2022 届本科生学士学位论文　　学校代码：10269

華東師範大學

East China Normal University

本科生毕业论文

微生物浓度监测方法

Monitor Method of Microorganism Concentration

姓　　名：刘正达

学　　号：10195000499

学　　院：统计学院

专　　业：统计学

指导教师：项冬冬

职　　称：教授

2022 年 4 月

华东师范大学学位论文诚信承诺

本毕业论文是本人在导师指导下独立完成的，内容真实、可靠。本人在撰写毕业论文过程中不存在请人代写、抄袭或者剽窃他人作品、伪造或者篡改数据以及其他学位论文作假行为。

本人清楚知道学位论文作假行为将会导致行为人受到不授予/撤销学位、开除学籍等处理（处分）决定。本人如果被查证在撰写本毕业论文过程中存在学位论文作假行为，愿意接受学校依法作出的处理（处分）决定。

承诺人签名：　　日期：　　年　　月　　日

华东师范大学学位论文使用授权说明

本论文的研究成果归华东师范大学所有，本论文的研究内容不得以其它单位的名义发表。本学位论文作者和指导教师完全了解华东师范大学有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅；本人授权华东师范大学可以将论文的全部或部分内容编入有关数据库进行检索、交流，可以采用影印、缩印或其他复制手段保存论文和汇编本学位论文。

保密的毕业论文（设计）在解密后应遵守此规定。

作者签名：　　导师签名：　　日期：　　年　　月　　日

Contents

1 Problem Formulation

In the paper, we collect the data of the concentrations of some different kinds of microorganism at multiple depths in the water. Our goal is to identify critical points at which the concentration of one microorganism confronts a huge change. Such points are called change points. Our procedure contains an initial screening approach using likelihoods to find candidates of potential change points and a further sift procedure to determine our final choices of change points by control charts.

In the paper studying the trends of data in ecology, the stochastic differential equation (SDE) is commonly used to model a series of data. Suppose there is an index set I, I is usually chosen as time or other geological measures like depths. The series of data we care of is $\{X(t), t \in I\}$. We say $\{X(t), t \in I\}$ follows a SDE model of parameters $\boldsymbol{\theta} = (\alpha, s, r, n, \sigma) \in \Theta$ if

$$dX(t) = \left[\alpha - sX(t) + r\frac{X(t)^n}{X(t)^n + 1}\right] dt + \sigma X(t) dW, \tag{1}$$

where W represents Wiener process. $\Theta = \{(\alpha, s, r, n, \sigma) : \alpha \in \mathbb{R}, s \in \mathbb{R}, r \in \mathbb{R}, n \in \mathbb{Z}^+, \sigma \in \mathbb{R}^+\}$ is the parameter space. In the future, we write $X(t) \sim \text{SDE}(\boldsymbol{\theta})$ in short. Figure 1 shows a simulated trace of $X(t) \sim \text{SDE}(\boldsymbol{\theta})$ with $\alpha = 0.1, s = 1, r = 1, n = 8, \sigma = 0.25$.

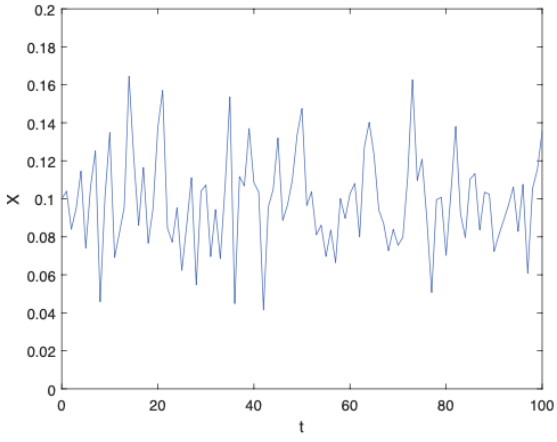


Figure 1: Trace of a SDE process

Next, we will specify our change point model. Suppose we have collected $X(t_0), X(t_1), \cdots, X(t_T)$ and there are k change points in total. They are denoted as $t_{r_1}, t_{r_2}, \cdots, t_{r_k}$ respectively.

Then the data can be modeled as

$$\begin{aligned} &X(t) \sim \text{SDE}(\boldsymbol{\theta}_0), \quad t = t_0, \cdots, t_{r_1}, \\ &X(t) \sim \text{SDE}(\boldsymbol{\theta}_1), \quad t = t_{r_1+1}, \cdots, t_{r_2}, \\ &\vdots \\ &X(t) \sim \text{SDE}(\boldsymbol{\theta}_k), \quad t = t_{r_k+1}, \cdots, t_T, \end{aligned} \tag{2}$$

where $\boldsymbol{\theta}_i \neq \boldsymbol{\theta}_j, i \neq j$. That's to say the series are cut into $k+1$ different fragments with respect to various parameters of SDE process. Our goal is to detect these k change points and get their estimates $\hat{t}_{r_1}, \cdots, \hat{t}_{r_k}$.

2 Change Point Detection Procedure

2.1 Parameters Estimation in SDE Model

To implement our change point detection procedure, an necessary tool is the estimations of parameters of the SDE model given observations. A frequently used method in statistics is maximum likelihood estimation (MLE). However, it is not easy to derive the oracle likelihood function of a SDE model directly due to the differential parts. To solve the problem, a discretized version of (1) is used to get a approximate likelihood function. Given observations $\boldsymbol{X} = (X(t_0), X(t_1), \cdots, X(t_T))' \sim \text{SDE}(\boldsymbol{\theta})$, (1) can be discretized as

$$X(t_i + \Delta t_i) = X(t_i) + \left[\alpha - sX(t_i) + r\frac{X(t_i)^n}{X(t_i)^n + 1}\right]\Delta t_i + \sigma X(t_i) Z\sqrt{\Delta t_i}, \tag{3}$$

where $\Delta t_i = t_{i+1} - t_i$, $i = 0, 1, \cdots, T-1$, and Z is a standard normal random variable. As a result, the condition distribution of $X(t_{i+1})$ given $X(t_i)$ is

$$X(t_{i+1})|X(t_i), \boldsymbol{\theta} \sim N\left(X(t_i) + \left[\alpha - sX(t_i) + r\frac{X(t_i)^n}{X(t_i)^n + 1}\right]\Delta t_i, X(t_i)^2\sigma^2\Delta t_i\right).$$

The corresponding density function is denoted as $f_{\boldsymbol{\theta}}(X(t_{i+1})|X(t_i))$. The approximate likelihood function using all $T+1$ observations is given as

$$\ell(\boldsymbol{\theta}; \boldsymbol{X}(0:T)) = \sum_{i=0}^{T-1} \log(f_{\boldsymbol{\theta}}(X(t_{i+1})|X(t_i)). \tag{4}$$

The estimations of parameters are $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} l(X;\boldsymbol{\theta})$. Note that $\boldsymbol{X}(l_1 : l_2), 0 \leq l_1 < l_2 \leq T$ is a slice of given vector $\boldsymbol{X}$.

