Информационные процессы, Том 19, № 2, 2019, стр. 170–177

Letter to the editors: Statistical simulation of Change Point detection

M. Malyutov

Mathematics Dept., Northeastern University, 360 Huntington Ave., Boston, MA 02115, USA
E-mail: m.malioutov@neu.edu

Received 05.05.2019

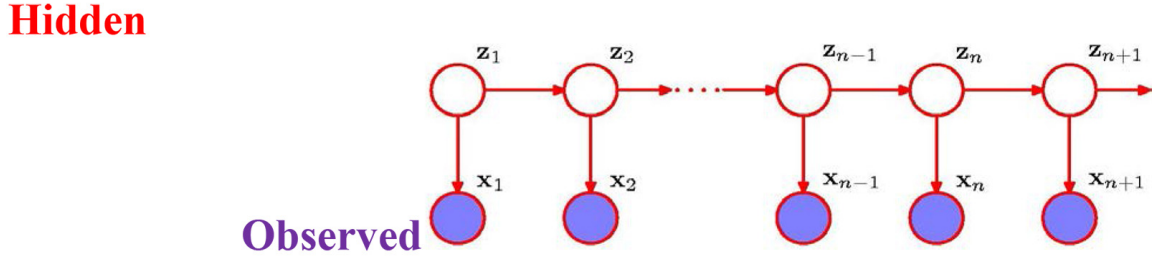
Abstract—My paper IP, **19:1** 2019 contained the penultimate version of Appendix 2, Statistical simulation of Change Point detection, instead of the last version of Appendix 2 submitted on December 13, 2018. Below is the version of December 13 extended with the case of estimating parameters of the same SCOT emissions model.

KEYWORDS: Change Point detection online and offline, quadratic risk.

1. INTRODUCTION

Stochastic COntext Tree (abbreviated as SCOT) is a m-Markov Chain (m-MC) with every state of a string independent of the symbols in its more remote past than the **context** of **length** determined by the preceding symbols of this state. A parallel super-fast fitting and asymptotically optimal inference in a **sparse** SCOT model including the nonparametric homogeneity test are described in our previous papers. [1, 4] and finally [2] established the equivalence of a **perfect memory sparse** SCOT to 1-MC with state space consisting of the collection of m-MC contexts which we consider as the new **alphabet** $\mathcal{A}$ of cardinality A. For not perfect memory sparse SCOT, its perfect memory sparse *envelope* (also studied in [2]) plays this role.

The evaluation of log-likelihoods under SCOT requires **sophisticated software** and **cumbersome calculations**. Thanks to the above SCOT perfect memory reduction to a 1-MC with enlarged state space, its statistical theory ([3]) simplifies. **Statistical simulation** simplifies, if the SCOT memory structure is established up to probabilities involved. This is done in our last section for the HMM's 'Change Point' (change of state z) detection. Observed random variables x_i **depend only on current hidden state** z_i modeled as a Markov Chain. If the SCOT memory structure is unknown, then we would need to apply the methodology of [3, 4] in full strength.



Plot 1. HMM-SCOT scheme

We model all regimes as long SCOT strings. We call HMM z_i **SLOW**, if the mean time that HMM keeps staying in the same state is proportional to a large parameter l in all states, while the **sample size** is $kl, k \to \infty$. Emissions shown in dark in the above Fig. are modeled as **strings of MC over the space of contexts**, transition matrix depending on the current HMM state. The emissions x_i over the alphabet of SCOT contexts are assumed **ergodic, different for all states of HMM**, expectations are taken everywhere under their stationary distributions. Our segmentation method is a combination of preliminary online change point (CP) detection with its subsequent offline Maximal Likelihood update.

2. CP DETECTION FOR SCOT MODELS KNOWN BOTH BEFORE AND AFTER CP

Simulation was made by PhD student Jiewei Feng.

i) Let z_t indicate the state of Hidden Markov Model with two states 0 or 1 at time t.

The transition probability of this Hidden Markov Model is

$$P(z_1 = 0) = 1, P(z_t = 1|z_{t-1} = 0) = 0.001, P(z_t = 1|z_{t-1} = 1) = 1.$$

So the Hidden Markov Model starts in state 0 and then has one change point after some time (the state changes from 0 to 1) and will never go back to state 0 again.

