequences in high resolution. It has been identified that abrupt changes in the human genome play a vital role in the progression and development of many complex diseases. In this study we propose two distinct hybrid algorithms that combine efficient sequential change-point detection procedures (the Shiryaev-Roberts procedure and the cumulative sum control chart (CUSUM) procedure) with the Cross-Entropy method, which is an evolutionary stochastic optimization technique to estimate both the number of change points and their corresponding locations in aCGH data. The proposed hybrid algorithms are applied to both artificially generated data and real aCGH experimental data to illustrate their usefulness. Our results show that the proposed methodologies are effective in detecting multiple change-points in biological sequences of continuous measurements.

Background

Change-point problems (or disorder problems, break-point problems) are used to model heterogeneity in sequences of observations. This is essential in order to understand the underlying properties of a process as a part of the statistical diagnosis of data. Primarily it serves the purpose of checking and validating the homogeneity assumption of the data, which is one of the main assumptions in statistical modelling. Thus, accounting for these changes facilitates more improved and reliable estimates for unknown parameters. This is an imperative step in statistical modelling directly associated with the decision making process. Change-point detection problem has received increasing attention due to these reasons and has attracted wide range of applications in many scientific streams. These change-point models are employed in health informatics, financial and economic data analysis, signal processing, oceanographic studies, quality control, surveillance analysis, etc.

In health informatics, detection and characterization of genomic structural variations are essential in identifying disease causing genes that have functional importance in exemplifying genome-wide complex diseases, such as cancer, autism, immune disorders, etc. These structural variations in the human genome can be acquired somatically in the lifespan as well as be inherited through germline. Copy number variation (CNV) is one of the common and major types of structural variations in the human genome. CNV is defined as a DNA segment that is 1kb or larger and present at variable copy number in comparison with a reference genome [9]. It is identified in multiple studies that CNV plays an important role in genetic susceptibility to common diseases [27, 38]. There are multiple platforms and procedures built to detect CNV in different perspectives [3, 12, 20, 54]. The array comparative genomic hybridization (aCGH) is a popular and a widely used methodology to detect CNVs in genome-wide studies. It is developed on the principles of the conventional comparative genomic hybridization (CGH) technique [1], which produces a map of DNA sequence copy number with respect to the chromosomal location. The CGH technique was firstly developed to detect copy number changes in solid tumors. In CGH experiments, the differentially labeled test and control genomes are hybridized to metaphase chromosomes. The fluorescent signal intensity of the test DNA relative to the reference DNA along the chromosome is linearly plotted to identify CNVs. The aCGH technique uses slides arrayed with small segments of DNA as the targets for analysis [20] in contrast to the use of metaphase chromosomes in CGH. The aCGH technique offers high resolution for CNV detection. Moreover, simultaneously detection of different alterations types is one of the advantages of the CGH technique [50]. Furthermore, it has been proven that aCGH is a powerful tool for detecting submicroscopic chromosomal abnormalities in individuals with idiopathic mental retardation and various birth defects.

There is a large amount of literature on CNV detection in aCGH data. A method based on fitting a mixture of three Gaussian distributions corresponding to gain, loss and normal regions is considered in [11]. Later, a test based on moving averages proposed in [26] to compute a threshold level to detect CNVs. In [24], a modified version of the circular binary segmentation [40] introduced. Their methodology is termed as circular binary segmentation (CBS) method. A test based on the maximum of a likelihood ratio is used in the CBS to detect CNVs. The method discussed in [24] is employed in the popular *DNAcopy* R package [35, 52]. Different methods based on hidden Markov models (HMMs) introduced in [38, 45]. Furthermore, a fast Bayesian change-point detection method based on the product partition models [16] introduced in [48] and it is deployed in the *bcp* R package [6]. A different approach for the problem discussed in [21] which uses the "lars" algorithm [5] and a generalized version [53] of the BIC [135] to estimate change-points in aCGH data. The methodology is freely available in the *cumSeg* R package [22]. Recently, a Pruned Exact Linear Time (PELT) method is introduced in [16]. The *changepoint* R package [17] employs the methods discussed in [16]. Readers are referred to [19] for a detailed review on the segmentation methods on aCGH data.

Detection of CNVs falls into the posteriori (retrospective or off-line) class of change-point problems. In the posteriori change-point problem the data set is fixed and it is not getting changed periodically as in the sequential (quickest or on-line) change-point problem. There exists an extensive literature on both of these main classes of change-point problems. Readers are referred to [15, 23, 27, 32, 34, 46, 47, 113] for a detailed review on some of the techniques. The quickest change-point problem, a sequence of random variables is observed on-line, that is, the future observations are not known. Initially, we assume that the sequence considered is in so-called "controlled" state. But at some unknown moment a breakage occurs and the sequence runs "out of control". The objective of sequential change-point analysis is to detect this breakage (change-point) as soon as possible with a minimum number of false alarms. There are two well-known sequential procedures discussed in the literature: the Shiryaev-Roberts (SR) procedure [36, 41–43] and the Cumulative Sum (CUSUM) procedure [25].

The process of change-point analysis in both the retrospective and sequential change-point methods deals with two main issues: detecting number of change-points and estimating their locations. In this chapter, we propose novel hybrid algorithms that combine sequential change-point techniques and the Cross-Entropy (CE), which is a model based stochastic optimization technique. We emphasize that the hybrid algorithm in [113] is based on a genetic algorithm and a local search procedure, whereas the proposed method uses sequential change-point techniques and the CE algorithm. Our method utilizes a sequential change-point detection methodology to provide initial estimates on the number as well as the locations of the change-points. Based on the initial estimates the CE method is initiated to optimize the solution to provide more accurate estimates of the number as well as their corresponding locations. We propose two new algorithms within this framework. The first approach, which combines the SR procedure and the CE method, will be referred as the "SR-CE". The second approach combines the CUSUM procedure and the CE method. We will refer to this method as the "CUSUM-CE". In this study, we apply the proposed algorithms to aCGH data in order to detect CNVs. Notice that the new hybrid algorithms can easily be extended or modified to solve change-point problems in other research fields.

This chapter is organized as follows. First, we describe the multiple change-point problem. Then we provide details on the proposed hybrid algorithms, quickest change-point detection methods and the CE method. In the numerical results section, we present results on simulated data and two publicly available real data sets. Finally, in the discussion and conclusions section, we consider the strengths and limitations of the proposed methodology and conclude the paper with future research directions.

Multiple Change-Point Problem

Let us consider a sequence of observations $\mathbf{X} = (x_1, x_2, \ldots, x_L)$ of length L, in which the x_i's are independently distributed Gaussian random variables. A segmentation of the sequence is specified by the number of change-points N and the corresponding locations of the change-points $\mathbf{C} = (c_1, c_2, \ldots, c_N)$, where $1 = c_0 < c_1 < \cdots < c_N < c_{N+1} = L + 1$. A change-point is defined as a boundary between two adjacent segments in this context. The value of c_i is the sequence position of the rightmost character of the segment to the left of the i-th change-point. The segments are numbered from 0 to N as there will be one or more segments than the number of change-points. The model assumes that within each segment the observations are distributed as normal with mean μ_i, $i = 0, 1, \ldots, N$ and variance σ^2. Both mean and variance are not known in advance and maximum likelihood method is used to obtain estimates. The joint distribution of $\mathbf{x}$ conditional on N, $\mathbf{C} = (c_1, c_2, \ldots, c_N)$, $\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots, \mu_N)$, and σ^2 is given by:

$$f(\mathbf{X} \mid N, \mathbf{C}, \boldsymbol{\mu}, \sigma^2) = \prod_{n=0}^{N} \left[\prod_{i=c_n}^{c_{n+1}-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x_i - \mu_n)^2}{2\sigma^2} \right\} \right]. \tag{4.1}$$

The corresponding log-likelihood of the model is

$$ll(\mathbf{X} \mid N, \mathbf{C}, \boldsymbol{\mu}, \sigma^2) = \sum_{n=0}^{N} \left[-\frac{\lambda_n}{2} \ln\left(2\pi\sigma^2\right) - \frac{1}{2} \sum_{i=c_n}^{c_{n+1}-1} \left(\frac{x_i - \mu_n}{\sigma} \right)^2 \right], \tag{4.2}$$

where the length of the n-th segment is defined as $\lambda_n = c_{n+1} - c_n$.

Framework of the Algorithms

The proposed algorithms combine a sequential procedure with the CE method to detect multiple change-points in biological sequences of continuous measurements. We consider both the SR procedure and CUSUM procedure to combine with the CE method to form a hybrid framework to detect multiple change-points. In general, the SR-CE and CUSUM-CE hybrid algorithms can be summarized as follows:

1. Run a sequential procedure (either SR or CUSUM) along the sequence of observations to obtain initial estimates for the number (N) as well as the locations ($\mathbf{C}$) of change-points.
2. Based on the estimates of N and $\mathbf{C}$, initiate the CE algorithm to obtain an optimized locations of change-points.
3. For all pairs of adjacent segments, perform a two sample t-test to identify the least important change-point that associated with the highest p-value with respect to the significance level (α). The p-values are adjusted on the Bonferroni correction [44] to control the family wise error rate in multiple hypothesis testing. Thus, we eliminate the least significant change-point from the solution and update the solution vector with the other estimates.
4. Initiate the CE algorithm with the new set of change-point locations.
5. Repeat steps 3 and 4 until all change-points found are significant. Return $\mathbf{C}$: the vector of change-point locations. The length of this vector is the number of change-points.

Quickest change-point detection

The sequential change-point problem can be described in mathematical terms as follows. Let $\{X_n\}_{n\geq 1}$ be independent random variables which are observed sequentially, one by one. Suppose that initially the sequence is in so-called "controlled" state for $n = 1, 2, \ldots, \tau - 1$, that is, the random variables are distributed with $f_0(x)$, a common normal probability density function with mean μ_0 and variance σ^2. At some unknown moment τ a breakage occurs and the observed sequence runs "out of control", which means that after the breakage (change-point) the probabilistic characteristics of the sequence are changed. From moment τ we observe random variables with $f_1(x)$, $f_1(x) \neq f_0(x)$, another normal probability density function with mean μ_1 and variance σ^2. Our objective is to detect the change-point as soon as feasible and with as few as possible false alarms. In other words, in the sequential change-point problem, we would like to detect the moment τ as quickly as possible after it has occurred and, at the same time, we would like to keep the rate of false alarms at a low predefined level.

There are two main cases in the sequential change-point problem [48]. In the simplest situation, we know the probability density functions before and after the breakage, which may be unrealistic assumption. In the second case, we assume that the $f_0(x)$ is known before the change-point, whereas the $f_1(x)$ is unknown. In what follows we assume that the μ_0 is known (it can be estimated from an archive of data) and the μ_1 is unknown and must be estimated from the data, the σ^2 is fixed. For the sake of simplicity of the formulas below (and without loss of generality), we can assume that $\mu_0 = 0$ and $\sigma^2 = 1$.

We have two statistical hypotheses: the null hypothesis H_0: there is no change-point versus the alternative hypothesis H_1 : a breakage happens at time $\tau = k \geq 0$. The sequential decision rule can be constructed as follows. Let $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$ be a vector of the first $n \geq 1$ values. The probability density functions of $\mathbf{X}_n$ under either of these hypotheses are given by

$$\begin{aligned} p(\mathbf{X}_n \mid H_0) &= \prod_{j=1}^{n} f_0(X_j), \\ p(\mathbf{X}_n \mid H_1) &= \prod_{j=1}^{k-1} f_0(X_j) \prod_{j=k}^{n} f_1(X_j), \quad k \leq n. \end{aligned}$$

Then we can calculate the likelihood ratio, which can be used to test H_0 versus H_1, as follows

$$LR_k = \prod_{j=k}^{n} \frac{f_1(X_j)}{f_0(X_j)} = \prod_{j=k}^{n} \frac{\frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(X_j - \hat{\mu}_1)^2\}}{\frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}X_j^2\}} = \exp\left\{ \frac{\left(\sum_{j=k}^{n} X_j\right)^2}{2(n-k+1)} \right\},$$

where $\hat{\mu}_1 = \sum_{j=k}^{n} X_j/(n-k+1)$, is an estimate of the μ_1 based on the last $n-k+1$ observations.

There are two common characteristics of a sequential detection procedure: the average run length (ARL) to false alarm (the expected number of values to an alarm assuming that there is no breakage) and the average delay to detection (the expected delay between a change and its detection). The objective is to find a sequential procedure that minimizes the average detection delay with restriction on the ARL to false alarm.

In this chapter, we consider two main procedures: the Shiryaev-Roberts (SR) procedure [36, 41–43] and the CUSUM procedure [25]. Various probabilistic properties of these methods are discussed in [28–30].

The SR procedure stops and raises an alarm at time

$$T_{A_{SR}} = \inf\{n \geq 1 : R_n \geq A_{SR}\},$$

where

$$R_n = \sum_{k=1}^{n} LR_k = \sum_{k=1}^{n} \exp\left\{ \frac{\left(\sum_{j=k}^{n} X_j\right)^2}{2(n-k+1)} \right\}, \quad n = 1, 2, \ldots$$

is the SR statistic, and A_{SR} is a positive threshold that controls the false alarm rate.

The stopping time of the CUSUM procedure is defined by

$$T_{A_C} = \inf\{n \geq 1 : W_n \geq A_C\},$$

where

$$W_n = \max_{1 \leq k \leq n} LR_k = \max_{1 \leq k \leq n} \exp\left\{ \frac{\left(\sum_{j=k}^{n} X_j\right)^2}{2(n-k+1)} \right\}, \quad n = 1, 2, \ldots$$

is the CUSUM statistic, and A_C is an unknown threshold that controls the false alarm rate in the CUSUM procedure.

In order to identify the thresholds A_{SR} and A_C we generate an artificial sequence with a single change-point and apply the SR and the CUSUM procedures for the sequence. We

TABLE 4.1: The average run length (ARL), CPU time and the probability of detecting the true change-point for different values of the threshold A_{SR}

A	700	800	900	1000	2000	3000	4000	5000	6000	7000
$\boldsymbol{p}$	0.411	0.432	0.472	0.493	0.629	0.679	0.715	0.739	0.765	0.779
ARL	108	107	109.3	109	110.5	112.8	113.2	115.7	113.5	116.7
CPU time	0.23	0.39	0.58	0.65	0.82	1.07	1.21	1.34	1.46	1.48

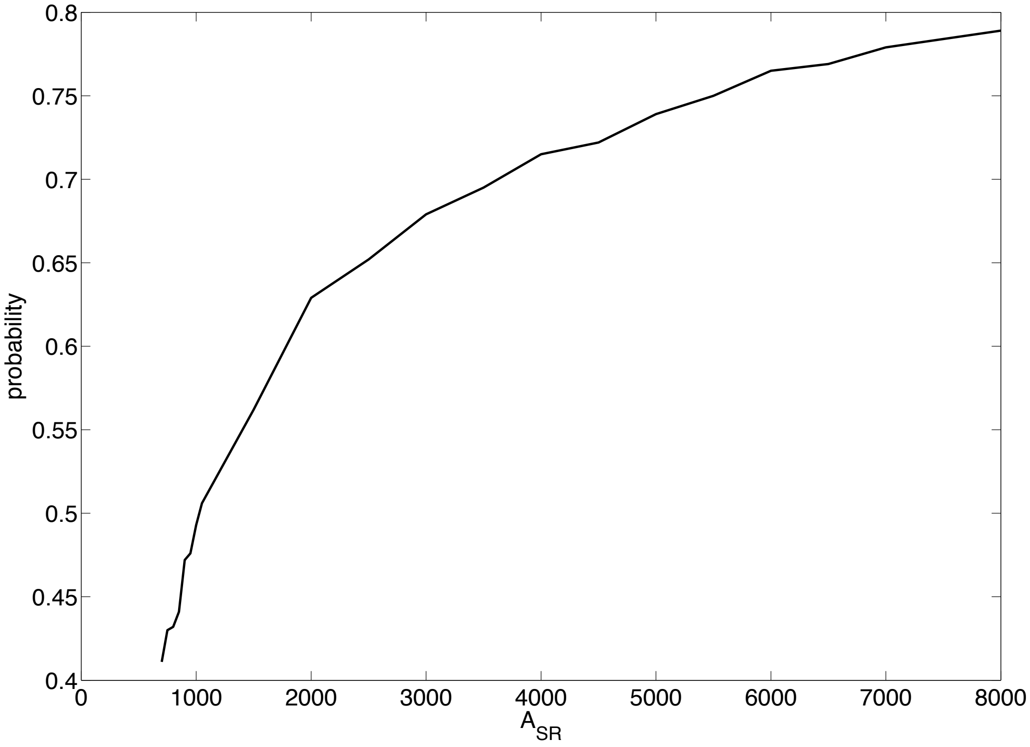


FIGURE 4.1: The probability of detecting the true change-point depending on the value of the threshold A_{SR}.

assume that the first several observations are in "controlled" state. Therefore, the estimates of the initial (unknown) parameters of the probability density function $f_0(x)$ can be obtained using these first observations. In this study, we consider random observations with normal distribution $\texttt{Normal}(\mu_0, 1)$ before the change-point and with $\texttt{Normal}(\mu_1, 1)$ after the change-point, where $\mu_0 = 0$, $\mu_1 = 1$. We use 2000 as the length of the sequence and 100 as the number of observations utilized for estimating the parameters of the initial distribution. After simulating this experiment 2000 times, it is clear that the threshold A_{SR} should be quite large (see Table 4.1). For instance, if $A_{SR} = 5000$, then the probability of detecting the true change-point is 0.739.

Note that p, the probability of detecting the true change-point, increases as the threshold A_{SR} increases (see Figure 4.1). We can conclude that for long sequences we should use

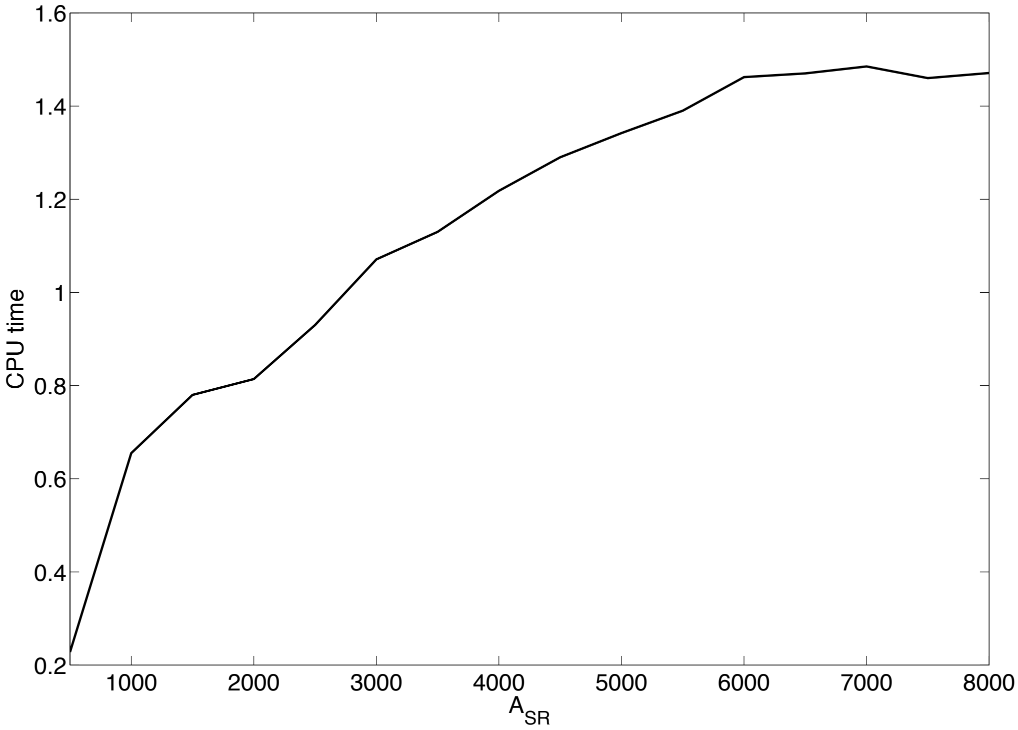


FIGURE 4.2: CPU time depending on the value of the threshold A_{SR}.

TABLE 4.2: The ARL and CPU time for different values of $\mu_1 - \mu_0$, $A_{SR} = 5000$

$\boldsymbol{\mu_1 - \mu_0}$	0.25	0.5	0.75	1	1.25	1.5	1.75	2	5	7
ARL	324.9	162.9	125.7	115.7	109.5	103.7	103.0	102.4	99.2	95.0
CPU time	2.10	1.94	1.65	1.39	1.13	1.02	0.92	0.94	0.87	0.85

large values of the threshold A_{SR}, for example, $A_{SR} > 5000$, in order to detect the change-point with a high probability.

TABLE 4.3: The ARL and CPU time for different values of $\mu_1 - \mu_0$, $A_{SR} = 7000$

$\boldsymbol{\mu_1 - \mu_0}$	0.25	0.5	0.75	1	1.25	1.5	1.75	2	5	7
ARL	351.8	172.7	129.7	116.7	108.7	107.4	106.2	105.5	102.1	99.0
CPU time	2.47	2.03	1.82	1.46	1.31	1.27	1.15	0.98	0.94	0.88

Since the estimates of the positions of the change-points will be used as the initial values for the CE method, we should emphasize that these estimates are found with some delay. Table 4.2 and Table 4.3 show how the ARL depends on the value of $\mu_1 - \mu_0$. These tables also demonstrate that the ARL is significantly large for small differences between μ_1 and μ_0. Note for rather long sequences this delay is not an issue whereas for relatively short sequences we should use a higher value of the A_{SR} and reduce the length of the region used for estimating unknown parameters of $f_0(x)$.

Table 4.4: The average run length (ARL) and the probability of detecting the true change-point for different values of the threshold A_C

A	1000	5000	7000	10000	12000	15000	18000	19000	20000	21000
$\boldsymbol{p}$	0.269	0.583	0.642	0.672	0.701	0.694	0.728	0.743	0.752	0.748
ARL	101.1	101.7	111.4	112.6	110.5	112.3	114.3	114.2	115.5	115.4
CPU time	0.63	1.17	1.26	6.047	1.34	1.54	1.48	1.48	1.49	1.53

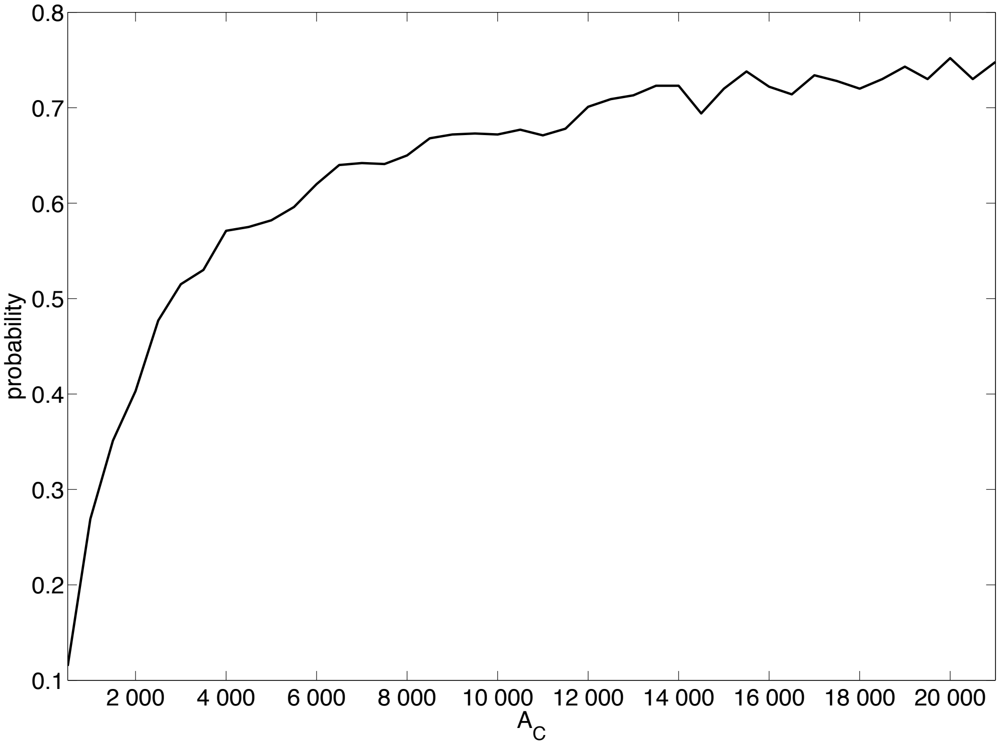


Figure 4.3: The probability of detecting the true change-point depending on the value of the threshold A_C.

We repeat the same simulation for the CUSUM procedure. Using the simulated sequences, we estimate the threshold A_C (note that A_C is significantly larger than A_{SR}) (see Table 4.4). For large values of A_C the probability of detecting the true change-point increases very slow. It is customary that the value of the False Discovery Rate (FDR) less than 0.25 is used as a popular threshold [49]. We use higher values of A_C and A_{SR}, since the probabilities of detecting the true change-point in the both cases are larger than 0.75.

The Cross-Entropy Method

The Cross-Entropy (CE) method [131] is a model-based evolutionary stochastic optimization framework which was originally developed as a method to estimate rare event

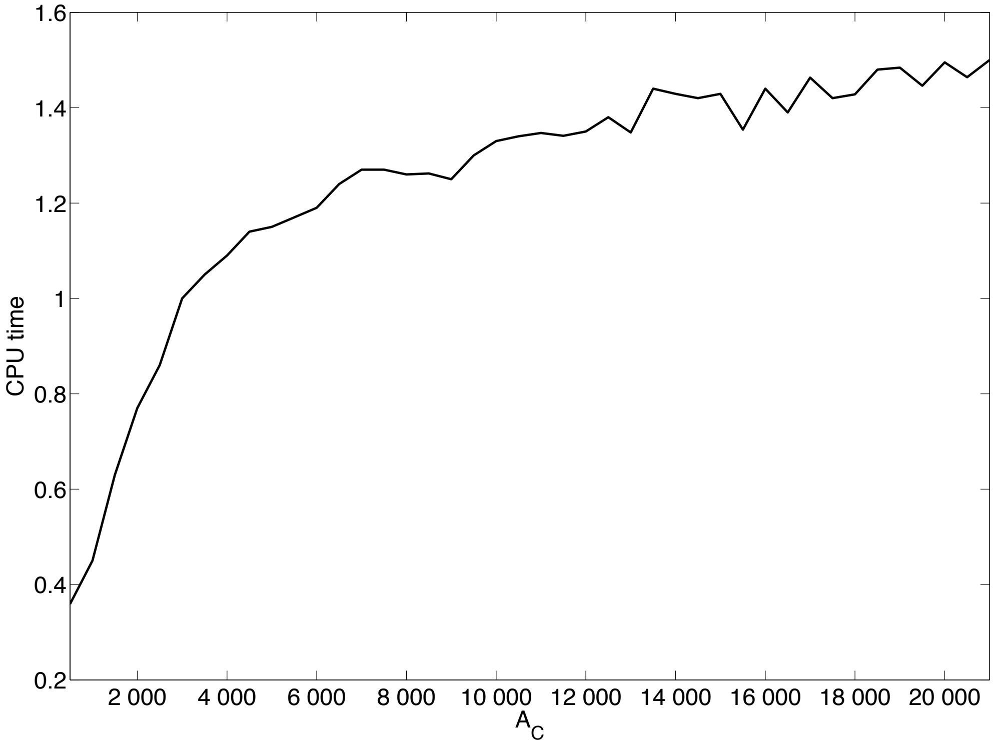


FIGURE 4.4: CPU time depending on the value of the threshold A_C.

probabilities. It can be used to solve both estimation and optimization problems. The CE method is developed on the basis of the Kullback-Leibler divergence [18]. The process of multiple change-point detection can be viewed as a combinatorial optimization problem. In the context of combinatorial optimization problems, the CE method is an iterative procedure that starts with a parametrized sampling distribution from which M number of random samples generated. Then, each combinatorial arrangement is scored for its performance using an objective function F. A fixed number of best performing combinatorial arrangements are selected based on the performance score and it is referred as the elite sample. We define the size of this elite sample as M_{elite}. Let us define $M_{elite} = \rho \times M$, where ρ is the elite sample fraction. The elite sample is used to update the parameters of the sampling distribution based on a smoothing rule. This process is iterated until a stopping criterion (SC) is met or user defined number of iterations. The sampling distribution eventually converges to a degenerate distribution about a locally optimal solution, which ideally will be globally optimal [38].

There are few parameters that have to be specified prior to the initialization of the CE method in the context of multiple change-point problem. They are the minimum aberration width (h), lower and the upper limit for the search space of number of change-points (let us define the lower limit as N_{min} and upper limit as N_{max}), sample size M, elite sample fraction ρ, smoothing parameter vector $\boldsymbol{\beta}$ and a cut-off value for the SC (ε). In this study, truncated normal distribution is utilized as the parametrized sampling distribution to simulate locations of the change-points based on the user defined minimum aberration width h. We simulate $M = 200$ number random solutions. The value of ρ is considered as 0.05, smoothing parameter values of β and γ [49] is used for μ and σ respectively and ε is set as 0.01. The performance function used in the study is the model log-likelihood based on the simulated change-point locations.