2.2 The Initial Screening of Change Points

2.2.1 General change points detection

We first derive a general framework for change points detection of SDE modeled data. The basic idea is maximum likelihood as well. Suppose we known the number of change points in $\boldsymbol{X} = (X(t_0), X(t_1), \cdots, X(t_T))'$ is k in advance, and the data can be cut into $k+1$ segments according to different $\boldsymbol{\theta}$s. For convenience, define $L_k = \{\boldsymbol{l} = (l_1, \cdots, l_k) : l_i \in \{1, \cdots, T-1\}, l_i < l_j \text{ for } i < j\}$, the set contains locations of all possible k change points, $\Gamma_k = \{\boldsymbol{\gamma} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_k) : \boldsymbol{\theta}_i \in \Theta, i = 0, \cdots, k\}$, the set includes all possible parameters of $k+1$ segments. Given $\boldsymbol{l} \in L_k$define the likelihood function of $\boldsymbol{X}$ as

$$\begin{aligned} Q(\boldsymbol{l}; \boldsymbol{X}) &= \max_{\boldsymbol{\gamma}\in\Gamma_k} \left(\sum_{j=0}^{k} \sum_{i=l_j}^{l_{j+1}-1} \log f_{\boldsymbol{\theta}_j}(X(t_{i+1})|X(t_i)) \right) \\ &= \sum_{j=0}^{k} \max_{\boldsymbol{\theta}_j\in\Theta} \ell(\boldsymbol{\theta}_j; \boldsymbol{X}(l_j : l_{j+1})), \end{aligned} \tag{5}$$

where $l_0 = 0 l_{k+1} = T$. Let

$$\hat{\boldsymbol{r}} = (\hat{r}_1, \cdots, \hat{r}_k) = \arg\max_{\boldsymbol{l}\in L_k} Q(\boldsymbol{l}; \boldsymbol{X}). \tag{6}$$

Then $(t_{\hat{r}_1}, \cdots, t_{\hat{t}_k})$ are possible change points.

In reality, the MLE of n is not accurate due to severe instability. As a result, the values of n are limited in $\{1, 2, \cdots, 10\}$. Let $\tilde{\Theta}_{n_0} = \{(\alpha, s, r, n_0, \sigma) : \alpha \in \mathbb{R}, s \in \mathbb{R}, r \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$ be the parameter space given $n = n_0$. Define

$$\tilde{Q}_{n_0}(\boldsymbol{l}; \boldsymbol{X}) = \sum_{j=0}^{k} \max_{\boldsymbol{\theta}_j\in\tilde{\Theta}_{n_0}} \ell(\boldsymbol{\theta}_j; \boldsymbol{X}(l_j : l_{j+1})), \tag{7}$$

$$\tilde{\boldsymbol{r}}_{n_0} = \arg\max_{\boldsymbol{l}\in L_k} \tilde{Q}_{n_0}(\boldsymbol{l}; \boldsymbol{X}). \tag{8}$$

Then Algorithm 1 can be used to find change points.

Algorithm 1: Theoretical change points detection

Input: number of change points k, data $\boldsymbol{X}$
Output: estimated change points $\{t_{\hat{r}_1}, \cdots, t_{\hat{t}_k}\}$
for *n=1:10* **do**
  Use (8) to obtain $\tilde{\boldsymbol{r}}_n$, and compute $\tilde{Q}_n(\tilde{\boldsymbol{r}}_n; \boldsymbol{X})$;
end
Let $n_0 = \arg\max_{n=1,2,\cdots,10} \tilde{Q}_n(\tilde{\boldsymbol{r}}_n; \boldsymbol{X})$;
Compute $\hat{\boldsymbol{r}} = (\hat{r}_1, \cdots, \hat{r}_k) = \arg\max_{\boldsymbol{l}\in L} \tilde{Q}_{n_0}(\boldsymbol{l}; \boldsymbol{X})$

There are two main problems of Algorithm 1. First, you have to know the number of change points k in advance to implement the procedure. However, k is usually unknown in reality. Second, if k is quite large, to find the optimal combination of change points locations through (8) is computationally tough. Though we can use certain selective algorithms to alleviate the burden, the results often turn out to be unsatisfactory. To solve the problems, we derive the screening and sifting procedure in the next section.

2.2.2 Screening Procedure

The main idea of screening and sifting procedure is that we first screen the whole series using likelihoods to initially find suspected change points, and then sift them by control charts. After the two steps, we can determine the change points in the data without an oracle k and heavily computational burden.

The screening procedure is a modified version of Algorithm 1. Instead of considering multiple change points directly, in this procedure we suppose $k = 1$. Given data $\boldsymbol{X} = (X(t_0), X(t_1), \cdots, X(t_T))'$, compute

$$\tilde{Q}_{n_0}(l; \boldsymbol{X}) = \max_{\boldsymbol{\theta}_1\in\tilde{\Theta}_{n_0}} \ell(\boldsymbol{\theta}_1; \boldsymbol{X}(0:l)) + \max_{\boldsymbol{\theta}_2\in\tilde{\Theta}_{n_0}} \ell(\boldsymbol{\theta}_2; \boldsymbol{X}(l+1:T)), \tag{9}$$

where $l = 2, \cdots, T-2$. (9) is the special case of (5) when $k = 1$. The first and last two observations are dropped out because there would not be enough data to accurately estimate the likelihoods.

Our goal is to make a rough decision that which points are possible change points based on the trace plot of $\tilde{Q}_{n_0}$. Intuitively speaking, if there is only one change point t_r, the trace plot of $\tilde{Q}_{n_0}$ should reach its peak at $l = t_r$. If there exist multiple change point, denoted as $t_{r_1}, \cdots, t_{r_k}$, some severe changes of $\tilde{Q}_{n_0}$ should occur at these points, such as a sudden rise or drop or being a local maximum. In practice, we give five decision rules for the initial screening of suspected change points.

