Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

23-09

Visually Identifying Potential Domains for Change Points in Generalized Bernoulli Processes: an Application to DNA Segmental Analysis

Yan-Xia Lin

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Visually Identifying Potential Domains for Change Points in Generalized Bernoulli Processes: an Application to DNA Segmental Analysis

Yan-Xia Lin
School of Mathematics and Applied Statistics
University of Wollongong, NSW 2522, Australia

1 Introduction

A Bernoulli process is a discrete-time stochastic process consisting of a finite or infinite sequence of i.i.d. random variables $Y_1, Y_2, Y_3, \cdots$, and Y_1 has Bernoulli distribution with mean p.

Following the generalization of binomial distribution given by Drezner and Farnum (1993), we name a process $\{Y_t\}$ a generalized Bernoulli process if, for all $t > 0$, Y_t has Bernoulli distribution with mean $p_t > 0$, where $Y_1, Y_2, Y_3, \cdots$ are not necessarily independent and p_t are not necessarily all the same.

In this paper, without further notice, we are only interested a special scenario of generalized Bernoulli processes, where all Y_t are mutually independent. A Bernoulli process is a special generalized Bernoulli process where all $p_t = p$.

If there is an integer τ such that $p_t = c_1$ for $t < \tau$ and $p_t = c_2$ for $t \geq \tau$, where $c_1 \neq c_2$, we say, the generalized Bernoulli process $\{Y_t\}$ has structure change at position τ. Simply the position is called a change point of the process. A generalized Bernoulli process might have more than one change points sometimes.

Many real life data can be well modelled by generalized Bernoulli processes, particularly the data of DNA sequences.

In molecular biology, two nucleotides on opposite complementary DNA or RNA strands that are connected via hydrogen bonds are called a base pair (often abbreviated bp). In the canonical Watson-Crick base pairing, adenine (A) forms a base pair with thymine (T), as does guanine (G) with cytosine (C) in DNA (see Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/Base_pair). In melting curve studies, the DNA is gradually denatured (the strands separated) by heating. The bonds of the G-C base-pairs are stronger, taking more heat to melt. Therefore, the proportion of G-C in a DNA sequence

is of interested.

A DNA sequence can be expressed as a sequence of $Y_1, \cdots, Y_n$ where Y_i takes one of DNA alphabet (A,C,G or T). Sometimes G-C base-pairs model, A-G base-pairs model or T-G base-pairs model are considered in DNA sequence analysis. For example, for G-C base-pairs model, bases G and C are classified into a same category. Thus, a DNA sequence can be modelled by a generalized Bernoulli process $\{Y_t\}$: $Y_t = 0$ if the tth observation in the sequence is G or C; otherwise $Y_t = 1$ (Braun and Muller, 1998; Fu and Curnow, 1990).

A subsequence of a DNA sequence is called a segment of the sequence if it is a subsequence between two consecutive change points of the DNA sequence (see Braun and Muller, 1998). In other words, a segment of a DNA sequence is a stationary subsequence of the DNA sequence. A DNA sequence might consist of several segments with different stationary probability structure. Early evidence of segmental genomic structure was provided by the phenomenon of chromosome banding. The distribution of the location and the length of segments is a useful characteristic on DNA sequence and the information can be considerable assistance to molecular biologists particularly when they incorporate the discrete nature of changes caused by evolutionary processes. The study on segmental structure in DNA sequence might helpful in understanding and discovering the secrete of DNA sequence and the secrete involved in DNA sequence evolution. The study on DNA segments may include detecting segments which are anomalous (in the sense that they are either mistakenly included in the sequence under consideration or perhaps derive from some other organizational scheme).

Detecting segments in a DNA sequence is equivalent to detecting structure changes in a process. A comprehensive overview of mathematical methods for DNA segmentation can be found from Braum and Muller (1998), which covers the maximum likelihood estimation of segments, hidden Markov approach, Bayesian approach, locally weighted split polynomial regression approach, scan statistics method, Global segmentation method and binary segmentation method. It is a fact that none of these methods are super powerful and none of them are universally efficient in terms of correctly identifying change points in processes. Hinkley and Hinkley (1970) investigated the inference on the change points in a sequence of binomial variable. The asymptotic distribution of the maximum likelihood estimator of the change point of a

sequence of binomial variables was derived. They realize that it will be less confident on results given by the ML method if the size of tested samples is small and the knowledge on the means of the tested sequence before and after change points are not available.

The accuracy of the estimations of change points can be significantly improved if the approach below is followed. (i) Firstly, to identify relatively shorter interval domains covering change points such that each interval only covers one change point. (ii) Then, to apply an inference method, say the ML method, to tested process on each the intervals and to obtain the estimation of the change point in the interval. Muller and Song (1997) named this approach as a two-step approach. We call the shorter intervals in this paper potential domains for change points.

Given a sample path of a process, it is of interested how potential domains of change points for the process can be identified. A simple way to identify the potential domains for change points in a process is by identifying the plot pattern changes in the plot of the process. Currently two type of plots can be used for the purpose, time series plot and moving average plot. An example of moving average method for DNA segmental analysis can be found from Braum and Muller (1998). However, the time series plot method is not suitable for generalized Bernoulli processes as the processes only take two difference values "0" and "1" and the mean changes of the process is difficultly observed through its plot. Moving average plot also has its weakness. Usually, the outputs of moving average plots are too sensitive to the window size used. Furthermore, the moving average plot used for observing purpose is not unique for each given sample path of a tested process.

In this paper, a new graphical method for generalized Bernoulli processes is developed for two purposes. (i) Given a sample drawn from a generalized Bernoulli process, the plot pattern produced by the new method has to be unique determined by the sample. (ii) By using the new method, the structure changes in generalized Bernoulli processes should be easily observed. By using this new graphical method, potential domains for changes points in generalized Bernoulli processes should be easily determined.

As an application of the new method, we demonstrate how to use the new method to locate potential domains for change points in generalized Bernoulli processes and subsequently to improve the ML estimations of change points.

This paper consists of 5 sections. An example is given in Section 2 to show that sometimes maximum likelihood estimations of change points might be misleading. The information of potential domains on change points will benefit the improvement of the ML estimations of change points. In Section 3, an associate process of a generalized Bernoulli process is introduced. The relationship between a Bernoulli process and its associate process is investigated. Four simulation examples are given to show how the plot patterns of an associate process are affected by the changes of the probability structure in its original Bernoulli process. Due to the nature of associated processes, the pattern changes in the time series plot of associate processes are more visible. Therefore, the information obtained from the plots of associate processes may benefit the estimations of change points in generalized Bernoulli processes. A second-layer process of a generalized Bernoulli process is defined in Section 4. The relationship between the probability structure of a generalized Bernoulli process and its second-layer process is given in the section. Our study shows that, in certain scenarios, the information from the second-layer processes may be helpful in detecting the structure changes in its root generalized Bernoulli process. This paper develops a new graphical approach to gain the information on change points in generalized Bernoulli processes and use the information to improve the estimation of change points the processes. Applications of the procedure are presented in Section 5.

2 The ML Estimations of Change Points

The Maximum likelihood (ML) method has been widely used in detecting structure changes of discrete stochastic processes (see Fu and Curnow, 1990). A simple example of using the ML method to estimate change points in a generalized Bernoulli process is given below.

Consider a generalized Bernoulli process $Y_1, Y_2, \cdots, Y_T$, where $Y_t \sim B(1, p)$ for $t < \tau$; $Y_t \sim B(1, p+\delta)$ for $\tau \leq t \leq T$ and $\delta \neq 0$. The position τ is unknown and needs to be estimated. In the following discussion, assume that the values of p and δ are known, although it is not always the case in practice.

