WCCI 2012 IEEE World Congress on Computational Intelligence
June, 10-15, 2012 - Brisbane, Australia IEEE CEC

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

Madawa Priyadarshana, W.J.R.
Department of Statistics
Faculty of Science
Macquarie University
Sydney NSW 2109 Australia
Email: madawa.weerasinghe@mq.edu.au

Georgy Sofronov
Department of Statistics
Faculty of Science
Macquarie University
Sydney NSW 2109 Australia
Email: georgy.sofronov@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 criterions are also discussed. The proposed CE method applies on over-dispersed count data, in which the observations are distributed as independent Negative Binomial. Furthermore, we incorporate the Bayesian Information Criterion 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.**

***Index Terms*—change point problem, Cross-Entropy method, combinatorial optimization, DNA count data, stochastic optimization**

I. 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 methods developed to segment binary sequences as well as continuous data. However, in literature there exist only a handful of resources concentrating mainly on change point detection in count data and especially on Deoxyribonucleic Acid (DNA) count data.

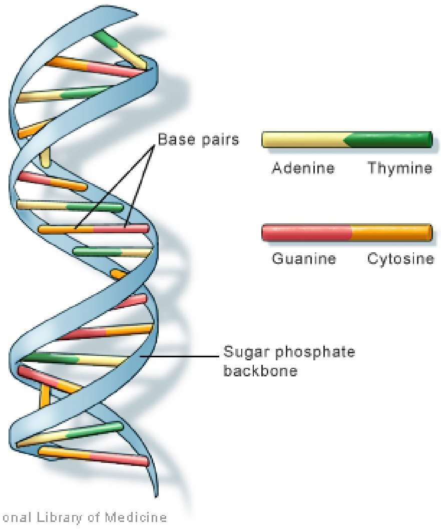

Fig. 1. The DNA structure. Source: US National Library of Medicine (http://ghr.nlm.nih.gov/handbook/basics/dna)

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, Braun and Müller [1] 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 modeling in DNA count data. On the more recent advances, Tibshirani and Wang [2] applied fused lasso method to the hot spot detection in comparative genomic hybridization (CGH) data. Where, CGH [3] is a technique for measuring DNA copy number of selected genes on the genome. Erdman and Emerson [4] introduced an improved version of the computing package on change point modeling based on Product Partition Models (PPM) introduced by Barry and Hartigan [5]. Zhang *et al.* [6] 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. Furthermore, various approaches to detect multiple change points in DNA sequences were discussed in [7], [8], [9] and [10]. More recently with the development of next-generation sequencing data, Ivakhno *et al.* [11] proposed a novel approach called CNAsegbased on number of reads in order to identify the copy number abnormalities (CNAs).

The above literature on change point modeling 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 modeling has been discussed extensively by many authors mainly in the context of Generalized Linear Modeling (GLM). This comprises of analysis when the over-dispersion is present or not and with many other attributes [12]. 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 reveals the true nature of the underlying structure of the observations. Li and Lund [9] have recently 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 count data, considering it as a combinatorial problem and discusses two competing stopping criterions. 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 [13] in order to identify the number of change points as well as the locations in DNA count data. Change point modeling with the use of CE concept was first utilized in [14] 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 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. The Bayesian Information Criterion (BIC) [15], [16] is used to identify the number of change points in the count data. Yao [16] shows that for normally distributed data the estimate on the number of change points obtained by the BIC weakly converges to the true number of change points. Finally, the study will compare the results with the general approach proposed as in [14] and discuss two stopping criterions 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, the 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.

II. 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 i-th 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

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

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

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

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

A. Four parameters Beta distribution (Beta4 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_0 and L_M. Then, the probability density function of the four parameter beta distribution is given by,

$$f(y \mid \alpha, \beta, L_0, L_M) \\ = \frac{1}{B(\alpha, \beta)}(y - L_0)^{\alpha-1}\frac{(L_M - y)^{\beta-1}}{(L_M - L_0)^{\alpha+\beta-1}}.$$

The method-of-moment estimates (e.g., [17]) of the shape parameters are

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

$$\hat{\beta} = (1-\bar{y})\left[\frac{\bar{y}(1-\bar{y})}{s^2} - 1\right]. \tag{3}$$