(I) The point of maximum value of $\tilde{Q}_{n_0}(l; \boldsymbol{X})$, $l = 2, \cdots, T-2$;

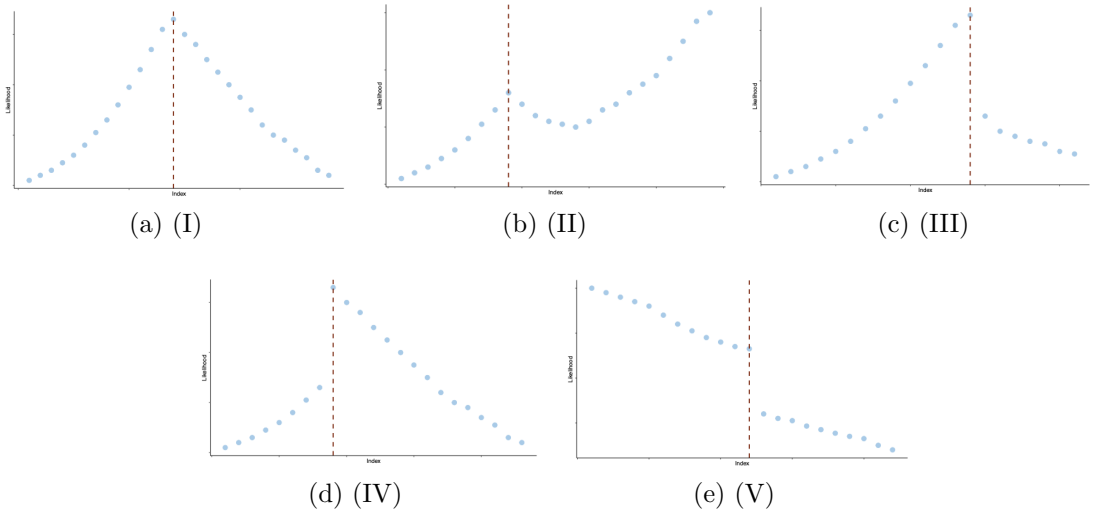
(II) The local maximum point, shaped like a peak, two sides are relatively continuous;

(III) The point that locates at the end of a continuous increasing process, followed by a steep drop;

(IV) The point that locates at a steep rise, followed by a continuous decreasing process or a platform;

(V) The point that locates at the end of a continuous slow dropping process, followed by a steep dive and then the likelihood values start to drop continuously again.

Examples of these five rules are shown in Figure 2.



(a) (I) (b) (II) (c) (III)

(d) (IV) (e) (V)

Figure 2: Examples of the five decision rules for the initial screening of suspected change points

It is appropriate to assume that the change points are sparse, or it is hard to apply above rules. As a result, if the distance between two suspected change points is less than 5, only one of them is taken into consideration.

What's more, under the scenario of multiple change points, several kinds of patterns among the five rules may appear in the trace plot likelihoods of data $\boldsymbol{X}$. Due to randomness of the data, such patterns might exist where no changes are going through. We call such points outliers or flickering points. However, it is possible to identify such points because such patterns of outliers are less obvious than those of real change points. As a result, it is reasonable to discard those trivial patterns and focus on truly apparent patterns in rules (I)-(V).

Another thing worthy to mention is that the tails of the trace plot of likelihoods are less credible due to the poor parameter estimation since the data left at the beginning and ending of the data series are few. If patterns of rules (I)-(V) appear at the tails, we can optionally ignore them if they are less pivotal compared with the patterns located at the main domain of the trace plot.

Then the screening procedure is summarized in Algorithm 2.

Algorithm 2: Screening procedure for change points detection

Input: data $\boldsymbol{X}$
Output: roughly estimated change points $\{t_{\hat{r}_1}, \cdots, t_{\hat{t}_k}\}$

for *n=1:10* **do**
  Obtain $\tilde{r}_n = \arg\max_{l \in L_1} \tilde{Q}_n(l; \boldsymbol{X})$;
  Compute $\tilde{Q}_n(\tilde{r}_n; \boldsymbol{X})$ through (9);
end
Let $n_0 = \arg\max_{n=1,2,\cdots,10} \tilde{Q}_n(\tilde{r}_n; \boldsymbol{X})$;
Construct the trace plot of $\tilde{Q}_{n_0}(l; \boldsymbol{X}),\ l = 2, \cdots, T-2$;
Use the decision rules (I)-(V) to do initial screening of suspected change points and get $\{t_{\hat{r}_1}, \cdots, t_{\hat{t}_k}\}$.

2.3 The Sifting Procedure of Change Points

After the initial screening of suspected change points, further sifting procedure is necessary to take out fake change points. A useful tool is the control charts in statistic process control, which are used to judge whether a process has experienced a dramatic change, or

being out of control. Specially for a parameterized process, the dramatic change means the change of certain parameter, which is of our interest. For one suspected change point, we can use the data before the point to estimate the parameters of a SDE model and construct control chart based on the estimations. If this point is actually a change point, the control charts should raise an alarm. It is of great importance to note that the start point of the control chart of a suspected change point is the suspected change point right before the current point.

Firstly, we introduce the construction of a control chart. Given data $\boldsymbol{X}$, consider a change point t_{r_0}. We use the EWMA control charts to monitor the changes of mean and variance of the process. First, get the MLE of the SDE parameters using the data before t_{r_0}, that is $(\alpha_0, s_0, r_0, \sigma_0) = \arg\max_{\boldsymbol{\theta}\in\tilde{\Theta}_{n_0}} \ell(\boldsymbol{\theta}; \boldsymbol{X}(0, t_{r_0}))$. Define

$$Z_{r_0}(t_i) = \frac{X(t_i) - \left[X(t_{i-1}) + \left(\alpha_0 - s_0 X(t_{i-1}) + r_0 \frac{X(t_{i-1})^{n_0}}{X(t_{i-1})^{n_0}+1}\right)\right]\Delta t_{i-1}}{X(t_{i-1})\sigma_0\sqrt{\Delta t_{i-1}}};$$

$$W_{r_0}(t_i) = \frac{\sqrt{|Z_{r_0}(t_i)|} - 0.822}{0.349},$$

where $i = 1, 2, \cdots, T$. Then the monitor statistics for mean and variance monitoring are defined as

$$\begin{aligned} C_{r_0}(t_i) &= \lambda Z_{r_0}(t_i) + (1-\lambda)C_{r_0}(t_{i-1}),\ C_{r_0}(t_0) = 0; \\ V_{r_0}(t_i) &= \lambda W_{r_0}(t_i) + (1-\lambda)V_{r_0}(t_{i-1}),\ V_{r_0}(t_0) = 0. \end{aligned} \tag{10}$$

The control limits for the EWMA control charts are

$$\begin{cases} U = \rho\sqrt{\frac{\lambda}{2-\lambda}} \\ C = 0 \\ L = -\rho\sqrt{\frac{\lambda}{2-\lambda}} \end{cases}. \tag{11}$$

Here we choose $\lambda = 0.1$, and $\mathrm{ARL}_0 = 370\rho = 2.701$.