Based on these user defined set of parameters and the initial estimates from a sequential method, the CE algorithm can be summarized as below:

1. Set the change-point locations obtained from a sequential procedure as the initial values for the mean vector $\boldsymbol{\mu^0}$ and set all components of the standard deviation vector $(\boldsymbol{\sigma}^2)^0$ as 5^2 in order to simulate locations from the truncated normal distribution. Both vectors of parameters are N-dimensional. Set $t = 0$.

2. Increase t by 1. Simulate a random sample $\mathbf{C}^{(1)}, \mathbf{C}^{(2)}, \ldots, \mathbf{C}^{(M)}$ from $\texttt{Normal}(\boldsymbol{\mu}^{t-1}, (\boldsymbol{\sigma}^2)^{t-1})$ distribution, where $\mathbf{C}^{(i)} = (c_1^{(i)}, c_2^{(i)}, \ldots, c_N^{(i)})$, $i = 1, 2, \ldots, M$.

3. For each $i = 1, 2, \ldots, M$ order $c_1^{(i)}, \ldots, c_N^{(i)}$ from smallest to largest and set $\mathbf{C}^{(i)} = (c_1^{(i)}, c_2^{(i)}, \ldots, c_N^{(i)})$, where $\mathbf{C}^{(i)}$ is the change-point vector as defined earlier.

4. Evaluate the log-likelihood function (the performance score) of each $\mathbf{C}^{(1)}, \mathbf{C}^{(2)}, \ldots, \mathbf{C}^{(M)}$. Obtain the elite sample, which is the best performing combinations of the change-point locations. M_{elite} is the size of the elite sample.

5. For all $j = 1, 2, \ldots, N$ calculate maximum likelihood estimates of the mean and the standard deviation $\hat{\boldsymbol{\mu}}^{\boldsymbol{t}} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_N)^t$, $(\tilde{\boldsymbol{\sigma}}^{\mathbf{2}})^t = (\tilde{\sigma}_1^2, \tilde{\sigma}_2^2, \ldots, \tilde{\sigma}_N^2)^t$ by using the elite sample. Based on the smoothing rule update the parameters in the truncated normal distribution as below,

$$\boldsymbol{\mu}^{\boldsymbol{t}} = \beta\hat{\boldsymbol{\mu}}^t + (1-\beta)\hat{\boldsymbol{\mu}}^{t-1}, \quad (\boldsymbol{\sigma}^2)^t = \gamma(\tilde{\boldsymbol{\sigma}}^2)^t + (1-\gamma)(\tilde{\boldsymbol{\sigma}}^2)^{t-1}.$$

6. If the stopping criterion (SC) is met, then stop the process and identify the combination of the locations of change-points $\mathbf{C}^{(i)}$ that optimizes the performance function. Otherwise set $t = t + 1$ and iterate from step 2. In this study, we use a SC based on the Mean Absolute Deviation (MAD) [10], which is a robust measurement on dispersion.

$$\text{SC is : Stop the process if } \max_j \text{MAD}_j < \varepsilon, \text{ for all } j = 1, 2, \ldots, N.$$

where

$$\text{MAD}_j = \underset{i=1,2,\ldots,M}{\text{Median}} \left| c_j^{(i)} - \text{Median}\left(c_j^{(1)}, c_j^{(2)}, \ldots, c_j^{(M)}\right) \right|$$

for all $j = 1, 2, \ldots, N$.

Bonferroni Correction for Multiple Hypothesis Testing

The Bonferroni correction is a conservative method that can be used to control the overall significance level (α) or the family wise error rate (FWER) when conducting multiple hypotheses tests. If $T_1, T_2, \ldots, T_n$ is a set of n statistics with corresponding p-values $P_1, P_2, \ldots, P_n$ for testing hypotheses $H_1, H_2, \ldots, H_n$, the general Bonferroni multiple test procedure is performed by rejecting $H_i : i = 1, \ldots, n$ if the p-value (P_i) is less than or equals to α/n [44]. Thus, the Bonferroni inequality,

$$P\left\{\bigcup_{i=1}^{n}\left(P_i \leq \frac{\alpha}{n}\right)\right\} \leq \alpha \quad (0 \leq \alpha \leq 1)\,,$$

ensures that the probability of rejecting at least one hypothesis when all are true is no greater than the significance level α, which is the type I error rate.

Numerical Results

We include results of numerical experiments to validate and assess the proposed hybrid algorithms. First, an artificially generated data set is considered with different signal-to-noise ratio (SNR) values as well as with different segment widths. The SNR is defined as the segment mean divided by the standard deviation of the Gaussian noise in the process as considered in [19]. Finally, two of the well-known publicly available real aCGH data sets are considered to further demonstrate the effectiveness of the proposed methodology.

In order to assess the performance of the proposed SR-CE and CUSUM-CE algorithms over the standard CE method [33] a comparison study is carried out; which is the primary focus of this study. The variant of the CE method discussed in this study utilizes a multi-core architecture based parallel implementation in the R statistical software [35]. Furthermore, for the completeness of the study, we compare the results obtained through the proposed methodology with another four well established change-point detection methods in the literature: *DNAcopy* [52], *bcp* [6], *changepoint* [17] and *cumSeg* [22]. In all of these methodologies we consider the default parameter values in the respective algorithms [19], as most user will be exercising.

Results on Artificially Generated Data

Let us consider a random sequence of length 3500 with 10 abrupt change-points which results in having 11 segments. The standard deviation of the Gaussian noise is set as 1 in all the segments. Table 4.5 shows the parameter values used for the simulation study.

We follow the general work flow discussed in the framework of algorithms. First, based on a sequential procedure initial estimates of the change-point locations are obtained. Second, the CE algorithm is initiated based on these pre-estimates. We utilize the parameter values for the CE algorithm as described earlier with the smoothing parameters $\beta = 1$ and $\gamma = 1$. A significance level (α) of 0.001 [13] is considered in the two sample t-test to assess the statistical significance of the identified change-points. The Bonferroni correction is used to control the family wise error rate in multiple hypothesis testing [44]. In the standard CE method, we set N_{min} as 1 and N_{max} as 20 as the search space for the number of change-points.

Table 4.5: Parameter values for the simulation study

	Segment										
	1	2	3	4	5	6	7	8	9	10	11
Length	200	550	150	250	500	250	400	600	200	150	250
SNR	0	2	4	2.5	0	2	3	4	2.5	3.5	1
Mean[*]	0	2	4	2.5	0	2	3	4	2.5	3.5	1

[*] SNR=Mean/S.D., Standard Deviation is set as 1

Table 4.6 shows the initial estimates for the locations obtained by the two sequential procedures. It is observed that both methods have over estimated the true number of change-points as expected, even though the processing time for both methods are less than a second. The mean profile plots of the original data, hybrid algorithms and the other methods are shown in Figure 5.2. It is observed that the proposed SR-CE and CUSUM-CE procedures have correctly identified the true number of change-points ($N = 10$) as compared to the marginal over estimation ($N = 12$) given by the CE method. Except for the *cumSeg* method all other methods have over-estimated number of change-points, where the former method has under estimated the number of change-points. We observed that the over-estimation problem in *changepoint* and in *bcp* is severe than the other competing methods. However, when considering the average root mean square error (RMSE) both the proposed procedures and the general CE method have effectively segmented data with almost overlapping mean profiles. Table 5.1 shows the summary statistics on the performance of the proposed methods with the standard CE method.

Table 4.6: Initial estimates of locations and processing time of SCE-SR and SCE-CUSUM methods

Sequential Procedure	**Initial Estimates for Locations**	**Avg. Proc. time (s)**
SR	207, 422, 585, 755, 908, 1058, 1154, 1509, 1653, 1910, 2094, 2310, 2749, 2914, 3107, 3205, 3259	0.394
CUSUM	209, 755, 910, 1154, 1578, 1654, 1926, 2308, 2914, 3107, 3207, 3262	0.923

Both of the proposed methods and the CE method have resulted with lower average RMSE rates. The SR-CE gives the lowest average RMSE rate with an approximate average improvement of 10% over the CE method. The notable performance achievement in the proposed procedures when compared to the CE method is on the overall processing time. It is observed that both SR-CE and CUSUM-CE have performed significantly better than the general CE method, which even utilizes a multi-core architecture based parallel implementation. Among the two proposed methods, CUSUM-CE procedure performs better than the SR-CE when considering the overall processing time. The SR-CE procedure gives a significant improvement of around 91% ,while CUSUM-ce the improvement is more than two-folds (233%) as compared to the processing time of the CE method.

Table 4.7: Summary statistics on the performance of hybrid frameworks and the CE method

	SR-CE	**CUSUM-CE**	**CE method**
Average RMSE	0.083	0.094	0.092
Median Processing Time(s)*	19.868	11.419	37.977

*Relative to a 2.3 GHz Intel Core i7 processor (Mac OS X 10.9) with 8GB RAM.

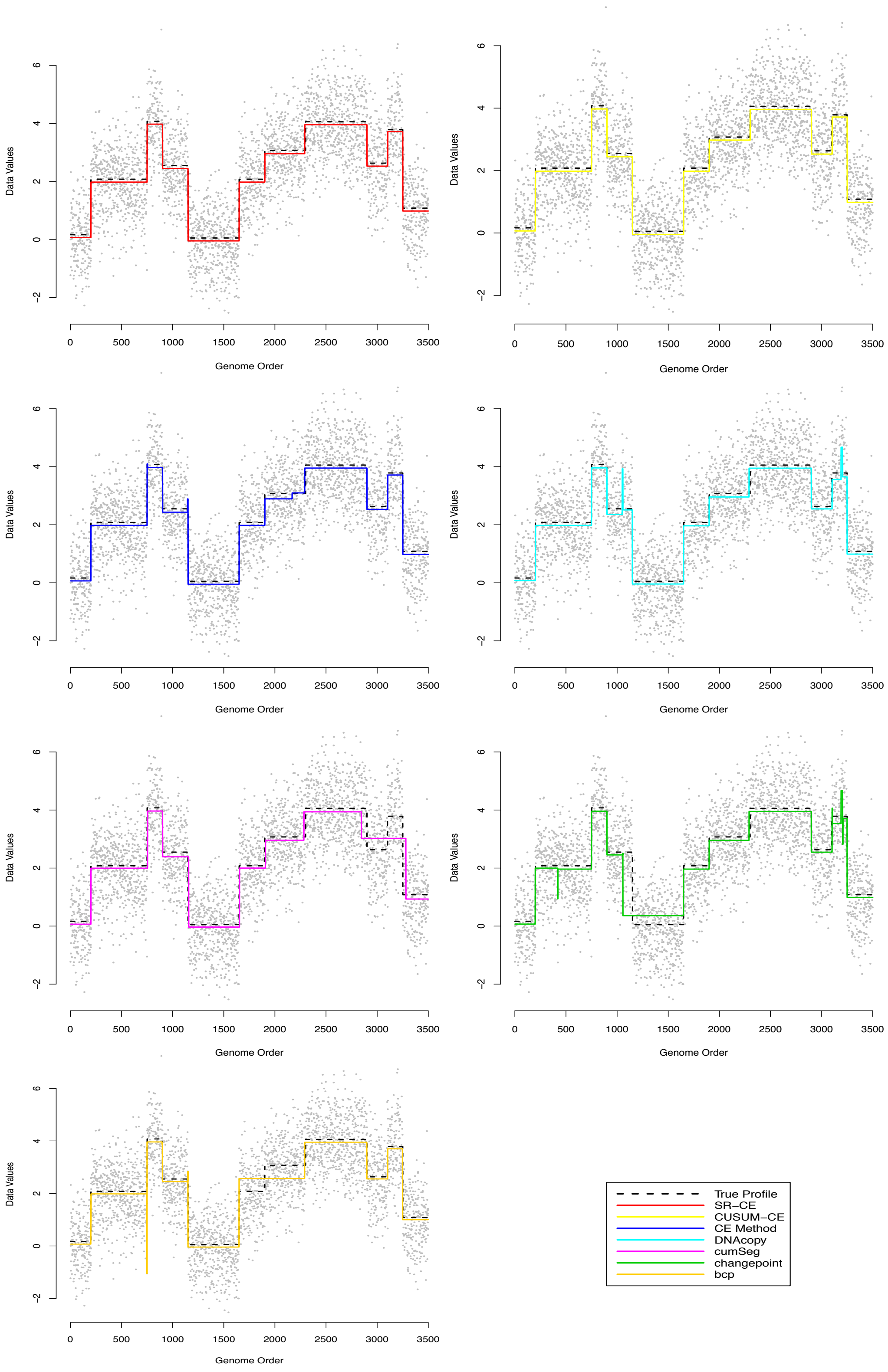


Figure 4.5: Mean profile plots of the proposed algorithms and the other methods for the artificial data.

Results on Real Data

Fibroblast cell lines data

This example considers a cDNA microarray-based CGH data of fibroblast cell lines which was originally discussed in [45]. The data set it is freely available to download from `http://www.nature.com/ng/journal/v29/n3/suppinfo/ng754_S1.html`. The data set consists of a single experiment on 15 fibroblast cell lines and it has already been discussed by several authors [21, 52] in the literature. We analyze the data in the fibroblast cell line GM03563 with respect to the chromosomes $1, 3$ and 7. By spectral karyotyping, real alteration (a single change-point) is only found in chromosome 3 out of the considered chromosomes.

Table 4.8: Initial estimates of change-point locations for the chromosomes 1, 3 and 7 of the GM03563 cell line data

Procedure	**Chromosome**		
	1	3	7
SR	122	57	131
CUSUM	81	59	161

By utilizing the SR and CUSUM procedures, we obtain initial estimates of the change-point locations for the chromosomes $1, 3, 7$ as in Table 4.6 separately. We initiate the CE algorithm with the same set of parameters considered in the artificial data example for all three chromosomal level data of GM03563 cell line. The standard CE method is initiated with the default parameter set with $N_{min} = 1$ and $N_{max} = 10$. Figure 4.6 shows the array CGH profiles for the three chromosomes based on the proposed methods as well as the other competing methods. In chromosome 1, both the proposed hybrid methods have correctly identified the true number of change-points as “zero”, as opposed to a single change-point estimation given by the CE method. For chromosomes 3 and 7, both the proposed algorithms and the CE method have given the same results estimating the correct number as well as the locations of the change-points. Considering the performances of the other methods, all other methods except for the *changepoint* have failed to estimate the number of change-points as zero for the chromosome 1. In fact all of them have over-estimated the true number. In general, the *bcp* method tends to over-estimate the true-number of change-points.

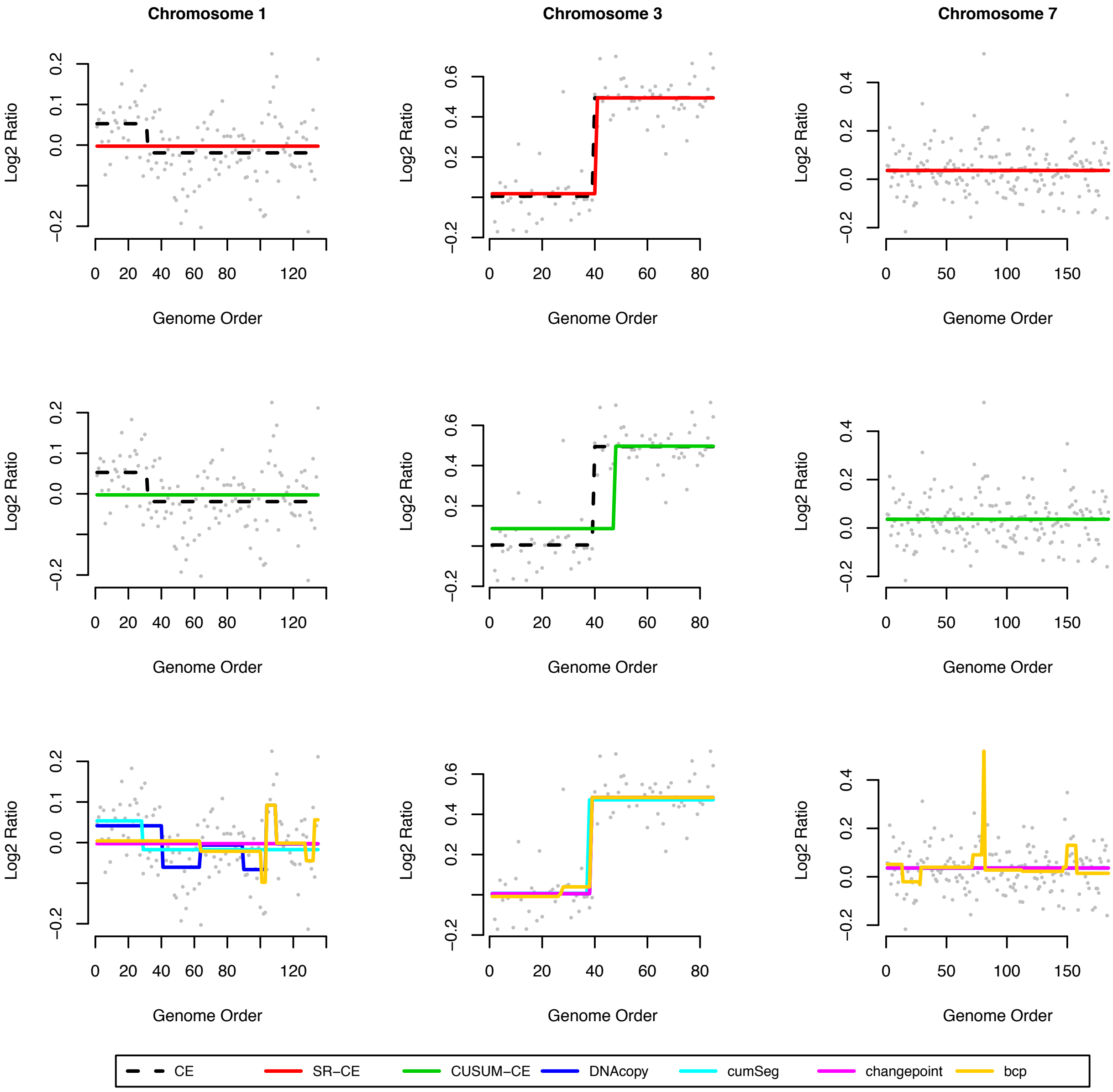


FIGURE 4.6: Array CGH profiles for the chromosomes 1, 3 and 7 in GM03563 cell line.

Table 4.9 shows the overall processing time. It is observed that both the proposed procedures are highly computationally efficient than the CE method. Furthermore, on average CUSUM-CE procedure is faster than the SR-CE procedure.

Breast tumor data

In this example we consider the breast cancer cell line (MDA157) data which was originally discussed in [27]. The cDNA microarray CGH was profiled across 6691 mapped human genes in 44 breast tumor samples and 10 breast cancer cell lines. This dataset is

Table 4.9: Processing time (s) for the GM03563 cell line data

Method	**Chromosome**		
	1	3	7
SR-CE	0.079	0.126	0.112
CUSUM-CE	0.124	0.072	0.069
CE	2.533	2.435	3.151

discussed in the [159] and [21] and can be downloaded from `http://www.pnas.org/content/99/20/12963/suppl/DC1`.

Table 4.10: Initial estimates of change-point locations for the chromosomes 3, 5, 9 and 13 of MDA157 cell line data

Procedure	**Chromosome**			
	3	5	9	13
SR	83, 174, 280, 387	84, 180, 250,328	126,181	97
CUSUM	82, 146, 326	60, 144, 234, 328	129,169	86

We apply our proposed algorithms as well as the other methods on chromosomes 3, 5, 9 and 13 data to estimate the underlying copy number variations. Figure 4.7 shows the aCGH profile plots for all the chromosomes. We observe that the proposed two procedures have behaved in a similar way in all chromosomes. Also, except for the chromosome 5, in all other cases CE method has also performed similar to SR-CE and CUSUM-CE procedures. In chromosome 9, *changepoint* method has not detected any change-points, whereas *bcp* has highly over-estimated the number of change-points. In general, our methods have similar profiles to *cumSeg* and *DNAcopy* procedures.

Table 4.11: Processing time (s) for MDA157 cell line data

Method	**Chromosome**			
	3	5	9	13
SR-CE	1.145	1.108	0.184	0.071
CUSUM-CE	0.643	0.994	0.251	0.071
CE	4.426	3.922	4.462	3.036

Discussion and Conclusions

We have proposed two novel hybrid algorithms (SR-CE, CUSUM-CE) that utilize powerful sequential change-point detection techniques (SR and CUSUM procedures) and a model based stochastic optimization technique (CE method) to estimate both the number and the locations of change-points in biological data of continuous measurements. This

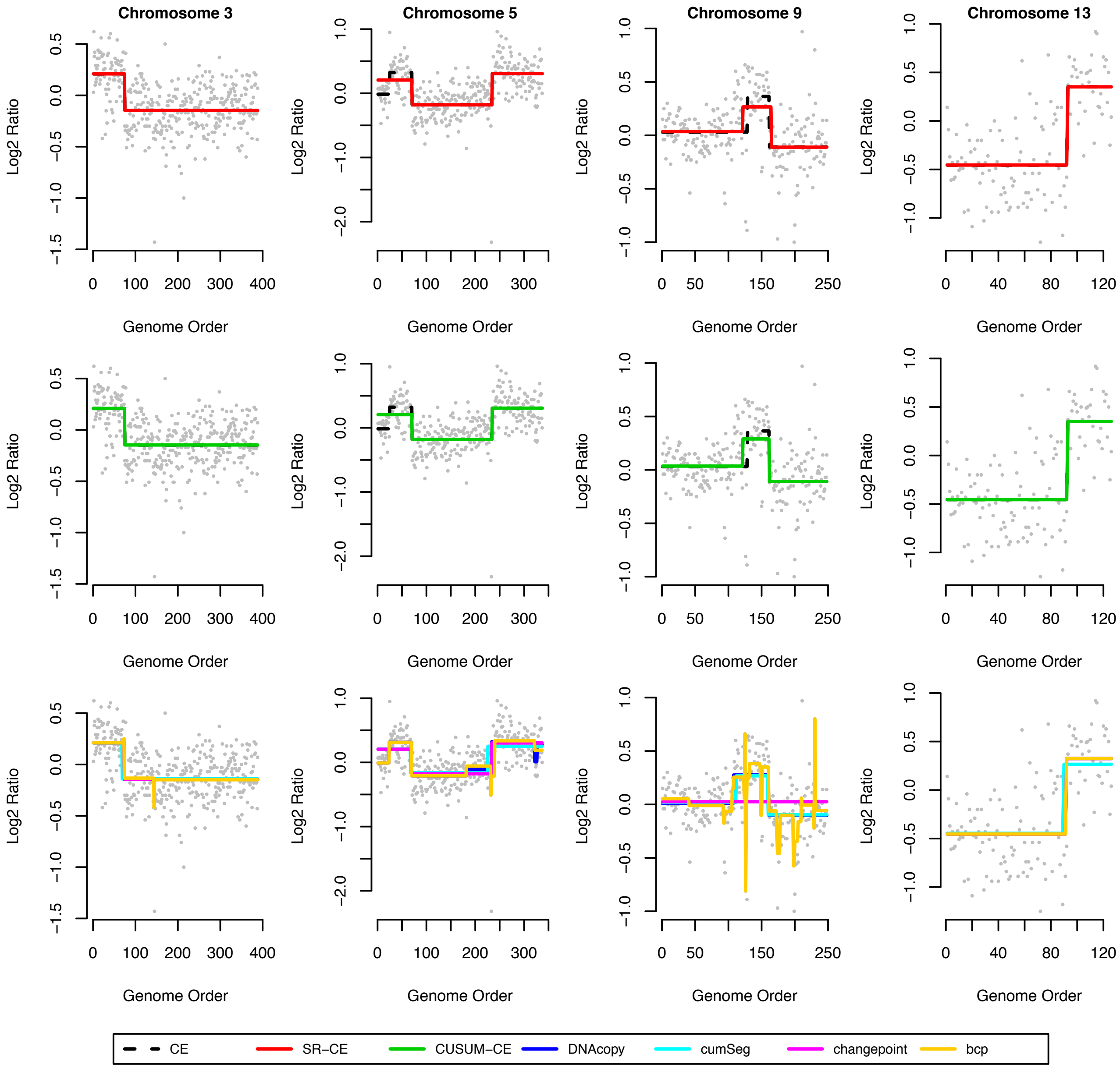


Figure 4.7: Array CGH profiles for the chromosomes 3, 5, 9 and 13 in MDA157 cell line

is the first-of-its kind implementation in the change-point literature that utilize on-line change point detection techniques to obtain initial estimates for a posteriori change-point problem and merge them with a model based stochastic optimization method (CE) to further improve the estimates on both the number and their corresponding locations. We compare the performance of the proposed hybrid algorithms with the standard CE algorithm, which does not use results from the sequential techniques as an input. Furthermore, for the completeness we have further compared our procedures with four other established change-point techniques. The effectiveness of the proposed methodology is assessed both in terms of artificially generated and real data. In all of the studies, it was found that the hybrid methods perform significantly better than the standard CE method both in terms of the precision and the processing time. In the standard CE method processing time is considered as one of the drawbacks in its implementation. Thus, incorporating sequential techniques has solved not only this critical issue for a greater extent, but also it has improved the detection power as well. Furthermore, use of the sequential techniques provide an upper limit for the search space for the number of change-points in the CE method. In the standard CE method user has to define these lower and upper search limit unknowingly. Thus, sequential techniques provide an important support for the standard CE method to perform more efficiently. While the results of this work are encouraging, there are plenty of avenues available as future research directions. In our study, it was identified that the sequential procedure is sensitive to the aberration width (i.e., segment width) resulting to favour analysis of longer sequences over the shorter sequences. Therefore, a versatile implementation of the methodology is worth for probing, which will work effectively in short as well as long sequences of data. Finally, the proposed procedures only developed to detect changes in mean levels of continuous measurements. We hope to extend these procedures to detect changes in the variance as well.

Acknowledgement

W. J. R. M. Priyadarshana acknowledges the funding received towards his PhD from the International Macquarie University Research Excellence (iMQRES) scholarship. The authors acknowledge the anonymous referees for their useful comments.

Bibliography

[1] D. Barry, J. A. Hartigan. A Bayesian analysis for change point problems. *Journal of the American Statistical Association.* 88:309-319, 1993.

[2] J. V. Braun, H. G. Müller. Statistical methods for DNA sequence segmentation. *Statistical Science.* 13:142-162, 1998.

[3] N. P. Carter. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet..* 39:S16-S21, 2007.

[4] A, Costa, O. D. Jones, D. Kroese. Convergence Properties of the Cross-Entropy Method for Discrete Optimization. *Operations Research Letters.* 35:573-580, 2007.

[5] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani. Least Angle Regression. *The Annals of Statistics.* 32:407-451, 2004.

[6] C. Erdman, J. W. Emerson. bcp: An R Package for Performing a Bayesian Analysis of Change Point Problems. *Journal of Statistical Software.* 23:1-13, 2007.

[7] C. Erdman, J. W. Emerson. A fast Bayesian change point analysis for the segmentation of microarray data. *Bioinformatics.* 24:2143-2148, 2008.

[8] G. E. Evans, G. Y. Sofronov, J. M. Keith, D. P. Kroese. Identifying change-points in biological sequences via the cross-entropy method. *Annals of Operation Research.* 189:155-165, 2011.

[9] L. Feuk, A. R. Carson, S. W. Scherer. Structural variation in the human genome. *Nature Reviews Genetics.* 7:85-97, 2006.

[10] D. C. Hoaglin, F. Mosteller, J. W. Tukey. Understanding Robust and Exploratory Data Analysis. *John Wiley and Sons Inc., New York*, 1983.

[11] G. Hodgson, J. H. Hager, S. Volik, S. Hariono, M. Wernick, D. Moore, N. Nowak, D. G. Albertson, D. Pinkel, C. Collins, D. Hanahan, J. W. Gray. Genome scanning

with array CGH delineates regional alterations in mouse islet carcinomas. *Nat. Genet.*. 29:459-464, 2001.

[12] S. Ivakhno, T. Royce, A. J. Cox, D. J. Evers, R. K. Cheetham, S. Tavare. CNAseg-a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics*. 26:3051-3058, 2010.

[13] V. E. Johnson. Revised standards for statistical evidence. *Proc. of the National Academy of Science*. doi:10.1073/pnas.1313476110, 2013.

[14] A. Kallioniemi, O. P. Kallioniemi, D. Sudar, D. Rutovitz, J. W. Gray, F. Waldman, D. Pinkel. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*. 258:818-821, 1992.

[15] J. M. Keith. Segmenting Eukaryotic Genomes with the Generalized Gibbs Sampler. *Journal of Computational Biology*. 13:1369-1383, 2006.

[16] R. Killick, P. Fearnhead, I. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*. 107:590-598, 2012.