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

$$\bar{y} = \frac{\bar{y} - L_0}{L_M - L_0}$$

and

$$s^2 = \frac{s^2}{(L_M - L_0)^2}.$$

Furthermore, mean and the variance of the four parameter Beta distribution are:

$$\text{Mean} = \frac{\alpha L_M + \beta L_0}{\alpha + \beta}, \tag{4}$$

$$\text{Variance} = \frac{(\alpha-1)L_M + (\beta-1)L_0}{\alpha + \beta - 2}. \tag{5}$$

III. Modified Cross-Entropy Method for Multiple Change Point Problem

A. The standard Cross-Entropy method

The Cross- Entropy (CE) method [13] 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 N_{elite} of best of these 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.

Let N 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-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 [14] considers truncated independent normal distributions 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.

B. 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, N beta distributions on the support $[L_0, L_M]$ are used to simulate the locations of change points. We denote the set of these beta distributions by $\mathsf{Beta}(\boldsymbol{\alpha}, \boldsymbol{\beta})$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_N)$ are the parameters of each component. In each iteration, the parameters of the $\mathsf{Beta}(\alpha_i, \beta_i)$ distributions, $i = 1, 2, \ldots, N$, are updated until a stopping criterion is met. Finally, the performance function F in this study is the BIC [15], [16] 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 vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ such that $\alpha_i = \beta_i = 1$, $i = 1, 2, \ldots, N$, which are the uniform distributions on the interval $[L_0, L_M]$, since we are dealing with four parameter beta distributions. Where L_0 and L_M are the lower and upper bound of the count data sequence, that is, $L_0 = 0$, $L_M = L$. 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, the BIC, that we wish to minimize is

$$F = -2ll(N, \mathbf{C}, \mathbf{p}) + k\ln(L), \tag{6}$$

where $ll(N, \mathbf{C}, \mathbf{p})$ is the log likelihood as in (1) of the count data sequence, $k = 2(N + 1)$.

In each of the iterations N_{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 criterions are discussed and evaluated. The first criterion is based on the [14] and the other is based on the original CE method as in [13]. 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 distributions.

The CE-Beta algorithm can be summarized as below:

1) Choose initial values for $\boldsymbol{\alpha}^{\mathbf{0}} = (1, 1, \ldots, 1)$ and $\boldsymbol{\beta}^{\mathbf{0}} = (1, 1, \ldots, 1)$. 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 $\mathbf{C}^{(1)}, \mathbf{C}^{(2)}, \ldots, \mathbf{C}^{(N_1)}$ from the $\mathsf{Beta}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\beta}^{t-1})$ distribution, where $\mathbf{C}^{(i)} = (c_1^{(i)}, c_2^{(i)}, \ldots, c_N^{(i)})$, $i = 1, 2, \ldots, N_1$, is the change point vector as defined earlier.

3) For each $i = 1, 2, \ldots, N_1$ order $c_1^{(i)}, \ldots, c_N^{(i)}$ from smallest to biggest and set $\mathbf{C}^{(i)} = (c_1^{(i)}, c_2^{(i)}, \ldots, c_N^{(i)})$.

4) Evaluate the performance of each $\mathbf{C}^{(1)}, \mathbf{C}^{(2)}, \ldots, \mathbf{C}^{(N_1)}$ using (6). Define the elite sample, which is the best performing combinations of the change points. Let $N_{\text{elite}} = \rho N_1$ be the size of the elite sample.

5) For all $j = 1, 2, \ldots, N$ estimate the two shape parameters $\boldsymbol{\alpha}^t = (\alpha_1^t, \alpha_2^t, \ldots, \alpha_N^t)$, $\boldsymbol{\beta}^t = (\beta_1^t, \beta_2^t, \ldots, \beta_N^t)$ as in (2) and (3) using the 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. Otherwise set $t = t + 1$ and iterate from step 2.

Note that the steps of the CE-Normal algorithm is identical to the CE-Beta except for the use of normal distribution to simulate the locations of change points. In order to start the two algorithms with a common basis, we initialized the normal distribution parameters with the corresponding mean and standard deviation of $\mathsf{Beta}(1, 1)$ on the support $[L_0, L_M]$ by using (4) and (5).