Given a series of suspected change points $\{t_{\hat{r}_1}, \cdots, t_{\hat{t}_k}\}$, the main idea for the sifting procedure is to construct control charts for each of them one by one. There are three possible situations for the control charts of a suspected change point $t_{\hat{r}_j}$. Firstly, if the

control chart do not raise an alarm at all or it raises an alarm after $t_{\hat{r}_{j+1}}$, the next suspected change point, we make the decision that it is not a real change point. It is worthy to note that the second case is included because we can't tell whether the alarm is caused by $t_{\hat{r}_j}$ or $t_{\hat{r}_{j+1}}$. Secondly, if the control chart raises an alarm before $t_{\hat{r}_{j+1}}$ but after $t_{\hat{r}_j}$, $t_{\hat{r}_j}$ is determined as a change point. Thirdly, if the control chart raises an alarm even before $t_{\hat{r}_j}$, the point $t_{\hat{r}_j}$ are regarded as a flickering point. The term 'flickering' means that the point may locate at an unstable phase, for example, a gradual change process between two different statuses.

With the tools of control charts, we can derive the whole procedure of change point detection under SDE model in the following Algorithm 3.

Algorithm 3: Screening and sifting procedure for change points detection

Input: data $\boldsymbol{X}$

Output: estimated change points $\{t_{\hat{r}_1}, \cdots, t_{\hat{t}'_k}\}$

Apply Algorithm 2 and get roughly estimated change points $\{t_{\hat{r}_1}, \cdots, t_{\hat{t}_k}\}$;

for *j=1:k* **do**

  Estimate the parameters $(\alpha_j, s_j, r_j, \sigma_j) = \arg\max_{\boldsymbol{\theta}\in\bar{\Theta}_{n_0}} \ell(\boldsymbol{\theta}; \boldsymbol{X}(t_{\hat{r}_{j-1}} : t_{\hat{r}_j}))$. (Let $t_{\hat{r}_0} = t_0$); use (10) and (11) to construct control charts for $t_{\hat{r}_j}$;

  if *the control chart raises no alarm or alerts after* $t_{\hat{r}_{j+1}}$ **then**

    $t_{\hat{r}_j}$ is not a valid change point

  else if *the control chart raises an alarm before* $t_{\hat{r}_{j+1}}$ *but after* $t_{\hat{r}_j}$ **then**

    $t_{\hat{r}_j}$ is a valid change point

  else

    $t_{\hat{r}_j}$ is a flickernig point;

  end

end

Get the final decision of change points $\{t_{\hat{r}_1}, \cdots, t_{\hat{t}'_k}\}$.

3 Simulation Study

This section is aimed to verify the rationality of our screening and sifting procedure and the superiority of our methods compared with several existing monitoring methods based on simple statistics that are usually used in relevant ecology papers.

In the papers discussed such questions in the filed before, they hardly directly use SDE model to establish methods for data monitoring or change point detection. One naive

method is to use simple descriptive statistics, such as standard deviation, lag-1 autocorrelation, kurtosis, skewness and so on, to monitor the change of target data series. The monitoring statistics are the value of these descriptive statistics based on a moving window of data with window size h. When you slide the window, a series of monitoring statistics are received to reflect the status of the data at a certain point. Intuitively speaking, when such methods detect a change point, the mean, variance or the trend of the monitoring statistics should experience a dramatic change. Unfortunately, there doesn't exist a credible rule to identify those changes. As a result, we can only rely on our eyes when using these descriptive based methods.

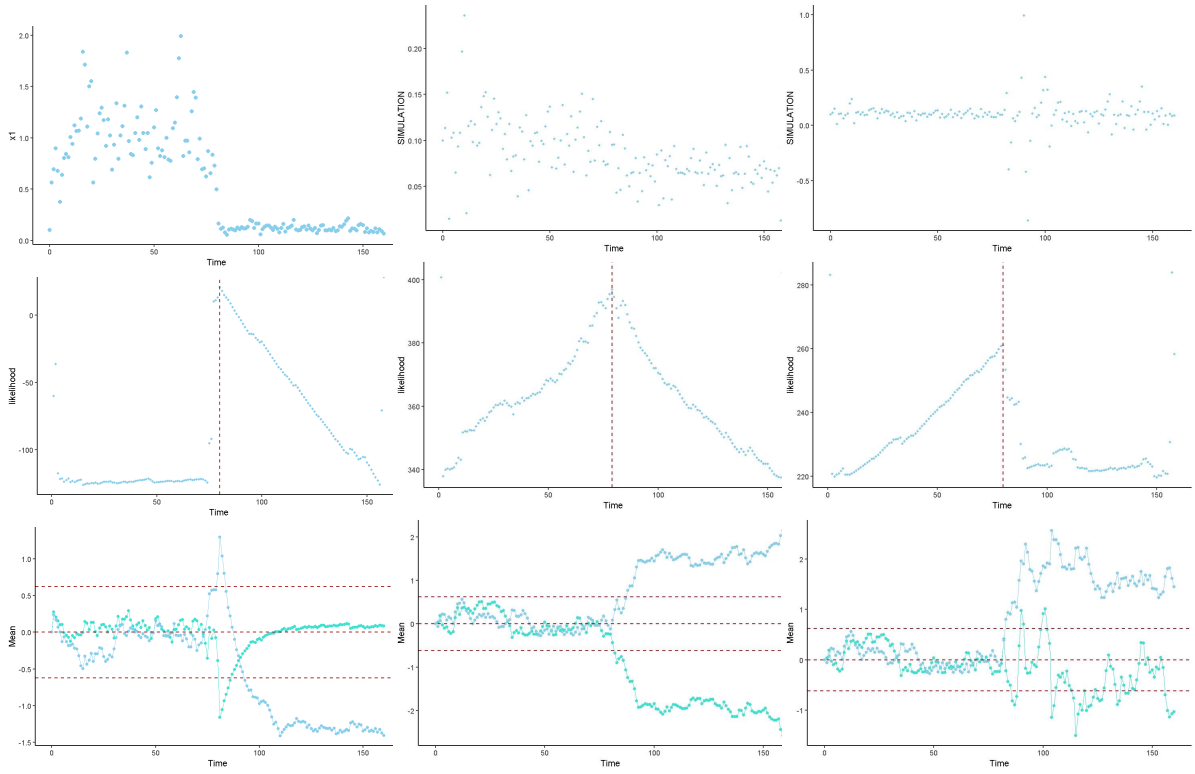
3.1 One Change Point Simulation

This section studies the performance of our screening and sifting method on the data that only exists one change point.

We generate 160 observations in total and the change point is located at 80. The change of parameters α, s and σ are studied respectively. The results are shown in Figure 3. In (a), before 80, we generate data with SDE parameters $(\alpha, s, r, \sigma, n) = (0.5, 1, 1, 0.25, 8)$. After 80, the parameters of SDE model are changed to be $(\alpha, s, r, \sigma, n) = (0.1, 1, 1, 0.25, 8)$.In (b) and (c), we generate data with SDE parameters $(\alpha, s, r, \sigma, n) = (0.1, 1, 1, 0.25, 8)$ before 80 and $(\alpha, s, r, \sigma, n) = (0.1, 1.5, 1, 0.25, 8)$, $(\alpha, s, r, \sigma, n) = (0.1, 1, 1, 0.75, 8)$ after 80 to show the impact of s and σ respectively.