[17] R. Killick, I. Eckley. changepoint: An R package for changepoint analysis. R package version 1.1. http://CRAN.R-project.org/package=changepoint, 2013.

[18] S. Kullback, R. A. Leibler. On Information and Sufficiency. *Annals of Mathematical Statistics*. 22:79-86, 1951.

[19] W. R. Lai, M. D. Johnson, R. Kucherlapati, P. J. Park. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*. 21:3763-3770, 2005.

[20] R. Lucito, J. Healy, J. Alexander, A. Reiner, D. Esposito, M. Chi, L. Rodgers, A. Brady, J. Sebat, J. Troge, J. A. West, S. Rostan, K. C. Q. Nguyen, S. Powers, K. Q. Ye, A. Olshen, E. Venkatraman, L. Norton, M. Wigler. Representational Oligonucleotide Microarray Analysis: A High-Resolution Method to Detect Genome Copy Number Variation. *Genome Research*. 13:2291-2305, 2003.

[21] M. R. V. Muggeo, G. Adelfio. Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics*. 27:161-166, 2011.

[22] V. M. R. Muggeo. cumSeg: Change point detection in genomic sequences. R package version 1.1. http://CRAN.R-project.org/package=cumSeg, 2012.

[23] J. Oliver, P. Bernaola-Galvan, P. Carpena, R. Roman-Roldan. Isochore chromosome maps of eukaryotic genomes. *Gene*. 276:47-56, 2001.

[24] A. B. Olshen, E. S. Venkatraman, R. Lucito, M. Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*. 5:557-572, 2004.

[25] E. S. Page. Continuous inspection schemes. *Biometrika*. 41:100-115, 1954.

[26] J. R. Pollack, C. M. Perou, A. A. Alizadeh, M. B. Eisen, A. Pergamenschikov, C. F. Williams, S. S. Jeffrey, D. Botstein, P. O. Brown. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.*. 23:1061-4036, 1999.

[27] J. R. Pollack, T. S. órlie, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning, R. Tibshirani, D. Bo, D. Botstein, A. L. Bórresen-Dale, P. O. Brown. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. of the National Academy of Sciences USA*. 99:12963-12968, 2002.

[28] M. Pollak, A. G. Tartakovsky. Exact optimality of the Shiryaev-Roberts procedure for detecting changes in distributions. *In: Information Theory and its Applications, ISITA 2008 International Symposium*. 1-6, 2008.

[29] M. Pollak, A. G. Tartakovsky. Optimality properties of the Shiryaev-Roberts procedure. *Statistica Sinica*. 19:1729-1739, 2009.

[30] A. Polunchenko, G. Sokolov, W. Du. Quickest Change-Point Detection: A Bird's Eye View. *In: Joint Statistical Meeting (JSM)*, 2013.

[31] T. Polushina, G. Sofronov. Change-point detection in biological sequences via genetic algorithm. *In: Proceedings IEEE Congress on Evolutionary Computation (CEC)*. 1966-1971, 2011.

[32] T. V. Polushina, G. Y. Sofronov. A hybrid genetic algorithm for change-point detection in binary biomolecular sequences. *In: Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2013)*. 1-8, 2013.

[33] W. J. R. M. Priyadarshana, G. Sofronov. A modified cross entropy method for detecting multiple change points in DNA Count Data. *In: WCCI 2012 IEEE World Congress on Computational Intelligence (CEC)*. 1020-1027, 2012.

[34] W. J. R. M. Priyadarshana, G. Sofronov. GAMLSS and Extended Cross-Entropy Method to Detect Multiple Change-Points in DNA Read Count Data. *In: Muggeo VMR, Capursi V, Boscaino G, Lovison G (Eds.), Proceedings of the 28th International Workshop on Statistical Modelling.* 1:453-457, 2013.

[35] R Core Team:R: A Language and Environment for Statistical Computing. (2013) R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

[36] S. W. Roberts. A comparison of some control chart procedures. *Technometrics.* 8:411-430, 1966.

[37] R. Rubinstein, D. P. Kroese. The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning. *Springer-Verlag, New York.* 2004.

[38] J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Maner, H. Massa, M. Walker, M. Chi, N. Navin, R. Lucito, J. Healy, J. Hicks, K. Ye, A. Reiner, T. C. Gilliam, B. Trask, N. Patterson, A. Zetterberg, M. Wigler. Large-scale copy number polymorphism in the human genome. *Science.* 305:525-528, 2004.

[39] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics.* 6:461-464, 1978.

[40] A. Sen, M. Srivastava. On tests for detecting a change in mean. *The Annals of Statistics.* 3:98-108, 1975.

[41] A. N. Shiryaev. The problem of the most rapid detection of a disturbance in a stationary process. *Soviet Mathmatics. Dokl..* 2:795-799, 1961.

[42] A. N. Shiryaev. On optimum methods in quickest detection problems. *Theory Probability and Its Applications.* 8:22-46, 1963.

[43] A. N. Shiryaev. Optimal stopping rules. *Springer, New York*, 1978.

[44] R. J. Simes. An Improved Bonferroni Procedure for Multiple Tests of Significance. *Biometrika.* 73:751-754, 1986.

[45] A. M. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Con-roy, G. Hamilton, A. K. Hindle, B. Huey, K. Kimura, S. Law, K. Myamboo, J. Palmer, B. Ylstra, J. P. Yue, J. W. Gray, A. N. Jain, D. Pinkel, D. G. Albertson. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics.* 29:263-264, 2001.

[46] G. Y. Sofronov, G. E. Evans, J. M. Keith, D. P. Kroese. Identifying change-points in biological sequences via sequential importance sampling. *Environmental Modeling and Assessment.* 14:577-584, 2009.

[47] G. Sofronov. Change-Point Modelling in Biological Sequences via the Bayesian Adaptive Independent Sampler. *In: International Proceedings of Computer Science and Information Technology.* 5:122-126, 2011.

[48] G. Sofronov, T. Polushina, W. J. R. M. Priyadarshana. Sequential Change-Point Detection via the Cross-Entropy Method. *In: The 11th Symposium on Neural Network Applications in Electrical Engineering (NEUREL'12).* 185-188, 2012.

[49] A. Subramanian, H. Kuehn, J. Gould, P. Tamayo, J. P. Mesirov. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics.* 23:3251-3253, 2007.

[50] A. Theisen. Microarray-based comparative genomic hybridization (aCGH). *Nature Education.* 1:45, 2008.

[51] R. Tibshirani, P. Wang. Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics.* 9:18-29, 2008.

[52] E. S. Venkatraman, A. Olshen. DNAcopy: DNA copy number data analysis. R package version 1.34.0, 2013.

[53] H. Wang, B. Li, C. Leng. Shrinkage Tuning Parameter Selection with a Diverging Number of Parameters. *Journal of the Royal Statistical Society. Series B (Statistical Methodology).* 71:671-683, 2009.

[54] C. Xie, M. T. Tammi. A new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics.* 10:80, 2009.

Chapter 5

The Cross Entropy Method for Detecting Multiple Change Points in DNA Read Count Data

This chapter constitutes a peer-reviewed full conference paper published in the proceedings of the 2012 IEEE Conference on Evolutionary Computation (CEC), Brisbane, Australia. The content of the paper including notation, plot sizes and wording has been slightly changed to improve the flow and the overall coherence of the thesis. The citation for the paper is:

Priyadarshana, W. J. R. M. and Sofronov, G. (2012). A Modified Cross- Entropy Method for Detecting Multiple Change-Points in DNA Count Data. In Proc. of the IEEE Conference on Evolutionary Computation (CEC), 1020-1027, DOI: 10.1109/ CEC.2012.6256470.
Writing: 95% , analysis: 100% , conception and design: 25%

Specific contribution of joint authors
Georgy Sofronov: Overall supervision of data analysis and interpretation. Provided continual feedback and suggestions on both analysis and writing.

Reprinted with permission from Priyadarshana, W. J. R. M. and Sofronov, G. (2012) IEEE Conference on Evolutionary Computation (CEC), Copyright © 2012 IEEE.

5.1 Summary

In the previous two chapters we discussed the application of the CE method along with different modelling techniques to detect change-points in aCGH data. In this chapter we further investigate the applicability of the CE method to detect change-points in next generation sequencing data. In this study we apply the Cross-Entropy (CE) method as described in Chapter 2 to detect multiple change-points (break-points) in biological sequences of discrete measurements. Particularly we applied the CE method to detect CNVs in DNA read count data obtained through next generation sequencing (NGS) techniques. To our knowledge, this is the first application of the CE method to detect multiple change-points in DNA read count data.

The motivation of this work was to model DNA read count data as it is, without conducting any transformation on the read counts, to find change-points with the use of the CE method and other model selection techniques. By the time of the publication, in the literature there were few methods available to detect CNVs in read count data, and among them, the majority have used transformation techniques on read counts to perform analysis. Furthermore, most of the methods were developed with a direct comparison of a reference sequence. Our aim in this work was to introduce a direct modelling approach to detect CNVs in read counts that does not need to be compared with a reference sequence.

The standard practice in statistical modelling is to model count data either by using Poisson or negative binomial distributions. Due to the observed over-dispersion in the read counts we ruled out the use of the Poisson distribution. Thus, we modelled the DNA read count data with a negative binomial distribution. Also, for the first time in the application of the CE method, we have proposed using the four-parameter beta distribution as described in Chapter 2 to simulate the change-point locations in the CE algorithm for the multiple change-point problem. Earlier in [49], the truncated normal distribution was used in the CE method to detect change-points in binary data. We carried out a detailed simulation study to illustrate the performance differences of the CE method with the use of the four-parameter beta distribution and with the truncated normal distribution. It was observed that the CE method with the four-parameter beta distribution performs better than the CE method with the truncated normal distribution. Furthermore, we compared the performance differences of two stopping criteria that are described in Chapter 2. A detailed simulation study was carried out to obtain best parameters for the CE algorithm and to assess the performance of the proposed method.

Finally, the proposed technique was applied on real DNA read count data obtained through Illumina TruSeq exome capture of patients with celiac disease. The data were provided by Dr. Vincent Plagnol (UCL Genetics Institute, University College London, Grower Street, London, UK. An exome is simply defined as the protein-coding content of the DNA, which comprises 1% - 2% of the genome. Targeted exome sequencing is a cost-effective sequencing technology as compared to whole genome sequencing methods. It enables researchers to obtain a closer look at a specific region of the genome to discover variants for many complex human diseases. We analyzed the DNA read counts corresponding to chromosome 2 of a patient. It was revealed that the proposed CE procedure with negative binomial modelling approach is an effective way of segmenting DNA read count data. A further extension of this work with zero-inflated modelling approach with a parallel implementation of the CE method was proposed in [117].

A Modified Cross Entropy Method for Detecting Multiple Change-Points in DNA Count Data

W. J. R. M. Priyadarshana[1,*], Georgy Sofronov[1]

[1] Department of Statistics, Faculty of Science, Macquarie University, Sydney NSW 2109, Australia.

*E-mail: madawa.weerasinghe@mq.edu.au

Abstract

We model DNA count data as a multiple change-point problem, in which the data are divided in to different segments by an unknown number of change-points. Each segment is supposed to be generated by unique distribution characteristics inherent to the underlying process. In this paper, we propose a modified version of the Cross-Entropy (CE) method, which utilizes Beta distribution to simulate locations of change-points. Several stopping criteria are also discussed. The proposed CE method applied on over-dispersed DNA read count data, in which the observations are distributed as independent Negative Binomial. Furthermore, we incorporate the Bayesian Information Criterion (BIC) to identify the optimal number of change-points within the CE method while not fixing the maximum number of change-points in the data sequence. We obtain estimates for the artificial data by using the modified CE method and compare the results with the general CE method, which utilizes normal distribution to simulate locations of the change-points. The methods are applied to a real DNA count data set in order to illustrate the usefulness of the proposed modified CE method.

Introduction

Change-point models are utilized to detect heterogeneity in many scientific fields to give an improved and more detailed interpretation of the properties inherent to the process. These models can be employed in many areas like biomedical sequences, financial and economic time series, quality control, signal processing, etc. There are two broader classes of change-point models: retrospective (off-line methods) and sequential (on-line) methods. Many authors have addressed the change-point problem both in terms of Bayesian and frequentist point of view. There is a rich class of literature available in the

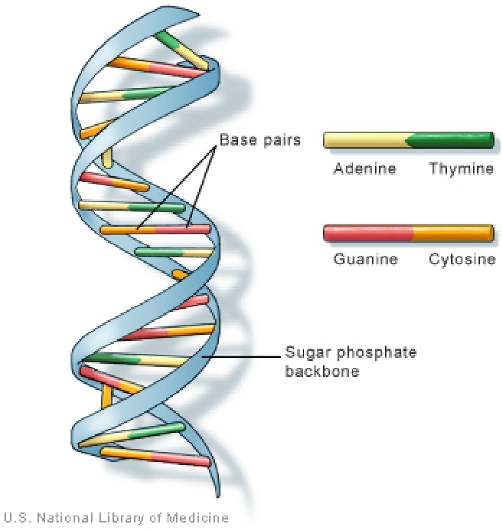


FIGURE 5.1: The DNA structure. Source: US National Library of Medicine (http://ghr.nlm.nih.gov/handbook/basics/dna)

methods developed to segment binary sequences as well as continuous data. However, in the literature there exists only a handful of resources concentrating mainly on change-point detection in count data and especially on deoxyribonucleic acid (DNA) read count data.

DNA is the heredity material or the information carrier in humans and almost all the living organisms. DNA consists of two long polymers of nucleotides. The information in DNA is stored as a code made up of four chemical bases known as Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). The order or the sequence of these chemical bases determines the information available for building and maintaining a living organism.

Reviewing the literature on change-point modelling in DNA sequences,[27] reviewed some of the methodologies that were used to segment DNA sequences. They have proposed and discussed a local segmentation method called split polynomial fitting. However, they have not addressed methodologies related to change-point modelling in DNA count data. On the more recent advances, [159] applied fused lasso method to the "hot spot" detection in comparative genomic hybridization (CGH) data. Where, CGH [1] is a technique for measuring DNA copy number of selected genes on the genome. Reference [48] introduced an improved version of the computing package on change-point modelling based on Product Partition Models (PPM), which was introduced by [16]. A scan statistic based on summing a chi-squared statistic for each individual sample was proposed in [173] to simultaneously detect change-points in multiple sequences. Furthermore, [151], [152], [85] and [113] discuss different approaches to detect multiple change-points in DNA sequences. More recently with the development of next-generation sequencing data, [14] proposed a novel approach named “CNAseg” to identify the copy number abnormalities (CNAs) based on the number of reads.

The above literature on change-point modelling related to DNA sequences data considered competing methodologies on segmenting binary sequences and do not consider the problem as a count data process. In the literature, count data modelling has been discussed extensively by many authors mainly in the context of GLM. This comprises of analysis when the over-dispersion is present or not and with many other attributes [4]. However, the usage of change-point analysis on DNA count data within the GLM context has not been addressed by many. The change-point analysis within the GLM framework adds more information to the outcome of the study, as it better explains the true nature of the underlying structure of the observations. Recently, [85] discussed a genetic algorithm approach to model multiple change-points in count data. They have considered count data related to a meteorology study in which the data are assumed to be distributed as independent Poisson random variables. However, they have not discussed any issues on over-dispersion of the data.

This paper contributes to the literature mainly in two aspects. Firstly, this proposes an efficient methodology to detect multiple change-points in DNA read count data, considering it as a combinatorial problem and discusses two competing stopping criteria. Secondly, this models the data in each segment by utilizing the negative binomial distribution while addressing the over-dispersion issue.

This paper utilizes a modified version of the Cross-Entropy (CE) method originally proposed in [131] in order to identify the number of change-points as well as the locations in DNA count data. Change-point modelling with the use of CE concept was first utilized in [49] to detect multiple change-points in DNA binary sequences. They have proposed a CE method using a normal distribution to simulate change-points in binary sequences. However they have fixed the maximum number of change-points in advance and did not search for an optimal combination of change-points that maximizes their proposed performance function.

This paper proposes the four parameter beta distribution to simulate the locations of the change-points within the CE method and does not place a restriction on the maximum number of change-points. In each segment of the count data sequence is modeled by using the negative binomial distribution. Bayesian Information Criterion (BIC) [7],[172] is used to identify the number of change-points in the count data. Reference [172] shows that the estimate on the number of change-points obtained by the BIC weakly converges to the true number of change-points. However, they have developed the methodology on normally distributed data. Finally, the study will compare the results with the general approach proposed as in [49] and discuss two stopping criteria that can be used to optimize the process.

The paper is structured as follows. Section 2 introduces the multiple change-point problem in mathematical terms. In Section 3, we explain the modified CE method, underlying distribution properties, BIC and the estimation of the parameters. Section 4 presents the results of numerical experiments. Finally, Section 5 will conclude the paper with future research directions.

The Multiple Change-Point Problem

Let us formulate the multiple change-point problem in mathematical terms. A count data sequence $\mathbf{y} = (y_1, y_2, \ldots, y_L)$ of length L is given.

A segmentation of the sequence is specified by the number of change-points N and the positions of the change-points $\mathbf{C} = (c_1, c_2, \ldots, c_N)$, where $0 = c_0 < c_1 < \cdots < c_N < c_{N+1} = L$. In this context, a change-point is a boundary between two adjacent segments. The value of c_i is the sequence position of the rightmost character of the segment to the left of the ith change-point. Segments are numbered from 0 to N as there will be one or more segment than number of change-points. The model assumes that within each segment, the observations are distributed as independent negative binomial with probability p_n and fixed dispersion parameter (size) of r, where $0 \leq p_n \leq 1$ for $n = 0, \ldots, N$. The dispersion parameter r can either be pre-specified or estimated from the data. Then the joint distribution of $\mathbf{y} = (y_1, y_2, \ldots, y_L)$ conditional on N, $\mathbf{C} = (c_1, c_2, \ldots, c_N)$, and $\mathbf{p} = (p_0, p_1, \ldots, p_N)$ is given by

$$\begin{aligned} & f(y_1, y_2 \ldots, y_L \mid N, \mathbf{C}, \mathbf{p}) \\ & = \prod_{n=0}^{N} \Bigg[\prod_{i=c_n+1}^{c_{n+1}} \frac{\Gamma(r + y_i)}{y_i! \Gamma(r)} (1 - p_n)^r p_n^{y_i} \Bigg]. \end{aligned} \tag{5.1}$$

Note that this is one of the forms of negative binomial distribution, which is also known as the gamma-poisson mixture distribution. The corresponding log likelihood of the model is

$$\begin{aligned} & ll(N, \mathbf{C}, \mathbf{p}) \\ & = \sum_{n=0}^{N} \Bigg[\sum_{i=c_n+1}^{c_{n+1}} \ln \Gamma(r + y_i) - \sum_{i=c_n+1}^{c_{n+1}} \ln(y_i!) \\ & -\lambda \ln \Gamma(r) + \lambda r \ln(1 - p_n) + \sum_{i=c_n+1}^{c_{n+1}} y_i \ln(p_n) \Bigg]. \end{aligned} \tag{5.2}$$

where $\lambda = (c_{n+1} - c_n - 1)$ is the length of the segment.

Four parameter Beta distribution

The standard beta distribution with two shape parameters ($\alpha > 0, \beta > 0$) is supported on the range $[0, 1]$. In this study the location of the change-points may vary based on the length of the data set. Therefore, two further parameters have to be introduced to obtain beta random values in the specified range. Let us consider the minimum and the maximum values of the distribution of beta values as L_L and L_U. Then, the probability density function of the four parameter beta distribution is given by,

$$\begin{aligned} & f(y \mid \alpha, \beta, L_L, L_U) \\ & = \frac{1}{\boldsymbol{B(\alpha, \beta)}} (y - L_L)^{\alpha - 1} \frac{(L_U - y)^{\beta - 1}}{(L_U - L_L)^{\alpha + \beta - 1}}. \end{aligned} \tag{5.3}$$

The method-of-moment estimates of the shape parameters are

$$\hat{\alpha} = \bar{y} \left[\frac{\bar{y}(1 - \bar{y})}{s^2} - 1 \right]. \tag{5.4}$$

$$\hat{\beta} = (1 - \bar{y}) \left[\frac{\bar{y}(1 - \bar{y})}{s^2} - 1 \right]. \tag{5.5}$$

Note that since we have two additional parameters specifying the range of the beta values, the $\bar{y}$ (sample mean) and s^2 (sample variance) values are replaced with

$$\bar{y} = \frac{\bar{y} - L_L}{L_U - L_L}.$$

and

$$s^2 = \frac{s^2}{(L_U - L_L)^2}.$$

Modified Cross-Entropy Method for Multiple Change-Point Problem

The standard Cross-Entropy method

The Cross- Entropy (CE) method [131] can be used for two types of problems:

1. Estimation
2. Optimization

In general the process of multiple change-point detection can be considered as either a minimization or a maximization problem based on the nature of the performance function (F). Let X be a finite set of states and F be a real valued performance function on X. We wish to find the optimum (minimum or maximum) of F over X and the state(s) corresponding to this value.

The CE method is an iterative optimization procedure that starts with a parameterized sampling distribution from which a random sample is generated. Then, each observation or the combinatorial arrangement is scored for its performance as the solution to a specified optimization problem. A fixed number of best of these combinatorial arrangements are referred to as the *elite sample* (M_{elite}). This elite sample is subsequently used to update the parameters for the sampling distribution. Thus, adaptive parameters are utilized in each iteration. The sampling distribution eventually converges to a degenerate distribution about a locally optimal solution which ideally will be globally optimal.

Let N_{max} is the maximum number of change-points in this study that we wish to find. We can represent the position of the change-points as a non decreasing N_{max} − dimensional vector. When the number of change-points is less than the maximum number of change-points, some of the components of the vector will be repeated, indicating the same change-point. The CE method in [49] considers truncated independent normal distributions in order to simulate the locations of change-points. They have used the likelihood function as the performance function F to identify change-points in DNA binary sequences. In each iteration the initial parameters are updated based on the standard CE method until a convergence state is achieved. A variance based stopping criterion is used to measure the fit of the combinations of change-points in each iteration.

Modified Cross-Entropy Method

The proposed modified CE method differs from the standard CE method mainly in three aspects. Firstly, this considers over-dispersed count data and each segment of the sequence are assumed to be distributed as independent negative binomial distribution with dispersion parameter r and probability p_n. The dispersion parameter is estimated from the data and held constant for each segment and the other parameter is estimated for each of the segments. Secondly, $Beta(\alpha, \beta)$ distribution on the support $[L_L, L_U]$ is used to simulate the locations of change-points. In each iteration the parameters of the $Beta(\alpha, \beta)$ distribution is updated until a stopping criterion is met. Finally, the performance function F in this study is the Bayesian Information Criterion (BIC) [7],[172] which is calculated for all the simulated combinations of change-points. The combination which minimizes F under the corresponding N is considered as the optimum solution. Therefore, a minimization problem is considered.

We choose initial values for both the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ such that $\alpha_i = \beta_i = 1$, which is the standard uniform distribution on the interval $[L_L, L_U]$, since we are dealing with a four parameter beta distribution. Where L_L and L_U are the lower and upper bound of the count data sequence. For each change-point vector $(\mathbf{C})$ in the sample we obtain the maximum likelihood estimate of p_n with respect to the each of the segments and evaluate the performance function F.

The performance function BIC that we wish to minimize is

$$F(\mathbf{C}) = -2 * ll(N, \mathbf{C}, \mathbf{p}) + k * \ln(L) \tag{5.6}$$

Where, $ll(N, \mathbf{C}, \mathbf{p})$ is the log likelihood as in (2) of the count data sequence which is distributed as negative binomial with (r, p) and $k = 2 * (N + 1)$. The performance function score is calculated in each iteration with respect to the change-point vector $\mathbf{C}$ and L is the length of the count data sequence.

In each of the iterations M_{elite} sample is calculated considering the best performing combinations of change-points with respect to the performance function score. The process is carried out until a convergence or a specific stopping criterion is achieved. In this study two stopping criteria are discussed and evaluated. The first criterion is based on the [49] and the other is based on the original CE method as in [131]. In each step, the initial parameters of the beta distribution are updated accordingly. Then, locations of the change-points are generated randomly according to the updated beta distribution.

The algorithm can be summarized as below:

1. Choose initial values for $\boldsymbol{\alpha}^{\mathbf{0}}$ and $\boldsymbol{\beta}^{\mathbf{0}}$. Set $t = 1$. (In this case we have set both parameters equal to one and both parameter vectors are N dimensional). Where, $\boldsymbol{\alpha}^{\mathbf{0}} = (1, 1, \ldots, 1)$ and $\boldsymbol{\beta}^{\mathbf{0}} = (1, 1, \ldots, 1)$.

2. Generate a random sample $\mathbf{C}^{(1)}$,$\mathbf{C}^{(2)}$,$\ldots$,$\mathbf{C}^{M}$ from the Beta$\boldsymbol{(\alpha^{t-1}}$, $\boldsymbol{\beta^{t-1}}$) distribution. Where M is the sample size and $\mathbf{C} = (c_1, c_2, \ldots, c_N)$ is the change-point vector as defined earlier.

3. For each $i = 1, \ldots, M$ order $c_1^{(i)}, \ldots, c_N^{(i)}$ from smallest to biggest and set $\mathbf{C}^{(\mathbf{i})} = (c_1^{(i)}, \ldots, c_N^{(i)})$.

4. Evaluate the performance of each $\mathbf{C}^{(1)}$,$\mathbf{C}^{(2)}$,$\ldots$,$\mathbf{C}^{M}$ using (6). Let's define the M_{elite} sample, which is the best performing combinations of the change-points as $M_{elite} = \rho * M$.

5. For all $j = 1, 2, \ldots, N$ estimate the two beta shape parameters as in (4) and (5) from the M_{elite} sample and update the current parameter set.

6. If the stopping criterion (SC) is met, then stop the process and identify the combination of the locations of change-points ($\mathbf{C}^{(i)}$) that minimizes the BIC, which is the optimal number of change-points. Otherwise set $t = t + 1$ and iterate from step 2.

The two stopping criteria (SCs) considered in this study are

SC1: Stop the process if $max_j(\sigma_j^2)^t < \epsilon$

SC2: Stop the process if for some $t \geq k$, say $k = 4$,
$F_t = F_{t-1} = \cdots = F_{t-k}$

The final solution will be a single vector of change-points.

Results

In this section, we include results of numerical experiments that illustrate the performance of the modified CE method. First, we consider an artificial count data sequence with a known distribution, in which observations of each segment are generated from a negative binomial process. We carried out the analysis based on the two stopping criteria distinctly under the standard CE method which utilizes a normal distribution (CE-Normal) and the modified CE method which uses a beta distribution (CE-Beta) to simulate the locations of change-points. The BIC criterion, which is the performance function, is then used to identify the optimal combination of the change-points. This will allow us to carry out direct comparison of the methods in terms of the Root Mean Squared Error (RMSE) and running time.

Finally, a real DNA count data set is considered. We continue the process until a convergence in the performance function is achieved or a stopping criterion is met. Since we do not know the number of change-points in advance, an agreement between the methods is considered by looking at the mean profile plots followed by a comparison study on the processing time.

Example 1: Artificial Data Set

Let $(y_1, y_2, \ldots, y_{20000})$ be a sequence of independent negative binomial random variables with the parameters given in the Table 5.1, where the dispersion parameter of the distribution is held constant at 10. We generated 200 random sequences using these parameters and carried out the analysis based on the CE algorithms with different stopping criteria.

First we carried out the analysis varying number of change-points (N) from 1 to 20 for both CE-Beta and CE-Normal algorithms with respect to the two stopping criteria. Then we obtained the best solution in each of the N situations which minimizes (6). Figure

5.2 shows the BIC values for each of the N cases (from 8 to 20) for both algorithms. Table II shows the average processing times on CE-Beta and CE-Normal with respect to the stopping criteria.

Figure 5.2 shows that in both the algorithms with two stopping criteria, BIC score is minimized when N equals to 9. More importantly, when considering the processing time as in Table 5.2, there is a significant improvement in the proposed CE-Beta algorithm when compared to the competing CE algorithms based on normal assumption.

The CE-Beta SC1 algorithm can be identified as the optimal CE algorithm on the basis of processing time when compared with the other three algorithms considered in the study. The processing time is considered as one of the most important aspects in combinatorial studies especially when dealing with change-point modelling. Furthermore, Table 5.2 shows that the running time(s) in CE-Beta is significantly less than that of the competing CE-Normal method with the two SCs. Note, that this study is carried out in a corei3 first generation 2.27GHz processor with 4GB RAM. Therefore, the processing time is relative to this operation conditions.

Table 5.3 shows the average Root Mean Squared Error (RMSE) for each algorithm CE-Beta and CE-Normal with two SCs under the optimal change-point numbers detected (i.e. N equals to 9). The RMSE values indicate that even though the computing time is highly superior under SC1 it gives less precision when compared with the SC2. Moreover, it is noted that the RMSE value is lower in the proposed CE- Beta under SC1 method than the competing CE- Normal method. Figure 3 shows the fit of the change-points with the average counts over the sequence. It is noted that both methods under the two stopping criteria correctly captured the major regions in the over dispersed count data series.