Given a sequence of observations $y_1, y_2, \cdots, y_T$ drawn from the process, the likelihood function of $(y_1, y_2, \cdots, y_T)$ is

$$\Pi_{i=1}^{\tau-1} p^{y_i}(1-p)^{1-y_i} \Pi_{i=\tau}^{T} (p+\delta)^{y_i}(1-p-\delta)^{1-y_i}$$

and the log likelihood conditional on $\tau = k$, is

$$\begin{aligned} L(k) &= \sum_{i=1}^{k-1}(y_i \log(p) + (1-y_i)log(1-p)) \\ &\quad + \sum_{i=k}^{T}(y_i \log(p+\delta) + (1-y_i)\log(1-p-\delta)) \end{aligned} \tag{1}$$

(see Fu and Curnow, 1990; Braun and Muller, 1998). The ML estimator of τ, denoted $\hat{\tau}$, satisfies

$$L(\hat{\tau}) = \max_{0<k<T}\{L(k)\}.$$

Theoretically, it is acceptable to use ML estimator to estimate the unknown change point τ. However, sometimes the ML estimation of τ, especially when sample size T is small or the values of p and δ are unknown or lack of information on the possible range of the change point, is misleading (see Hinkley and Hinkley, 1970). This fact is also found from Example 1 below.

Example 1. Consider 1000 independent samples simulated from a generalized Bernoulli process $Y_1, Y_2, \cdots, Y_T$, where Y_t has Bernoulli distribution with parameter p when $t < T/2 = \tau$; with parameter $p+\delta$ when $T/2 \leq t \leq T$. For different values of p, δ and T, the statistics descriptions on the ML estimations of τ given by the 1000 independent samples are presented in Table 1.

Table 1: Statistics description of $\hat{\tau}$

p	δ	T	True τ	sample mean of $\hat{\tau}$	sample variance of $\hat{\tau}$
0.2	0.2	200	100	101.366	484.7488
		80	40	39.899	219.6965
		60	30	29.629	150.7301
0.2	0.1	60	30	30	285.9659

The sample mean of the estimations of τ is reasonably close to the true value of τ. This reflects that $\hat{\tau}$ is unbiased. However, the sample variance is larger, which means some estimations of τ are far way from the true value τ (See Figure 1, for example.).

Figure 1 shows that 284 out of the 1000 estimations of τ are less than 20 and 289 out of 1000 are greater than 40. Among the 1000 ML estimations of τ,

only 42.7% of them are between 20 and 40. If one randomly assigns a position between $[0, 60]$ as an estimation of τ, it will be a 30% chance to have the estimation between 20 and 40. Comparing 42.7% with 30%, the performance of the ML method does not sound well.

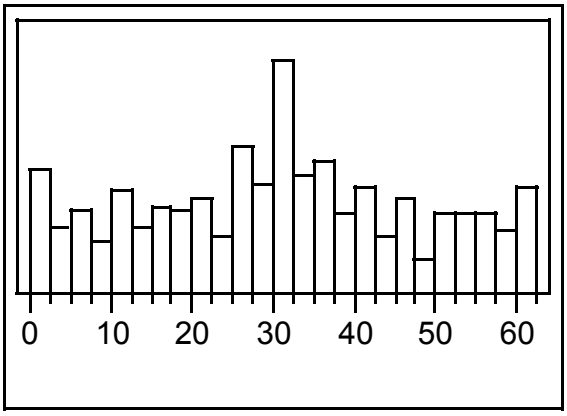

Figure 1: The histogram of the estimations of change point given by 1000 independent samples drawn from a generalized Bernoulli process with $p = 0.2$, $\delta = 0.1$, $\tau = 30$ and sample size 60.

Obviously, the accuracy of the ML estimation of τ can be significantly improved, if the sample size before and after the change point are large (See Example 2 below), which means more information before and after the change point are provided.

Example 2. Consider the same model in Example 1. Simulate 1000 independent samples from the model with $p = 0.2$, $\delta = 0.1$, $\tau = 100$ and $T = 200$. Apply the ML method to these 1000 samples respectively and obtain the ML estimations of τ. The histogram of the estimations is given by Fig. 2.

Clearly, the shape of the histogram in Fig. 2 has been significantly improved in terms of that the sample variance is significantly reduced. But, there is still approximately 36% of the estimations of τ outside interval $[90, 110]$.

In Example 1, the ML method is applied to the data by assuming that all the positions k in $[1, 60]$ have an equal chance to be the change point. Therefore, the values of the log likelihood function $L(k)$, $k \in [1, 60]$, have been weighted equally. However, this fundamental assumption is wrong. It is impossible that all positions in $[1, 60]$ have an equal chance to be the change

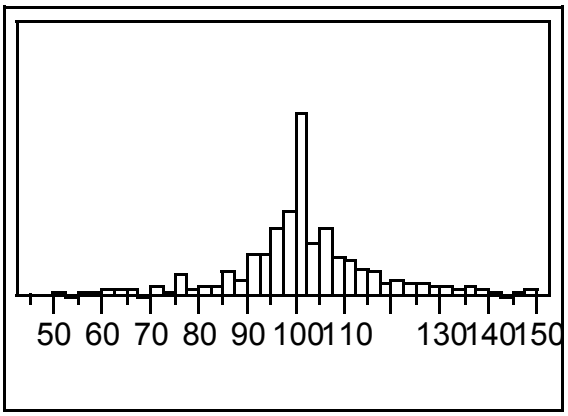

Figure 2: The histogram of the estimations of change point given by the 1000 independent samples drawn from a generalized Bernoulli process with $\mu = 0.2$, $\delta = 0.1$, $\tau = 100$ and sample size 200.

point. It is not surprising that there are so many misleading estimations of τ produced by the ML method.

As mentioned in Section 1, to improve the ML estimations of change points in a tested process, two steps are needed: (i) Identify potential domains for change points such that each true change point is covered by a domain. (ii) Calculate the log likelihood $L(k)$ for all k in potential domains only and determined the ML estimations of change points based on the values of these log likelihood $L(k)$s. Correctly to determine the potential domains for change points is the key step.

In following sections, we develop two types of processes from generalized Bernoulli processes. The information of structure changes in generalized Bernoulli processes can be easily observed from the plots of the processes. Subsequently potential domains for change points in tested generalized Bernoulli processes might be easily identified.

3 Associate Process

If there is a structure change in a generalized Bernoulli process, obviously any sequences of observations of the process must carry the information on the change. The information sometime may be easily identified from the time series plots of the sequences of observations, but sometimes it may not be. See

examples below.

Example 3. Let us consider the 994th independent sample path in Example 1 with $p = 0.2$, $\delta = 0.1$, $T = 60$ and $\tau = 30$. The ML estimation of τ given by this sample is 9, which is far away from the true change point position 30. By observing the time series plot of the sample path (see Fig 3), one is able to claim that there might be a structure change around position 30 as the distributions of 1's positions before and after position 30 are different. For this example, the information on structure change is relatively easy to be observed through the time series plot of the sample. Therefore, a potential domain for the change point τ can be easily located and the ML estimation of τ can be improved if the ML method is apply to the potential domain only.

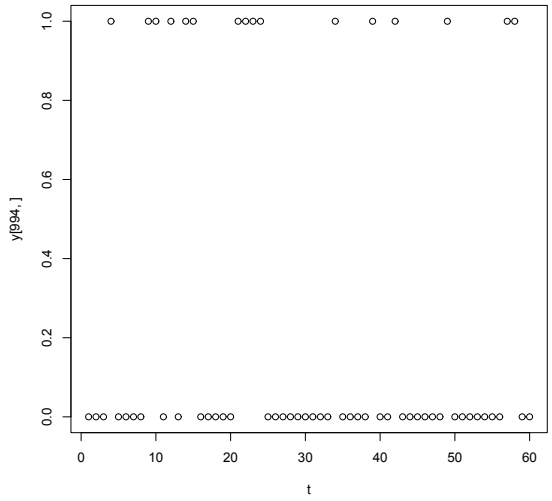

Figure 3: The scatter plot of the 994th sample.