Denote CP as the least t such that $z_t = 1$. Then

$$P(CP = i) = 0.999^{i-1} * 0.001 \text{ for } i > 1.$$

ii) Define SCOT under state 0 of HMM (model 2ii in [1], p. 86): Let x_t act under the rule of 'increasing SCOT' if $z_t = 0$, that is

$$x_0 = -1, x_1 = 0.$$

If $x_{t-1} = -l$ where $-l$ is the left boundary, then

$$x_t = -l + 1.$$

If $x_{t-1} = l$ where l is the right boundary (we assume l is large enough such that we will not reach the right boundary in limited time), then

$$x_t = l - 1.$$

If for the greatest $k < t$ such that $x_k \neq x_{k-1}$, we have $x_k = x_{k-1} + 1$ and $x_{t-1} \neq 0$, then

$$x_n = \begin{cases} x_{t-1} + 1 & \text{with probability } 0.8, \\ x_{t-1} & \text{with probability } 0.1, \\ x_{t-1} - 1 & \text{with probability } 0.1. \end{cases}$$

If for the greatest $k < t$ such that $x_k \neq x_{k-1}$, we have $x_k = x_{k-1} - 1$ and $x_{t-1} \neq 0$, then

$$x_t = \begin{cases} x_{t-1} - 1 & \text{with probability } 0.8, \\ x_{t-1} & \text{with probability } 0.1, \\ x_{t-1} + 1 & \text{with probability } 0.1. \end{cases}$$

iii) Define 'decreasing SCOT' under state 1 in Hidden Markov Model:

Use the same model as 2 ii) with probabilities $(0.4, 0.3, 0.3)$ (in the 'increasing SCOT' probabilities are $(0.8, 0.1, 0.1)$).

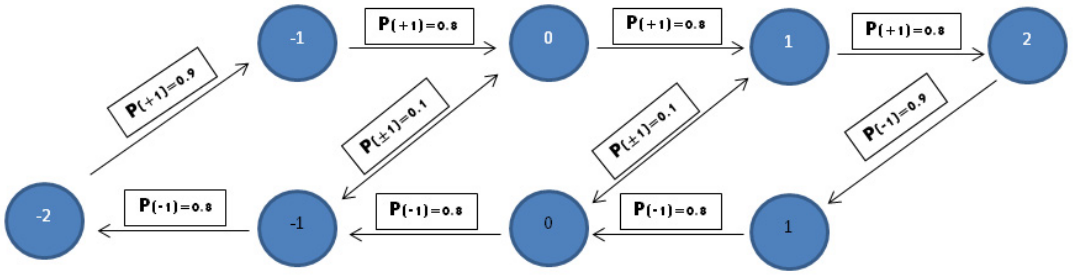


Fig. 2. Model 2(ii) with $l = 5$.