Parameter smoothing: Rho (ρ)

We have considered smoothing up the parameter Rho (ρ), which is used to obtain the $Nelite$ sample. The RMSE and processing time(s) is obtained for the Rho values from 0.01 to 0.1 with the bin of 0.01 when N equals to 9. We have obtained the average results based on 100 simulations under each of the Rho values.

Figure 5.4 indicates that the RMSE for the SC2 is lower than the SC1 algorithms both in CE-Beta and CE-Normal cases. Furthermore, CE-Beta algorithms have lower RMSE on average than that of the competing CE-Normal algorithms. Also, by looking at the

Table 5.1: Negative binomial "p" parameters for the artificial sequence with fixed size of 10.

Positions	Negative Binomial Parameter(p)
1—2000	$p_0 = 0.05$
2001—4000	$p_1 = 0.15$
4001—6000	$p_2 = 0.40$
6001—8000	$p_3 = 0.02$
8001—10000	$p_4 = 0.20$
10001—12000	$p_5 = 0.50$
12001—14000	$p_6 = 0.10$
14001—16000	$p_7 = 0.85$
16001—18000	$p_8 = 0.18$
18001—20000	$p_9 = 0.90$

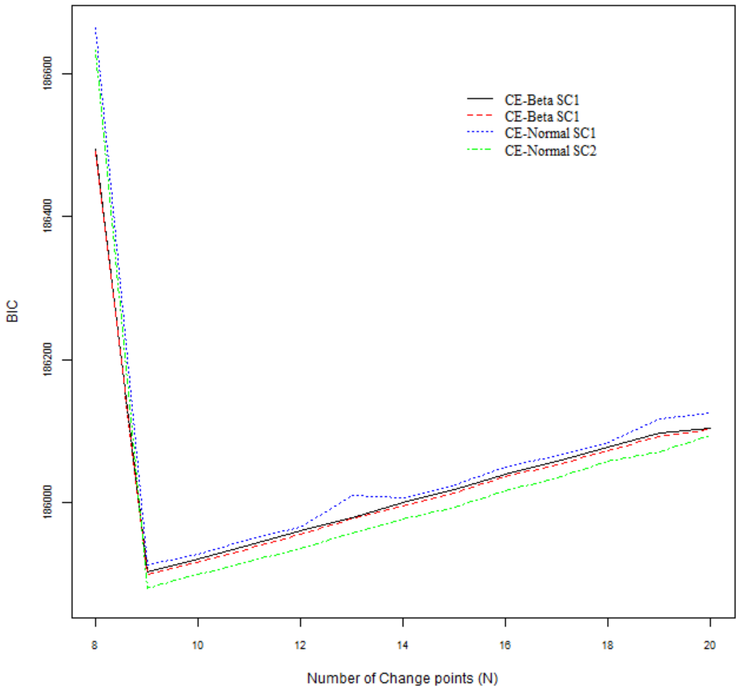


Figure 5.2: BIC vs. N for CE-Beta and CE-Normal with two SCs

Figure 5.4 it can be noted that the RMSE tends to scatter around 4 for the CE-Beta cases and around 0 for the CE-Normal cases after Rho value of 0.05 .

However, based on the processing time (Figure 5.5) the SC1 algorithms outperform the SC2 algorithms in both CE-Beta and CE-Normal cases. On average the CE-Normal

Table 5.2: Total running time of CE-Beta and CE-Normal with two SCs.

Algorithm	Running Time(s)	
	SC1	SC2
CE-Beta	5322.01	12916.66
CE-Normal	8546.73	27460.3

Table 5.3: Average RMSE for both Beta and Normal with two SCs when N=9.

Algorithm	RMSE	
	SC1	SC2
CE-Beta	3.6603	0.0665
CE-Normal	4.9598	0.6885

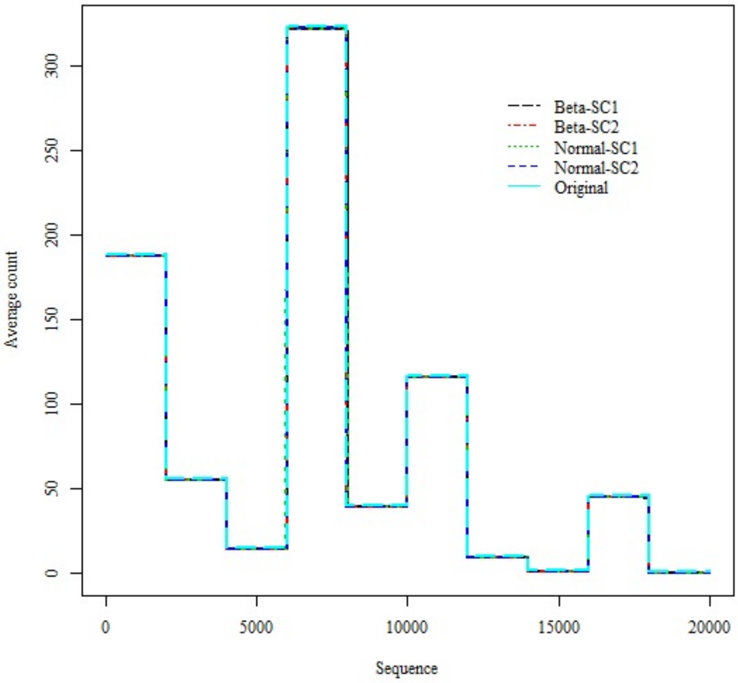


Figure 5.3: Average count vs. sequence position for CE-Beta and CE-Normal with two SCs

algorithms take more processing time than the CE-Beta cases.Therefore, we have to consider a Rho value that will balance the trade-off between the RMSE and the processing time. We have used the Rho value as 0.05 in this study to obtain the results. This is mainly based on the average RMSE results as discussed above.

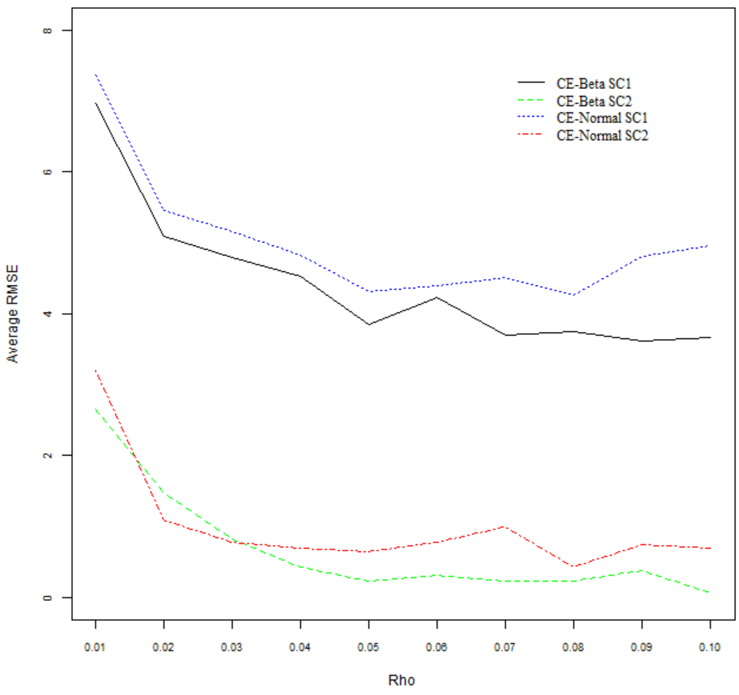


FIGURE 5.4: Plot of Average RMSE vs. Rho (ρ)

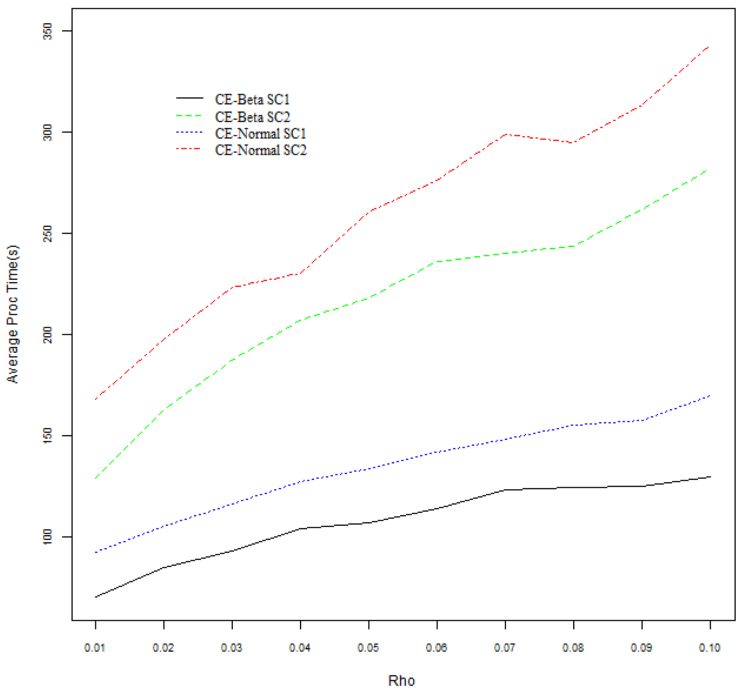


FIGURE 5.5: Plot of Average Processing time (s) vs. Rho (ρ)

Example 2: Real Data

This example considers a real DNA count data. The data correspond to the chromosome 2 of a subject in the study. Due to this being real data we do not know the true

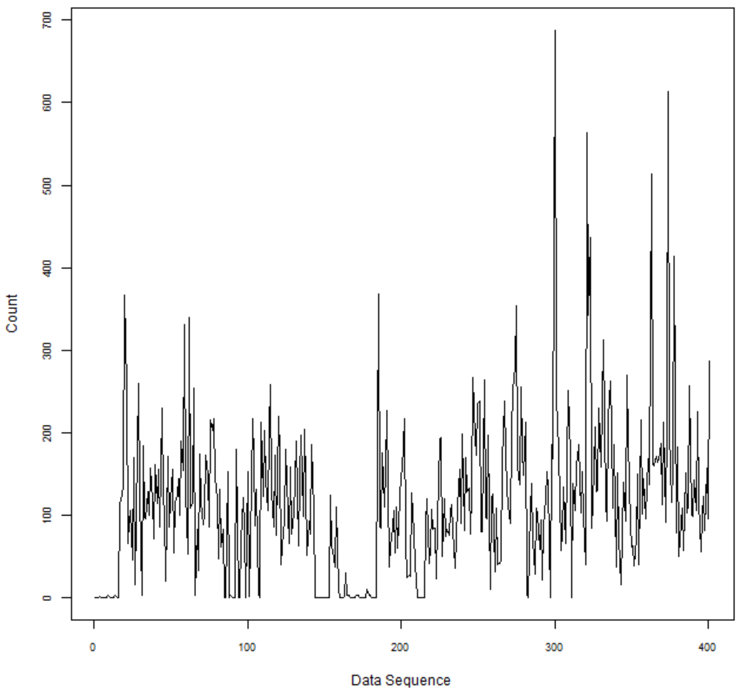


FIGURE 5.6: Part of the DNA count data

number of change-points in advance. Therefore, we look for agreement between the two methodologies. We have considered the proposed CE- Beta method and the CE- Normal method under the SC1 to compare the results. In order to calculate the $Nelite$ fraction of samples ρ value of 0.05 is used.

Figure 5.6 shows a portion of the DNA count data set that we have used in our study. Since, the data are highly over-dispersed; negative binomial distribution will model the process more informatively and accurately. Figure 5.7 shows the iterations results for the CE-Beta and CE-Normal under SC1. The optimum number of change-points is obtained by considering the combination of change-points that minimizes the BIC value. The BIC is minimized when N equals to 28 for the CE-Beta and 24 for the CE-Normal under the SC1. Table IV shows the running time for each of the cases under SC1 with number of change-points equal to 28 and 24 respectively.

Figure 5.8, the mean profile plot shows the agreement of the two methods in identifying the number of change-points in the DNA count data. In addition to the major regions that has also been captured by the CE-Normal algorithm, the proposed method has also identified few more small regions as well. Furthermore, as in Table IV the proposed CE-Beta method is computationally efficient compared to the CE-Normal method in detecting the locations of change-points of the DNA count data as well.

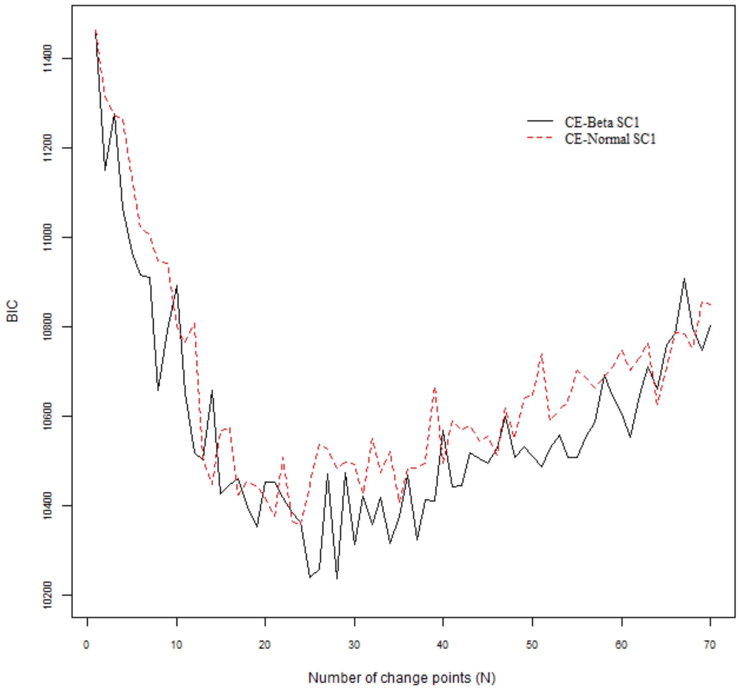


FIGURE 5.7: BIC vs. Number of change-points (N)

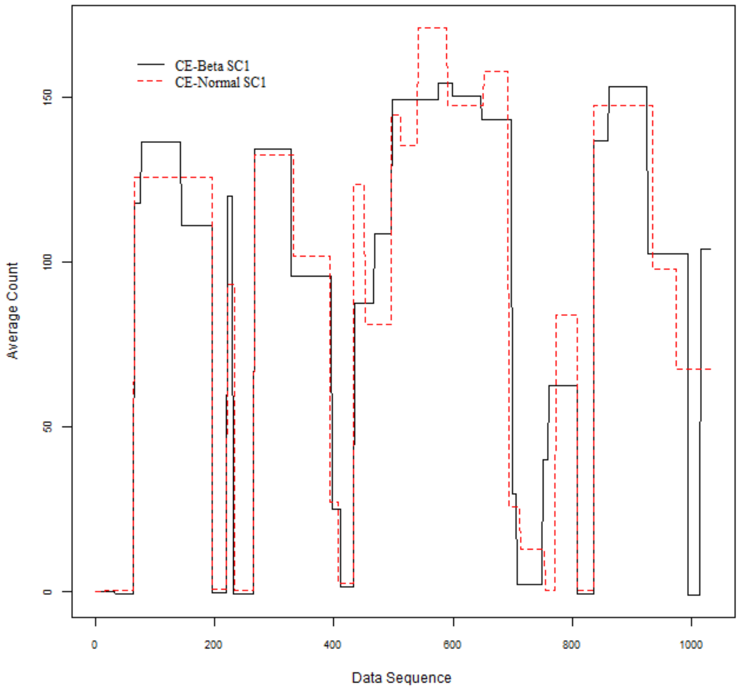


FIGURE 5.8: Average count vs. sequence position for CE-Beta and CE-Normal under SC1 of DNA count data

Table 5.4: Running time for the DNA count data with CE-Beta and CE-Normal under SC1

Algorithm	Running time (s)
CE-Beta	229.17
CE-Normal	555.54

Conclusion

A modified CE method is proposed with different stopping criteria. This proposed method utilizes beta distribution to simulate location of change-points in over dispersed count data. It was identified that the processing time under the proposed CE method is significantly less than the original CE method with respect to the two stopping criteria. However, CE algorithm with SC2 produced lower RMSE both in the proposed CE-Beta as well as the CE-Normal at the cost of high processing time.

While the results of this work are encouraging, there are plenty of avenues available for future research work, especially on smoothing up the CE algorithms and fine-tuning its parameters. In addition to that it would be helpful to investigate the possibilities of fine-tuning the penalty term in the BIC in the case of number of change-points is not known. A modified BIC will certainly help to obtain more smooth results and will more effectively address the dimension of the models with the increase of number of change-points.

Bibliography

[1] A.Kallioniemi, O.P. Kallioniemi, D. Sudar, D.Rutovitz, J.W. Gray, F. Waldman and D. Pinkel, “Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors, ” *Science*, vol. 258, pp. 818-821, 1992.

[2] C. Erdman and J. W. Emerson, “A fast Bayesian change point analysis for the segmentation of microarray data,” *Bioinformatics*, vol. 24, pp. 2143-2148, 2008.

[3] D. Barry and J. A. Hartigan,“A Bayesian analysis for change point problems,” *Journal of the American Statistical Association*, vol. 88, pp. 309-319, 1993.

[4] F. J. Anscombe, “The Statistical Analysis of Insect Counts Based on the Negative Binomial Distribution,” *Biometrics*, vol. 5, pp. 165-173, 1949.

[5] H. Cramer, *Mathematical methods of statistics*, Princeton University Press, Princeton, 1999.

[6] G. E. Evans, G. Y. Sofronov, J. M. Keith, and D. P. Kroese, “Identifying change-points in biological sequences via the cross-entropy method,” *Annals of Operation Research*, vol. 189, pp. 155-165, 2011.

[7] G. Schwarz, “Estimating the dimension of a model,” *The Annals of Statistics*, vol. 6, pp. 461-464, 1978.

[8] G. Y. Sofronov, G. E. Evans, J. M. Keith and D. P. Kroese, “Identifying Change-points in Biological Sequences via Sequential Importance Sampling,” *Environmental Modeling and Assessment*, vol. 14, pp. 577-584, 2009.

[9] J. V. Braun and H. G. Müller, “Statistical methods for DNA sequence segmentation,” *Statistical Science*, vol. 13, pp. 142-162, 1998.

[10] G. Sofronov, “Change-Point Modelling in Biological Sequences via the Bayesian Adaptive Independent Sampler,” In Proc. Computer Science and Information Technology, vol. 5, pp. 122-126, 2011.

[11] N. R. Zhang and D. O. Seigmund and H. Ji and J. Z. Li, “Detecting simultaneous changepoints in multiple sequences,” *Biometrica*, vol. 93, 631-645, 2010.

[12] R. Rubinstein and D. Kroese, *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning'*, Springer-Verlag, New York, 2004.

[13] R. Tibshirani and P. Wang, “Spatial smoothing and hot spot detection for CGH data using the fused lasso,” *Biostatistics*, vol. 9, pp. 18-29, 2008.

[14] S. Ivakhno, T. Royce, A. J. Cox, D.J. Evers, R.K. Cheetham and S.Tavar, “CNGseg-a novel framework for identification of copy number changes in cancer from second-generation sequencing data,” *Bioinformatics*, vol. 26, pp. 3051-3058, 2010.

[15] S. Li and R. Lund, “Multiple Changepoint Detection via Genetic Algorithms,” *Journal of Climate*, note=”doi: http://dx.doi.org/10.1175/2011JCLI4055.1”.

[16] T. Polushina and G. Sofronov, “Change point detection in biological sequences via genetic algorithm,” In Proc.IEEE Congress on Evolutionary Computation (CEC’2011), pp. 1966-1971, 2011.

[17] Y.Yao, “Estimating the number of change-points via Schwarz criterion,” *Statistics & Probability Letters*, vol. 6, pp. 181-189, 1988.

[18] N. R. Zhang and D. O. Siegmund, “A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data,” *Biometrics*, vol. 63, pp. 22-32, 2007.

Chapter 6

GAMLSS and Extended Cross-Entropy Method to Detect Multiple Change-Points in DNA Read Count Data

This chapter constitutes a peer-reviewed conference paper published in the proceedings of the 2013 International Workshop on Statistical Modelling (IWSM 2013), Palermo, Italy. The content of the paper including notation, plot sizes and wording has been slightly changed to improve the flow and the overall coherence of the thesis. The citation for the paper is:

Priyadarshana, W. J. R. M. and Sofronov, G. (2013). GAMLSS and Extended Cross-Entropy Method to Detect Multiple Change-Points in DNA Read Count Data, In: Muggeo VMR, Capursi V, Boscaino G, Lovison G (Eds.), Proceedings of the 28th International Workshop on Statistical Modelling, vol.1, 453-457, ISBN 978-88-96251-47-8.
Writing: 100% , analysis: 100% , conception and design: 90%

Specific contribution of joint authors
Georgy Sofronov: Overall supervision of data analysis and interpretation. Provided continual feedback and suggestions on both analysis and writing.

6.1 Summary

This chapter describes a further extension of the method discussed in Chapter 5. We explored the feasibility of incorporating auxiliary information into the process of detecting change-points with the CE method in DNA read count data. Generalized additive models for location, scale and shape (GAMLSS) [127] were considered to model DNA read count data. Apart from the issue of over-dispersion in the DNA read count data, we further observed an abnormally high frequency of zero values (a significant clump at zero). In order to account for this excess of zero observations and over-dispersion, we applied the zero-inflated negative binomial distribution (ZINB) to model DNA read count data in the GAMLSS framework. The GAMLSS modelling approach was considered in this work because of its extended capabilities in modelling more distributional parameters of the response variable. Furthermore, the GAMLSS model relaxes the exponential family assumption for the response variable by replacing it with a more general class of distributions, including highly kurtotic and/or skew continuous and discrete distributions. Therefore to model the extra variation and excess zeros we utilized the ZINB distribution in the GAMLSS model. We denote this model as ZINB-GAMLSS in this work. The exon length was considered as an explanatory variable when constructing the regression model. To our knowledge, this is the first implementation that utilizes the GAMLSS modelling framework with the CE method to detect change-points in NGS data that also incorporates auxiliary information. The proposed procedure was applied to the celiac disease data set as described in Chapter 5. We applied the method to DNA read count data corresponding to chromosome 15 of a patient. A comparison study was carried out with the negative binomial (NB), NB-GAMLSS, ZINB and ZINB-GAMLSS. It was observed that the ZINB-GAMLSS model more smoothly detected change-points in the read counts than the other methods.

GAMLSS and Extended Cross-Entropy Method to Detect Multiple Change-Points in DNA Read Count Data

W. J. R. M. Priyadarshana[1,*], Georgy Sofronov[1]

[1] Department of Statistics, Faculty of Science, Macquarie University, Sydney NSW 2109, Australia.

*E-mail: madawa.weerasinghe@mq.edu.au

Abstract

We model DNA read count data obtained through next generation sequencing (NGS) technologies as a multiple change-point process. This means that the data are divided into different segments based on the number of change-points. Each segment of the process is modeled by utilizing the zero-inflated negative binomial (ZINB), as well as the negative binomial (NB) distribution in the Generalized additive models for location, scale and shape (GAMLSS) framework. It is observed that ZINB and NB based models, fit the data better than the competing Poisson model, in which the observed read counts are highly over-dispersed as well as zero-inflated. Moreover, we have considered incorporating auxiliary information to further improve the change-point modelling process by utilizing the GAMLSS framework. The extended Cross-Entropy (CE) method which uses a four-parameter beta distribution is used to estimate the number of change-points as well as their corresponding genome locations. Furthermore, parallel implementation of the procedure results a significant improvement in total running time, in which the procedures are highly computationally intensive. We apply the proposed methodology to find change-points in DNA read count data obtained through Illumina TruSeq exome capture of patients with celiac disease. Our results suggest that the proposed GAMLSS based CE method is an effective methodology to detect change-points in genome-wide data.

Introduction

Discovering chromosomal aberrations in the genomic DNA is a widely discussed issue that has been addressed through various scientific techniques based on different perspectives.

It is an established fact that the variations in DNA copy number is a source of genetic variation [Campbell et al., 2008] even though the full understanding of the effect of these is still on probe. Recent studies based on microarray technology have identified around 12% of the human genome and thousands of genes are variable in copy number. It is predicted that the emerging technologies with high sensitivity level of data will further expand this knowledge.

Prior to the advent of next generation sequencing (NGS) technologies, number of methodologies have been developed to detect multiple change-points mainly based on the array-comparative genomic hybridization (aCGH) data. Analysis on aCGH data aims to find changes in the mean of the fluorescence color ratios, usually on logarithm scale to detect copy number variations in the human genome. See [Lai et al., 2005] for a review of the aCGH based segmentation methods. However, the introduction of the next generation DNA sequencing technologies and the resulted excess amount of data has increased the complexity level of the process of partitioning the genome in to homogeneous segments to a higher level.

Reviewing the literature on change-point modelling of NGS data, [Xie and Tammi, 2009] proposed a method called CNVseq to identify CNVs on the data generated through shotgun sequencing. Later [Magi et al., 2012] reviewed some of the existing methodologies to detect CNV in read count data. They have normalized the raw read counts and conducted the segmentation based on the techniques mainly developed on aCGH data. They also mentioned that there exist only few statistical procedures that utilize the raw read counts to detect CNVs. In fact, most of the prevailing methods transform the raw read counts by different normalization techniques to a stage, where they can utilize the existing aCGH based segmentation methods. They have not considered utilizing auxiliary information in the generalized linear models (GLM) context to detect multiple change-points in the read count data.

In order to fill this gap in the literature of direct usage of the DNA read counts generated by the NGS platforms, we propose a procedure which utilizes generalized additive models for location, scale and shape (GAMLSS) statistical framework [Rigby and Stasinopoulos, 2005] to incorporate auxiliary information into the modelling process, and extended Cross-Entropy method [Priyadarshana and Sofronov, 2012] to estimate the number of change-points as well as their corresponding genome locations. We observe that the DNA read counts we analyzed are highly over-dispersed as well as zero-inflated. Therefore, the response variable is modelled by utilizing the zero-inflated negative binomial (ZINB) as well as the negative binomial (NB) distribution in the GAMLSS framework.

Multiple Change-Point Problem, GAMLSS Framework and CE Method

Generalized additive models for location, scale and shape (GAMLSS)

The GAMLSS are a type of semi-parametric regression models use to model univariate response with a set of covariates. It allows for modelling not only the expected mean but other parameters of the distribution (e.g. location, scale and shape) of the response variable as well. Therefore, it gives more flexibility in modelling process than the generalized additive models, and GLMs.

Multiple Change-Point Problem

Let us formulate the multiple change point problem in mathematical terms. A count data sequence $y = (y_1, y_2, ..., y_L)$ of length L is given. A segmentation of the sequence is specified by the number of change points N and the positions of the change points $c = (c_1, c_2, ..., c_N)$, where $1 = c_0 < c_1 < ... < c_N < c_{N+1} = L + 1$. In this context, a change point is a boundary between two adjacent segments. The value of c_i is the sequence position of the rightmost character of the segment to the left of the i^{th} change-point. Segments are numbered from 0 to N as there will be one or more segments than number of change points. We model each segment of the DNA read count data utilizing ZINB and NB distribution in the GAMLSS regression framework with the use of exon length as the covariate.

Extended Cross-Entropy Method

The CE method [Rubinstein and Kroese, 2004] is a new generic approach to combinatorial and multi-extremal optimization and rare event simulation. Broadly it can be used to solve estimation and optimization problems. In this study, the process of multiple change point detection is considered as a minimization problem. The CE method is an iterative optimization procedure that starts with a parameterized sampling distribution from which a random sample is generated. Then, each observation or the combinatorial arrangement is scored for its performance, as the solution to a specified optimization problem. A fixed number of best performing combinatorial arrangements are referred to as the elite sample. This elite sample is subsequently used to update the parameters for the sampling distribution. Thus, adaptive parameters are utilized in each iteration. The sampling distribution eventually converges to a degenerate distribution about a locally optimal solution, which ideally will be globally optimal.

In this study we utilize the extended version of the standard CE method, proposed in [Priyadarshana and Sofronov, 2012] with further modifications. We use a stopping criterion (SC) based on Median Absolute Deviation (MAD) as opposed to the variance based SC proposed in the original paper. Furthermore, a multi-core architecture based parallel implementation of the algorithm is implemented in order to carry out calculations more efficiently.

Results and Conclusions

In this section, we include results of numerical experiments that illustrate the performance of the proposed method. This example considers a DNA read count data obtained from a study of celiac disease patients. All data were obtained from the Illumina TruSeq exome capture technology. We analyze DNA read count data with respect to chromosome 15 of a patient.

We compare the change-point modelling process with and without the effect of auxiliary information utilizing the extended CE method. In the case without any predictor variables, we model the read count data based on both NB and ZINB distributions. In the process of utilizing auxiliary information and the GAMLSS implementation, we consider natural logarithm of the exon length as a predictor variable. In the GAMLSS framework, we carry out the change-point modelling procedure considering the distribution of the response variable as zero-inflated negative binomial (ZINB-GAMLSS) as well as negative binomial (NB-GAMLSS).