In Figure 3, each column shows the scatter plot of data series, trace plot of likelihood and the control charts based on the suspected change point of one type of change. From Figure 3, we can see that our methods accurately detect the change point located at 80. In (a), the trace plot shows that the suspected change point is at 80 according to the decision rule (I) or (IV). And the corresponding control chart raises an alarm after 80. Similarly, the same conclusions can be drawn from (b) and (c) on the basis of rules (I) and (III). Note that in (b), there is a steep increase of likelihoods at the beginning of the trace plot. This is because the lack of data at the tails of the likelihoods. In (c), some flickers appear at around point 100. This is caused by high variance of the model.

The other thing we are interested in is the comparison between our methods and other

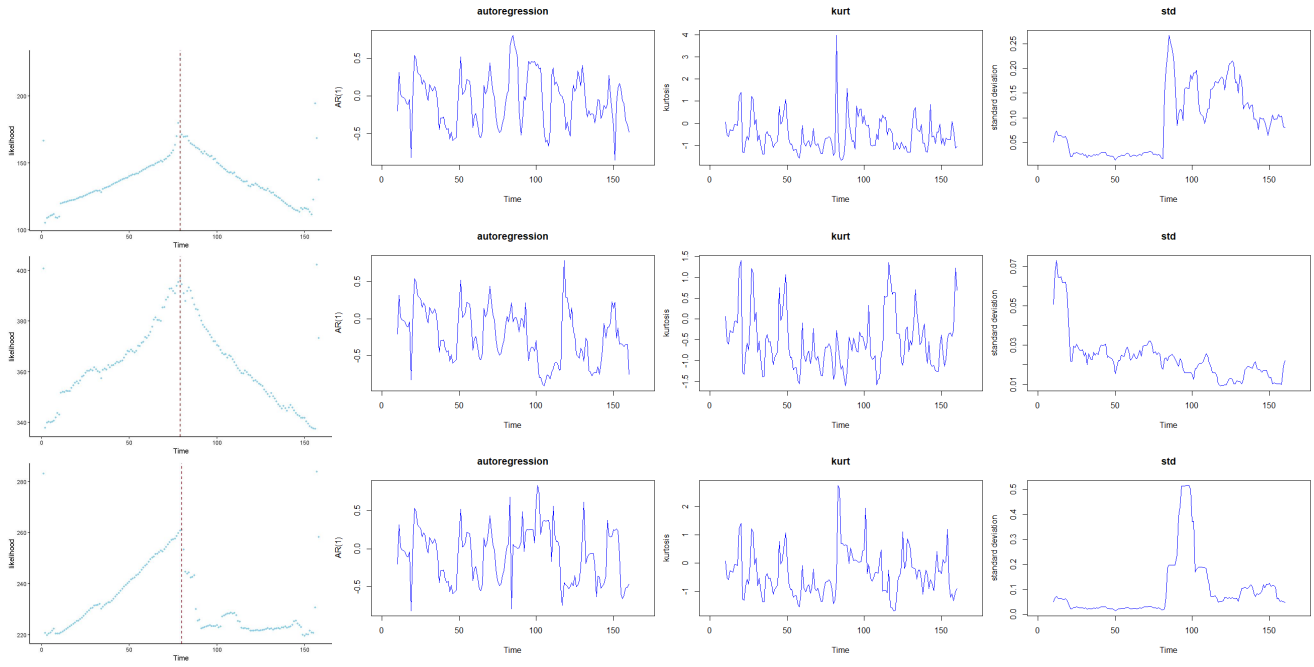


(a) Change α (b) Change s (c) Change σ

Figure 3: One Change Point Scenario

existing methods. For other methods that use descriptive statistics and moving window, we choose $h = 10$. For convenience, we only compare the trace plot of likelihood in our method with the trace plot of the descriptive statistics computed from the moving window in other methods, since the likelihood $\tilde{Q}_{n_0}(l; \boldsymbol{X})$ plays similar role in our method as in descriptive statistics based methods. In Figure 4, lag-1 autocorrelation, kurtosis and standard deviation based moving window methods are chosen as competitors against our method. Each row shows the results of four different methods applied in three types of change in one change point scenario.

From Figure 4, we can see that our method has the most obvious sign of change points even without the help of control charts. On the other hands, lag-1 autocorrelation has no significant sign of a change point, such as a steep change of mean or variance of the statistics. Although kurtosis and standard deviation based methods has signs of a change point in some scenarios, they are weak compared with our methods.



(a) Our method (b) Lag-1 autocorrelation (c) Kurtosis (d) Standard deviation

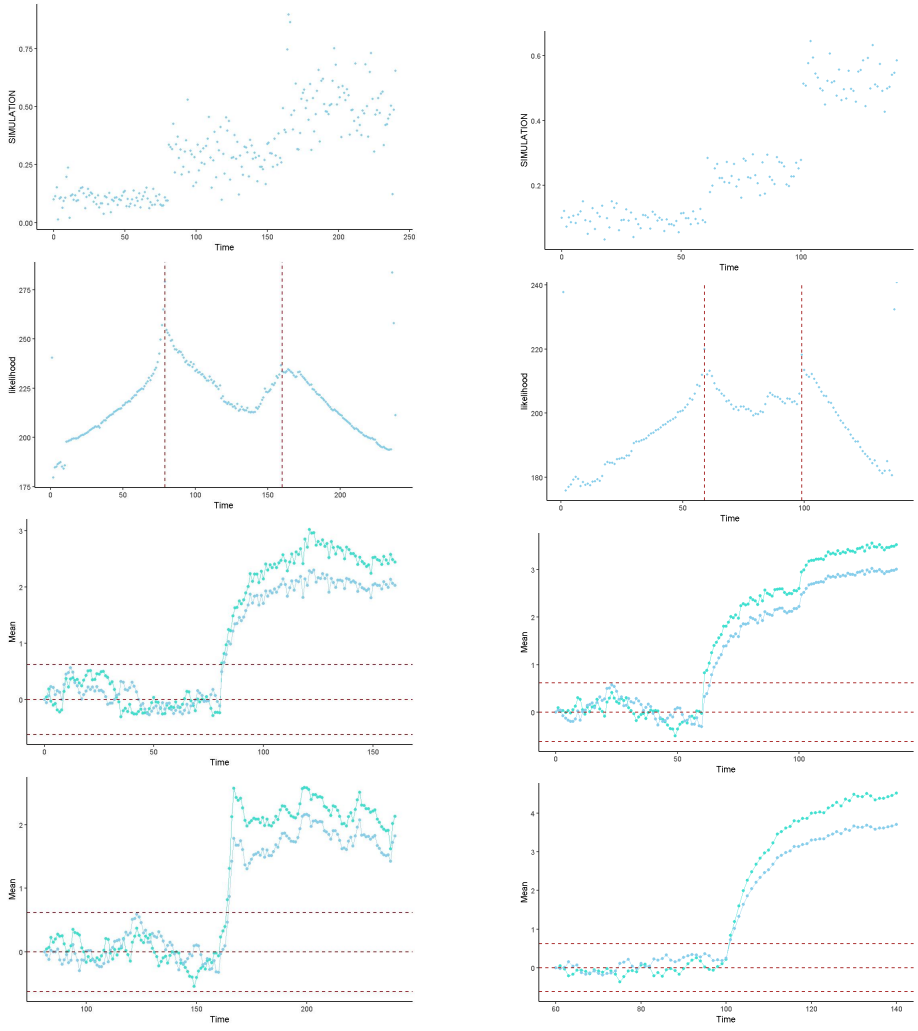
Figure 4: Comparisons of methods in one change point scenario