The two stopping criterions (SCs) that are considered in this study are

SC1: Stop the process if $\max_j (\sigma_j^2)^t < \varepsilon$.
SC2: Stop the process if for some $t \geq k$, say $k = 4$, $F_t = F_{t-1} = \cdots = F_{t-k}$.

Finally, the solution will be a single vector of change points.

IV. 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 criterions 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, 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.

A. *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 I, 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 criterions.

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 criterions. We have considered a sample size N_1 of 1000 and an elite proportion value ρ of 0.1 in all of the algorithms. In CE-Beta and CE-Normal under SC1 ε value of 0.5 considered and under SC2 k value of 4 used as the stopping criterions parameters. Then we obtained the best solution in each of the N situations which minimizes (6). Figure 6 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 criterions.

Figures 2, 3, 4 and 5 show the underlying process behind the CE-Beta SC1 algorithm when identifying the locations of the change points at $N = 5$. Figure 2 is drawn from the updated Beta parameters after running the algorithm with the initial parameters of $\boldsymbol{\alpha} = \boldsymbol{\beta} = 1$. Consequently, Figures 3, 4 and 5 are obtained from the updated Beta parameters after the corresponding iteration. It is noted that, as the iteration number increases the shape of the Beta distribution changes, such that the process converges to the mode of the Beta distributions. This shows the adaptive nature of the CE algorithm, where in each iteration the parameters are updated in order to obtain better estimates of the locations of change points. Finally, at iteration 19 the stopping criteria is met and it is almost converged to a point mass in the count data sequence.

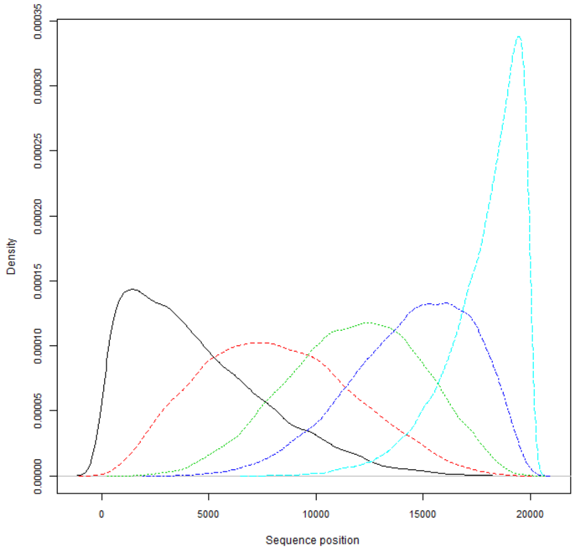

Fig. 2. The shape of the beta distributions after 1 iteration of the CE-Beta under SC1 at $N = 5$.

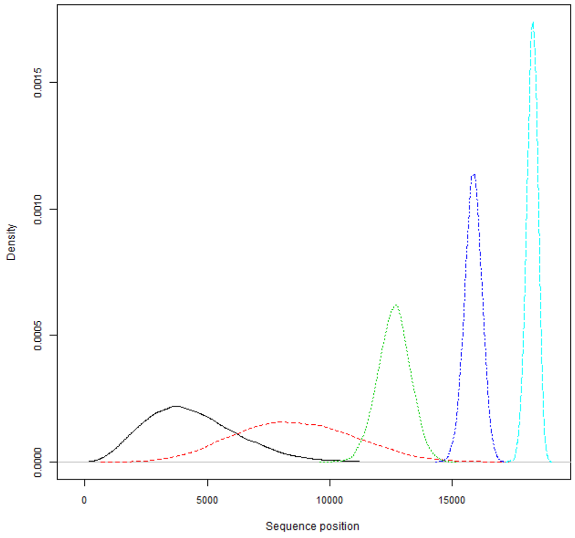

Fig. 3. The shape of the beta distributions after 4 iterations of the CE-Beta under SC1 at $N = 5$.