Figure 6.1 shows the mean profile plots of results for the case, with and without any predictor variables. It is visible that in the ZINB set up, the GAMLSS approach and the ZINB results have a higher level of concordance of estimated change-points, when compared to the NB results. This may be due to the fact that ZINB better models the observed read counts than the NB. It can be further noticed that in general NB based models have estimated more change-points than the ZINB based models for this particular DNA read count data. While the results of this work are encouraging, there are plenty of avenues available for future research work, especially on the implementation of GAMLSS framework and the incorporation of more predictor variables to the modelling process. Furthermore, cluster level implementation of the methodology will certainly improve the processing time, in which all these processes are highly computationally intensive.

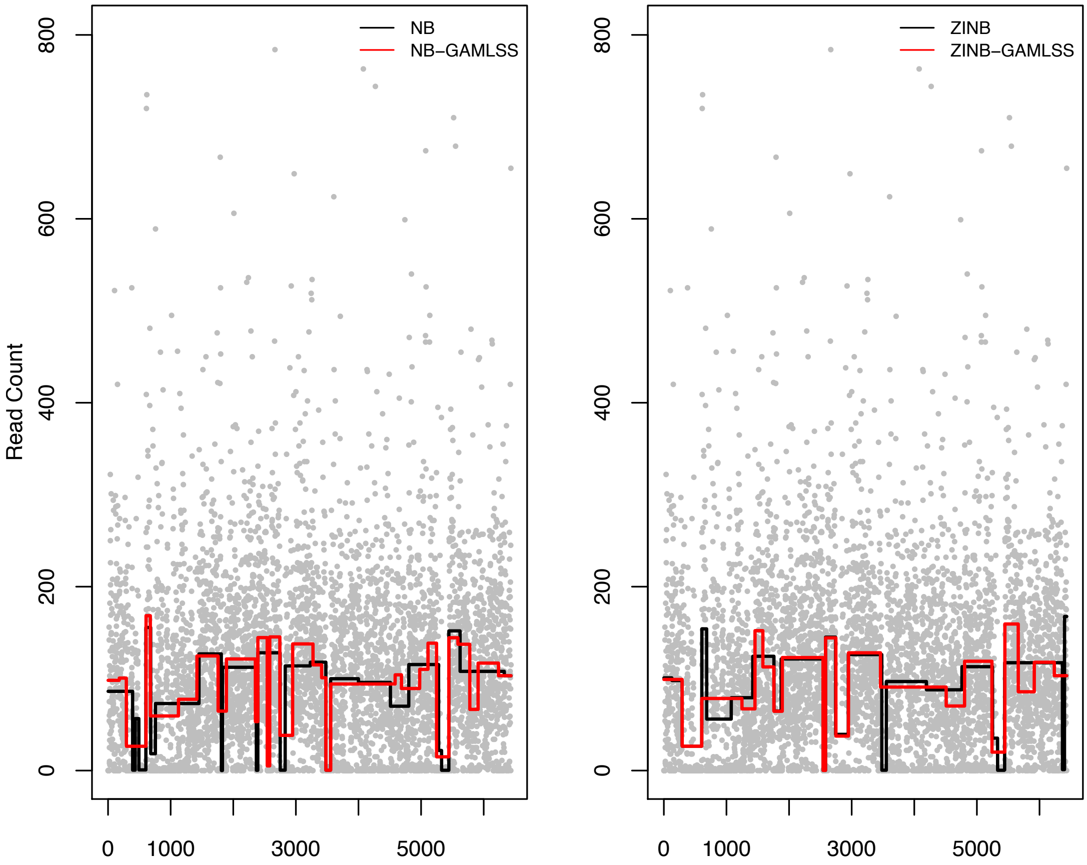


FIGURE 6.1: Mean profile plots for NB, NB-GAMLSS, ZINB, ZINB-GAMLSS.

Acknowledgment

We thank Dr. Vincent Plagnol (UCL Genetics Institute, University College London, Grower Street, London, UK) for providing celiac disease patients read count data for the analysis.

References

Campbell, P.J., et al. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genetics*, **6**, 722 – 729.

Lai, W.R., et al. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **19**, 3763 -– 3770.

Magi, A., et al. (2012). Read count approach for DNA copy number variants detection. *Bioinformatics*, **4**, 470-–478.

Priyadarshana, W. J. R. M. and Sofronov, G. (2012). A modified Cross Entropy Method for Detecting Multiple Change Points in DNA Count Data. In: *Proceedings of the IEEE World Congress on Computational Intelligence (IEEE CEC 2012)*, Brisbane, pp. 1020 – 1027.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Applied Statistics*, **54**, 507 – 554.

Rubinstein, R. and Kroese, D. (2004). *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning.* New York: Springer-Verlag.

Xie, C. and Tammi, M.T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, **10**, 80.

Chapter 7

breakpoint: An R Package to Detect Multiple Change-Points via the CE Method

This chapter illustrates the use of an R package named "breakpoint" developed to detect multiple change-points in continuous and discrete (count) data via the variants of the CE method proposed in Chapters 3, 4 and 5. It can be freely obtained from the CRAN [122] (http://cran.r-project.org/web/packages/breakpoint/index.html). The package manual is included as an appendix (see Appendix B). The citation for the R package is:

Priyadarshana, W. J. R. M. and Sofronov, G. (2014). breakpoint: An R Package for Multiple Break-Point Detection via the Cross-Entropy Method. R package version 1.1.
Coding: 100% , writing: 100% , conception and design: 95%

Specific contribution of joint authors
Georgy Sofronov: Overall supervision of the project. Provided continual feedback and suggestions.

7.1 Summary

In this thesis we have developed several algorithms based on the CE method as described in Chapters 3 to 6 to detect the number of change-points and their locations, particularly in biological sequences. These algorithms cover a broad spectrum of applications that can be utilized to detect multiple change-points in continuous (e.g., aCGH data) and discrete measurements (e.g., DNA read count data). The current version of the package (version 1.1.) is capable of analyzing both continuous and discrete measurements. The package was originally prepared (version 1.0) along with the publication in Chapter 3, which discusses the use of CE method in detecting multiple change-points in aCGH data. More recently an update to the package (version 1.1) was submitted with improved capabilities and additional functions. All the calculations of the CE algorithm can also be carried out with parallel computation techniques as described in Chapter 3. Additionally, the breakpoint package implements the concept of smooth parameter update in the CE algorithm as described in Chapter 2.

The normal distribution is used to model continuous data. The R function named "CE.Normal" performs all the calculations of the CE algorithm to estimate the number and the locations of the change-points. In the package there are two options available for simulation of change-point locations in the CE algorithm. They are the four-parameter beta distribution and the truncated normal distribution, as described in Chapter 2. The median absolute deviation is used as the stopping criterion in the CE algorithm. The modified BIC [174], described in Chapter 2, is used as the model selection criterion to obtain the optimal solution for the number of change-points. Finally, a list containing the number of break-points and the locations are reported.

The other two main functions "CE.NB" and "CE.ZINB" can be used to model count data with the negative binomial distribution and zero-inflated negative binomial distribution respectively. A model with zero-inflated negative binomial is preferred over the negative binomial model when there exists an excess amount of zero observations beyond those explained by over-dispersion. The general BIC [135] is used as the performance function in the CE algorithm and the median absolute deviation is used as the stopping criterion. These functions are also supported with the arguments and capabilities described in the "CE.Normal" function. The mean profile plot of the segmented data can be obtained by using the "profilePlot" function in the package.

In this chapter, we provide case studies to illustrate the capabilities of the functions in the breakpoint R package. A series of extensive simulation studies and comparative analyses were performed in [115–117, 119, 120, 167] to assess the performance of the proposed procedures. The manual of the package is included in Appendix B.

7.2 Case studies on analyzing continuous data

Two case studies are discussed in this section. The first study involves artificially generated data and the second study involves a real aCGH data set that is publicly available. A detailed simulation study was performed in [167] to assess the overall effectiveness of the CE algorithm.

7.2.1 Artificially generated data: Normal

The following R script was used to simulate data from the normal distribution. Let $\mathbf{y} = (y_1, y_2, \ldots, y_L)$ be a sequence of independent normal random variables with varying mean values and a common standard deviation. Let the sequence have length $L = 10000$ and five change-points.

Data generation

```r
segs <- 6 # Number of segments
M <- c(1500, 2200, 800, 2500, 1000, 2000) # Length of the segments
#chpnt.locs <- c(1501, 3701, 4501, 7001, 8001) # True change-point
 locations
seg <- NULL
mu.vals <- c(10, 5, 0, 3, 8, 5) # Mean value for each segment
sd.val <- 2 # Value of the standard deviation for each segment

for(j in 1:segs){
 seg <- c(seg, rnorm(M[j], mean = mu.vals[j], sd = sd.val))
}

simdata<- as.data.frame(seg)
```

Analysis with the four-parameter beta distribution

The “CE.Normal” function is used to perform the necessary calculations. We use parallel computation and alter the aberration width to 10, but otherwise use default values.

The option “distyp = 1” instructs the CE algorithm to use the four-parameter beta distribution to simulate change-point locations.

```
library(breakpoint) # Load the installed R package

obj1 <- CE.Normal(simdata, distyp = 1, parallel = TRUE)
obj1


## $No.BPs
## [1] 5
##
## $BP.Loc
##
## 1501 3700 4508 7001 8002


profilePlot(obj1, simdata) # Mean profile plot
```

Figure 7.1 shows the simulated data, along with the mean profile plot depicting the estimated change-points and means.

Analysis with the truncated normal distribution

The same “CE.Normal” function is used as earlier. We use parallel computation and alter the aberration width to 10, but otherwise use default values. The option “distyp = 2” instructs CE algorithm to use the truncated normal distribution to simulate change-point locations.

```
library(breakpoint) # Load the installed R package

obj2 <- CE.Normal(simdata, distyp = 2, h = 10, parallel = TRUE)
obj2


## $No.BPs
## [1] 6
##
## $BP.Loc
```

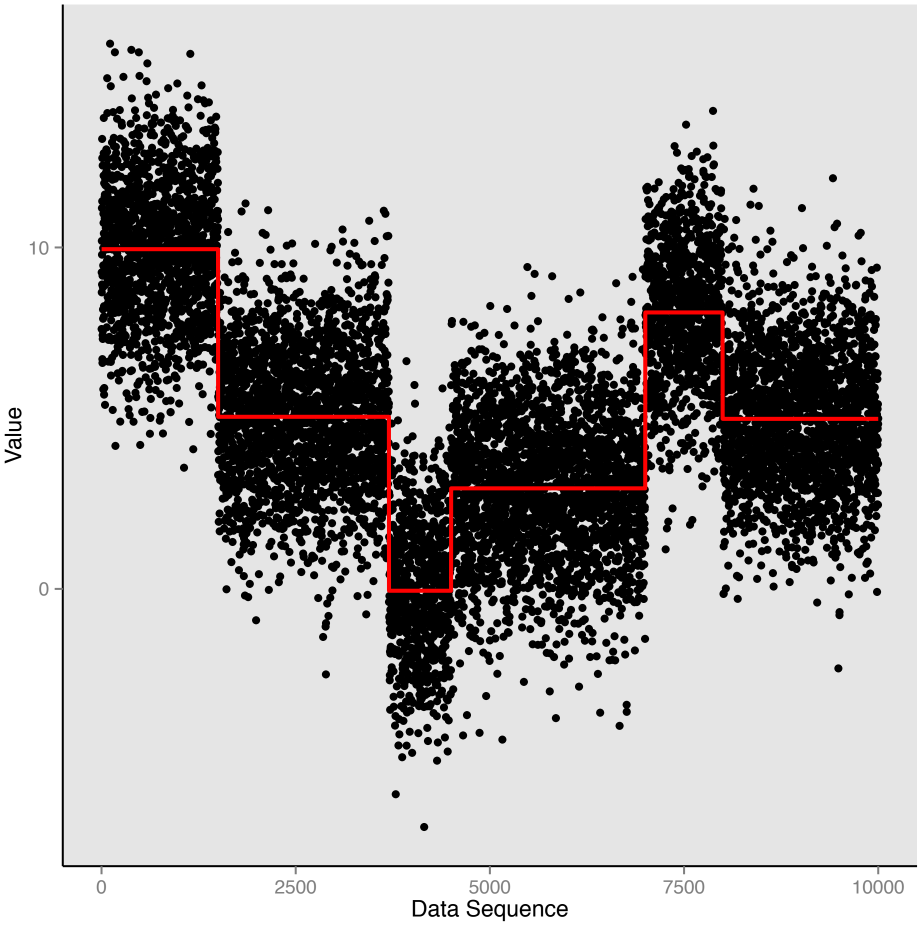


Figure 7.1: Mean profile plot of the simulated data: CE with the four parameter beta

```
##
## 1501 3701 4501 4513 7001 7997


profilePlot(obj2, simdata)
```

Figure 7.2 shows the mean profile plot of the simulated data with the estimated change points. It is observed that the "CE.Normal" with the truncated normal distribution has slightly over-estimated the true change-point number. We have discussed the performance differences of the CE method with the four parameter beta distribution and the truncated normal distribution in Chapter 5. It was found that the CE method with the four parameter beta distribution performs better than the CE method with the truncated normal distribution both in terms of efficiency and accuracy [119].

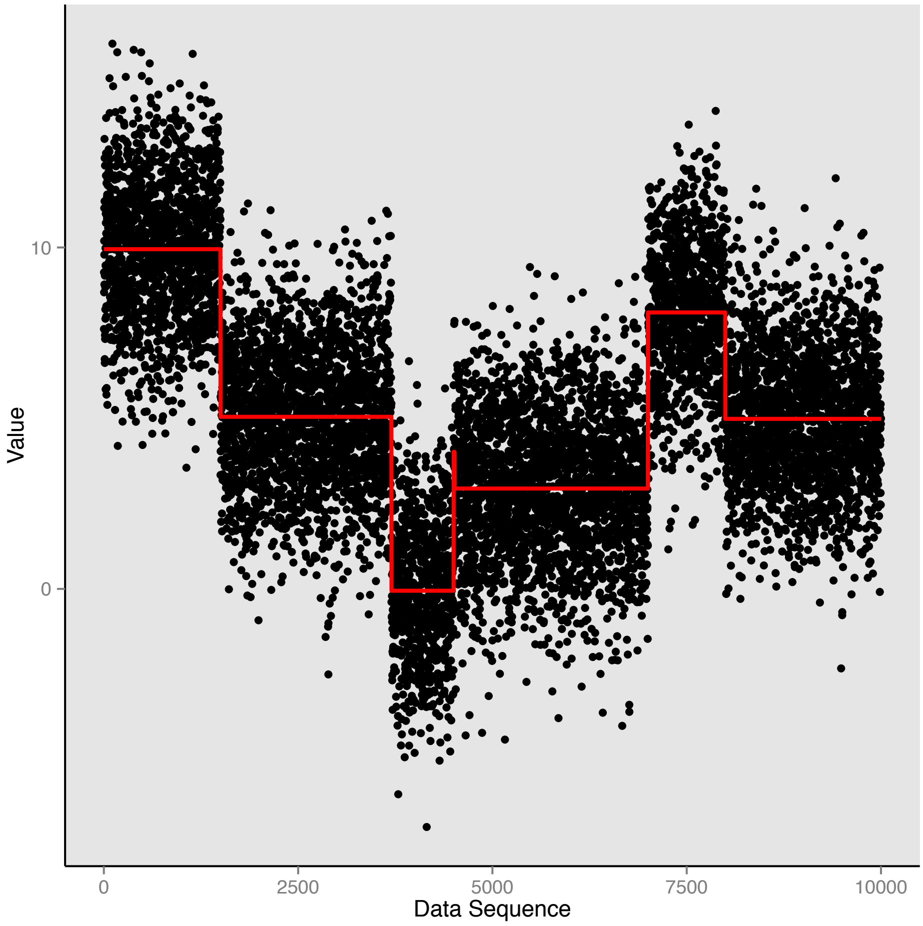


Figure 7.2: Mean profile plot of the simulated data: CE with truncated normal

7.2.2 Real data

This example aCGH data set was first discussed in [168]. The paper analyzed aCGH data of nine squamous cell carcinomas (SCCs), seven adenocarcinomas (AdCAs) and four human papillomavirus (HPV)-immortalized keratinocyte cell lines. Here, we consider a subset of this data set. We analyze log-ratio data corresponding to one AdCA and three SCC cell lines. They are AdCA10, SCC27, SCC36 and SCC39. In those cell lines, we particularly analyze the chromosome 11 data.

The "CE.Normal" function with the four-parameter beta distribution to simulate change-point locations is used here to obtain the estimates of the change-points. First we shall load the data file into the R environment. Then the "CE.Normal" function is used to estimate the number of change-points and their locations. Figure 7.3 gives the combined profile plots for the four cell lines.

```
library(breakpoint) # Load the installed R package

wilting.data <- read.table(file=paste("/Users/MKMM/Desktop/Wilting.txt"),
sep="\t", header=T)

wilting.AdCA10.ch11 <- subset(wilting.data[,c(2,3,4)], wilting.data[,2]==11)
wilting.AdCA10.ch11.log2 <- as.data.frame(wilting.AdCA10.ch11[,3])

wilting.SCC27.ch11 <- subset(wilting.data[,c(2,3,5)], wilting.data[,2]==11)
wilting.SCC27.ch11.log2 <- as.data.frame(wilting.SCC27.ch11[,3])

wilting.SCC36.ch11 <- subset(wilting.data[,c(2,3,7)], wilting.data[,2]==11)
wilting.SCC36.ch11.log2 <- as.data.frame(wilting.SCC36.ch11[,3])

wilting.SCC39.ch11 <- subset(wilting.data[,c(2,3,8)], wilting.data[,2]==11)
wilting.SCC39.ch11.log2 <- as.data.frame(wilting.SCC39.ch11[,3])

obj1 <- CE.Normal(wilting.AdCA10.ch11.log2, distyp = 1,  parallel = TRUE)
obj1


## $No.BPs
## [1] 2
##
## $BP.Loc
##
##  18 110


profilePlot(obj1, wilting.AdCA10.ch11.log2, x.label="Genomic Position",
y.label="Log2ratio Data")
```

```
obj2 <- CE.Normal(wilting.SCC27.ch11.log2, distyp = 1,  parallel = TRUE)
obj2


## $No.BPs
## [1] 3
##
```

```
## $BP.Loc
##
##  63  90 120

profilePlot(obj2, wilting.SCC27.ch11.log2, x.label="Genomic Position",
y.label="Log2ratio Data")
```

```
obj3 <- CE.Normal(wilting.SCC36.ch11.log2, distyp = 1,  parallel = TRUE)
obj3

## $No.BPs
## [1] 2
##
## $BP.Loc
##
## 61 83

profilePlot(obj3, wilting.SCC36.ch11.log2, x.label="Genomic Position",
y.label="Log2ratio Data")
```

```
obj4 <- CE.Normal(wilting.SCC39.ch11.log2, distyp = 1, parallel = TRUE)
obj4

## $No.BPs
## [1] 4
##
## $BP.Loc
##
##  65  88 109 117

profilePlot(obj4, wilting.SCC39.ch11.log2, x.label="Genomic Position",
y.label="Log2ratio Data")
```

It was reported in [168] that there exists chromosomal abnormalities in the cell lines considered in this example. Particularly, in the cell line AdCA10 chromosome 11, a

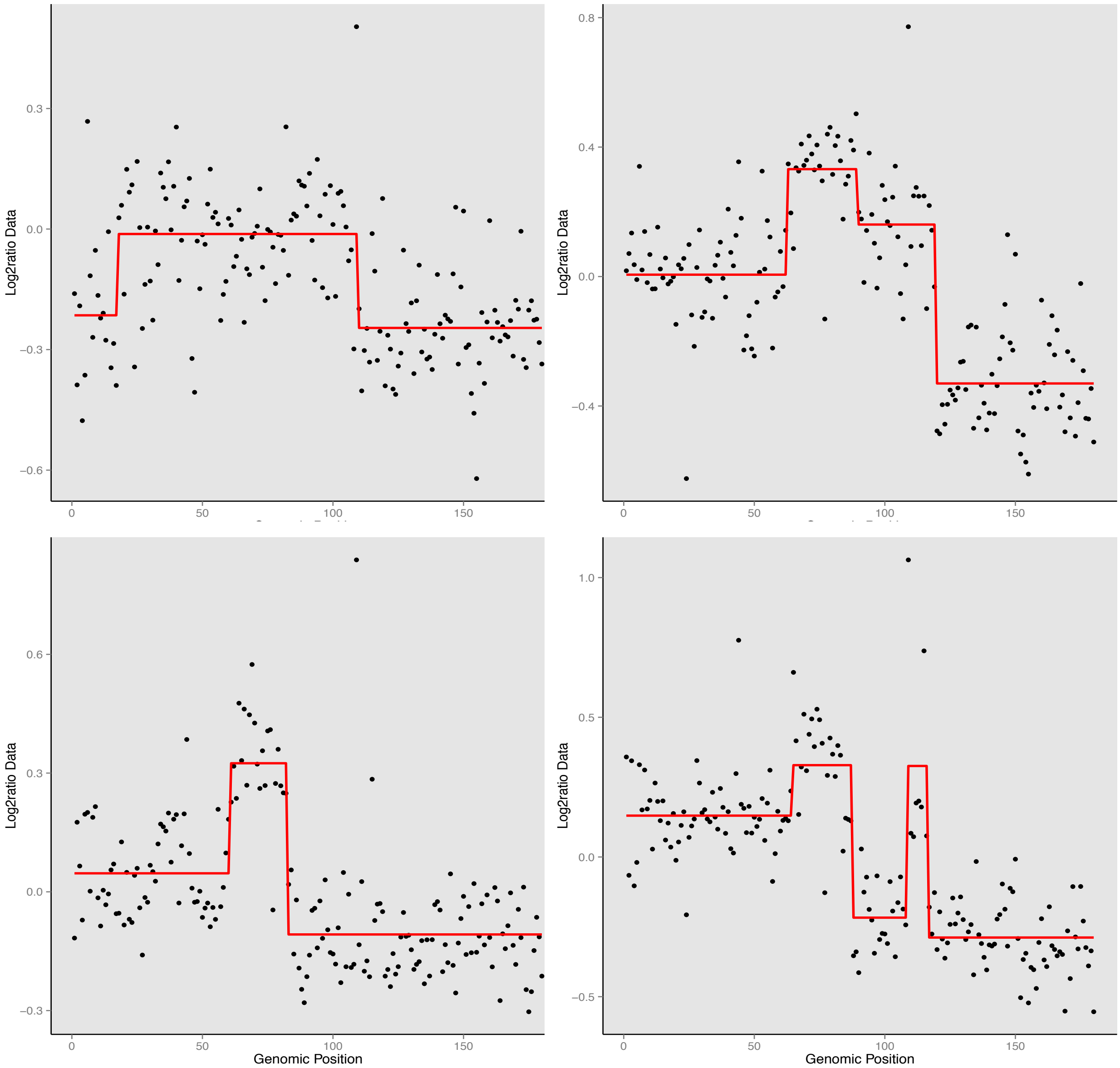


FIGURE 7.3: Mean profile plots of chromosome 11 data. A: AdCA10, B: SCC27, C: SCC36 and D: SCC39.

region of reduced copy numbers was identified. The CE method has also estimated two change-points referring to these losses in copy number. In cell lines SCC27, SCC36 and SCC39 both lost and gained regions of the chromosome 11 were observed. Our results also support this conclusion (see Figure 7.3).

7.3 Case studies on analyzing count data

In this section we provide two case studies to illustrate the usage of the functions "CE.NB" and "CE.ZINB" in the breakpoint R package, which can be used to detect change-points in discrete (count) data.

7.3.1 Artificially generated data: Negative Binomial Distribution

The negative binomial distribution, which is also known as the gamma-Poisson mixture distribution, is used to model count data when the equi-dispersion assumption is violated. More specifically it is used in the literature to account for over-dispersion [8]. We consider the particular parameterisation of the negative binomial distribution discussed in Chapter 5.

We use the following R script to simulate the data set. Let $\mathbf{y} = (y_1, y_2, \ldots, y_L)$ be a sequence of independent negative binomial random variables. The length of the sequence is denoted by L. There are five change-points in the simulated data sequence.

Data generation

```
segs <- 6 # Number of segments
M <- c(1500, 2200, 800, 2500, 1000, 2000) # Length of each segment
#chpnt.locs <- c(1501, 3701, 4501, 7001, 8001) # True change-point locations
seg <- NULL
p <- c(0.45, 0.25, 0.4, 0.2, 0.3, 0.6) # Specification of the probabilities

for(j in 1:segs){
 seg <- c(seg, rnbinom(M[j], size = 10, prob = p[j]))
}

simdata <- as.data.frame(seg)
```

Analysis with the four-parameter beta distribution

The "CE.NB" function is used to perform the calculations. We use parallel computation and alter the aberration width to 10, but otherwise use default values. The option "distyp = 1" instructs the CE algorithm to use the four parameter beta distribution to simulate change-point locations.

```
library(breakpoint) # Load the installed R package

obj1 <- CE.NB(simdata, distyp = 1, h = 10,  parallel = TRUE)
obj1


## $No.BPs
## [1] 5
##
## $BP.Loc
##
## 1501 3699 4501 7001 8001


profilePlot(obj1, simdata)
```

Figure 7.4 shows the mean profile plot of the simulated data with the estimated change-points. It is observed that the CE algorithm has correctly identified the number of change-points and estimated the locations accurately.

Analysis with the truncated normal distribution

The "CE.NB" function is used as earlier. The option "distyp = 2" instructs the CE algorithm to use the truncated normal distribution to simulate change-point locations.

```
library(breakpoint) # Load the installed R package

obj2 <- CE.NB(simdata, distyp = 2, h = 10, parallel = TRUE)
obj2


## $No.BPs
## [1] 4
##
## $BP.Loc
##
## 1501 4501 7001 8001


profilePlot(obj2, simdata)
```

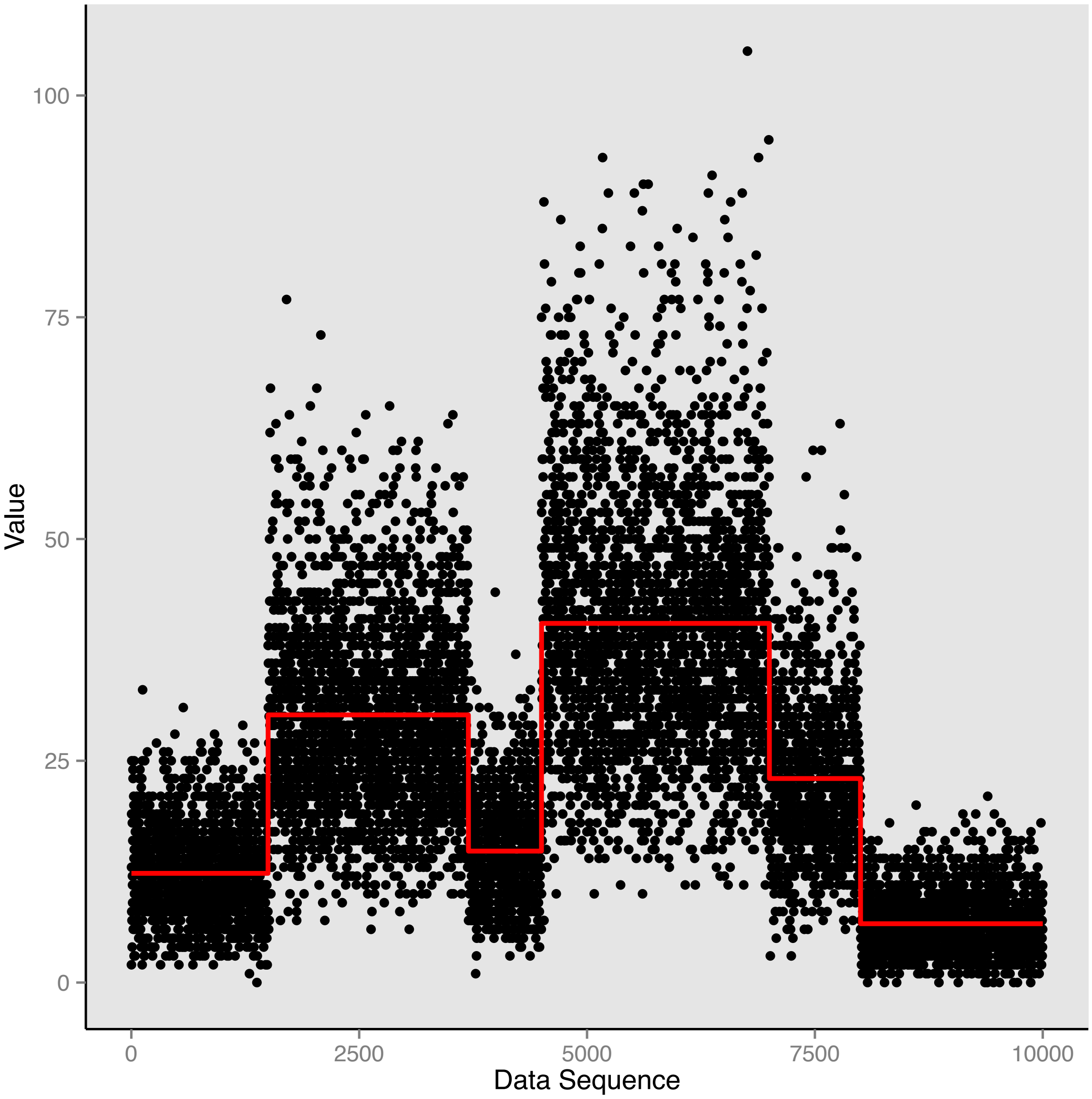


FIGURE 7.4: Mean profile plot of the simulated data: CE with the four parameter beta

Figure 7.5 depicts the mean profile plot. The “CE.NB” with the truncated normal distribution has slightly under-estimated the true number of change-points. We observe a similar behaviour of the CE algorithm with the truncated normal distribution as seen in the continuous data example.