However, to directly identify plot pattern changes from the time plot of a generalized Bernoulli process is rather difficult in many cases. For example, see Fig. 4.

Why is it difficult to observe the structure change in Fig. 4? The changes in the mean and variance of a stochastic process is usually observed from its time series plot through the horizontal movement of the observations of the process and the vertical movement of the spread range of the observations of the process. However, a generalized Bernoulli process only takes two possible values "0" and "1". All observations of the process are located on two horizon-

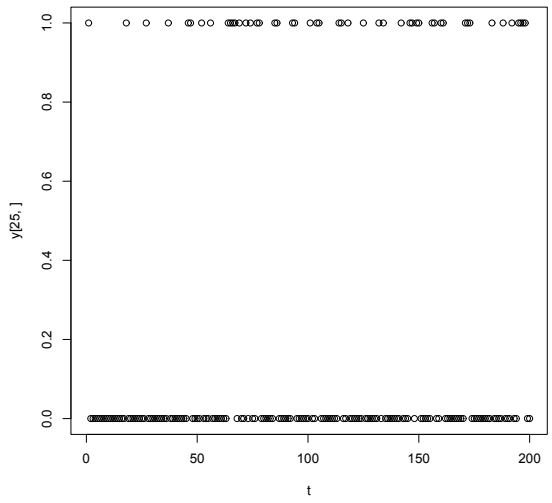

Figure 4: The scatter plot of $x[25, t]$ with $p = 0.2$, $\delta = 0.1$, $T = 200$ and $\tau = 100$.

tal levels, value "0" level and value "1" level. The information on horizontal and vertical movements of the process cannot be presented by the time series plot. It is desirable to have a new way to visually observe structure changes in generalized Bernoulli processes.

3.1 Associate Processes and Their Time Series Plots

The issue of the structure changes in a generalized Bernoulli process is related to the issue of the mean changes in the process. Given a sequence of observations drawn from a Bernoulli process, the larger the mean of the process is, the higher proportion of 1s in the sequence of observations will be, or the shorter the average length of the gaps between consecutive 1s in the sequence of observations will be. Using Fig.3 as an example, we might be able to check if there are any structure changes in a sequence of observations by observing if there are any the pattern changes in the lengths of the gaps between consecutive 1s in the sequence. This is the key idea used in the following study.

In the following, we develop new processes from generalized Bernoulli processes. Each of the processes is used to record the lengths of the gaps between consecutive 1s in their relevant generalized Bernoulli processes. Thus, by observing the time series plot of such processes, we might be able to check if there

are any structure changes in their relevant generalized Bernoulli processes.

Let $\{Y_t\}_{0<t}$ be a generalized Bernoulli process in a probability space $(\Omega, \mathcal{F}, P)$. Define $Y_0 \equiv 1$. For $\omega \in \Omega$, $\{Y_t(\omega)\}$ gives a path of realizations of the process,

$$y_0(\omega) = 1, y_1(\omega), y_2(\omega), \cdots, y_T(\omega), \cdots.$$

Let v_s be the position of the sth 1 in the path. v_s is a random variable mapping from Ω to the positive integer space. Define a sequence of observations $W(\omega, v_s(\omega)) \triangleq W_{v_s}(\omega)$ as follows: $W(\omega, v_s(\omega)) = k$, if there are k consecutive 0s following the sth 1 at position of $v_s(\omega)$. For each ω, the sequence $\{W_{v_s}\}$ is only defined at positions $v_1(\omega), v_2(\omega), \cdots$, which is related to the sample path

$$y_0(\omega) = 1, y_1(\omega), y_2(\omega), \cdots, y_T(\omega), \cdots.$$

We name $\{W_{v_s}\}$ an *associate process* of $\{Y_t\}$.

Theorem 1 *Let $\{Y_t\}_{t\geq 1}$ be a Bernoulli process with $EY_1 = p$, and $Y_0 = 1$. Then, W_{v_s}, $s = 1, 2, \cdots$, are i.i.d., having geometric distribution with parameter p.*

The proof of Theorem 1 is given in Appendix A.

Remarks of Theorem 1

(i) If a Bernoulli process has mean p, its associate process will be stationary with mean $(1-p)/p$ and variance $(1-p)/p^2$. It can prove that, any two Bernoulli processes have the same probability structure (i.e. their means are the same) if and only if their associate processes have the same mean and variance. Therefore, the probability structure of a Bernoulli process is unique determined by its associate process.

(ii) Instead of taking only two values 0 and 1, associate processes will take any nonnegative integer values. Therefore, vertical dispersion of the observations of an associate process as well as the horizontal movement of the process can be easily observed from its time series plots. More examples can be found in this paper.

(iii) If the associate process of a generalized Bernoulli process is not stationary, neither is the generalized Bernoulli process.

A segment of a generalized Bernoulli process Y_t is a subsequence of a Bernoulli process. Based on Theorem 1, each segment of a generalized Bernoulli

process Y_t corresponds to a stationary subsequence of its associate process W_{v_s}. Thus, if a generalized Bernoulli process $\{Y_t\}_{0<t\leq T}$ has a change point $\tau \in (0, T]$, the mean of the process before and after the change point τ will be changed. This information will be reflected from the time series plot of its associate process. Since associate process is defined on random positions, for each sample of $\{Y_t\}$, an interval $[B, A] \subset [0, T]$ should be able to be determined such that: (i) the associate process is stationary in $[0, B]$ and $[A, T]$, but with different probability structure; (ii) the interval $[B, A]$ covers the change point τ. Therefore, this interval $[B, A]$ can be served as a *potential domain* for the change point τ.

3.2 Simulation Studies on the Plots of Associate Processes

In this section, we firstly use a generalized Bernoulli process as an example to show the plots of its associate process and moving average processes with different moving window sizes, and explain the advantage of using the plot of associate processes in detecting the structure changes in the processes. Then more simulation examples are given to visually show the relationship between the mean changes in generalized Bernoulli processes and the plot pattern changes in the plot of their associate processes.

1. The plot of an associate process vs the plot of moving averages

The moving averages process of a process is very sensitive to the size of moving window used. If the size of moving window is smaller, the plot of the moving averages process will show too much variation of the original process; if the size of moving window is bigger, the information on the structure changes in the original process will be difficultly identified from the time series plot of the moving average process. There is no standard criteria for choosing the size of moving window and the determination on the size is subject to tested sample and the experience of data analyzer. Given a process, many different moving averages sequences can be produced from the process. Thus, it will be a question which moving averages sequence is an appropriate one for the purpose of checking structure changes in the process.

Opposite to moving averages processes, a generalized Bernoulli process can only produce one associate process. Therefore only one time series plot of

associate process can be provided for the generalized Bernoulli process.

Example 4: Simulate a sample with size 60 from a generalized Bernoulli process $\{Y_t\}_{0<t\leq 60}$, where Y_t has a Bernoulli distribution with mean $p = 0.2$ for $1 \leq t \leq 29$; with mean $p + \delta = 0.2 + 0.1$ for $30 \leq t \leq 60$. Compare the plot of associate process and the plots of moving averages of $\{Y_t\}$ with window sizes 5 and 10 respectively. The plots are presented in Fig. 5.

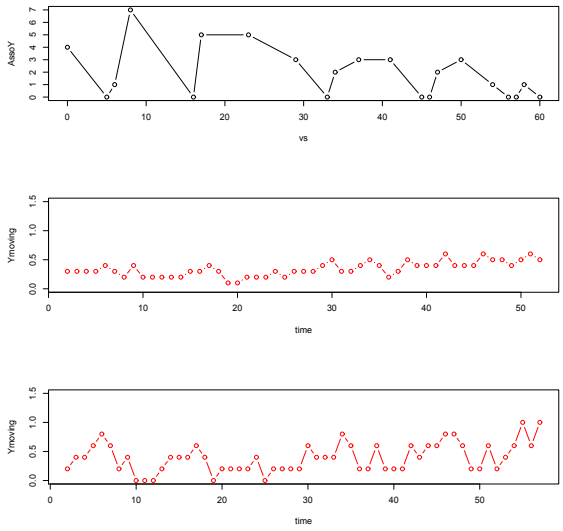

Figure 5: The first plot is given by associate process; the second and the third are the plots of moving averages with window size 10 and 5 respectively.