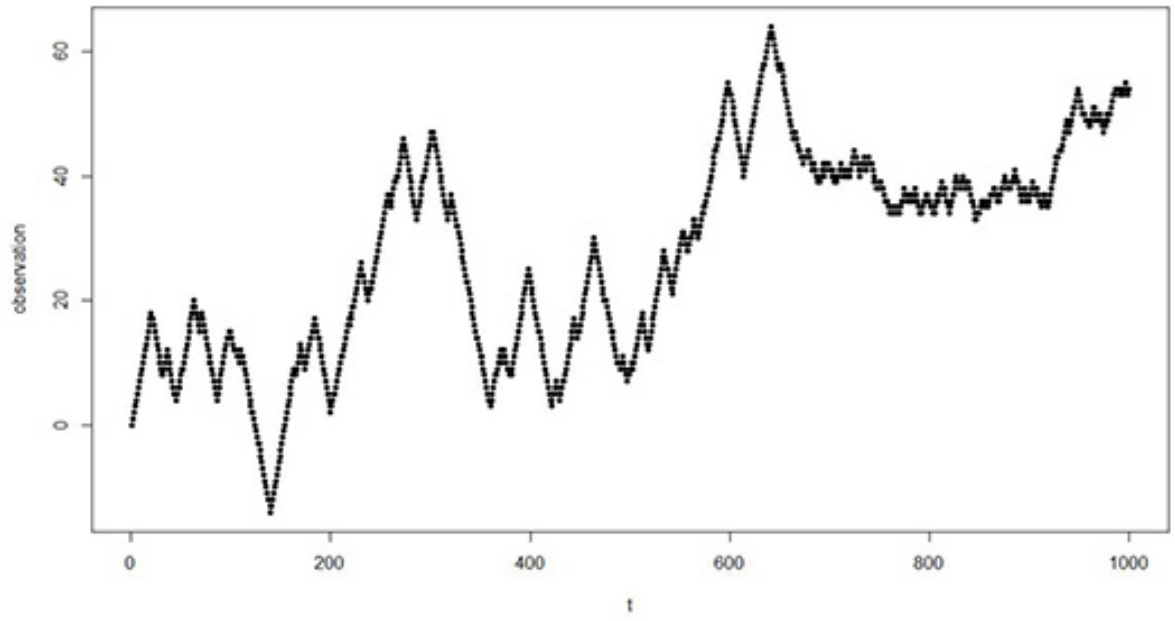


Fig 3. Simulated data.

We simulate HMM z_t and SCOT emissions, detect the CP online and offline following the algorithm of [3] on the simulated data set and find how close our estimates are to the real CP. In this simulation, the sample size is 1000 (i.e. $t = 1, ..., 1000$).

The next picture shows the generated observations x_t. The actual change point is 662.

Online change point detection:

a) The probability of getting $x_1, x_2, ..., x_t$:

$$P(x_1, x_2, ..., x_t|CP = i) = \prod_{n=3}^{t} P(x_n|x_{n-1}, CP = i),$$

$$P(x_1, x_2, ..., x_t) = \sum_{i=2}^{t} P(x_1, x_2, ..., x_t|CP = i)P(CP = i) + P(x_1, x_2, ..., x_t|CP > t)P(CP > t).$$

Then the log-likelihood of $x_1, x_2, ..., x_t$ is $l_t = \log P(x_1, x_2, ..., x_t)$.

b) Average log-likelihood:

Choose the window size of 10 points, then the average log-likelihood from the window $(x_t, \dots x_{t+9})$

is

$$(\bar{l_t}) = 1/10 \sum_{k=0}^{9} l_{(t+k)}.$$

c) Calculating the trend L_t of the data:
We use the Least Squares estimate for model $l_k = ak + b$ with data points $l_1, l_2, ..., l_t$, then $L_t = a$.
d) getting critical point:
Let C denote the critical point, by using the first hundred data points, we have first 100 $L_t's$.
Define

$$V(t) = |(\bar{l_t}) - Lt|. \tag{1}$$

Then let

$$C = 1.2 \max_{51 \le t \le 150} V(t). \tag{2}$$

e) Online change point estimate:
The estimator is the least t such that $|(\bar{l_t}) - Lt| > C$.
Simulation result