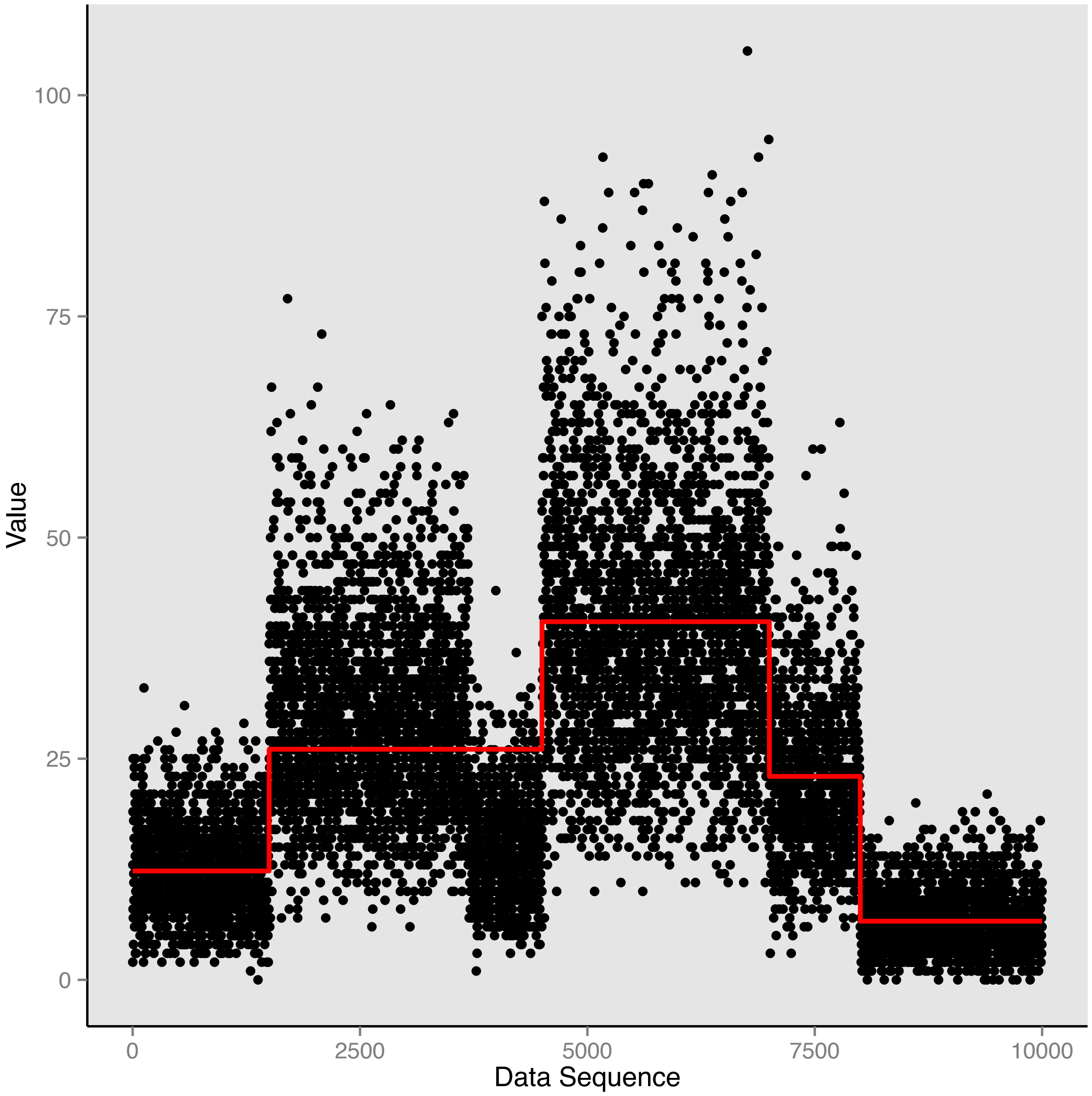


Figure 7.5: Mean profile plot of the simulated data: CE with truncated normal

7.3.2 Artificially generated data: Zero-Inflated Negative Binomial

The Zero-Inflated Negative Binomial (ZINB) is used to model over-dispersion and excess zero values in the observed count data. We use the "gamlss" R package [127] to simulate data from the ZINB distribution.

The following R script is used to generate data from the zero-inflated negative binomial distribution by using the gamlss R package.

Data generation

```
library(gamlss) # Load gamlss package

segs <- 6 # Number of segments
M <- c(1500, 2200, 800, 2500, 1000, 2000) # Length of each segment
#chpnt.locs <- c(1501, 3701, 4501, 7001, 8001) # True change-point locations
seg <- NULL
p <- c(0.6, 0.1, 0.3, 0.05, 0.2, 0.4) # Specification of the probabilities
sigma.val <- c(1,2,3,4,5,6) # Specification of sigma values

for(j in 1:segs){
 seg <- c(seg, rZINBI(M[j], mu = 300, sigma = sigma.val[j], nu = p[j]))
}

simdata <- as.data.frame(seg)
```

Analysis with the four-parameter beta distribution

The “CE.ZINB” function is used to perform the calculations. We use parallel computation and alter the aberration width to 10, but otherwise use default values. The option “distyp = 1” instructs the CE algorithm to use the four parameter beta distribution to simulate change-point locations.

```
library(breakpoint) # Load the installed R package

obj1 <- CE.ZINB(simdata, distyp = 1, h = 10,  parallel = TRUE)
obj1

## $No.BPs
## [1] 5
##
## $BP.Loc
##
## 1476 3701 4494 7010 8023
```

```
profilePlot(obj1, simdata)
```

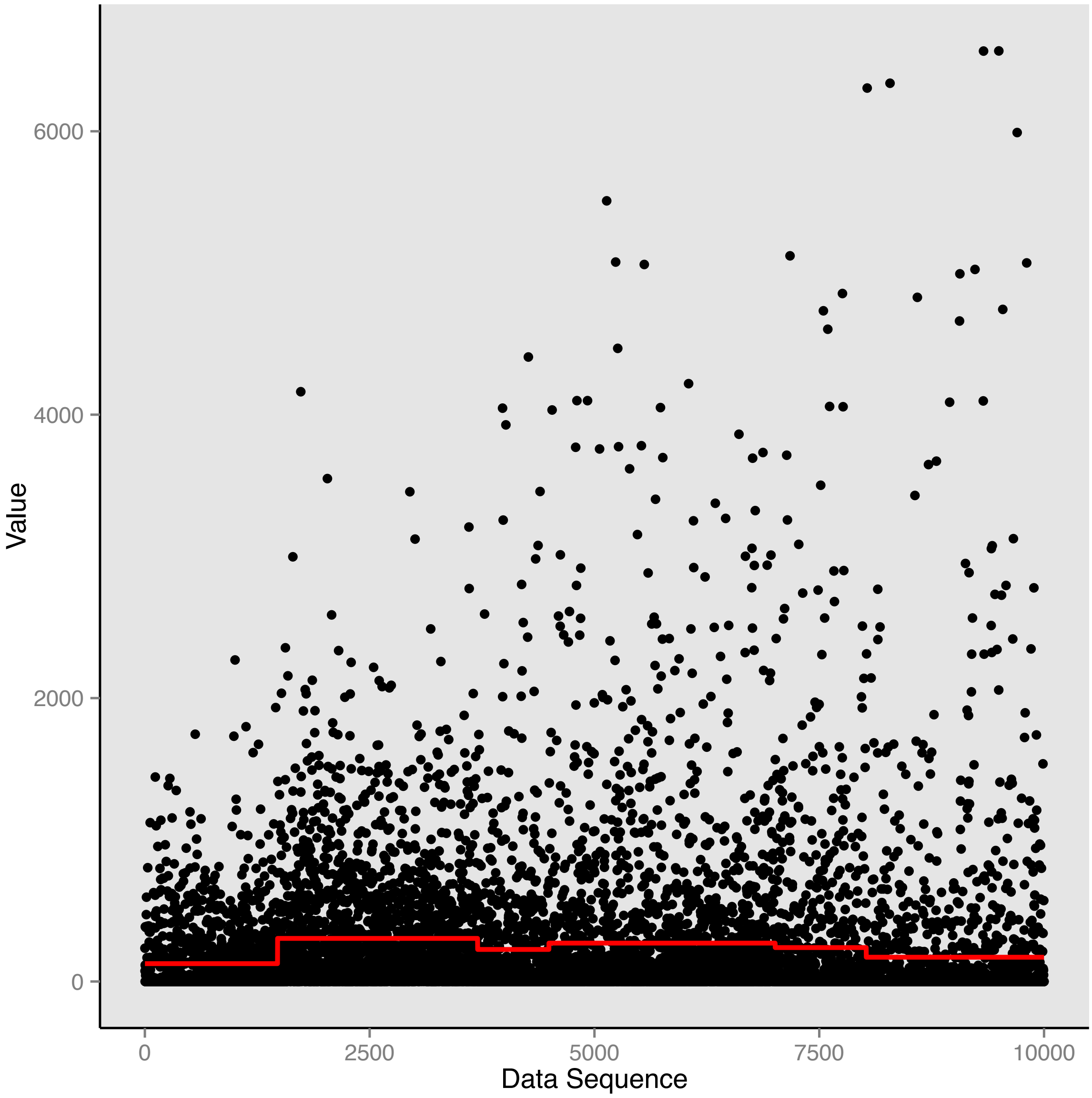


Figure 7.6: Mean profile plot of the simulated data: CE with the four parameter beta

Figure 7.6 shows the mean profile plot of the simulated data. The CE method has correctly identified the number of change-points with fairly accurate estimates for their locations.

Analysis with the truncated normal distribution

The "CE.ZINB" function is used as earlier. The option "distyp = 2" instructs the CE algorithm to use the truncated normal distribution to simulate change-point locations.

```
library(breakpoint) # Load the installed R package

obj2 <- CE.ZINB(simdata, distyp = 2, h = 10, parallel = TRUE)
obj2

## $No.BPs
## [1] 5
##
## $BP.Loc
##
## 1494 3701 5136 7010 7774

profilePlot(obj2, simdata)
```

Figure 7.7 shows the mean profile plot of the simulated data. The CE method has correctly identified the number of change-points with fairly accurate estimates for their locations.

In this Chapter, we illustrated the main usage of the functions in the breakpoint R package. In all the case studies, we execute the variants of the CE methods only once. We refer to [115–117, 119, 120, 167] for detailed information about the performance of the proposed procedures.

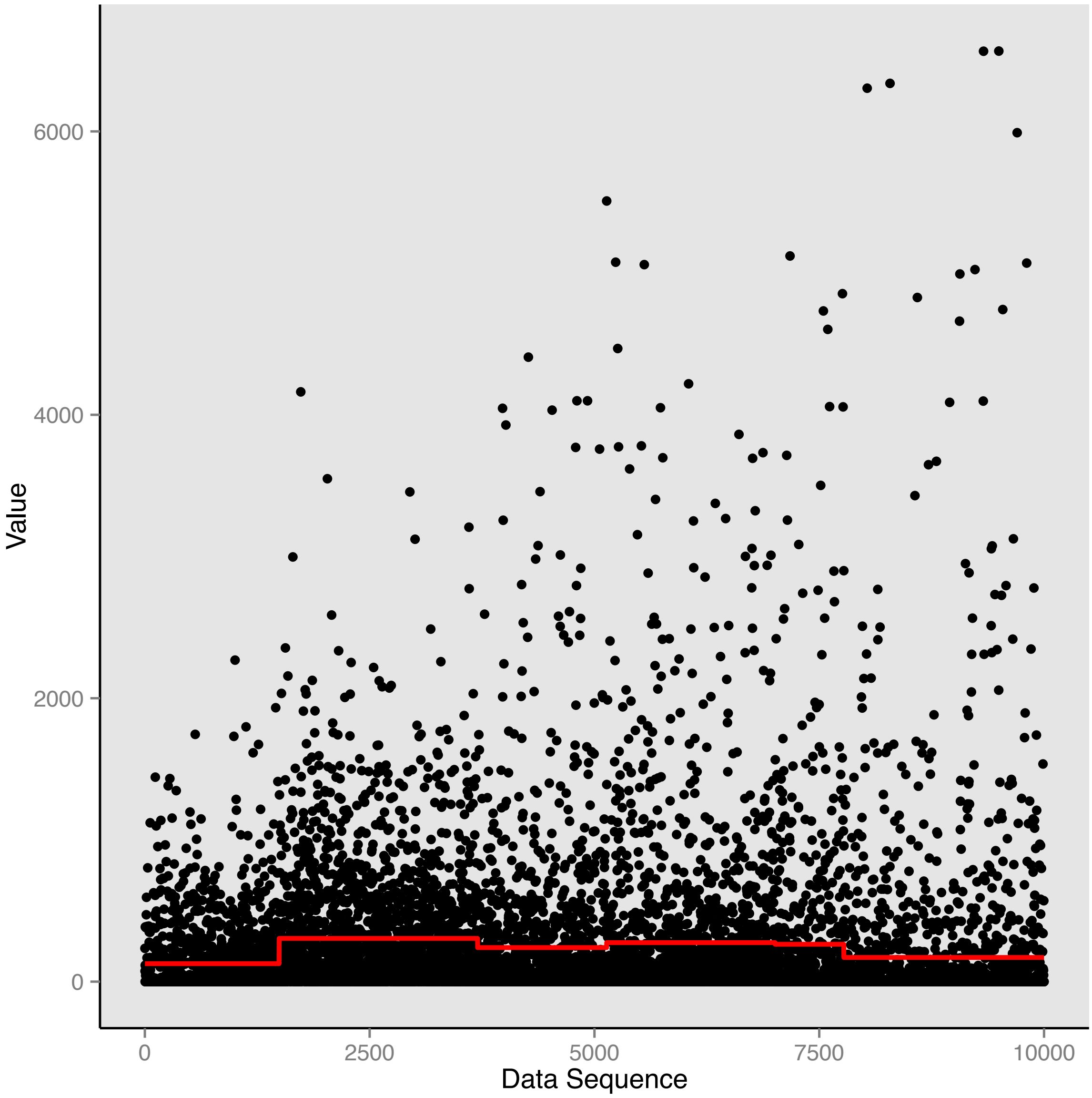


FIGURE 7.7: Mean profile plot of the simulated data: CE with truncated normal

Chapter 8

Discussion and Future Directions

This chapter briefly recapitulates the proposed methods in the thesis with a particular focus on implications for future research directions.

8.1 Summary

This thesis is divided into eight chapters. It contains information on four peer-reviewed publications and one R package. In Chapter 1, a gentle introduction to the change-point problem is given. It discusses the main classes of the general change-point problem, namely the retrospective and sequential. It further provides an overview of the main branches of segmentation methods in the literature. Lastly, it briefly introduces the main biological concepts and methods discussed in the thesis.

In Chapter 2, we describe the methods that are used in the subsequent chapters. It elaborates the briefly discussed concepts in multiple publications to obtain an overview of the scope of the thesis. Particularly, it discusses in detail the CE algorithm and its characteristics, the usage of the CE method in detecting multiple change-points and the applicability of sequential techniques. We conclude Chapter 2 with an introduction to the parallel computing techniques that can be performed in the R statistical computing environment. It is known that evolutionary computing methods, including the CE method, consume more computational resources, due to the nature of their implementation. In the CE method, the final solutions are obtained through a model-based stochastic approach. It continually simulates a large number of sample solutions from a statistical distribution and assesses their performance through a model selection criterion until an optimal solution is reached. Because of this it takes significantly large amount of computational resources in comparison to dynamic programming algorithms. In this thesis, we propose parallel computing techniques and hybrid algorithms to ameliorate this efficiency issue of the general CE algorithm.

Chapter 3 contains a peer-reviewed journal article, proposing a variant of the CE method to detect multiple change-points in aCGH data. The four parameter beta distribution is used in the CE method to simulate change-point locations, and the modified BIC is used as the performance function in the CE method to estimate the number of change-points. Extensive simulation studies are carried out to obtain best parameter values for the CE algorithm and to compare its performance with other publicly available methods. Finally, we apply the proposed method to three real aCGH experimental data sets to further illustrate its usefulness.

Chapter 4 includes a peer-reviewed book chapter, extending the work in Chapter 3. We propose two novel hybrid algorithms that merge powerful sequential techniques (i.e., the cumulative sum method and the Shiryaev-Roberts procedure) with the CE method to effectively estimate both the number and locations of the change-points. A pair-wise hypothesis test is used with an improved Bonferroni correction to test for the significance of the estimated change-points. We found that the proposed hybrid procedures significantly increase the efficiency of the CE method.

Chapter 5 and 6, two peer-reviewed full conference papers, are devoted to the problem of identifying change-points in discrete (count) data with the use of the CE method. In Chapter 5, we propose a variant of the CE method to detect change-points in DNA read count data. Several stopping criteria and simulation distributions for the CE algorithm are discussed in this work. Chapter 6 extends the procedure proposed in Chapter 5 to further incorporate auxiliary information in the CE algorithm to detect change-points more effectively. A generalized additive model for location, scale and shape (GAMLSS) is used when modelling DNA read count data. The procedures proposed in these chapters are used to detect multiple change-points in DNA read count data of patients with celiac disease, obtained through the Illumina TruSeq exome capture method.

In Chapter 7, we provide details of the R package that is developed based on some of the methods discussed in this thesis. The R package is freely available from the CRAN. This chapter includes several case studies to exemplify its functions and capabilities.

8.2 Discussion

In this thesis we have introduced the use of the Cross-Entropy method, a model-based evolutionary computing technique, in estimating the number and location of change-points in genomic sequences. Particularly we applied the proposed variants of the CE method to detect copy number variations in aCGH and DNA read count data. In the literature there are few studies that have used different evolutionary computing methods (e.g. genetic algorithms, simulated annealing) to solve both the estimation and optimization issues in the general change-point problem. The use of the CE method is still at its inception in the change-point analysis literature, except for the work in [49], which used the CE method to estimate change-point locations in binary data with a fixed number of change-change points. However, we found that in the literature, the characteristics of the CE method and its performance with respect to the change-point problem were never discussed in detail. In this thesis, we aim to fill this gap by proposing several variants of the CE method and discussing its characteristics. We contribute in multiple ways to the existing literature of the use of the CE method in change-point analysis with a special focus on analysing genomic sequences. The proposed variants of the CE method not only estimate the change-point locations, but also estimate the number of change-points.

In general, the change-point analysis can be considered as a mixture of estimation and optimization problem. It contains two major steps in finding a final solution. Firstly, one has to estimate the optimal number of change-points. Secondly, optimal change-point locations have to be estimated based on the number obtained in the first step. Because of the complexity it involves, the majority of the earlier methods were developed to address the second step of the change-point problem, with a known number of change-points, or simply considered the single change-point problem. However, during the last few decades, with the exponential development in the technological infrastructure, a number of methods have been developed to address both the optimization and estimation problems simultaneously in different paradigms. The variants of the CE method proposed in this thesis were also developed to address both the estimation and optimization problems in the general change-point problem.

Being a stochastic approach, the CE method is an ideal candidate to represent the uncertainty associated with the number and locations of the change-points. The model-based approach in the CE method further enhances its capabilities, in which it provides flexibility to incorporate many statistical modelling concepts to obtain best estimates. All the major steps of the proposed variants of the CE method in this thesis are governed by a particular statistical concept. A parametric statistical distribution is used in the stage of simulating change-point locations. Then a performance function, which is also based on a model selection procedure, is used to obtain best solutions and to update the parameters of the simulation distribution. The stopping criteria discussed in the thesis are also based on statistical concepts of variance minimization. Overall, it provides a sound statistical framework to detect multiple change-points in observed data sequences. However, we observed that the general CE method consumes more computational resources, as is reported in the literature. In this thesis we proposed enhancements to ameliorate this problem. The first approach is to use multi-core parallel computation techniques, and the other approach is to amalgamate sequential detection techniques with the CE method. These approaches have greatly improved the efficiency of the proposed procedures as compared to the general CE method.

In our work, we observed that the performance of the CE method is mainly affected by the internal parameters discussed in Chapter 2, especially the simulation sample size (M) and the elite sample fraction (ρ), as shown in Chapters 3 and 5. The impact of the cut-off value for the stopping criterion (ε) appeared marginal in the context of the analysis performed in Appendix A. In the literature, the concept of parameter estimate smoothing is discussed in detail. We incorporated smoothing into the CE algorithms in the proposed R package [114]. The proposed procedures were tested in a series of simulation studies conducted in [115–120, 167] and applied to real aCGH data and to DNA read count data. Furthermore, in [115, 116, 167] we further compared the effectiveness of the proposed procedures with other popular publicly available methods.

The applications of the proposed procedures are not limited to analysing biological sequences. They can be easily extended and applied to other data sequences in different scientific streams. For instance, we have applied the CE method to detect change-points in the All Share Price Index (ASPI) data of the Sri Lankan stock market [118]. Thus, the proposed methods are versatile in their applications.

8.3 Future directions

The studies presented in this thesis could lead to advancements in many directions. We conclude our discussion by listing potential extensions and applications.

Check for robustness to model assumptions violations

The statistical models considered in this thesis depend on certain assumptions. It would be worth investigating the performance of the proposed procedures when the model assumptions are violated. A similar study that was carried out in [98] could be adapted here. The results would indicate the level of robustness of the proposed procedures.

Further investigation on hybrid methods

In Chapter 4, we proposed two novel hybrid algorithms that merged two sequential detection techniques with the CE approach to obtain better estimates with a fraction of the processing time observed in the general CE method. We propose to further explore this initiative in two directions. Firstly, to incorporate the proposed procedures into the R package as separate functions. Secondly, to further investigate the issues of developing sequential techniques that can be applied to both short and long sequences as discussed in [116].

Furthermore, it will be also beneficial to investigate the amalgamation of other faster detection methods such as binary segmentation [136, 161] to obtain a preliminary estimate for the number of change-points to initialize the CE algorithm.

Further investigations on incorporating auxiliary information to the process of detecting change-points

We have initiated the work on incorporating auxiliary information to obtain improved estimates of the change-points by using the GAMLSS modelling procedure in [120]. We hope to extend this approach by adding more significant variables to the system and assess their performance. Furthermore, it would be beneficial to estimate not only the number of change-points and their locations, but also the parameter estimates of the GAMLSS or any other model for each segment.

Extensive study to compare the performances of different evolutionary computing methods in the change-point problem

In the literature there exists a rich class of applications of evolutionary computing methods to solve different optimization and estimation problems. However, we found that most of these methods are rarely used in the change-point literature (genetic algorithms and simulated annealing being the main exceptions). It would be beneficial to carry out an extensive study to compare the performances of these different evolutionary computing approaches in the context of change-point problem to obtain a better overview of their effectiveness.

Cluster level implementation

The excessive processing time is considered as the major drawback of the CE method. This issue is already being discussed in the literature [23] and we also considered it in our papers [117, 119]. Multi-core parallel computational techniques are one solution to the problem, as we have proposed in [117, 167]. However, further processing efficiency can be achieved through a cluster level implementation of the CE method. Thus, a comparative analysis can be conducted to identify the merits and drawbacks of different implementations of the the CE method.

Extensions to the R package

The current implementation (version 1.1) of the "breakpoint" R package contains functions to perform segmentation on both continuous and count data. We hope to incorporate further functions to validate the estimated change-points, as proposed in [116], with the use of improved Bonferroni correction. Furthermore, we are in the process of adding non-parametric tests (e.g., the Kolmogorov-Smirnov and Mann-Whitney tests) and resampling-based multiple tests to further validate the estimated change-points.

Appendix A

Supplementary Material to "Multiple Break-Points Detection in array CGH Data via the Cross-Entropy Method

The supplementary material is structured as follows. Section 1 provides detailed results of the second simulation study carried out in the '**Selection of best parameter values for the CE algorithm**' section to justify the choice of parameter value for the cut-off value (ε) in the SC. In Section 2, a comparative analysis is carried out to identify the advantages of the use of MAD in the CE algorithm over the use of SD. Section 3 concludes the supplementary material with the results of processing time (in seconds) for the three real aCGH data examples considered in the paper.

A.1 Choice of parameter value for the cut-off value (ε) in the SC

In this simulation study, we set the M and the ρ values at 200 and 0.06 in the CE algorithm as suggested in the first simulation study. A range of values from 0.001 to 0.1 were considered for the ε. For each value of ε 100 simulations were carried-out. We consider the Root Mean Square Error (RMSE) value as the performance measurement in this study. Table 1 shows the average RMSE values for each of the ε values. It can be identified that the impact of the ε is less critical to the performance of the CE algorithm. Therefore, in the CE algorithm we set 0.01 as the default value for the cut-off value in the SC.

Table A.1: Average RMSE based on 100 simulations for ε values

	cut-off value (ε)	Average RMSE
1	0.001	0.0357
2	0.002	0.0341
3	0.003	0.0353
4	0.004	0.0364
5	0.005	0.0357
6	0.006	0.0364
7	0.007	0.0370
8	0.008	0.0352
9	0.009	0.0354
10	0.01	0.0357
11	0.02	0.0356
12	0.03	0.0366
13	0.04	0.0353
14	0.05	0.0359
15	0.06	0.0369
16	0.07	0.0366
17	0.08	0.0350
18	0.09	0.0352
19	0.10	0.0379

A.2 Performance of the CE method with different stopping criterion

In the paper, we have introduced the median absolute deviation (MAD, [1]) which is a more robust dispersion measure to utilize in the CE algorithm as the stopping criterion, as opposed to the standard deviation (SD) considered in [2]. We carried out a detailed simulation study to explore the advantages of using the MAD over the SD in the CE method. We set the parameters of the data generation process to be the same as in the simulation study to obtain the best set of parameters to the CE method (Refer section 3 in the paper). We repeat the process for 100 times separately for the two stopping criterion to obtain the average results. Table A.2 shows the average processing time and average root mean square error rates for the CE method with the use of MAD and SD as the stopping criterion. It is observed that the processing time of the CE method is improved around 6% with the use of the MAD as compared to the use of SD. In terms of the accuracy both criterion perform similarly, where the use of SD marginally better than the MAD.

Table A.2: Average processing time (s) and Average RMSE values for the CE method with the use of MAD and SD

	MAD	SD
Average processing time	1.3950	1.4922
Average RMSE	0.0392	0.0371

A.3 Processing time for the real aCGH data examples

A.3.1 Processing time for fibroblast cell lines (GM03563) data

Table A.3 shows the processing time (s) of the CE method and other four methods for the fibroblast cell lines (GM03563) data.

Table A.3: Processing time (s) for the proposed CE method and other methods for the fibroblast cell lines (GM03563) data

Method	Ch 1	Ch 3	Ch 9	Ch 11
CE	2.264	1.596	2.007	3.150
changepoint	0.018	0.011	0.014	0.014
cumSeg	0.048	0.130	0.127	0.028
DNAcopy	0.070	0.029	0.060	0.168
bcp	0.157	0.148	0.145	0.228

A.3.2 Processing time for GBM data

Table A.4 shows the processing time (s) of the CE method and other four methods for the GBM data.

A.3.3 Processing time for MDA157 data

Table A.5 shows the processing time (s) of the CE method and other four methods for the MDA157 data.

Table A.4: Processing time (s) for the proposed CE method and other methods for GBM data

Method	Ch 7	Ch 13	Ch 19	Ch 20
	GBM29	GBM31	GBM11	GBM12
CE	7.976	6.203	4.858	4.939
changepoint	0.029	0.012	0.023	0.023
cumSeg	0.105	0.088	0.094	0.094
DNAcopy	0.113	0.046	0.219	0.047
bcp	0.823	0.361	0.408	0.766

Table A.5: Processing time (s) for the proposed CE method and other methods for MDA157 data

Method	Ch 6	Ch 7	Ch 10	Ch 19
CE	5.143	4.133	3.696	3.264
changepoint	0.239	0.014	0.015	0.016
cumSeg	0.039	0.046	0.036	0.045
DNAcopy	0.075	0.089	0.039	0.071
bcp	0.227	0.218	0.198	0.232

It is observed that the proposed CE method is not as computationally efficient with the other four methods considered in this study. This is mainly due to the evolutionary computing nature that the CE method inherits. However, in the simulation study (refer "Numerical Results" section in the paper) we have shown that in terms of the accuracy of the break-point location estimates it outperforms all other competing methods considered in the study.