3.2 Two Change Points Simulation

This section considers the performance of our screening and sifting method on the data that exists two change points.

We consider the situation that the two changes are both caused by the change of α as well as the situation that the changes are caused by multiple parameters. In the first case, we generate 240 observations in total and the two changes points are set at 80 and 160. The parameters of the three pieces cut by the two change points are $(\alpha, s, r, \sigma, n) = (0.1, 1, 1, 0.25, 8)$, $(\alpha, s, r, \sigma, n) = (0.3, 1, 1, 0.25, 8)$ and $(\alpha, s, r, \sigma, n) = (0.5, 1, 1, 0.25, 8)$ respectively. In the second case, we generate 140 observations and the change points are set at 60 and 100. The parameters of the three pieces are $(\alpha, s, r, \sigma, n) = (0.1, 1, 1, 0.25, 8)$, $(\alpha, s, r, \sigma, n) = (0.3, 1, 1, 0.5, 8)$ and $(\alpha, s, r, \sigma, n) = (0.5, 1, 1, 0.75, 8)$ respectively. The results are shown in Figure 5.

From (a) in Figure 5, the trace plot of likelihoods shows that two suspected change points may exist at 80 and 160 according to rules (I) and (II). The control chart of suspected change point 80 raises an alarm after 80 and before 160, and the control chart of point 160 also alerts after 160, which shows that the two suspected change points are valid. Although



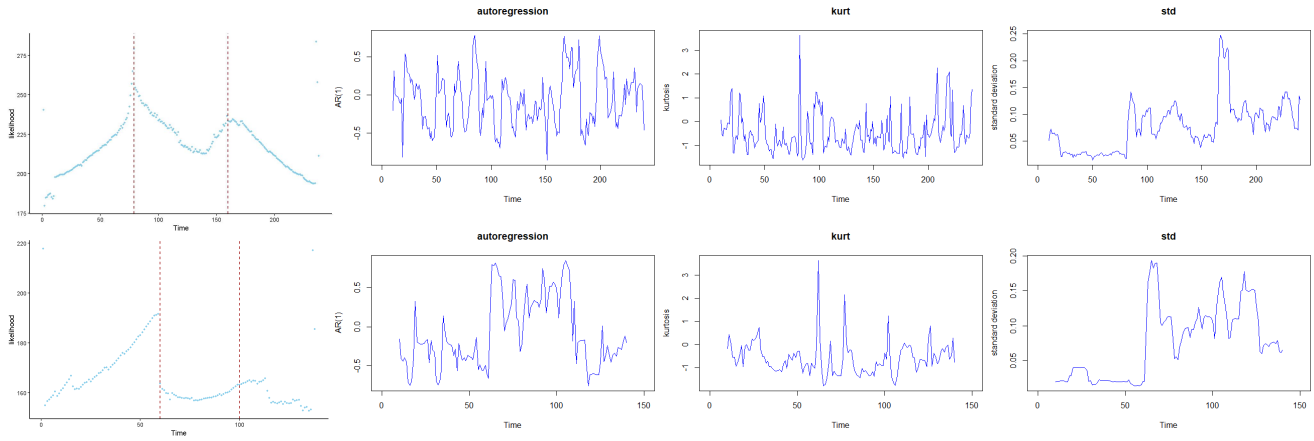
(a) Change α (b) Change s

Figure 5: Two change points scenario

a steep jump exists at the beginning of the trace plot of likelihoods, it is caused by the lack of data and it's minor compared to the real change. From (b) in Figure 5, 60 and 100 can be concluded as suspected change point according to rules (I) and (II). The control charts confirm that these two suspected change points are valid.

Still, we are interested in the comparisons between our methods and the descriptive statistics based methods. The comparison results are shown in Figure 6.

From the figure, we can see that our method has the most obvious signs of the two change points in two cases. However, all three descriptive statistics based methods have difficulties detecting the two real change points.