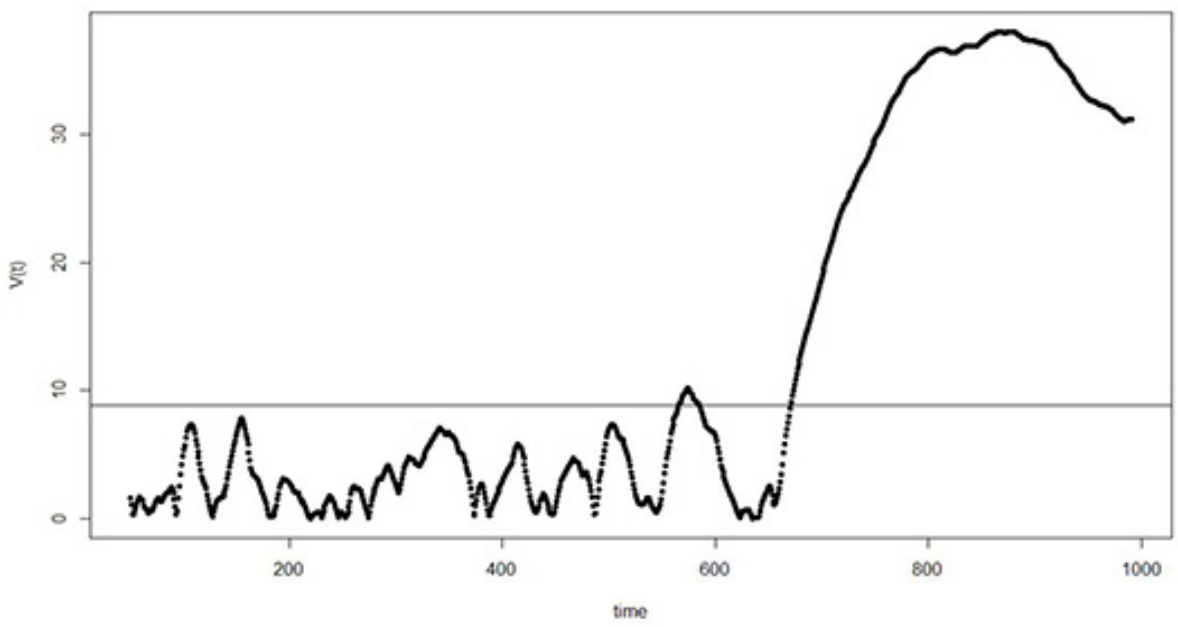


Fig 4. Online detection.

The above picture describes the statistic $V(t)$ as time t changes (refer to equation (1)). The horizontal line indicates the critical value(refer to equation(2), in this case 8.801). We find that the first point that exceeds this critical value is at $t = 566$ where the real change point is 662.

Offline change point detection:

we use maximum likelihood estimator (MLE) as our estimate.
Define

$$L(\theta) = L(x_1, ..., x_{1000}; \theta) = P(x_1, ..., x_{1000} | CP = \theta), \tag{3}$$

where $P(x_1, ..., x_{1000}|CP = \theta)$ can be deduced from similar procedure introduced in online change point detection. Then the MLE θ_0 is

$$\theta_0 = \arg \max_{1 \le \theta \le 1000} L(\theta).$$

Simulation result:

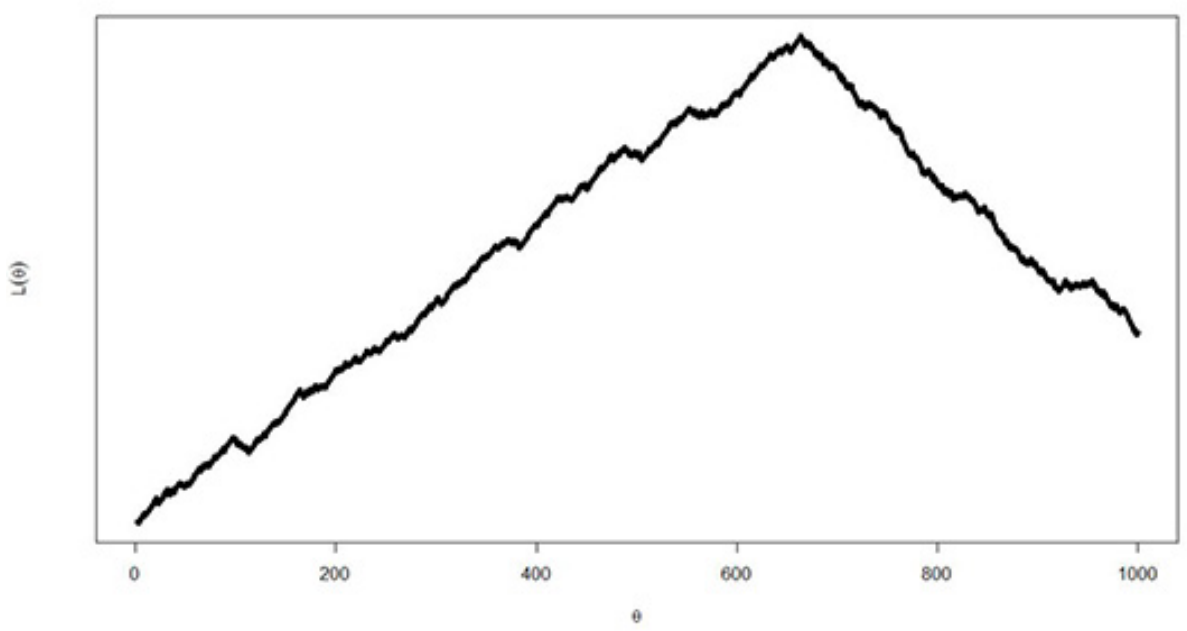


Fig. 5. Offline detection via ML method.