In Fig. 5, the moving averages plot given by window size 5 has more variation than the plot given by window size 10. Both of them indicate that there might have a change point around position 30 in the process. But the message is not shown as clearly as that in the plot of associate process.

2 The mean of a Bernoulli process vs the variation of its associate process

An associate process is used to record the gap distance between 1s in the original generalized Bernoulli process. From Theorem 1, if a Bernoulli process has mean p, the variance of its associate process will be $(1-p)/p^2$. It is a monotonic decreasing function of p. The larger the mean of a Bernoulli process is, the less the variation of its associate process will be. To visually observe the impact of the mean p on the variation of associate process in its time series plot, the following examples are presented.

Example 5: We simulate a sample from a generalized Bernoulli process

$\{Y_t\}$, where $Y_t \sim B(1, 0.4)$ for $t < 300$; $Y_t \sim B(1, 0.6)$ for $300 \leq t < 500$; $Y_t \sim B(1, 0.2)$ for $500 \leq t \leq 700$. The plots of Y_t and its associate process are presented in Fig. 6.

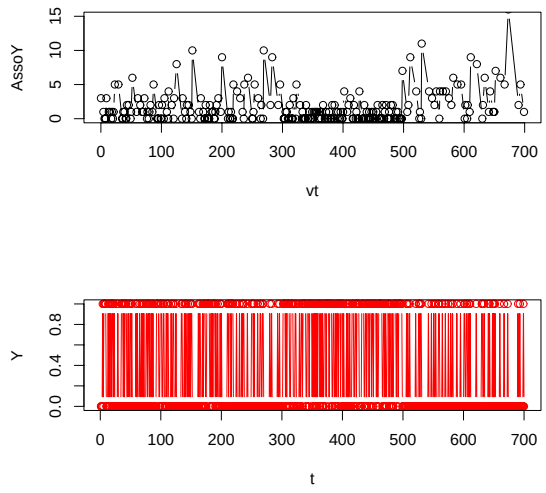

Figure 6: The plot of a generalized Bernoulli process with three different means of segments ($p = 0.4$, 0.6 and 0.2) and the plot of its associate process.

Example 6 below gives two more examples of associate processes. Their original generalized Bernoulli processes have three segments respectively. The value of p is increasing from 0.1 to 0.5 with increment 0.1.

Example 6: Simulate two independent samples from the following two processes respectively

(1) $Y_{1,t} \sim B(1, 0.1)$ for $t < 300$; $Y_{1,t} \sim B(1, 0.2)$ for $300 \leq t < 500$; $Y_{1,t} \sim B(1, 0.3)$ for $500 \leq t \leq 700$.

(2) $Y_{2,t} \sim B(1, 0.3)$ for $t < 300$; $Y_{2,t} \sim B(1, 0.4)$ for $300 \leq t < 500$; $Y_{2,t} \sim B(1, 0.5)$ for $500 \leq t \leq 700$.

The plots of their associate processes are given by Fig. 7.

The plots demonstrate that, as the value p increases, the variation of associate process decreases. When the value of p is close to 0.5 or greater than 0.5, the values of associate process will crowd around X axis. Different values of p, the plot patterns shown in the plots of associate processes are different in

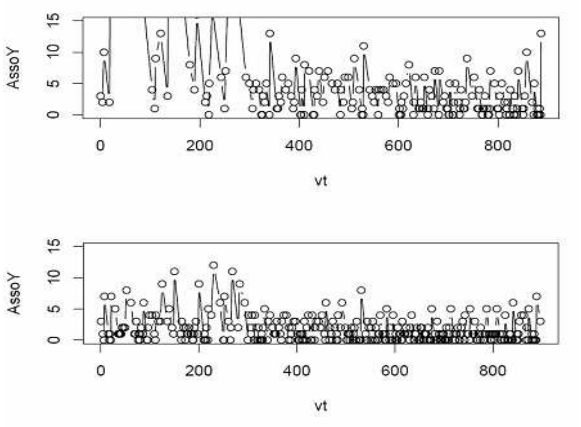

Figure 7: The first one is the plot of associate process of $\{Y_{1,t}\}$; the second one is the plot of associate process of $\{Y_{2,t}\}$.

terms of the level of dispersion. Therefore, the plot patterns given by different p can serve as references in identifying potential domains for change points in generalized Bernoulli processes.

3. The plots of associate processes with p greater than 0.5

As shown in Examples 5 and 6, when p is close to or greater than 0.5, it becomes difficult to distinguish the pattern changes in the plots of associate processes. In this scenario, it might be worth to consider a new process defined by $Y_t^* = 1 - Y_t$ and observe the plot of the associate process of Y_t^* instead of the plot of the associate process of Y_t. For example, by observing Fig. 8, one might lack of confidence to claim that there is a change point around 250 based on the plot produce by the associate process of Y_t, but one is able to see the plot pattern changed before 200 and after 300 from the plot of associate process Y_t^*.

4 The Second-Layer Processes

We have demonstrated that potential domains for change points in a generalized Bernoulli process can be easily located through observing the plot of its associate process in many cases. However, when all the value of p_t in a generalized Bernoulli process are close to 0.5, it become difficult to locate potential domains for change points through the plot of its associate process. An example is given below.

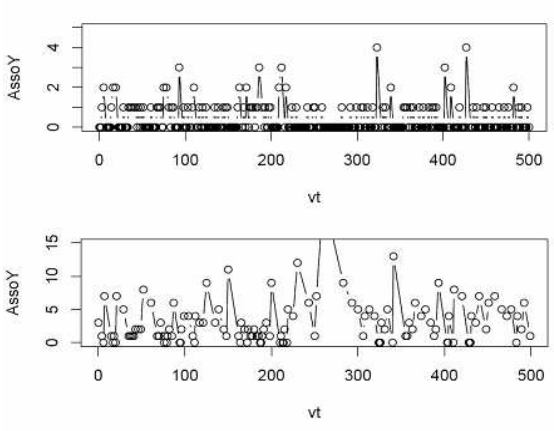

Figure 8: Generalized Bernoulli process Y_t follows model: $Y_t \sim B(1, 0.7)$ for $t < 250$; $Y_t \sim B(1, 0.8)$ for $250 \leq t < 500$. The first plot is the plot of associate process of Y_t and the second one is the plot of associate process of $Y_t^* = 1 - Y_t$.

Example 7: Simulate a sample from $\{Y_t\}$, where $Y_t \sim B(1, 0.5)$ for $t < 250$ and $Y_t \sim B(1, 0.54)$, for $250 \leq t \leq 500$. The plots of Y_t and its associate process are give by Fig. 9.

In the following, we consider another method to obtain the information on change points in generalized bernoulli processes. As mentioned before, when p_t is close to 0.5, the observations of associate processes are close to X axis. This means that the gaps between 1's in this scenario become very shorter. In others words, the information on 1 over crowds and over dominates other information in processes.

The aim of the method developed below is to construct a new process from a generalized Bernoulli process such that the new process does not involve over crowded message presented by the generalized Bernoulli processes, but the structure changes in the generalized Bernoulli process can be still reflected from the new process.

Given a generalized Bernoulli process Y_t, three steps are involved in the new method.