(a) Our method (b) Lag-1 autocorrelation (c) Kurtosis (d) Standard deviation

Figure 6: Comparisons of methods in two change points scenario

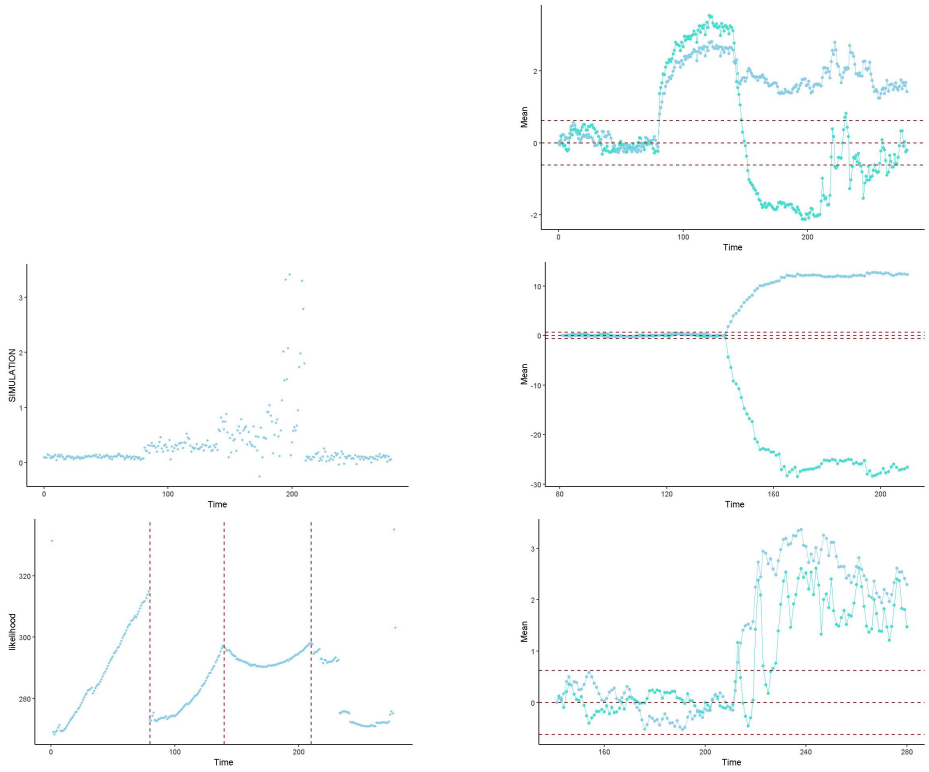
3.3 Three Change Points Simulation

This section considers the performance of our screening and sifting method on the data that exists three change points. In the simulation, we generate 280 observations in total. The change points are located at 80, 140 and 210 respectively. The parameters of the four pieces cut by the three change points are $(\alpha, s, r, \sigma, n) = (0.1, 1, 1, 0.25, 8)$, $(\alpha, s, r, \sigma, n) = (0.3, 1, 2, 0.25, 8)$, $(\alpha, s, r, \sigma, n) = (0.5, 1, 2, 0.5, 8)$ and $(\alpha, s, r, \sigma, n) = (0.1, 1, 1, 0.5, 4)$. The analyzing results are shown in Figure 7.

From (a) in Figure 7, we can see that the trace plot of likelihoods claims three suspected change points 80, 140 and 210 according to rules (III) and (II). From (b), the control charts conforms the speculation and conclude that they are change points. However, we notice that the likelihoods at point 232 has patterns similar to rules (III) or (V). As a result, it is necessary to construct the control chart for point 232 as well. Figure 8 shows the the trace plot of the likelihoods computed based on the observations after point 211 and the control charts of point 232. We can tell that point 232 is not a real change point because the control charts don't raise an alarm and the trace plot of likelihoods has no conspicuous sign of a change point.

Similarly, the comparisons between our methods and the descriptive statistics based methods are implemented.

From Figure 9, our methods again have apparent signs of suspected change point in



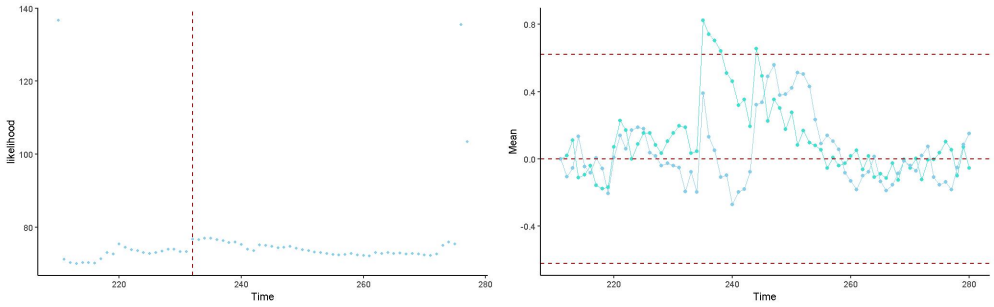
(a) Scatter plot and trace plot of likelihoods(b) Control charts for the three suspected points

Figure 7: Three change points scenario

the trace plot of likelihoods. Nevertheless, all the other three methods hardly show signs of change points.

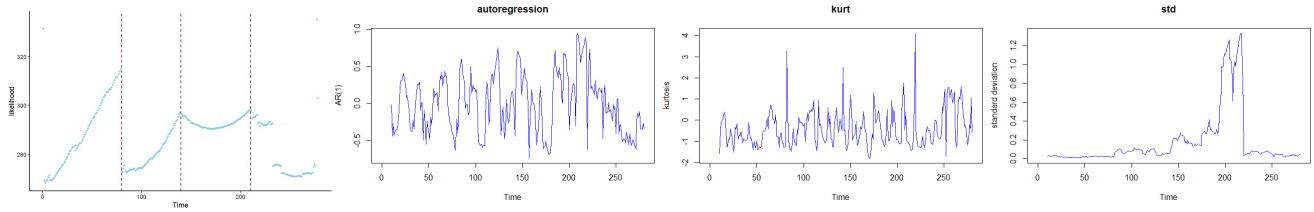
3.4 No Change Point Simulation

Instead of the scenarios that there exists change points in the data, we also interest in the situation that if there is no change point in the data series, our methods would make no signals of a change point. In this section, we generate 80 observations based on the SDE model parameters $(\alpha, s, r, \sigma, n) = (0.1, 1, 1, 0.25, 8)$. The analyzing results are shown in Figure 10. In the trace plot of likelihoods, we notice there are no significant sign of a change point. To verify the deduction, we pick 69 as a suspected change point since it has a relatively higher likelihood value. (c) in Figure 10 shows the control chart of point 69. We can see that the chart don't raise an alarm al all. As a result, our methods work when there is no change point at all.



(a) $\tilde{Q}_{n_0}(l; \boldsymbol{X}(211:280))$ trace plot (b) Control charts for point 232

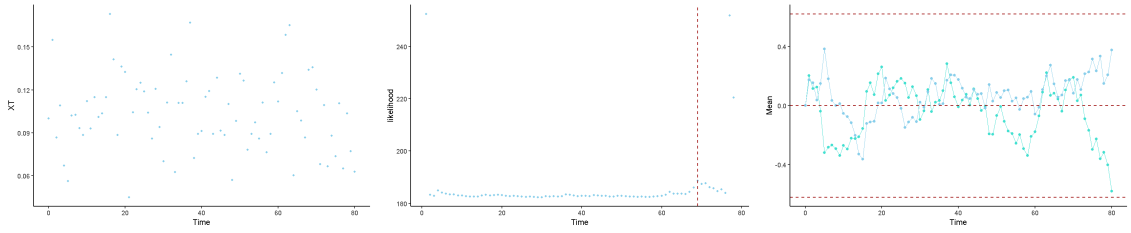
Figure 8: The analysis of point 232



(a) Our method (b) Lag-1 autocorrelation (c) Kurtosis (d) Standard deviation

Figure 9: Comparisons of methods in three change points scenario

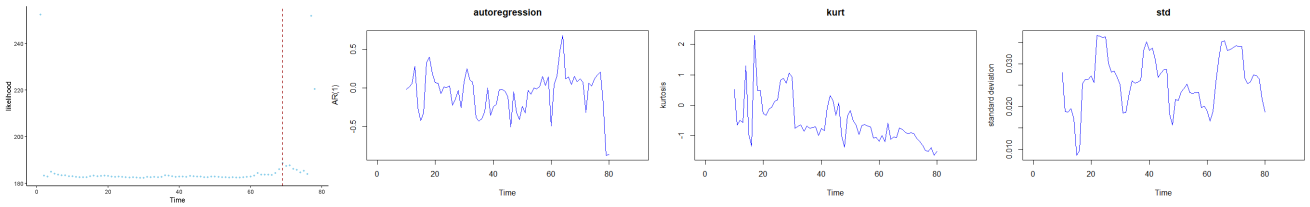
In this section, we also compare our methods with descriptive statistics based methods. The results are shown in Figure 11. We can see that the trace plot of likelihoods in our method has no sign of any change point at all. However, the trace plots of monitoring statistics in other methods have no significant difference compared with the situations that there exist change points. They go up and down all the time and it is hard to judge whether there exists any change point.