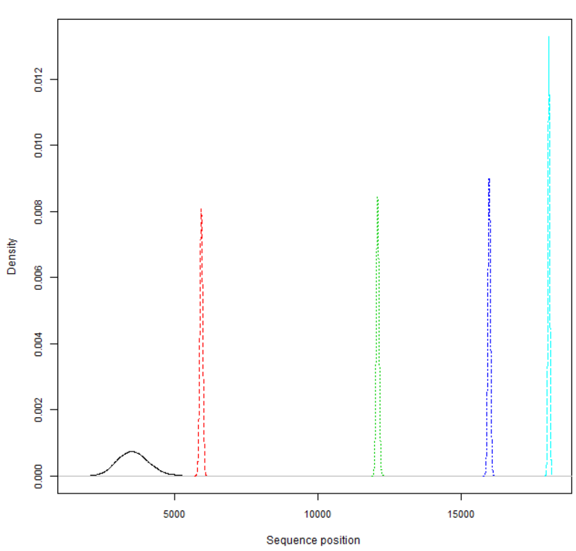

Fig. 4. The shape of the beta distributions after 10 iterations of the CE-Beta under SC1 at $N = 5$.

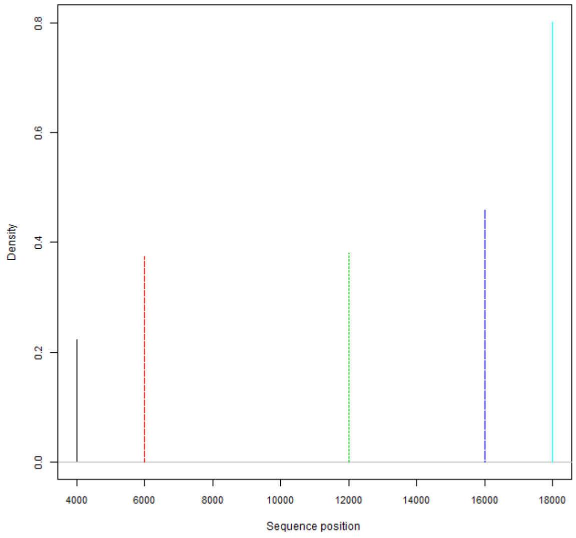

Fig. 5. The shape of the beta distributions after 19 iterations of the CE-Beta under SC1 at $N = 5$.

Figure 6 shows that in both algorithms with two stopping criterions, the BIC is minimized when $N = 9$. More importanlty, when considering the processing time as in Table II, there is a significant improvement in the proposed CE-Beta algortihm 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 II 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 processer with 4GB RAM. Therefore, the processing time is relative to this operation conditions.

Table III 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. $N = 9$). The RMSE values indicate that even though the computing time is highly superior under SC1 it gives less

TABLE I

NEGATIVE BINOMIAL p PARAMETERS FOR THE ARTIFICIAL SEQUENCE WITH FIXED SIZE OF $r = 10$.

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	$p_9 = 0.90$

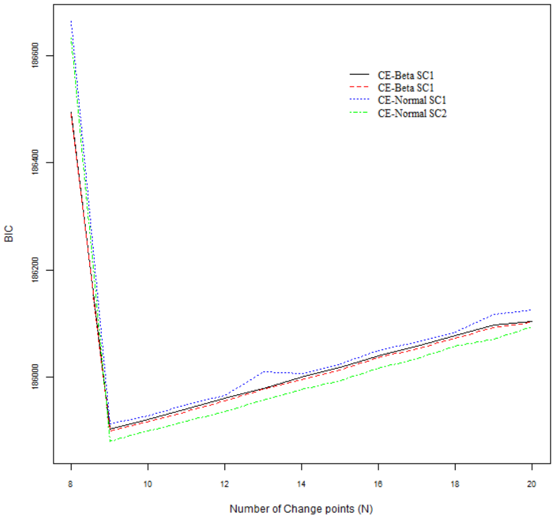

Fig. 6. BIC vs. N for CE-Beta and CE-Normal with two SCs

TABLE II

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	27460.30

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 criterions correctly captured the major regions in the over dispersed count data series.

B. Parameter smoothing: Rho (ρ)

We have considered smoothing up the paramter Rho (ρ), which is used to obtain the elite sample. The RMSE and