The above picture gives a plot of the likelihood with respect to all possible values of change points (i.e. $L(y_1, ..., y_{1000}; \theta)$) (refer to equation (3)). At $t = 664$ this likelihood achieves its maximum, thus it is our MLE. It is very close to the actual change point 662.

3. SIMULATION UNDER ESTIMATED PARAMETERS OF MODEL 2II

Consider the CP detection with unknown parameters of the model 2.ii. We still generate data using the same parameters (0.8, 0.1, 0.1) and (0.4, 0.3, 0.3) to first estimate these parameters by simulated data, then use this result to repeat the procedure introduced in the last model and find the change point.

The distribution of SCOT without CP (i.e. under permanent state of HMM) becomes:

$$x_0 = -1, x_1 = 0.$$

If $x_{t-1} = -l$ where $-l$ is the left boundary, then

$$x_t = -l + 1.$$

If $x_{t-1} = l$ where l is the right boundary (we assume l is large enough such that we will not reach the right boundary in limited time), then

$$x_t = l - 1.$$

If for the greatest $k < t$ such that $x_k \neq x_{k-1}$, we have $x_k = x_{k-1} + 1$ and $x_{t-1} \neq 0$, then

$$x_n = \begin{cases} x_{t-1} + 1 & \text{with probability } p, \\ x_{t-1} & \text{with probability } q, \\ x_{t-1} - 1 & \text{with probability } 1 - p - q. \end{cases}$$

If for the greatest $k < t$ such that $x_k \neq x_{k-1}$, we have $x_k = x_{k-1} - 1$ and $x_{t-1} \neq 0$, then

$$x_t = \begin{cases} x_{t-1} - 1 & \text{with probability } p, \\ x_{t-1} & \text{with probability } q, \\ x_{t-1} + 1 & \text{with probability } 1 - p - q. \end{cases}$$

In this process, we assume boundary l is sufficiently large so that x_i would not reach the boundary.

To estimate q, define sequence of random variables $w_t := x_{t+1} - x_t$, then $w_t \in \{-1, 0, 1\}$. Take $\{x_0, x_1, x_2, x_3, ..., x_{101}\}$ as our training data, we get the sequence $\{w_0 = 1, w_1, w_2, ...w_{100}\}$. Let n_1 be the number of times that $w_t = 0$, then our MLE for q is $\frac{n_1}{100}$ (similar to parameter estimation for the Bernoulli distribution)

Using the same sequence $\{w_0 = 1, w_1, w_2, ...w_{100}\}$, we delete all the data for which $w_t = 0$, and get a sub-sequence $\{w'_0 = 1, w'_1, w'_2, ...w'_{100-n_1}\}$. Further, introduce sequence $u_t := w'_t - w'_{t-1}$. We have $u_t \in \{-2, 0, 2\}$. Let n_2 be the number of times such that $u_t = 0$, then our MLE for q is $\frac{n_2}{100}$.

The way to understand the procedure is the following : q is the probability that x_t stays at its location , p gives the probability that x_t keep moving in the same direction as its most recent movement. And $1 - p - q$ gives the probability that it moves in the opposite direction as its most recent movement. These choices are independent. So the frequency of specific choice (stay, same direction, or opposite) is the maximum likelihood estimator of the parameter. The construction of sequence $\{w_t\}$ and $\{u_t\}$ extract the information of choices from the original data.