SL Step 1. Construct a process $\{Z_s\}$ from $\{Y_t\}$:

For each ω in probability space S, let

$$Z_1(\omega) \;=\; k_1 \qquad \text{if } Y_1(\omega) = \cdots = Y_{k_1}(\omega) \text{ but } Y_{k_1}(\omega) \neq Y_{k_1+1}(\omega),\ k_1 \geq 1$$

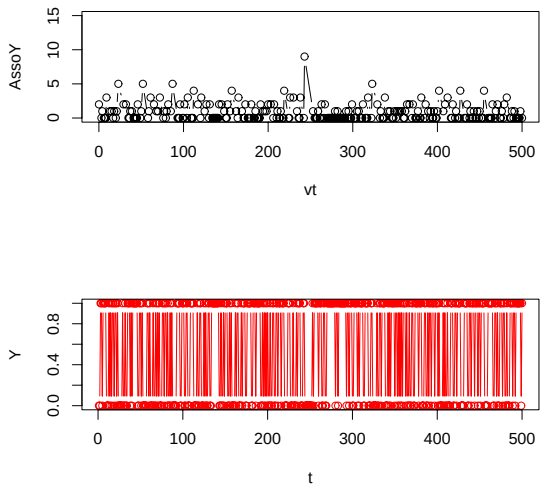

Figure 9: The second plot is the time series plot of a generalized Bernoulli process Y_t, where $Y_t \sim B(1, 0.5)$ for $t < 250$ and $Y_t \sim B(1, 0.54)$, for $250 \leq t \leq 500$. The first plot is the plot of Y_t's associate process. The plot of Y_t is helpless for checking whether there are any change points in the process.

$$
\begin{aligned}
Z_2(\omega) &= k_2 \quad \text{if } Y_{k_1+1}(\omega) = \cdots = Y_{k_1+k_2}(\omega) \\
&\qquad\quad \text{but } Y_{k_1+k_2}(\omega) \neq Y_{k_1+k_2+1}(\omega),\ k_2 \geq 1 \\
&\ \vdots \qquad \vdots \\
Z_s(\omega) &= k_s \quad \text{if } Y_{\sum_{j=1}^{s-1} k_j+1}(\omega) = \cdots = Y_{\sum_{j=1}^{s} k_j}(\omega) \\
&\qquad\quad \text{but } Y_{\sum_{j=1}^{s} k_j}(\omega) \neq Y_{\sum_{j=1}^{s} k_j+1}(\omega),\ k_s \geq 1 \\
&\ \vdots \qquad \vdots
\end{aligned}
$$

where $k_s \geq 1$, $s = 1, 2, \cdots$.

SL Step 2. For $i \geq 1$ we define $Z_s^{(i)}$ as follows:

$$Z_s^{(i)} = Z_s \qquad \text{if } Z_s = i; \qquad = 0 \qquad \text{if } Z_s \neq i\ .$$

SL Step 3. Define $Y_s^{(2)} = Z_s^{(2)}/2$, $s = 1, 2, \cdots$, and name $Y_s^{(2)}$ the second-layer (SL) process of Y_t.

Opposite to Y_t, Z_t may take any positive integer values. The probability structure of $\{Z_t\}$ is determined by the probability structure of $\{Y_t\}$. The plot of a sample path of $\{Y_t\}$ and the plot of $\{Z_t\}$ which is related to the sample path of $\{Y_t\}$ are given by Fig. 10 and Fig. 11 respectively.

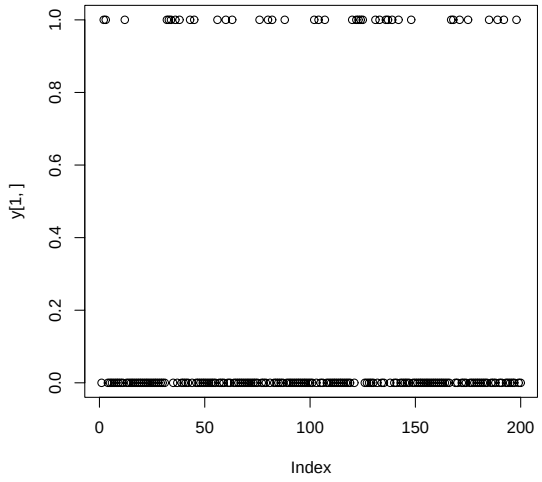

Figure 10: A sample plot of a Bernoulli process $\{y_t\}_{1\leq t\leq 200}$ with mean $p = 0.2$.

We call a subsequence

$$Y_t(\omega) = 0, Y_{t+1}(\omega) = 1 = \cdots = Y_{k_1}(\omega), Y_{k_1+1}(\omega) = 0$$

a 1's subsequence of $Y_t(\omega)$ and a subsequence

$$Y_t(\omega) = 1, Y_{t+1}(\omega) = 0 = \cdots = Y_{k_1}(\omega), Y_{k_1+1}(\omega) = 1$$

a 0's subsequence of $Y_t(\omega)$. Thus, process $\{Z_s\}$ is a process used to sequentially record the length of 1's subsequences and 0's subsequences in the process Y_t.

In Theorem 2 below, we prove that, if $\{Y_t\}$ is a Bernoulli process, so is $\{Y_s^{(2)}\}$. However, the means of $\{Y_t\}$ and $\{Y_s^{(2)}\}$ are different.

Theorem 2 *If Y_t is a Bernoulli process, i.e. $\{Y_t\}$ are i.i.d and Y_t has Bernoulli distribution $B(1, p)$, $0 < p < 1$, then $Y_s^{(2)}$ will form a new Bernoulli process with mean $p(1 - p)$.*

The proof of Theorem 2 is given in Appendix B. Theorem 2 points that (i) if the mean of a Bernoulli process is close to 0.5, the mean of its second-layer process will be much different from 0.5; (ii) if a second-layer process has change points, so does its original generalize Bernoulli process; (iii) if a generalized Bernoulli process has a change point τ and the sum of the means

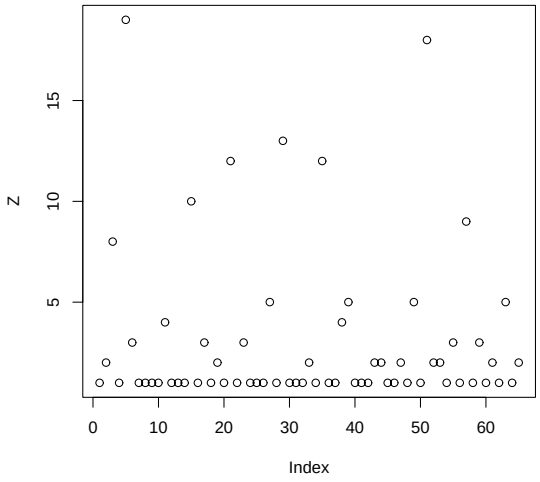

Figure 11: Plot of z_s, which is constructed from the sample of y_t given in Figure 10.

of the segments before τ and after τ is not equal to 1, then the time series of plot of its second-layer process should be able to reflect the mean changes in the generalized Bernoulli process.

Borrowing the notation of second-layer process, process $(Y_t^{(2)})^{(2)} \hat{=} Y_t^{(2^2)}$ denotes the second-layer process of $Y_t^{(2)}$. In general, $\{Y_t^{(2^k)}\}$ denotes the second-layer process of $\{Y_t^{(2^{k-1})}\}$, $k = 1, 2, \cdots$. Process $\{Y_t^{(2^k)}\}$, $k = 1, 2, \cdots$, are Bernoulli processes if its root process $\{Y_t\}$ is a Bernoulli process. The value of the mean of $\{Y_t^{(2^k)}\}$ exponentially decreases as k increases. The information of the root process can be quickly lost from $\{Y_t^{(2^k)}\}$ for $k \geq 2$. However, sometimes reasonably discarding some information of $\{Y_t\}$ might make structure changes in $\{Y_t\}$ to be visible through the time series plot of $\{Y_t^{(2^k)}\}$ for certain k, especially for $k = 2$. This can be seen from Example 8 below.