TABLE III

AVERAGE RMSE FOR 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

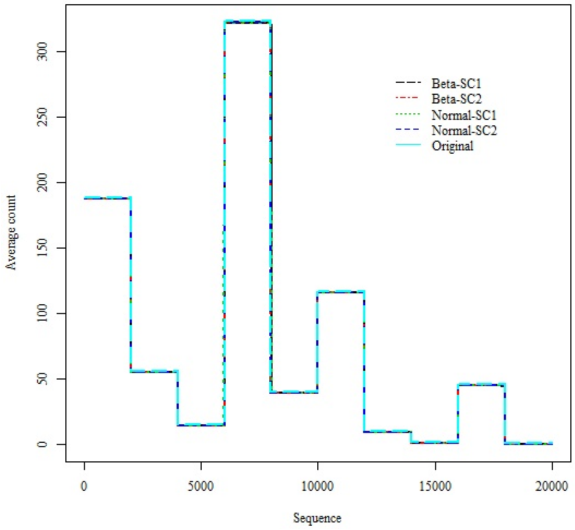

Fig. 7. Average count vs. sequence position for CE-Beta and CE-Normal with two SCs

processing time(s) is obtained for the Rho values from 0.01 to 0.1 with the bin of 0.01 when $N = 9$. We have obtained the average results based on 100 simulations under each of the Rho values.

Figure 8 indicates that the RMSE for the SC2 is lower than the SC1 algorithms both in CE-Beta and CE-Normal cases. Furthermore, CE-Beta algotrithms have lower RMSE on average than that of the competing CE-Normal algorithms. Also, by looking at the Figure 8 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 9) the SC1 algorithms outperform the SC2 algorithms in both CE-Beta and CE-Normal cases. On average the CE-Normal 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 disucssed above.

C. Example 2: Real Data

This example considers a real DNA count data. The data correspond to the chromosome 2 of a subject in a study. Due to this being real data we do not know the true number of change points in advance. Therefore, we look for agreement between the two methodologies. We have considered the proposed CE-

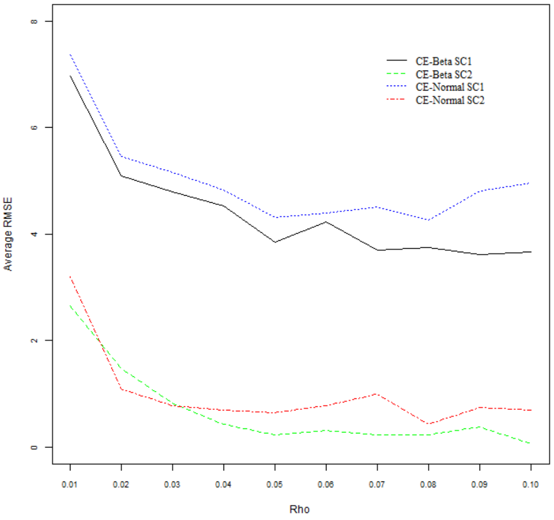

Fig. 8. Plot of Average RMSE vs. Rho (ρ)

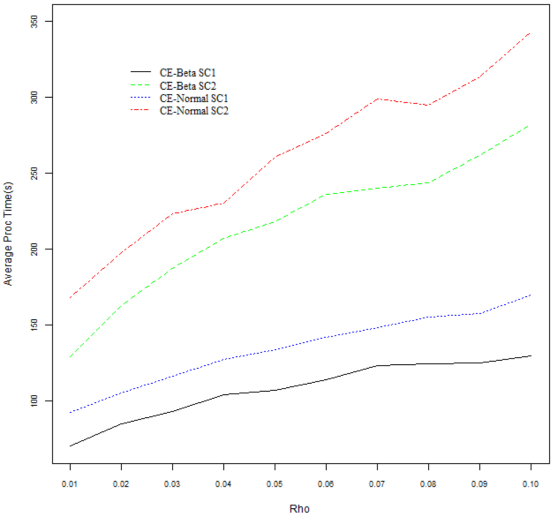

Fig. 9. Plot of Average Processing time (s) vs. Rho (ρ)

Beta method and the CE- Normal method under the SC1 to compare the results. In order to calculate the N_{elite} fraction of samples ρ value of 0.05 is used.