(a) Scatter plot (b) Trace plot of likelihoods (c) Control charts

Figure 10: No change point scenario

To conclude, our methods can accurately identify change points in all scenarios. Although there may exist some miss findings in the screening procedure, they can be fixed by



(a) Our method (b) Lag-1 autocorrelation (c) Kurtosis (d) Standard deviation

Figure 11: Comparisons of methods in no change point scenario

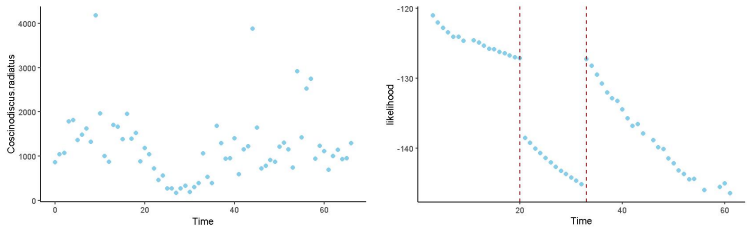
the sifting procedure using control charts. As for the descriptive statistics based methods, they lack credible criteria for identifications. You can't tell whether there exists a change point by any threshold or trend of the monitoring statistics. And there is no significant difference of the trace plots of monitoring statistics between the no change point scenario and others. Plus, the hyper parameter h in these methods may influence the trace plot of monitoring statistics dramatically. The choice of h in these methods is a problem as well.

4 Real Data Analysis

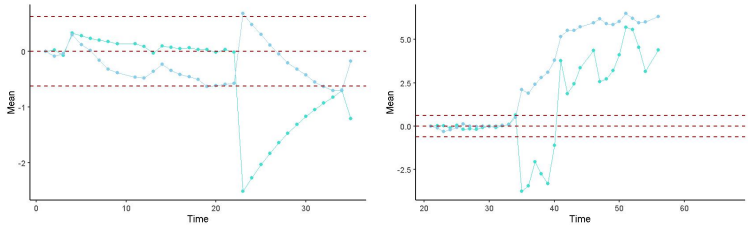
We collect the concentration of different microorganism in the river.Taken into consideration the scale of the original data, we use mean and variance scaling to preprocess the data before any analyzing. When analyzing each microorganism, the descriptive statistics based methods are also considered as contrasts. We choose the window size $h = 10$ in all cases. In consideration of confidentiality, we present only one sample here for data analysis and visualization.

4.1 Microorganism sample

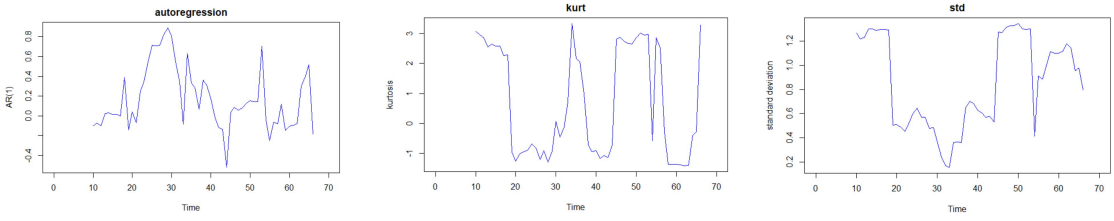
Figure 12 shows the analysis of Microorganism sample in River1. n_0 is chosen as 2. From (b), point 20 and 33 are deemed as suspected change points according to rule (V) and (IV). From (c) and (d), we find that the control charts raise an alarm after point 20 and 33. As a result, point 20 and 33 are determined as valid change points.



(a) Scatter plot of original data (b) Trace plot of likelihoods



(c) Control chart of point 21 (d) Control chart of point 33



(e) Lag-1 autocorrelation (f) Kurtosis (g) Standard deviation

Figure 12: The analysis of Microorganism sample(River1)

参考文献

[1] Dakos, Vasilis, et al. “Flickering as an Early Warning Signal.” *Theoretical Ecology*, vol. 6, no. 3, 28 Apr. 2013, pp. 309–317, https://doi.org/10.1007/s12080-013-0186-4. Accessed 2 Mar. 2023.

---. “Methods for Detecting Early Warnings of Critical Transitions in Time Series Illustrated Using Simulated Ecological Data.” *PLoS ONE*, vol. 7, no. 7, 17 July 2012, p. e41010, https://doi.org/10.1371/journal.pone.0041010.

[2] Hyvarinen, A. “Fast and Robust Fixed-Point Algorithms for Independent Component Analysis.” *IEEE Transactions on Neural Networks*, vol. 10, no. 3, May 1999, pp. 626–634, https://doi.org/10.1109/72.761722. Accessed 11 Oct. 2019.

[3] Kovářík, Martin. “Volatility Change Point Detection Using Stochastic Differential Equations and Time Series Control Charts.” *INTERNATIONAL JOURNAL of MATHEMATICAL MODELS and METHODS in APPLIED SCIENCES*, 1 Jan. 2013. Accessed 21 Apr. 2023.

[4] Scheffer, Marten, et al. “Early-Warning Signals for Critical Transitions.” *Nature*, vol. 461, no. 7260, Sept. 2009, pp. 53–59, www.nature.com/articles/nature08227, https://doi.org/10.1038/nature08227. Accessed 11 Jan. 2020.

[5] Wang, Rong, et al. “Flickering Gives Early Warning Signals of a Critical Transition to a Eutrophic Lake State.” *Nature*, vol. 492, no. 7429, 18 Nov. 2012, pp. 419–422, https://doi.org/10.1038/nature11655. Accessed 31 July 2020.

[6] Awty-Carroll, Katie, et al. “An Evaluation and Comparison of Four Dense Time Series Change Detection Methods Using Simulated Data.” *Remote Sensing*, vol. 11, no. 23, 25 Nov. 2019, p. 2779, https://doi.org/10.3390/rs11232779.

致谢

待补充