Example 8: Simulate a sample from a generalized Bernoulli process Y_t, where $Y_t \sim B(1, 0.4)$ for $t < 80$ and $Y_t \sim B(1, 0.5)$ for $80 \leq t \leq 150$. The plots of Y_t, $Y_t^{(2)}$ and $Y_t^{(4)}$ are presented in Fig.12.

By scrutinizing Fig. 12, one hardly judges whether there are any structure changes in $\{Y_t\}$ from the plot of $\{Y_t\}$. One might feel no enough information for making comments on the plot pattern changes in the plot of $\{Y_t^{(4)}\}$ either. However, the plot of $\{Y_t^{(2)}\}$ clearly demonstrates that at least a change point

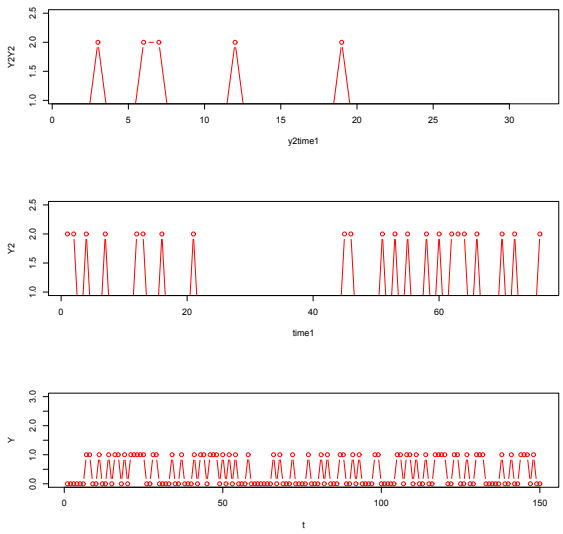

Figure 12: The plot of the root process $\{Y_t\}$ is in bottom. The middle one is given by $\{Y_t^{(2)}\}$ and the top plot is for $\{Y_t^{(4)}\}$.

of $\{Y_t^{(2)}\}$ is within or slightly beyond the interval $[25, 45]$.

Process $\{Y_t^{(2)}\}$ is used to record all 1's and 0's subsequences with length 2 in $\{Y_t\}$. Given a path of $\{Y_t\}$, each the starting positions of those 1s or 0s subsequences with length 2 will map to an unique subscription of $\{Y_t^{(2)}\}$. Therefore, if the plot of $\{Y_t^{(2)}\}$ shows structure changes and if a potential domain for the change is located from the domain of $\{Y_t^{(2)}\}$, due to the mapping between the subscripts of $\{Y_t^{(2)}\}$ and $\{Y_t\}$, a potential domain for the change point in $\{Y_t\}$ is able to be located (An program in R for mapping the subscripts of $\{Y_t^{(2)}\}$ into the subscripts of $\{Y_t\}$ is available from the author of this paper. For different sample of $\{Y_t\}$ the mapping between the subscripts of $\{Y_t^{(2)}\}$ and $\{Y_t\}$ is different.)

In Example 8, the plot patterns of $\{Y_t^{(2)}(\omega)\}$ before $t = 25$ and after $t = 50$ are different. Part of the mappings between the subscripts of $\{Y_t^{(2)}\}$ and $\{Y_t\}$ are listed in Table 2. The domain $(25, 50)$ for $\{Y_t^{(2)}(\omega)\}$ roughly corresponds to the domain $(44, 91)$ for $\{Y_t(\omega)\}$. Therefore, we may suspect that $\{Y_t(\omega)\}$ have a change point within $(44, 91)$. In fact, the true change point of $\{Y_t\}$ is $80 \in (44, 91)$.

Example 9: In this example, another independent sample was drawn from

Table 2: The corresponding between the subscripts of $Y_t^{(2)}(\omega)$ and $Y_t(\omega)$.

subscript of $Y_t^{(2)}$	subscript of Y_t	subscript of $Y_t^{(2)}$	subscript of Y_t
25	44	38	70
26	45	39	71
27	46	40	74
28	47	41	75
29	48	42	76
30	51	43	77
31	52	44	80
32	59	45	82
33	60	46	84
34	61	47	85
35	62	48	86
36	65	49	87
37	66	50	91

the same model used in Example 8. We compare the plots of a generalized Bernoulli process $\{Y_t\}$, its associate process, its second-layer process $\{Y_t^{(2)}\}$ and the associate process of $\{Y_t^{(2)}\}$. All the plots are given by Fig.13. We use this example to further show that sometime the associate process of a second-layer process may be helpful if structure changes in a generalized Bernoulli process is difficultly observed from the plots of its associate process as well as its second-layer process.

Fig.13 clearly demonstrates that the plot pattern changes in the plots of associate process of $Y_t^{(2)}$ and $Y_t^{(2)}$ itself are easy to be observed than those in the plots of associate process of Y_t and Y_t. Observing the plot of associate process $Y_t^{(2)}$, we consider interval $(30, 50)$ as a potential domain for the change point in $Y_t^{(2)}$. Based on the relationship between the subscripts of Y_t and $Y_t^{(2)}$ for this independent sample of Y_t, the domain $(30, 50)$ of $Y_t^{(2)}$ roughly corresponds to the domain $(57, 89)$ of Y_t. In fact, the true change point of Y_t is 80 and is covered by the potential domain $(57, 89)$.

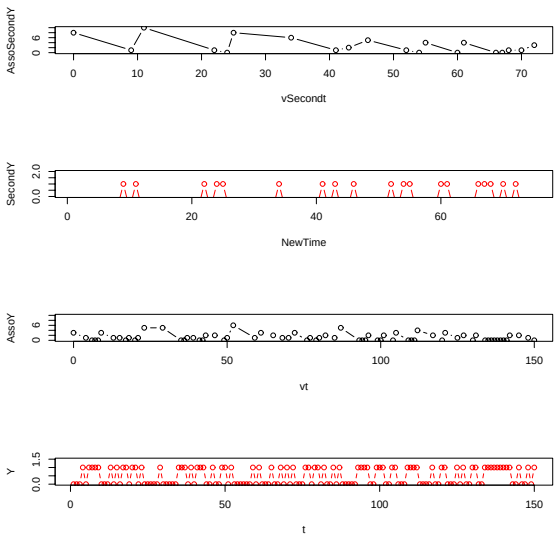

Figure 13: The plots of the original process (Y), its associate process (AssoY), the second-layer process (SecondY) and the associate process of the second-layer process (AssoSecondY).

5 Applications

In Section 2, we point out that the ML estimations of change points sometime are misleading. To improve the ML estimations of change points in a generalized Bernoulli process $\{Y_t\}$, two steps need to be involved.

Step 1: Identify potential domain for change points.

Step 2: Calculate the log likelihood function $L(k)$ for k within potential domains, then determine the ML estimations of change points.

In previous sections, we introduce how to construct associate processes and second-layer processes from generalized Bernoulli processes. Several examples have been demonstrated that potential domains of change points for generalized Bernoulli processes can be identify through observing the plots of their associate processes or second-layer processes.

Potential domain for a change point is not unique. Ideally, we would like to choose one with relatively shorter length of interval and at the same time it would be better to have the suspected change point roughly located at the center of the domain. Since the true position of the change point is unknown, the length of potential domain should not be too short. A decision on potential

domain is subject from person to person. However, No matter which potential domains is used, the ML estimation of change points will be improved.

The following examples demonstrate how to use the new method developed in this paper to improve the ML estimation of change points in Generalized Bernoulli processes.

Example 10. Use the same data in Example 4. The plot of logarithm likelihood function (see (1)), which is calculated based on the true values of $p = 0.2$ and $\delta = 0.1$, is given by the top plot in Fig. 14.