Figure 10 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 11 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 $N = 28$ 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

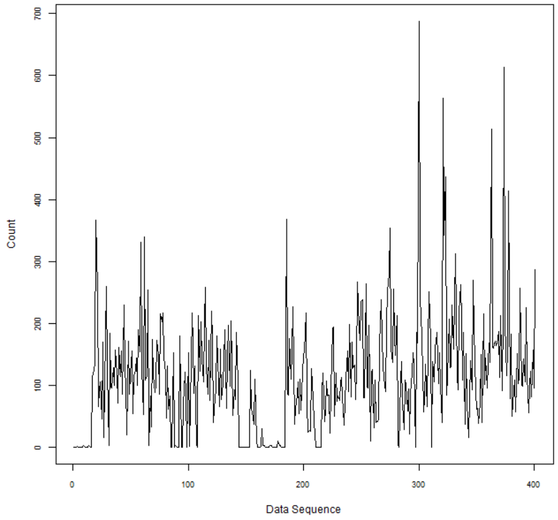

Fig. 10. Part of the DNA count data

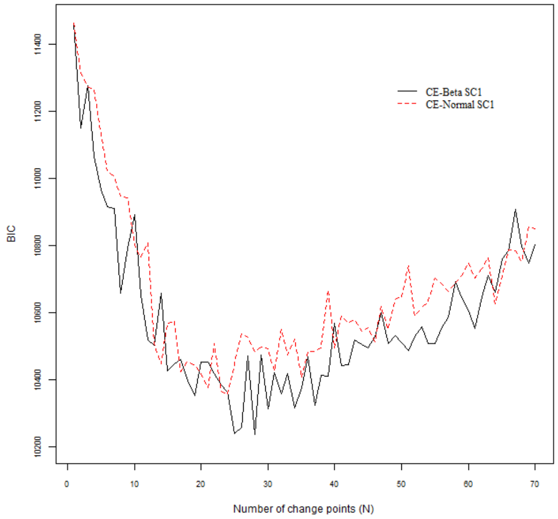

Fig. 11. BIC vs. Number of change points (N)

number of change points equal to 28 and 24 respectively.

Figure 12, 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.

V. Conclusion

A modified CE method is proposed with different stopping criterions. This proposed method utilizes beta distribution to

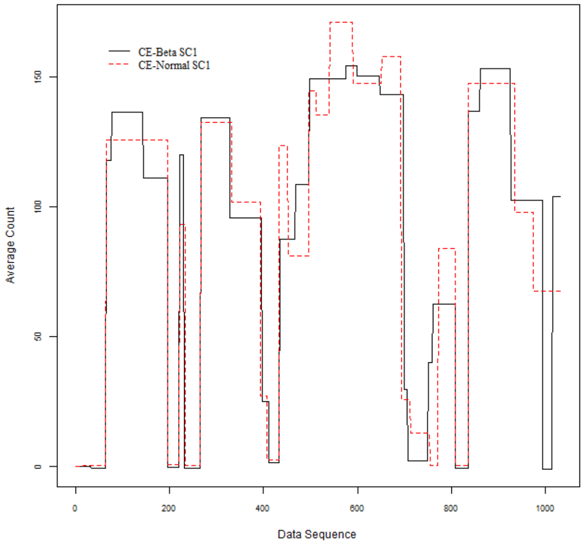

Fig. 12. Average count vs. sequence position for CE-Beta and CE-Normal under SC1 of DNA count data

TABLE IV
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

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 criterions. However, the CE algorithm with SC2 produced lower RMSE 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 for count data problems. A modified BIC [18] 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.

REFERENCES

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

[2] 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.

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

[4] 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.

[5] 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.

[6] N. R. Zhang, D. O. Seigmund, H. Ji, and J. Z. Li, “Detecting simultaneous changepoints in multiple sequences,” *Biometrica*, vol. 93, no. 3, pp. 631–645, 2010.

[7] G. Sofronov, “Change-point modelling in biological sequences via the bayesian adaptive independent sampler,” *International Proceedings of Computer Science and Information Technology*, vol. 5, pp. 122–126, 2011.

[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, no. 5, pp. 577–584, 2009.

[9] S. Li and R. Lund, “Multiple changepoint detection via genetic algorithms,” *Journal of Climate*, doi: http://dx.doi.org/10.1175/2011JCLI4055.1.

[10] 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.

[11] S. Ivakhno, T. Royce, A. J. Cox, D. Evers, R. 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.

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

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

[14] G. E. Evans, G. Y. Sofronov, J. M. Keith, and D. P. Kroese, “Identifying change-points in biological sequences via the coss-entropy method,” *Annals of Operation Research*, vol. 189, no. 1, pp. 155–165, 2011.

[15] G. Schwarz, “Estimating the dimension of a model,” *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

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

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

[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.