Example of the sequence described above:

$$\begin{gathered} \{x_0, x_1, x_2, x_3, ..., x_{101}\} = \{-1, 0, 0, 1, 0, 0, 1, 2, 3, 2, 1, ..., z_{101}\} \\ \downarrow \\ \{w_0 = 1, w_1, w_2, ..., w_{100}\} = \{1, 0, 1, -1, 0, 1, 1, 1, -1, -1, ..., w_{100}\} \\ \downarrow \\ \{w'_0 = 1, w'_1, w'_2, ..., w'_{100-n_1}\} = \{1, 1, -1, 1, 1, 1, -1, -1, ..., w'_{100-n_1}\} \\ \downarrow \\ \{u_1, u_2, u_3, ..., u_{100-n_1}\} = \{0, -2, 2, 0, 0, -2, 0, ..., u_{100-n_1}\} \end{gathered}$$

By assuming the change point only happens after considerable time (slow HMM!), we can take the first 200 data points from the simulated x-string used for change point detection to train the parameter under $z_t = 0$. Similarly we can take the last 200 data points in the sequence to estimate parameters under $z_t = 1$.

Note that in the second case, the sequence we have under $z_t = 1$ is

$$\{z_{k-n-1}, z_{k-n}, z_{k-n+1}, z_{k-n+2}, z_{k-n+3}, ..., z_k\},$$

where k is the size of the whole sequence used for change point detection and n is the number of data points for training the SCOT. Hence unless we know $z_0 = -1$ and $z_1 = 0$ which give us the initial direction, if the sequence begins with the condition $z_{k-n-1} = z_{k-n}$, we lose the information

about the initial direction. Because of this, we need to find when the sequence first moves in either of directions before time $k-n$.

We temporarily define $z_{k-n-1} := z_{k-n} - 1$ if the sequence first moves in either of directions before time $k-n$ is increasing, and $z_{k-n-1} := z_{k-n} + 1$ if the opposite happens. Then using the same procedure as discussed before we can estimate the parameters.

3.1. Simulation of CP detection under estimated SCOT parameters

By creating 200 data points for each of the SCOT before and after the change point with parameters (0.8, 0.1, 0.1) and (0.4, 0.3, 0.3), we get the following estimates: (0.805, 0.095, 0.1) and (0.415, 0.285, 0.3).

We use the above estimation to find the change point in the same sequence we used before and get the following result.