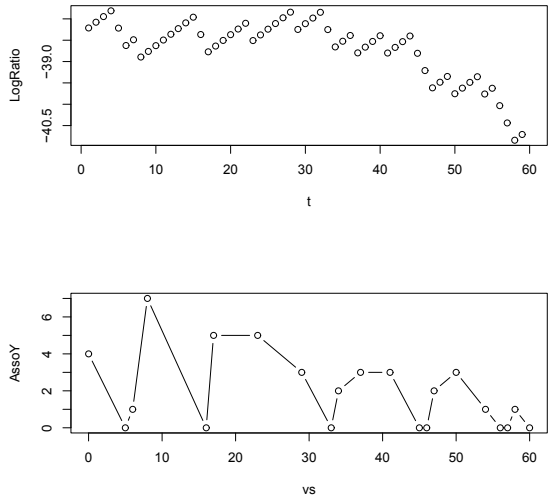

Figure 14: The scatter plot of the logarithm likelihood function and the plot of the associate process of Y_t, where $p = 0.2$, $\delta = 0.1$, $T = 60$, $\tau = 30$.

In the plot of the logarithm likelihood function, the ML estimation of the change point is less than 10.

Consider the plot of associate process of $\{Y_t\}$, which is the second plot in Fig. 14. There is a pattern change around position 30. We suggest $(20, 40)$ to be a potential domain of the change point for this sample, which in fact cover the true change point $\tau = 30$. After this potential domain is determined, we then calculate the log likelihood function for all $k \in (20, 40)$ and obtain the ML estimation of change point $\hat{\tau}$ such that $L(\hat{\tau}) = \max_{20 \le k \le 40}\{L(k)\}$. From the plot of log likelihood function, it clearly show the new ML estimation is more accurate than before.

Example 11 We consider the application of the new method to DNA

sequences.

Many different models have been used in DNA sequence segmental analysis (See Braun and Muller, 1998). Independent multinomial model is one of them and is adapted in this example.

Consider a base sequence of intron 7 of the chimpanzee α-fetoprotein gene (see Boys, Henderson and Wilkinson, 2000, or Fig.15) and conduct a generalized Bernoulli process Y_t from the base sequence based on G-C pairs: $Y_t = 1$ if the base is A or T; $Y_t = 0$ if the base is G or C.

Figure 15: Base sequence of intron 7 of the chimpanzee α-fetoprotein gene.

As an example, we consider a subsequence of $\{Y_t\}$ from base position 900 to 1140. Two change points in this subsequence has been reported from literature and suggested to be at bases 981 and 1072 (see Boys, Henderson and Wilkinson, 2000).

To check if there are any change points in this tested sequence, we firstly observe the plot of the associate process of the tested sequence. The plot of associate process is given by Fig.16. There are pattern changes in the plot and the plot clearly shows that there may have two change points in the tested sequence. Since the message obtained from the plot of associate process is such

clear, it is not necessarily to consider the second-layer process of the tested sequence.

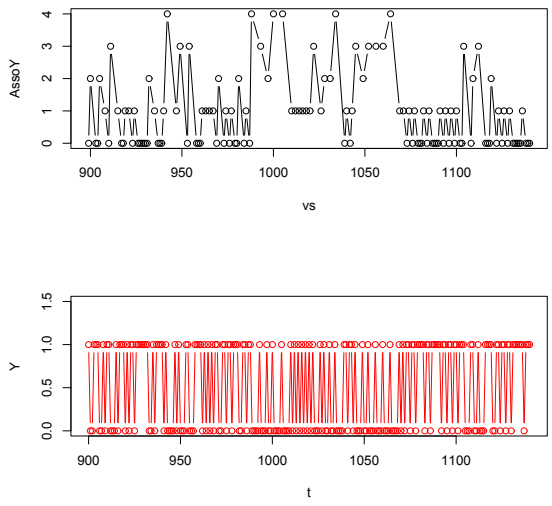

Figure 16: The plots of the subsequence and its associated process from position 900 to 1140.

Observing the plot in Fig. 16, we claim that the first change point may appear between 950 to 999 and the second change point may appear between 1050 and 1100. Thus, two potential domains for change points are determined. Given the two potential domains, the tested sequence is partitioned into five subsequences: $[900, 950]$, $[950, 999]$, $[999, 1050]$, $[1050, 1100]$ and $[1100, 1140]$. The plot indicates that the first and the last subsequences might have the same probability structure.

After the potential domains for change points are determined, the estimation of change points can be obtained by the ML method.

Base on our observation, we are able to believe that the process is stable between 900 to 950, 1000 to 1050 and 1100 to 1140. The means of the tested sequence in the three intervals $[900, 950]$, $[1000, 1050]$ and $[1100, 1140]$ can be evaluated by the data in each relevant interval respectively. They are $\hat{p}_1 = 0.5686275$ for sequence in $[900, 950]$, $\hat{p}_2 = 0.372549$ for $[1000, 1050]$ and $\hat{p}_3 = 0.6341463$ for $[1100, 1140]$.

Then we apply the ML method to the data in $[950, 999]$ by assuming that the means of segments before and after the change point within the potential

domain $[950, 999]$ are $\hat{p}_1$ and $\hat{p}_2$ respectively. Also apply the ML method to data in $[1050, 1100]$ by assuming that the means of segments before and after the change point within the potential domain $[1050, 1100]$ are $\hat{p}_2$ and $\hat{p}_3$ respectively; The plots of logarithm likelihood are given by Fig. 17. The ML estimations of the change points are 989 and 1070 respectively, which are very close to the estimations given by literature.

The new method works well for the above subsequence as well as for other subsequences of the base sequence of intron 7 of the chimpanzee α-fetoprotein gene. The applications are omitted.

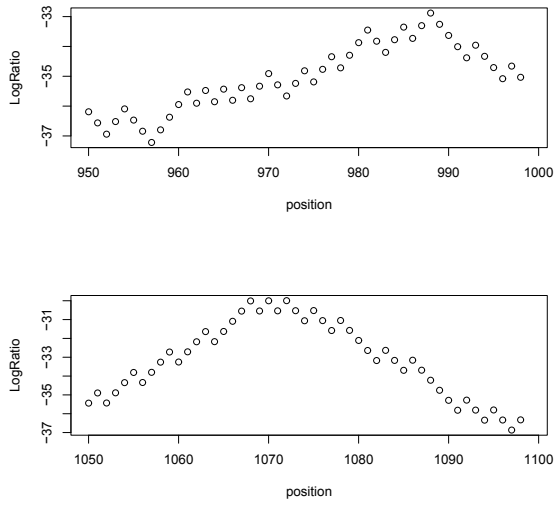

Figure 17: The first plot is the plot of logarithm of likelihood for $k \in [950, 999]$. The second plot is the plot of logarithm of likelihood for $k \in [1050, 1100]$.

To significantly improve the ML estimations of change points in generalized Bernoulli processes, information on potential domains for change points are necessary. Usually the locations of the potential domains are determined through observing the plot of tested processes. However, directly observing structure changes from the time series plots of generalized Bernoulli processes sometimes is not practicable. This paper develops a method to construct two types of processes from generalized Bernoulli processes and proves that the information on structure changes of a generalized Bernoulli process can be obtained from the plots of its associate process or second-layer process. Simulation studies and application to DNA sequence are presented. The method

discussed in this paper can be also applied to time series with categorical values, for example a sequence of opinion polls.

References

[1] J. V. Braum and H.-G. Muller (1998). Statistical methods for DNA sequence segmentation, *Statitical Science*, **13**, 142-162.

[2] R. J. Boys and D. A. Henderson (2004). A Bayesian approach to DNA sequence segmentation, *Biometrics*, **60**, 573-588.

[3] R.J. Boys, D. A. Henderson and D. J. Wilkinson (2000). Detecting homogeneous segments in DNA sequences by using hidden Markov models, *Appl. Statist.* , **49**, 269-285.