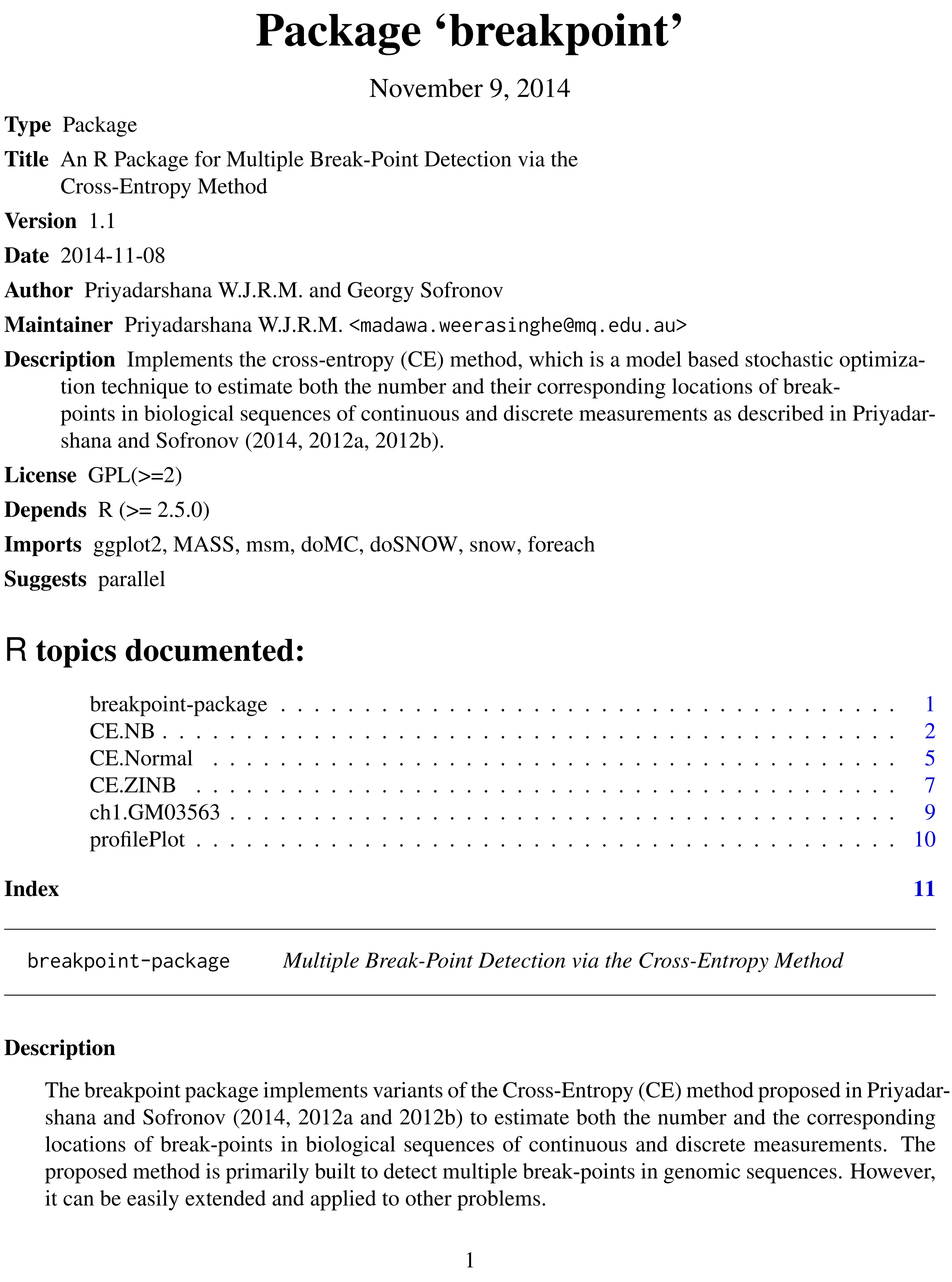
Bibliography

[1] D. C. Hoaglin, F. Mosteller and J. W. Tukey, *Understanding Robust and Exploratory Data Analysis*. John Wiley and Sons Inc., New York, 1983.

[2] W. J. R. M. Priyadarshana and G. Sofronov, "A Modified Cross Entropy Method for Detecting Multiple Change Points in DNA Count Data," In Proc. IEEE World Congress on Computational Intelligence (CEC'2012), pp. 1020–1027, 2012.

Appendix B

The manual of the “breakpoint” R package



Package ‘breakpoint’

November 9, 2014

Type Package

Title An R Package for Multiple Break-Point Detection via the Cross-Entropy Method

Version 1.1

Date 2014-11-08

Author Priyadarshana W.J.R.M. and Georgy Sofronov

Maintainer Priyadarshana W.J.R.M. <madawa.weerasinghe@mq.edu.au>

Description Implements the cross-entropy (CE) method, which is a model based stochastic optimization technique to estimate both the number and their corresponding locations of break-points in biological sequences of continuous and discrete measurements as described in Priyadarshana and Sofronov (2014, 2012a, 2012b).

License GPL(>=2)

Depends R (>= 2.5.0)

Imports ggplot2, MASS, msm, doMC, doSNOW, snow, foreach

Suggests parallel

R topics documented:

`breakpoint-package` *Multiple Break-Point Detection via the Cross-Entropy Method*

Description

The breakpoint package implements variants of the Cross-Entropy (CE) method proposed in Priyadarshana and Sofronov (2014, 2012a and 2012b) to estimate both the number and the corresponding locations of break-points in biological sequences of continuous and discrete measurements. The proposed method is primarily built to detect multiple break-points in genomic sequences. However, it can be easily extended and applied to other problems.

Details

Package:	breakpoint
Type:	Package
Version:	1.1
Date:	2014-11-08
License:	GPL 2.0

"breakpoint"" package provides estimates on both the number as well as the corresponding locations of break-points. The algorithms utilize the Cross-Entropy (CE) method, which is a model based stochastic optimization procedure to obtain the estimates on location. Model selection procedures based on penalized likelihood methods are used to obtain the number of break-points. In analyzing continuous data, it uses the modified BIC introduced by Zhang & Siegmund (2007). In discrete data analysis it uses the general BIC. Current implementation of the methodology works as an exact search method in estimating the number of break-points. A parallel implementation of the algorithm can be carried-out in Unix/Linux/MAC OS X and Windows operating systems with the use of "doMC", "parallel", "snow" and "doSNOW" packages.

Author(s)

Priyadarshana, W.J.R.M. and Sofronov, G.

Maintainer: Priyadarshana, W.J.R.M. <madawa.weerasinghe@mq.edu.au>

References

Priyadarshana, W. J. R. M., Sofronov G. (2014). Multiple Break-Points Detection in array CGH Data via the Cross-Entropy Method, IEEE/ACM Transactions on Computational Biology and Bioinformatics, no. 1, pp. 1, PrePrints, doi:10.1109/TCBB.2014.2361639, ISSN: 1545-5963.

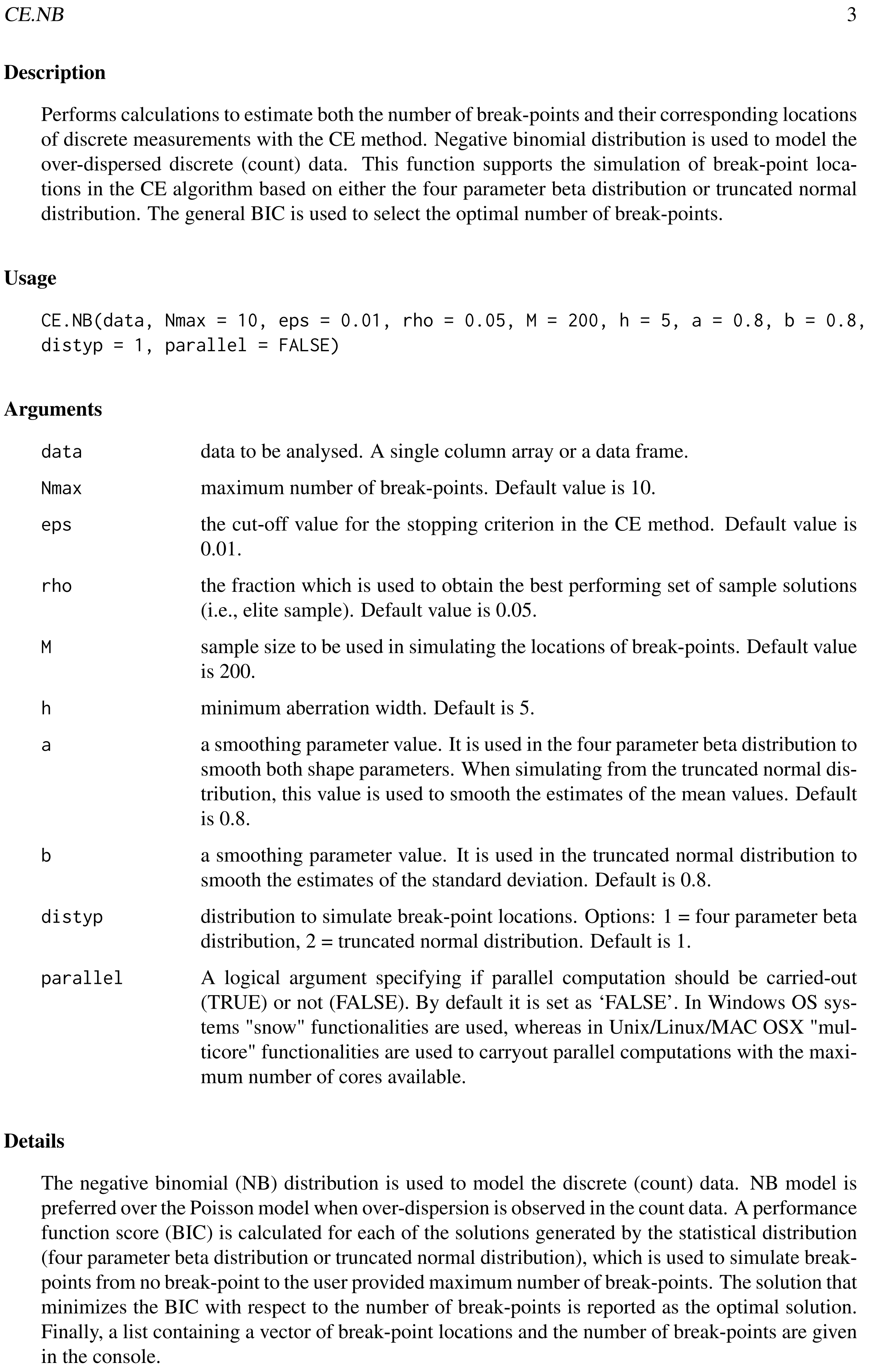
Priyadarshana, W. J. R. M. and Sofronov, G. (2012a). A Modified Cross- Entropy Method for Detecting Multiple Change-Points in DNA Count Data. In Proc. of the IEEE Conference on Evolutionary Computation (CEC), 1020-1027, DOI: 10.1109/CEC.2012.6256470.

Priyadarshana, W. J. R. M. and Sofronov, G. (2012b). The Cross-Entropy Method and Multiple Change-Points Detection in Zero-Inflated DNA read count data. In: Y. T. Gu, S. C. Saha (Eds.) The 4th International Conference on Computational Methods (ICCM2012), 1-8, ISBN 978-1-921897-54-2.

Rubinstein, R., and Kroese, D. (2004) The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning. Springer-Verlag, New York.

Zhang, N.R., and Siegmund, D.O. (2007) A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. Biometrics, 63, 22-32.

CE.NB *Multiple Break-point Detection via the CE Method with Negative Binomial Distribution*



Description

Performs calculations to estimate both the number of break-points and their corresponding locations of discrete measurements with the CE method. Negative binomial distribution is used to model the over-dispersed discrete (count) data. This function supports the simulation of break-point locations in the CE algorithm based on either the four parameter beta distribution or truncated normal distribution. The general BIC is used to select the optimal number of break-points.

Usage

```
CE.NB(data, Nmax = 10, eps = 0.01, rho = 0.05, M = 200, h = 5, a = 0.8, b = 0.8,
distyp = 1, parallel = FALSE)
```

Arguments

`data` data to be analysed. A single column array or a data frame.

`Nmax` maximum number of break-points. Default value is 10.

`eps` the cut-off value for the stopping criterion in the CE method. Default value is 0.01.

`rho` the fraction which is used to obtain the best performing set of sample solutions (i.e., elite sample). Default value is 0.05.

`M` sample size to be used in simulating the locations of break-points. Default value is 200.

`h` minimum aberration width. Default is 5.

`a` a smoothing parameter value. It is used in the four parameter beta distribution to smooth both shape parameters. When simulating from the truncated normal distribution, this value is used to smooth the estimates of the mean values. Default is 0.8.

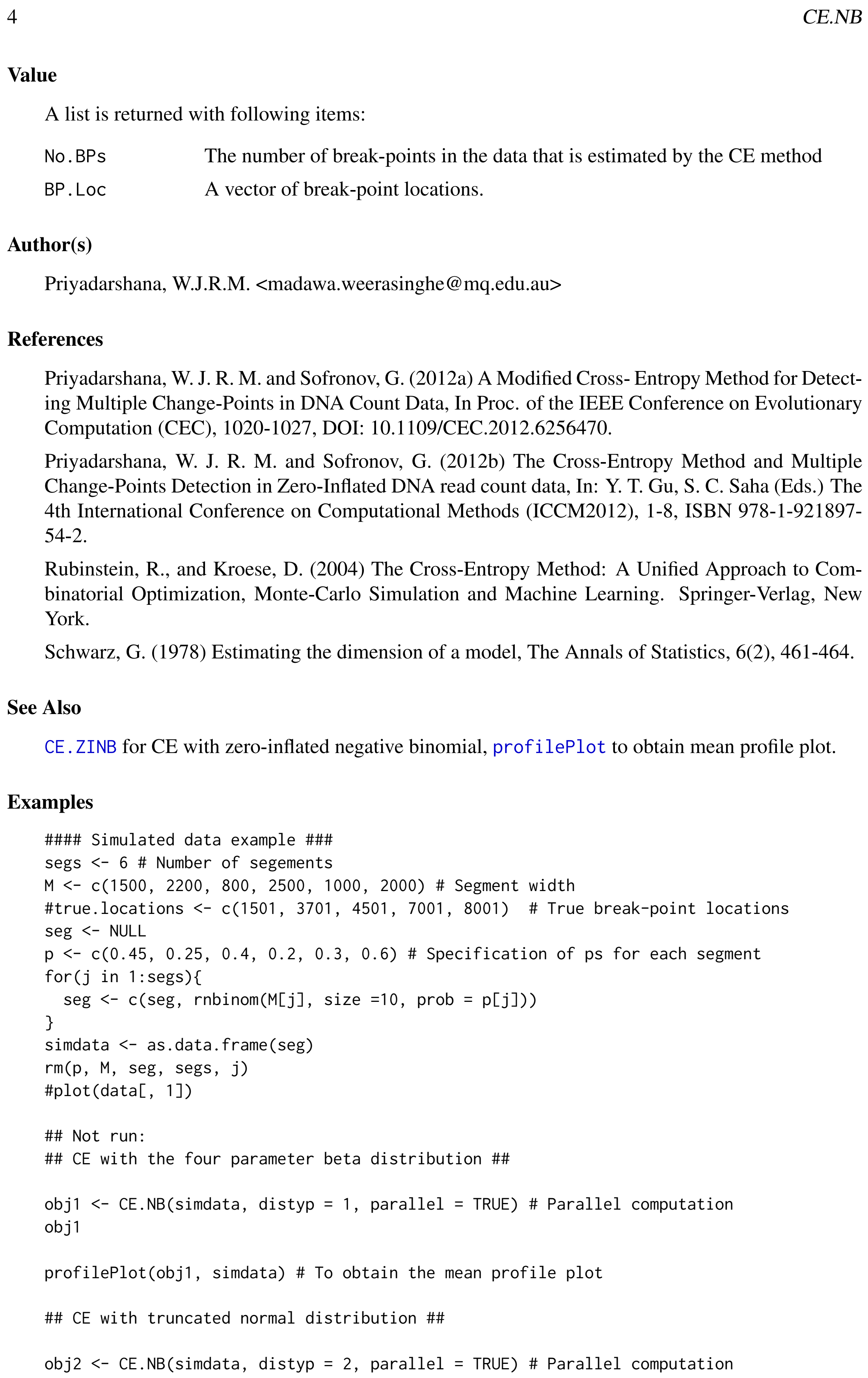
`b` a smoothing parameter value. It is used in the truncated normal distribution to smooth the estimates of the standard deviation. Default is 0.8.

`distyp` distribution to simulate break-point locations. Options: 1 = four parameter beta distribution, 2 = truncated normal distribution. Default is 1.

`parallel` A logical argument specifying if parallel computation should be carried-out (TRUE) or not (FALSE). By default it is set as 'FALSE'. In Windows OS systems "snow" functionalities are used, whereas in Unix/Linux/MAC OSX "multicore" functionalities are used to carryout parallel computations with the maximum number of cores available.

Details

The negative binomial (NB) distribution is used to model the discrete (count) data. NB model is preferred over the Poisson model when over-dispersion is observed in the count data. A performance function score (BIC) is calculated for each of the solutions generated by the statistical distribution (four parameter beta distribution or truncated normal distribution), which is used to simulate break-points from no break-point to the user provided maximum number of break-points. The solution that minimizes the BIC with respect to the number of break-points is reported as the optimal solution. Finally, a list containing a vector of break-point locations and the number of break-points are given in the console.



Value

A list is returned with following items:

`No.BPs` The number of break-points in the data that is estimated by the CE method

`BP.Loc` A vector of break-point locations.

Author(s)

Priyadarshana, W.J.R.M. <madawa.weerasinghe@mq.edu.au>

References

Priyadarshana, W. J. R. M. and Sofronov, G. (2012a) A Modified Cross- Entropy Method for Detecting Multiple Change-Points in DNA Count Data, In Proc. of the IEEE Conference on Evolutionary Computation (CEC), 1020-1027, DOI: 10.1109/CEC.2012.6256470.

Priyadarshana, W. J. R. M. and Sofronov, G. (2012b) The Cross-Entropy Method and Multiple Change-Points Detection in Zero-Inflated DNA read count data, In: Y. T. Gu, S. C. Saha (Eds.) The 4th International Conference on Computational Methods (ICCM2012), 1-8, ISBN 978-1-921897-54-2.

Rubinstein, R., and Kroese, D. (2004) The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning. Springer-Verlag, New York.

Schwarz, G. (1978) Estimating the dimension of a model, The Annals of Statistics, 6(2), 461-464.

See Also

`CE.ZINB` for CE with zero-inflated negative binomial, `profilePlot` to obtain mean profile plot.

Examples

```
#### Simulated data example ###
segs <- 6 # Number of segements
M <- c(1500, 2200, 800, 2500, 1000, 2000) # Segment width
#true.locations <- c(1501, 3701, 4501, 7001, 8001)  # True break-point locations
seg <- NULL
p <- c(0.45, 0.25, 0.4, 0.2, 0.3, 0.6) # Specification of ps for each segment
for(j in 1:segs){
  seg <- c(seg, rnbinom(M[j], size =10, prob = p[j]))
}
simdata <- as.data.frame(seg)
rm(p, M, seg, segs, j)
#plot(data[, 1])

## Not run:
## CE with the four parameter beta distribution ##

obj1 <- CE.NB(simdata, distyp = 1, parallel = TRUE) # Parallel computation
obj1

profilePlot(obj1, simdata) # To obtain the mean profile plot

## CE with truncated normal distribution ##

obj2 <- CE.NB(simdata, distyp = 2, parallel = TRUE) # Parallel computation
```

```
obj2

profilePlot(obj1, simdata) # To obtain the mean profile plot

## End(Not run)
```

`CE.Normal` *Multiple Break-point Detection via the CE Method for Continuous Data*

Description

This function performs calculations to estimate both the number of break-points and their corresponding locations of continuous measurements with the CE method. The normal distribution is used to model the observed continuous data. This function supports the simulation of break-point locations based on the four parameter beta distribution and truncated normal distribution. The modified BIC proposed by Zhang and Siegmund (2007) is used to select the optimal number of break-points.

Usage

```
CE.Normal(data, Nmax = 10, eps = 0.01, rho = 0.05, M = 200, h = 5, a = 0.8,
b = 0.8, distyp = 1, parallel = FALSE)
```

Arguments

`data` data to be analysed. A single column array or a data frame.

`Nmax` maximum number of break-points. Default value is 10.

`eps` the cut-off value for the stopping criterion in the CE method. Default value is 0.01.

`rho` the fraction which is used to obtain the best performing set of sample solutions (i.e., elite sample). Default value is 0.05.

`M` sample size to be used in simulating the locations of break-points. Default value is 200.

`h` minimum aberration width. Default is 5.

`a` a smoothing parameter value. It is used in the four parameter beta distribution to smooth both shape parameters. When simulating from the truncated normal distribution, this value is used to smooth the estimates of the mean values. Default is 0.8.

`b` a smoothing parameter value. It is used in the truncated normal distribution to smooth the estimates of the standard deviation. Default is 0.8.

`distyp` distributions to simulate break-point locations. Options: 1 = four parameter beta distribution, 2 = truncated normal distribution. Default is 1.

`parallel` A logical argument specifying if parallel computation should be carried-out (TRUE) or not (FALSE). By default it is set as ‘FALSE’. In Windows OS systems "snow" functionalities are used, whereas in Unix/Linux/MAC OSX "multicore" functionalities are used to carryout parallel computations with the maximum number of cores available.

Details

The normal distribution is used to model the continuous data. A performance function score (mBIC) is calculated for each of the solutions generated by the statistical distribution (four parameter beta distribution or truncated normal distribution), which is used to simulate break-points from no break-point to the user provided maximum number of break-points. The solution that maximizes the mBIC with respect to the number of break-points is reported as the optimal solution. Finally, a list containing a vector of break-point locations and the number of break-points are given in the console.

Value

A list is returned with following items:

`No.BPs`	The number of break-points in the data that is estimated by the CE method
`BP.Loc`	A vector of break-point locations.

Author(s)

Priyadarshana, W.J.R.M. <madawa.weerasinghe@mq.edu.au>

References

Priyadarshana, W. J. R. M., Sofronov G. (2014) Multiple Break-Points Detection in array CGH Data via the Cross-Entropy Method, IEEE/ACM Transactions on Computational Biology and Bioinformatics, no. 1, pp. 1, PrePrints, doi:10.1109/TCBB.2014.2361639, ISSN: 1545-5963.

Priyadarshana, W. J. R. M. and Sofronov, G. (2012) A Modified Cross- Entropy Method for Detecting Multiple Change-Points in DNA Count Data, In Proc. of the IEEE Conference on Evolutionary Computation (CEC), 1020-1027, DOI: 10.1109/CEC.2012.6256470.

Rubinstein, R., and Kroese, D. (2004) The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning. Springer-Verlag, New York.

Zhang, N.R., and Siegmund, D.O. (2007) A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. Biometrics, 63, 22-32.

See Also

`profilePlot` to obtain mean profile plot.

Examples

```
data(ch1.GM03563)
## Not run:
## CE with four parameter beta distribution ##
obj1 <- CE.Normal(ch1.GM03563, distyp = 1, parallel =TRUE)
profilePlot(obj1, simdata)

## CE with truncated normal distribution ##
obj2 <- CE.Normal(ch1.GM03563, distyp = 2, parallel =TRUE)
profilePlot(obj2, simdata)

## End(Not run)
```

CE.ZINB *Multiple Break-point Detection via the CE Method with Zero-Inflated Negative Binomial Distribution*

Description

Performs calculations to estimate both the number of break-points and their corresponding locations of discrete measurements with the CE method. Zero-inflated negative binomial distribution is used to model the excess zero observations and to model over-dispersion in the observed discrete (count) data. This function supports the simulation of break-point locations in the CE algorithm based on the four parameter beta distribution and truncated normal distribution. The general BIC is used to select the optimal number of break-points.

Usage

```
CE.ZINB(data, Nmax = 10, eps = 0.01, rho = 0.05, M = 200, h = 5, a = 0.8,
b = 0.8, distyp = 1, parallel = FALSE)
```

Arguments

`data` data to be analysed. A single column array or a data frame.

`Nmax` maximum number of break-points. Default value is 10.

`eps` the cut-off value for the stopping criterion in the CE method. Default value is 0.01.

`rho` the fraction which is used to obtain the best performing set of sample solutions (i.e., elite sample). Default value is 0.05.

`M` sample size to be used in simulating the locations of break-points. Default value is 200.

`h` minimum aberration width. Default is 5.

`a` a smoothing parameter value. It is used in the four parameter beta distribution to smooth both shape parameters. When simulating from the truncated normal distribution, this value is used to smooth the estimates of the mean values. Default is 0.8.

`b` a smoothing parameter value. It is used in the truncated normal distribution to smooth the estimates of the standard deviation. Default is 0.8.

`distyp` distribution to simulate break-point locations. Options: 1 = four parameter beta distribution, 2 = truncated normal distribution. Default is 1.

`parallel` A logical argument specifying if parallel computation should be carried-out (TRUE) or not (FALSE). By default it is set as ‘FALSE’. In Windows OS systems "snow" functionalities are used, whereas in Unix/Linux/MAC OSX "multicore" functionalities are used to carryout parallel computations with the maximum number of cores available.

Details

Zero-inflated negative binomial (ZINB) distribution is used to model the discrete (count) data. ZINB model is preferred over the NB model when both excess zero values and over-dispersion observed in the count data. A performance function score (BIC) is calculated for each of the solutions generated by the statistical distribution (four parameter beta distribution or truncated normal distribution), which is used to simulate break-points from no break-point to the user provided maximum number of break-points. The solution that minimizes the BIC with respect to the number of break-points is reported as the optimal solution. Finally, a list containing a vector of break-point locations and the number of break-points are given in the console.

Value

A list is returned with following items:

`No.BPs`	The number of break-points in the data that is estimated by the CE method
`BP.Loc`	A vector of break-point locations.

Author(s)

Priyadarshana, W.J.R.M. <madawa.weerasinghe@mq.edu.au>

References

Priyadarshana, W. J. R. M. and Sofronov, G. (2012a) A Modified Cross- Entropy Method for Detecting Multiple Change-Points in DNA Count Data, In Proc. of the IEEE Conference on Evolutionary Computation (CEC), 1020-1027, DOI: 10.1109/CEC.2012.6256470.

Priyadarshana, W. J. R. M. and Sofronov, G. (2012b) The Cross-Entropy Method and Multiple Change-Points Detection in Zero-Inflated DNA read count data, In: Y. T. Gu, S. C. Saha (Eds.) The 4th International Conference on Computational Methods (ICCM2012), 1-8, ISBN 978-1-921897-54-2.

Rubinstein, R., and Kroese, D. (2004) The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning. Springer-Verlag, New York.

Schwarz, G. (1978) Estimating the dimension of a model, The Annals of Statistics, 6(2), 461-464.

See Also

`CE.NB` for CE with negative binomial, `profilePlot` to obtain mean profile plot.

Examples

```
#### Simulated data example ###
# gamlss R package is used to simulate data from the ZINB.

## Not run:
library(gamlss)
segs <- 6 # Number of segements
M <- c(1500, 2200, 800, 2500, 1000, 2000) # Segment width
#true.locations <- c(1501, 3701, 4501, 7001, 8001)  # True break-point locations
seg <- NULL
p <- c(0.6, 0.1, 0.3, 0.05, 0.2, 0.4) # Specification of ps on each segment
sigma.val <- c(1,2,3,4,5,6) # Specification of sigma vlaues

for(j in 1:segs){
  seg <- c(seg, rZINBI(M[j], mu = 300, sigma = sigma.val[j], nu = p[j]))
}

simdata <- as.data.frame(seg)
```

```
rm(p, M, seg, segs, j, sigma.val)
#plot(data[, 1])

## CE with the four parameter beta distribution ##

obj1 <- CE.ZINB(simdata, distyp = 1, parallel = TRUE) # Parallel computation
obj1

profilePlot(obj1, simdata) # To obtain the mean profile plot

## CE with truncated normal distribution ##

obj2 <- CE.ZINB(simdata, distyp = 2, parallel = TRUE) # Parallel computation
obj2

profilePlot(obj2, simdata) # To obtain the mean profile plot

## End(Not run)
```

`ch1.GM03563` *Fibroblast cell line (GM03563) data*

Description

Chromosome 1 of cell line GM03563

Usage

```
data("ch1.GM03563")
```

Format

A single column data frame with 135 observations that corresponds to chromosome 1 of cell line GM03563.