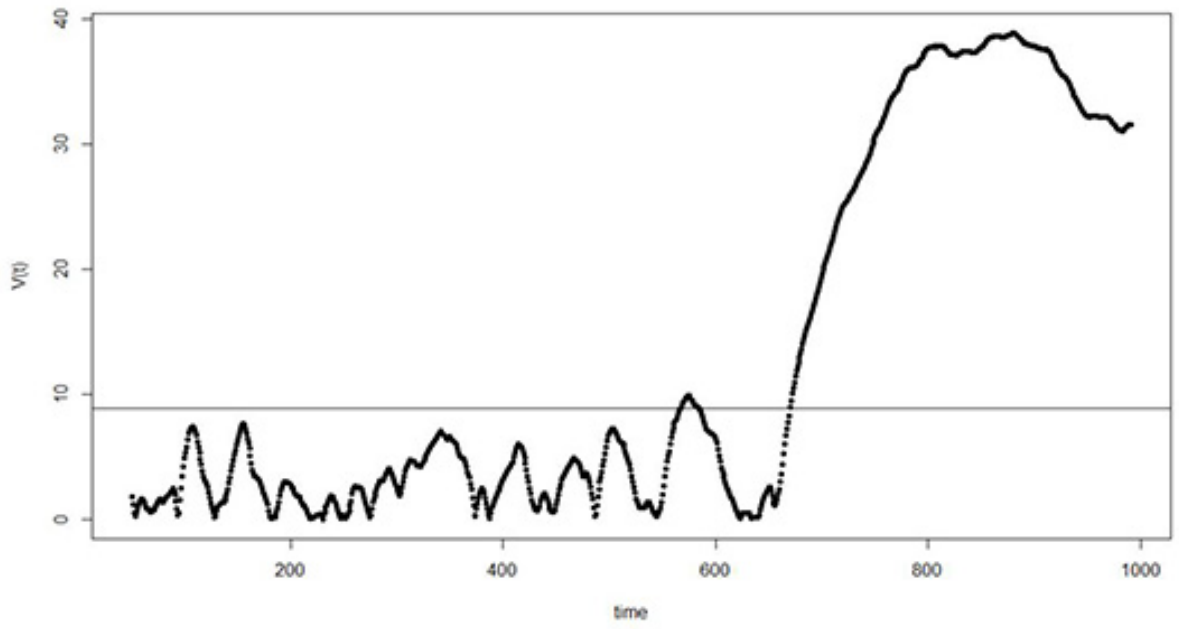


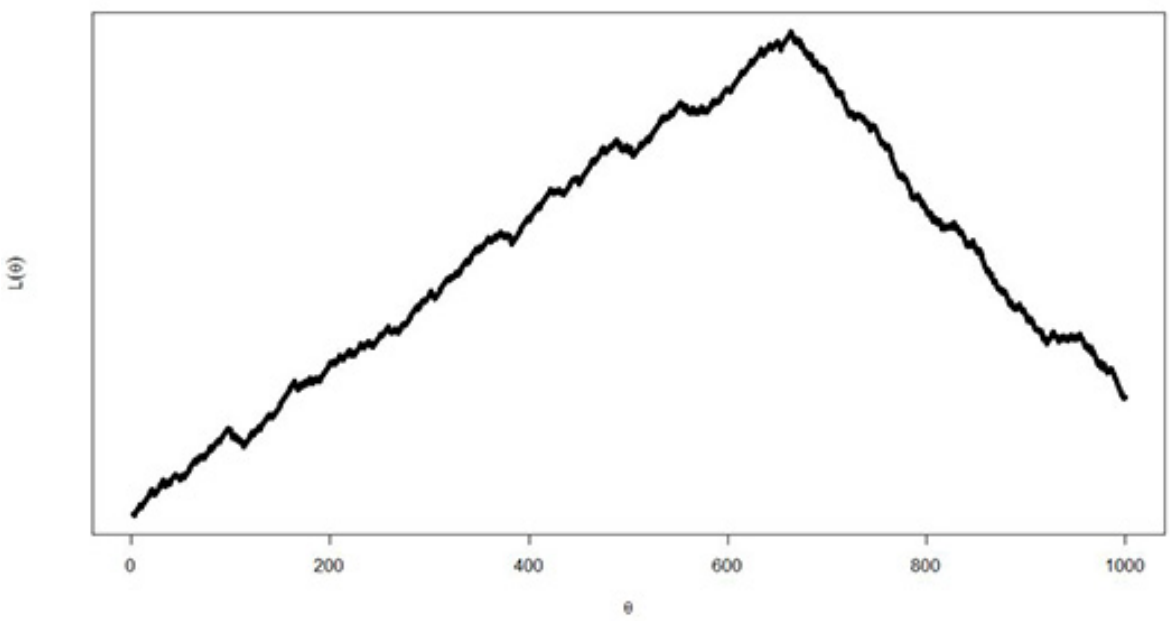
Fig. 6. Online detection with estimated parameters

In online change point detection, the above picture describes the statistic $V(t)$ as time t change (refer to equation (1)). The horizontal line indicates the critical value(refer to equation(2), in this case 8.8321)) .We find that the first point that exceeds this critical value is at $t = 566$ where the real change point is 662 (recall that our online CP-estimate under the true parameter was 568).

In offline change point detection, the above picture gives a plot of likelihood with respect to different values of change points (i.e. $L(y_1, ..., y_{1000}; \theta)$) (refer to equation (3)). At $t = 664$ this likelihood achieves its maximum, thus it is our MLE. It is very close to the actual change point 662. (Recall that our offline estimation using true parameter gives the same CP estimate 664).

REFERENCES

1. Ryabko B., Astola J. and Malyutov M. *Compression-Based Methods of Prediction and Statistical analysis of Time Series: Theory and Appllications*, Springer International, 2016.



Plot 7. Offline detection with estimated parameters.

2. Zhang T. Perfect Memory Context Trees in time series modeling, *Information Processes*, 2017, vol. 17, no. 1, pp. 70–81.

3. Malyutov M. B. Quadratic risk of Change Point detection, *Information Processes*, 2019, vol. 17, no. 1, pp. 60–77.

4. Malyutov M. and Grosu P. SCOT approximation, modeling and training. In: *Proceedings of Machine Learning Research (PMLR)*, June 2017, vol. 60, pp. 241–265.