[4] Z. Drener and N. Farnum (1993). A generalized Binomial distribution, *Communications in Statistics - Theory Meth.*, **22**, 3051-3063.

[5] Y.-X. Fu and R. N. Curnow (1990). Maximum likelihood estimation of multiple change points. *Biometrika*, **77**, 563-573.

[6] H.-G. Muller and K.-S. Song (1997). Two-stage change-point estimators in smooth regression models, Statistics & Probability Letters, **34**. 323-335.

[7] D. V. Hinkley and E. A. Hinkley (1970). Inference about the change-point in a sequence of binomal variables, Biometrika, **57**, 477-488.

Appendix A

Theorem 1: Let $\{Y_t\}_{t\geq 1}$ be a Bernoulli process with $EY_1 = p$, and $Y_0 = 1$. Then, W_{v_s}, $s = 1, 2, \cdots$, are i.i.d., having geometric distribution with parameter p.

Proof: Firstly, we prove that W_{v_s} has geometric distribution with parameter p for each s. Since $\{Y_t\}_{t\geq 1}$ are i.i.d. and $Y_0 = 1$, following the definition of W_{v_s}, for any integer $k \geq 0$, we have

$$P(W_{v_s} = k) = \sum_{k_1=1}^{\infty} \cdots \sum_{k_{s-1}=1}^{\infty} [P(Y_1 = 0 = \cdots = Y_{k_1-1}, Y_{k_1} = 1,$$

$$Y_{k_1+1} = 0 = \cdots = Y_{k_1+k_2-1}, Y_{k_1+k_2} = 1, \cdots,$$

$$Y_{\sum_{i=1}^{s-2} k_i+1} = 0 = \cdots = Y_{\sum_{i=1}^{s-1} k_i-1}, Y_{\sum_{i=1}^{s-1} k_i} = 1,$$
$$Y_{\sum_{i=1}^{s-1} k_i+1} = 0 = \cdots = Y_{\sum_{i=1}^{s-1} k_i+k}, Y_{\sum_{i=1}^{s-1} k_i+k+1} = 1)]$$
$$= (1-p)^k p^s \sum_{k_1=1}^{\infty} \cdots \sum_{k_{s-1}=1}^{\infty} [(1-p)^{k_1-1} \cdots (1-p)^{k_{s-1}-1}]$$
$$= (1-p)^k p^s (1/p)^{s-1} = p(1-p)^k$$

as required. Then, we prove that W_{v_s} is independent of W_{v_t} for any $s \neq t$. Without loss of generality, assume $s < t$. For any nonnegative integers l_s and l_t,

$$P(W_{v_s} = l_s, W_{v_t} = l_t)$$
$$= \sum_{k_1=1}^{\infty} \cdots \sum_{k_{s-1}=1}^{\infty} \sum_{k_{s+1}=1}^{\infty} \cdots \sum_{k_{t-1}=1}^{\infty} p^t (1-p)^{k_1-1} \cdots (1-p)^{k_{s-1}-1} (1-p)^{l_s} \times$$
$$(1-p)^{k_{s+1}-1} \cdots (1-p)^{k_t-1} (1-p)^{l_t}$$
$$= p^t (1-p)^{l_s} (1-p)^{l_t} \frac{1}{p^{t-2}} = p(1-p)^{l_s} p(1-p)^{l_t}$$
$$= P(W_{v_s} = l_s) P(W_{v_t} = l_t).$$

Therefore, $\{W_{v_s}\}$ are i.i.d., as required. □

Appendix B

Theorem 2: If Y_t is a Bernoulli process, i.e. $\{Y_t\}$ are i.i.d and Y_t has Bernoulli distribution $B(1,p)$, $0 < p < 1$, then $Y_s^{(2)}$ will form a new Bernoulli process with mean $p(1-p)$.

Proof: To prove $Y_s^{(2)}$ is a Bernoulli process with mean $p(1-p)$, we need to prove that

(1) $Y_s^{(2)}$ has Bernoulli distribution with mean $p(1-p)$ for all s.

(2) $\{Y_s^{(2)}\}$ are independent.

Firstly, we prove $Y_s^{(2)} \sim B(1, p(1-p))$ for all s.

Obviously,

$$P(Y_1^{(2)} = 1) = P(Z_1^{(2)} = 2) = P(Z_1 = 2) = P(Y_1 = Y_2 = 1, Y_3 = 0)$$
$$+P(Y_1 = Y_2 = 0, Y_3 = 1) = p^2(1-p) + (1-p)^2 p = p(1-p).$$

For any $k \geq 1$,

$$P(Y_{2k}^{(2)} = 1) = P(Z_{2k}^{(2)} = 2) = P(Z_{2k} = 2, Y_1 = 1) + P(Z_{2k} = 2, Y_1 = 0)$$

$$= \sum_{r_1=1}^{\infty} \sum_{r_2=1}^{\infty} \cdots \sum_{r_{2k-1}=1}^{\infty} P(Z_1 = r_1, Z_2 = r_2, \cdots, Z_{2k-1} = r_{2k-1}, Z_{2k} = 2, Y_1 = 1) +$$

$$\sum_{r_1=1}^{\infty} \sum_{r_2=1}^{\infty} \cdots \sum_{r_{2k-1}=1}^{\infty} P(Z_1 = r_1, Z_2 = r_2, \cdots, Z_{2k-1} = r_{2k-1}, Z_{2k} = 2, Y_1 = 0)$$

$$= \sum_{r_1=1}^{\infty} p^{r_1} \sum_{r_2=1}^{\infty} (1-p)^{r_2} \cdots \sum_{r_{2k-1}=1}^{\infty} p^{r_{2k-1}} (1-p)^2 p +$$

$$\sum_{r_1=1}^{\infty} (1-p)^{r_1} \sum_{r_2=1}^{\infty} p^{r_2} \cdots \sum_{r_{2k-1}=1}^{\infty} (1-p)^{r_{2k-1}} p^2 (1-p)$$

$$= (1-p)^2 p^2/(1-p) + p^2(1-p)^2/p = (1-p)p^2 + p(1-p)^2 = p(1-p).$$

Similarly, we are able to show that, for any $k \geq 1$,

$$P(Y_{2k+1}^{(2)} = 1) = P(Z_{2k+1}^{(2)} = 2) = P(Z_{2k+1} = 2, Y_1 = 1) + P(Z_{2k+1} = 2, Y_1 = 0)$$

$$= p^2(1-p) + (1-p)^2 p = p(1-p).$$

Therefore, $Y_s^{(2)}$, $s \geq 1$, has Bernoulli distribution with mean $p(1-p)$.

Now we prove that $\{Y_s^{(2)}\}$ are independent. We only need to prove that for any $1 \leq i_1 < i_2 < \cdots < i_k$;

$$P(Y_{i_1}^{(2)} = 1, \cdots, Y_{i_k}^{(2)} = 1) = P(Y_{i_1}^{(2)} = 1) \cdots P(Y_{i_k}^{(2)} = 1). \tag{2}$$

Without loss of generality, we only prove

$$P(Y_{2L}^{(2)} = 1, Y_{2M}^{(2)} = 1) = P(Y_{2L}^{(2)} = 1) P(Y_{2M}^{(2)} = 1).$$

for any $1 \leq L < M$.

After calculation, we have

$$P(Y_{2L}^{(2)} = 1, Y_{2M}^{(2)} = 1)$$

$$= P(Y_{2L}^{(2)} = 1, Y_{2M}^{(2)} = 1, Y_1 = 1) + P(Y_{2L}^{(2)} = 1, Y_{2M}^{(2)} = 1, Y_1 = 0)$$

$$= (1-p)^2 p^3 + (1-p)^3 p^2 = p^2(1-p)^2 = P(Y_{2L}^{(2)} = 1) P(Y_{2M}^{(2)} = 1).$$

as required. □