`log2ratio` normalized average of the log base 2 test over reference ratio data

Details

This data set is extracted from a single experiments on 15 fibroblast cell lines with each array containing over 2000 (mapped) BACs spotted in triplicate discussed in Snijders et al.(2001). Data corresponds to the chromosome 1 of cell line GM03563.

References

Snijders,A.M. et al. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. Nature Genetics, 29, 263-26.

Examples

```
data(ch1.GM03563)
## Not run:
## CE with four parameter beta distribution ##
obj1 <- CE.Normal(ch1.GM03563, distyp = 1, parallel =TRUE)
profilePlot(obj1, ch1.GM03563)

## CE with truncated normal distribution ##
obj2 <- CE.Normal(ch1.GM03563, distyp = 2, parallel =TRUE)
profilePlot(obj2, ch1.GM03563)

## End(Not run)
```

`profilePlot` *Mean profile plot*

Description

Plotting function to obtain mean profile plot of the data based on the estimates of the break-points through CE method. An R object created from the CE.Normal, CE.NB or CE.ZINB is required. User can alter the axes names.

Usage

```
profilePlot(obj, data, x.label = "Data Sequence", y.label = "Value")
```

Arguments

`obj`	R object created from CE.Normal, CE.NB or CE.ZINB.
`data`	data to be analysed. A single column array or a data frame.
`x.label`	x axis label. Default is "Data Sequence".
`y.label`	y axis label. Default is "Value".

Author(s)

Priyadarshana, W.J.R.M. <madawa.weerasinghe@mq.edu.au>

See Also

`CE.Normal`, `CE.NB`, `CE.ZINB`.

Examples

```
data(ch1.GM03563)
## Not run:
## CE with four parameter beta distribution ##
obj1 <- CE.Normal(ch1.GM03563, distyp = 1, parallel =TRUE)
profilePlot(obj1)

## CE with truncated normal distribution ##
obj2 <- CE.Normal(ch1.GM03563, distyp = 2, parallel =TRUE)
profilePlot(obj2)

## End(Not run)
```

Index

Bibliography

[1] Abbasi, M., Paquete, L., Liefooghe, A., Pinheiro, M., and Matias, P. (2013). Improvements on bicriteria pairwise sequence alignment: algorithms and applications. *Bioinformatics*, 29(8):996–1003.

[2] AbouRizk, S., Halpin, D., and Wilson, J. (1994). Fitting beta distributions based on sample data. *Journal of Construction Engineering and Management*, 120(2):288–305.

[3] Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723.

[4] Alon, G., Kroese, D., Raviv, T., and Rubinstein, R. (2005). Application of the Cross-Entropy method to the buffer allocation problem in a simulation-based environment. *Annals of Operations Research*, 134(1):137–151.

[5] Analytics, R. (2014). *doMC: Foreach parallel adaptor for the multicore package*. R package version 1.3.3.

[6] Analytics, R. and Weston, S. (2014a). *doSNOW: Foreach parallel adaptor for the snow package*. R package version 1.0.12.

[7] Analytics, R. and Weston, S. (2014b). *foreach: Foreach looping construct for R*. R package version 1.4.2.

[8] Anscombe, F. J. (1949). The statistical analysis of insect counts based on the negative binomial distribution. *Biometrics*, 5:165–173.

[9] Ansorge, W., Sproat, B., Stegemann, J., Schwager, C., and Zenke, M. (1987). Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic Acids Res.*, 15(11):4593–4602.

[10] Auger, I. and Lawrence, C. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, 51(1):39–54.

[11] Autio, R., Hautaniemi, S., Kauraniemi, P., Yli-Harja, O., Astola, J., Wolf, M., and Kallioniemi, A. (2003). CGH-Plotter: MATLAB toolbox for CGH-data analysis. *Bioinformatics*, 19(13):1714–1715.

[12] Autio, R., Saarela, M., Jarvinen, A. K., Hautaniemi, S., and Astola, J. (2009). Advanced analysis and visualization of gene copy number and expression data. *BMC Bioinformatics*, 10 Suppl 1:S70.

[13] Back, T., Fogel, D. B., and Michalewicz, Z., editors (1997). *Handbook of Evolutionary Computation*. IOP Publishing Ltd., Bristol, UK, UK, 1st edition.

[14] Backenroth, D., Homsy, J., Murillo, L. R., Glessner, J., Lin, E., Brueckner, M., Lifton, R., Goldmuntz, E., Chung, W. K., and Shen, Y. (2014). CANOES: detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Res.*, 42(12):e97.

[15] Bainbridge, M. N., Warren, R. L., Hirst, M., Romanuik, T., Zeng, T., Go, A., Delaney, A., Griffith, M., Hickenbotham, M., Magrini, V., Mardis, E. R., Sadar, M. D., Siddiqui, A. S., Marra, M. A., and Jones, S. J. (2006). Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics*, 7:246.

[16] Barry, D. and Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319.

[17] Basseville, M. and Nikiforov, I. V. (1993). *Detection of Abrupt Changes: Theory and Application*. Prentice Hall information and system sciences series. Prentice Hall.

[18] Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563.

[19] Bellman, R. (1953). An introduction to the theory of dynamic programming. Technical report, RAND Corp.

[20] Bhattacharya, P. K. (1994). *Some aspects of change-point analysis*, volume Volume 23 of *Lecture Notes–Monograph Series*, pages 28–56. Institute of Mathematical Statistics, Hayward, CA.

[21] Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268.

[22] Botev, Z. and Kroese, D. (2011). The generalized Cross Entropy method, with applications to probability density estimation. *Methodology and Computing in Applied Probability*, 13(1):1–27.

[23] Botev, Z. and Kroese, D. P. (2004). Global likelihood optimization via the Cross-Entropy method with an application to mixture models. In *Proceedings of the 36th Conference on Winter Simulation*, WSC '04, pages 529–535. Winter Simulation Conference.

[24] Boubezoul, A., Paris, S., and Ouladsine, M. (2008). Application of the Cross-Entropy method to the GLVQ algorithm. *Pattern Recogn.*, 41(10):3173–3178.

[25] Boyer, S., Brown, S. D., Collins, R. A., Cruickshank, R. H., Lefort, M. C., Malumbres-Olarte, J., and Wratten, S. D. (2012). Sliding window analyses for optimal selection of mini-barcodes, and application to 454-pyrosequencing for specimen identification from degraded DNA. *PLoS ONE*, 7(5):e38215.

[26] Braun, J., Braun, R., and Müller, H.-G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika*, 87(2):301–314.

[27] Braun, J. V. and Müller, H. G. (1998). Statistical methods for DNA sequence segmentation. *Statistical Science*, 13:142–162.

[28] Bredel, M., Bredel, C., Juric, D., Harsh, G. R., Vogel, H., Recht, L. D., and Sikic, B. I. (2005). High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. *Cancer Res.*, 65(10):4088–4096.

[29] Brodsky, E. and Darkhovsky, B. S. (2000). *Non-parametric statistical diagnosis: problems and methods*, volume 509. Springer Netherlands.

[30] Campbell, P. J., Stephens, P. J., Pleasance, E. D., O'Meara, S., Li, H., Santarius, T., Stebbings, L. A., Leroy, C., Edkins, S., Hardy, C., Teague, J. W., Menzies, A., Goodhead, I., Turner, D. J., Clee, C. M., Quail, M. A., Cox, A., Brown, C., Durbin, R., Hurles, M. E., Edwards, P. A., Bignell, G. R., Stratton, M. R., and Futreal, P. A. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, 40(6):722–729.

[31] Carnahan, J. V. (1989). Maximum likelihood estimation for the 4-parameter Beta distribution. *Communications in Statistics - Simulation and Computation*, 18(2):513–536.

[32] Carson, A. R., Feuk, L., Mohammed, M., and Scherer, S. W. (2006). Strategies for the detection of copy number and other structural variants in the human genome. *Hum. Genomics*, 2(6):403–414.

[33] Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.*, 39(7 Suppl):16–21.

[34] Chambers, L. D. (1995). *Practical Handbook of Genetic Algorithms.* CRC Press, Inc., Boca Raton, FL, USA.

[35] Chepuri, K. and Homem-de Mello, T. (2005). Solving the vehicle routing problem with stochastic demands using the Cross-Entropy method. *Annals of Operations Research*, 134(1):153–181.

[36] Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *The Annals of Mathematical Statistics*, 35(3):999–1018.

[37] Consortium, T. W. T. C. C. (2010). Genome-wide association study of copy number variation in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464:pp. 713–720.

[38] Costa, A., Jones, O. D., and Kroese, D. (2007). Convergence properties of the Cross-Entropy method for discrete optimization. *Operations Research Letters*, 35(5):573 – 580.

[39] Cramér, H. (1999). *Mathematical Methods of Statistics.* Princeton landmarks in mathematics and physics. Princeton University Press.

[40] D. G. Albertson, C. Collins, F. M. and Gray, J. W. (2003). Chromosome aberrations in solid tumors. *Nature Genetics*, 34:369–376.

[41] Darkhovski, B. S. (1994). *Nonparametric methods in change-point problems: a general approach and some concrete algorithms*, volume Volume 23 of *Lecture Notes–Monograph Series*, pages 99–107. Institute of Mathematical Statistics, Hayward, CA.

[42] Davis, L., editor (1991). *Handbook of Genetic Algorithm.* Van Nostrand Reinhold, New York, USA.

[43] de Boer, P., Kroese, D., and Rubinstein, R. (2004). A fast Cross-Entropy method for estimating buffer overflows in queueing networks. *Management Science*, 50(7):883–895. Imported from research group DACS (ID number 280).

[44] de Boer, P.-T., Kroese, D., Mannor, S., and Rubinstein, R. (2005). A tutorial on the Cross-Entropy method. *Annals of Operations Research*, 134(1):19–67.

[45] Dongarra, J., Foster, I., Fox, G., Gropp, W., Kennedy, K., Torczon, L., and White, A., editors (2003). *Sourcebook of Parallel Computing*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[46] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

[47] Erdman, C. and Emerson, J. W. (2007). bcp: An R package for performing a Bayesian analysis of change point problems. *Journal of Statistical Software*, 23(3):1–13.

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

[49] Evans, G. E., Sofronov, G. Y., Keith, J. M., and Kroese, D. P. (2011). Identifying change-points in biological sequences via the Cross-Entropy method. *Annals of Operation Research*, 189(1):155–165.

[50] Fanciulli, M., Norsworthy, P. J., Petretto, E., Dong, R., Harper, L., Kamesh, L., Heward, J. M., Gough, S. C., de Smith, A., Blakemore, A. I., Froguel, P., Owen, C. J., Pearce, S. H., Teixeira, L., Guillevin, L., Graham, D. S., Pusey, C. D., Cook, H. T., Vyse, T. J., and Aitman, T. J. (2007). FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nature Genetics*, 39:721–723.

[51] Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural variation in the human genome. *Nat. Rev. Genet.*, 7(2):85–97.

[52] Fielitz, B. D. and Myers, B. L. (1975). Estimation of parameters in the Beta distribution. *Decision Sciences*, 6(1):1–13.

[53] Fisher, R. A. (1922a). The goodness of fit of regression formulae and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, 85:597–612.

[54] Fisher, R. A. (1922b). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, A: 222:309–368.

[55] Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G., and Jain, A. N. (2004). Hidden markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis*, 90(1):132 – 153. Special Issue on Multivariate Methods in Genomic Data Analysis.

[56] Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281.

[57] Gallagher, M., Wood, I., Keith, J., and Sofronov, G. (2007). Bayesian inference in estimation of distribution algorithms. In *Evolutionary Computation, 2007. CEC 2007. IEEE Congress on*, pages 127–133.

[58] Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S., and Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Res.*, 17(6):669–681.

[59] Giegerich, R. (2000). A systematic approach to dynamic programming in bioinformatics. *Bioinformatics*, 16(8):665–677.

[60] Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. *Computers Operations Research*, 13(5):533 – 549. Applications of Integer Programming.

[61] Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R. J., Freedman, B. I., Quinones, M. P., Bamshad, M. J., Murthy, K. K., Rovin, B. H., Bradley, W., Clark, R. A., Anderson, S. A., O'Connell, R. J., Agan, B. K., Ahuja, S. S., Bologna, R., Sen, L., Dolan, M. J., and Ahuja, S. K. (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, 307(5714):1434–1440.

[62] Grada, A. and Weinbrecht, K. (2013). Next-generation sequencing: methodology and application. *J. Invest. Dermatol.*, 133(8):e11.

[63] Guo, X., Zhu, S. X., Brunner, A. L., van de Rijn, M., and West, R. B. (2013). Next generation sequencing-based expression profiling identifies signatures from benign stromal proliferations that define stromal components of breast cancer. *Breast Cancer Res.*, 15(6):R117.

[64] Gutjahr, W. J. (2002). ACO algorithms with guaranteed convergence to the optimal solution. *Information Processing Letters*, 82(3):145 – 153.

[65] Hinkley, D. V. and Hinkley, E. A. (1970). Inference about the change-point in a sequence of binomial variables. *Biometrika*, 57(3):477–488.

[66] Hoaglin, D., Mosteller, F., and Tukey, J. (1983). *Understanding robust and exploratory data analysis*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

[67] Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems*. MIT Press, Cambridge, MA, USA.

[68] Hyman, E., Kauraniemi, P., Hautaniemi, S., Wolf, M., Mousses, S., Rozenblum, E., Ringner, M., Sauter, G., Monni, O., Elkahloun, A., Kallioniemi, O. P., and Kallioniemi, A. (2002). Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Res.*, 62(21):6240–6245.

[69] Ivakhno, S., Royce, T., Cox, A. J., Evers, D. J., Cheetham, R. K., and Tavare, S. (2010). CNAseg–a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics*, 26(24):3051–3058.

[70] Jann, A. (2000). Multiple change-point detection with a genetic algorithm. *Soft Computing*, 4(2):68–75.

[71] Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502.

[72] Kallioniemi, A., Kallioniemi, O., Sudar, D., Rutovitz, D., Gray, J., Waldman, F., and Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, 258(5083):818–821.

[73] Keith, J. and Kroese, D. P. (2002). Rare event simulation and combinatorial optimization using Cross Entropy: Sequence alignment by rare event simulation. In *Proceedings of the 34th Conference on Winter Simulation: Exploring New Frontiers*, WSC '02, pages 320–327. Winter Simulation Conference.

[74] Killick, R. and Eckley, I. (2014). changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, 58(3):1–19.

[75] Killick, R., Eckley, I. A., Ewans, K., and Jonathan, P. (2010). Detection of changes in variance of oceanographic time-series using changepoint analysis. *Ocean Engineering*, 37(13):1120–1126.

[76] Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

[77] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

[78] Knight, S. J., Regan, R., Nicod, A., Horsley, S. W., Kearney, L., Homfray, T., Winter, R. M., Bolton, P., and Flint, J. (1999). Subtle chromosomal rearrangements in children with unexplained mental retardation. *Lancet*, 354(9191):1676–1681.

[79] Kroese, D., Porotsky, S., and Rubinstein, R. (2006). The Cross-Entropy method for continuous multi-extremal optimization. *Methodology and Computing in Applied Probability*, 8(3):383–407.

[80] Krogh, A., Brown, M., Mian, I. S., Sjölander, K., and Haussler, D. (1994). Hidden Markov models in computational biology: applications to protein modeling. *Journal of Molecular Biology*, 235:1501–1531.

[81] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

[82] Lai, T. L. (2001). Sequential analysis: some classical problems and new challenges. *Statist. Sinica*, 11(2):303–408.

[83] Lampert, C. H., Blaschko, M., and Hofmann, T. (2008). Beyond sliding windows: Object localization by efficient subwindow search. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8.

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

[85] Li, S. and Lund, R. (2012). Multiple changepoint detection via genetic algorithms. *Journal of Climate*, 25:674–686.

[86] Li, W., Bernaola-Galvan, P., Haghighi, F., and Grosse, I. (2002). Applications of recursive segmentation to the analysis of DNA sequences. *Comput. Chem.*, 26(5):491–510.

[87] Lorden, G. (1971). Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908.

[88] Lucito, R., Healy, J., Alexander, J., Reiner, A., Esposito, D., Chi, M., Rodgers, L., Brady, A., Sebat, J., Troge, J., West, J. A., Rostan, S., Nguyen, K. C., Powers, S., Ye, K. Q., Olshen, A., Venkatraman, E., Norton, L., and Wigler, M. (2003). Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res.*, 13(10):2291–2305.

[89] Ma, T.-Y. (2012). A comparative study of the Cross Entropy approach with the state–of-the-art simulation-based traffic assignment algorithms. *Procedia - Social and Behavioral Sciences*, 54(0):749 – 757. Proceedings of EWGT2012 - 15th Meeting of the EURO Working Group on Transportation, September 2012, Paris.

[90] Margolin, L. (2005). On the convergence of the Cross-Entropy method. *Annals of Operations Research*, 134(1):201–214.

[91] Martin, J. A. and Wang, Z. (2011). Next-generation transcriptome assembly. *Nat. Rev. Genet.*, 12(10):671–682.

[92] Merlet, N. and Zerubia, J. (1996). New prospects in line detection by dynamic programming. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(4):426–431.

[93] Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.*, 11(1):31–46.

[94] Morin, R. D., O'Connor, M. D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A. L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., Eaves, C. J., and Marra, M. A. (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.*, 18(4):610–621.

[95] Moustakides, G. V. (1986). Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, 14(4):1379–1387.

[96] Moustakides, G. V., Polunchenko, A. S., and Tartakovsky, A. G. (2009). Numerical comparison of CUSUM and Shiryaev–Roberts procedures for detecting changes in distributions. *Communications in Statistics - Theory and Methods*, 38(16-17):3225–3239.

[97] Muggeo, V. M. (2012). *cumSeg: Change point detection in genomic sequences*. R package version 1.1.

[98] Muggeo, V. M. and Adelfio, G. (2011). Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics*, 27(2):161–166.

[99] Oliver, J. L., Carpena, P., Hackenberg, M., and Bernaola-Galvan, P. (2004). IsoFinder: computational prediction of isochores in genome sequences. *Nucleic Acids Res.*, 32(Web Server issue):W287–292.

[100] Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572.

[101] Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115.

[102] Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527.

[103] Pal, S., Bandyopadhyay, S., and Ray, S. (2006). Evolutionary computation in bioinformatics: a review. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 36(5):601–615.

[104] Pearson, H. (2006). Genetics: What is a gene? *Nature*, 441:398–401.

[105] Pearson, K. (1936). Method of moments and method of maximum likelihood. *Biometrika*, 28(1/2):34–59.

[106] Pennisi, E. (2007). Genomics. DNA study forces rethink of what it means to be a gene. *Science*, 316(5831):1556–1557.

[107] Peshkin, L. and Gelfand, M. S. (1999). Segmentation of yeast DNA using hidden Markov models. *Bioinformatics*, 15(12):980–986.

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

[109] Pollak, M. (1985). Optimal detection of a change in distribution. *The Annals of Statistics*, 13(1):206–227.

[110] Pollak, M. and Siegmund, D. (1985). A diffusion process and its applications to detecting a change in the drift of Brownian motion. *Biometrika*, 72(2):267–280.

[111] Pollak, M. and Tartakovsky, A. G. (2009). Optimality properties of the Shiryaev-Roberts procedure. *Stat. Sin.*, 19(4):1729–1739.

[112] Polunchenko, A. S. and Tartakovsky, A. G. (2010). On optimality of the Shiryaev–Roberts procedure for detecting a change in distribution. *The Annals of Statistics*, 38(6):3445–3457.

[113] Polushina, T. and Sofronov, G. (2011). Change-point detection in biological sequences via genetic algorithm. In *Evolutionary Computation (CEC), 2011 IEEE Congress on*, pages 1966–1971.

[114] Priyadarshana, W. and Sofronov, G. (2014). *breakpoint: An R Package for Multiple Break-Point Detection via the Cross-Entropy Method.* R package version 1.1.

[115] Priyadarshana, W. J. R. M., Polushina, T., and Sofronov, G. (2013). A hybrid algorithm for multiple change-point detection in continuous measurements. *AIP Conference Proceedings*, 1559(1):108–117.

[116] Priyadarshana, W. J. R. M., Polushina, T., and Sofronov, G. (2015). Hybrid algorithms for multiple change-point detection in biological sequences. In Sun, C., Bednarz, T., Pham, T., Vallotton, P., and Wang, D., editors, *Signal and Image Analysis for Biomedical and Life Sciences*, volume 823 of *Advances in Experimental Medicine and Biology*. Springer International Publishing.

[117] Priyadarshana, W. J. R. M. and Sofronov, G. (2012a). The Cross-Entropy method and multiple change-points detection in zero-inflated DNA read count data. In *The 4th International Conference on Computational Methods (ICCM2012)*, pages 1–8.

[118] Priyadarshana, W. J. R. M. and Sofronov, G. (2012b). A modified Cross-Entropy method for detecting change-points in the sri-lankan stock market. In *IASTED International Conference on Engineering and Applied Science (EAS2012).*

[119] Priyadarshana, W. J. R. M. and Sofronov, G. (2012c). A modified Cross-Entropy method for detecting multiple change points in DNA count data. In *Evolutionary Computation (CEC), 2012 IEEE Congress on*, pages 1–8.

[120] Priyadarshana, W. J. R. M. and Sofronov, G. (2013). GAMLSS and extended Cross-Entropy method to detect multiple change-points in DNA read count data. In Muggeo, V., Capursi, V., Boscaino, G., and Lovison, G., editors, *The 28th International Workshop on Statistical Modelling*, volume 1, pages 453–457.

[121] Proutski, V. and Holmes, E. (1998). SWAN: sliding window analysis of nucleotide sequence variability. *Bioinformatics*, 14(5):467–468.

[122] R Core Team (2014). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

[123] Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[124] Rebecca Killick, I. E. and Haynes, K. (2014). *changepoint: An R package for changepoint analysis.* R package version 1.1.5.

[125] Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., Gonzalez, J. R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., and Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, 444(7118):444–454.

[126] Rehm, H. L. (2013). Disease-targeted sequencing: a cornerstone in the clinic. *Nat. Rev. Genet.*, 14(4):295–300.

[127] Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, 54:507–554.

[128] Robert, C. (1995). Simulation of truncated normal variables. *Statistics and Computing*, 5(2):121–125.

[129] Roberts, S. W. (1966). A comparison of some control chart procedures. *Technometrics*, 8(3):411–430.

[130] Rubinstein, R. (1999). The Cross-Entropy method for combinatorial and continuous optimization. *Methodology And Computing In Applied Probability*, 1(2):127–190.

[131] Rubinstein, R. and Kroese, D. (2004). *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning.* Springer-Verlag, New York.

[132] Rubinstein, R. Y. (1997). Optimization of computer simulation models with rare events. *uropean Journal of Operational Research*, 99:89–112.

[133] Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, 74(12):5463–5467.

[134] Schmidberger, M., Morgan, M., Eddelbuettel, D., Yu, H., Tierney, L., and Mansmann, U. (2009). State of the art in parallel computing with R. *Journal of Statistical Software*, 31(1):1–27.

[135] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

[136] Scott, A. J. and Knott, M. (1974). A Cluster Analysis Method for Grouping Means in the Analysis of Variance. *Biometrics*, 30(3):507–512.

[137] Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., Leotta, A., Pai, D., Zhang, R., Lee, Y. H., Hicks, J., Spence, S. J., Lee, A. T., Puura, K., Lehtimaki, T., Ledbetter, D., Gregersen, P. K., Bregman, J., Sutcliffe, J. S., Jobanputra, V., Chung, W., Warburton, D., King, M. C., Skuse, D., Geschwind, D. H., Gilliam, T. C., Ye, K., and Wigler, M. (2007). Strong association of de novo copy number mutations with autism. *Science*, 316(5823):445–449.

[138] Sen, A. and Srivastava, M. S. (1975). On tests for detecting change in mean. *The Annals of Statistics*, 3(1):98–108.

[139] Shen, J., Gallagher, C. M., and Lu, Q. (2014). Detection of multiple undocumented change-points using adaptive Lasso. *Journal of Applied Statistics*, 41(6):1161–1173.

[140] Shewhart, W. A. (1930). Economic quality control of manufactured product. *Bell System Technical Journal*, 9(2):364–389.

[141] Shiryaev, A. (1963). On optimum methods in quickest detection problems. *Theory of Probability amp; Its Applications*, 8(1):22–46.

[142] Shiryaev, A. N. (1961). The problem of the most rapid detection of a disturbance in a stationary process. *Dokl. Math.*, 2:795–799.

[143] Shiryaev, A. N. (1978). *Optimal stopping rules*. Springer, New York.

[144] Siegmund, D. (1988). Confidence sets in change-point problems. *International Statistical Review / Revue Internationale de Statistique*, 56(1):31–48.

[145] SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.

[146] Smith, A. F. M. (1975). A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, 62(2):407–416.

[147] Smith, D. R., Quinlan, A. R., Peckham, H. E., Makowsky, K., Tao, W., Woolf, B., Shen, L., Donahue, W. F., Tusneem, N., Stromberg, M. P., Stewart, D. A., Zhang, L., Ranade, S. S., Warner, J. B., Lee, C. C., Coleman, B. E., Zhang, Z., McLaughlin, S. F., Malek, J. A., Sorenson, J. M., Blanchard, A. P., Chapman, J., Hillman, D., Chen, F., Rokhsar, D. S., McKernan, K. J., Jeffries, T. W., Marth, G. T., and Richardson, P. M. (2008). Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res.*, 18(10):1638–1642.

[148] Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B., and Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071):674–679.

[149] Smyth, P. (1994). Hidden Markov models for fault detection in dynamic systems. *Pattern Recognition*, 27(1):149 – 164.

[150] Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J. P., Gray, J. W., Jain, A. N., Pinkel, D., and Albertson, D. G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.*, 29(3):263–264.

[151] Sofronov, G. (2011). Change-point modelling in biological sequences via the Bayesian Adaptive Independent Sampler. *International Proceedings of Computer Science and Information Technology*, 5:122–126.

[152] Sofronov, G. Y., Evans, G. E., Keith, J. M., and Kroese, D. P. (2009). Identifying change-points in biological sequences via Sequential Importance sampling. *Environmental Modeling and Assessment*, 14(5):577–584.

[153] Srivastava, M. S. and Wu, Y. (1993). Comparison of EWMA, CUSUM and Shiryayev-Roberts procedures for detecting a shift in the mean. *The Annals of Statistics*, 21(2):645–670.

[154] Stratonovich, R. (1960). Conditional Markov processes. *Theory of Probability amp; Its Applications*, 5(2):156–178.

[155] Tajima, F. (1991). Determination of window size for analyzing DNA sequences. *Journal of Molecular Evolution*, 33(5):470–473.

[156] Theisen, A. (2008). Microarray-based comparative genomic hybridization (aCGH). *Nature Education*, 1(1):45.

[157] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

[158] Tibshirani, R. (2011). Regression shrinkage and selection via the Lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282.

[159] Tibshirani, R. and Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, 9:18–29.

[160] Tierney, L., Rossini, A. J., Li, N., and Sevcikova, H. (2013). *snow: Simple Network of Workstations*. R package version 0.3-13.

[161] Venkatraman, E. S. and Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6):657–663.

[162] von Heijne, G. (1986). Mitochondrial targeting sequences may form amphiphilic helices. *EMBO J.*, 5(6):1335–1342.

[163] Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186.

[164] Wang, H., Nettleton, D., and Ying, K. (2014). Copy number variation detection using next generation sequencing read counts. *BMC Bioinformatics*, 15:109.

[165] Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., Hakonarson, H., and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.*, 17(11):1665–1674.

[166] Watson, J. and Crick, F. (1953). A structure for deoxyribose nucleic acid. *Nature*, 171:737–738.

[167] Weerasinghe, M. and Sofronov, G. (2014). Multiple break-points detection in array cgh data via the cross-entropy method. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, PP(99):1–1.

[168] Wilting, S., Snijders, P., Meijer, G., Ylstra, B., van den IJssel, P., Snijders, A., Albertson, D., Coffa, J., Schouten, J., van de Wiel, M., Meijer, C., and Steenbergen, R. (2006). Increased gene copy numbers at chromosome 20q are frequent in both squamous cell carcinomas and adenocarcinomas of the cervix. *The Journal of Pathology*, 209(2):220–230.

[169] Wu, J. and Chung, A. (2005). Cross Entropy: A new solver for Markov random field modeling and applications to medical image segmentation. In Duncan, J. and Gerig, G., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2005*, volume 3749 of *Lecture Notes in Computer Science*, pages 229–237. Springer Berlin Heidelberg.

[170] Xie, C. and Tammi, M. T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, 10:80.

[171] Xuan, J., Yu, Y., Qing, T., Guo, L., and Shi, L. (2013). Next-generation sequencing in the clinic: promises and challenges. *Cancer Lett.*, 340(2):284–295.

[172] Y.Yao (1988). Estimating the number of change-points via Schwarz criterion. *Statistics & Probability Letters*, 6:181–189.

[173] Zhang, N. R., Seigmund, D. O., Ji, H., and Li, J. Z. (2010). Detecting simultaneous changepoints in multiple sequences. *Biometrica*, 93(3):631–645.

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