Transferring biological sequence analysis tools to break-point detection for on-line monitoring: A control chart based on the Local Score

Sabine Mercier

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Transferring biological sequence analysis tools to break-point detection for on-line monitoring: A control chart based on the Local Score. †

Sabine Mercier*

[1]Toulouse Mathematics Institute, France

Correspondence
*Sabine MERCIER. Email: sabine.mercier@univ-tlse2.fr

Present Address
Dpt MI UFR SES, Université Toulouse Jean Jaurès 5 allées Antonio Machado 31058 Toulouse cedex 9

Abstract

The Lindley process defined for the queuing file domain is equivalent to the CUSUM process used for break-point detection in process control. The maximum of the Lindley process, called Local Score, is used to highlight atypical regions in biological sequences and its distribution has been established by different manners. I propose here to use the Local Score, and also a partial maximum of the Lindley process over the immediate past, to create control charts. Stopping time corresponds to the first time where the statistic achieves a statistical significance less than a given threshold α in $]0,1[$, the instantaneous first error rate. The Local Score p-value is computed using existing theoretical results. I establish here the exact distribution of the partial maximum of the Lindley process. Performance of the control charts are evaluated by Monte Carlo estimation of the Average Run Lengths for an in-control process (ARL_0) and for an out-of-control process (ARL_1). I also use the Standard Deviation of the Run Length ($SdRL$) and the Extra Quadratic Loss (EQL). Comparison with the usual and recent control charts present in the literature shows that the Local Score control chart out-performs the others with a much larger ARL_0, and ARL_1 smaller or of the same order.

Many interesting openings exist for the Local Score chart: not only Gaussian model but any of them, Markovian dependance of the data, both location and dispersion monitoring at the same time can be considered.

KEYWORDS:
Average run length (ARL) ; control charts ; cumulative sum (CUSUM) ; exponentially weighted moving average (EWMA) ; statistical process control (SPC) ; high-quality process monitoring ; Local Score.

1 | STATE OF THE ART INTRODUCTION

Statistical quality control is a branch of industrial statistics which is also largely present in the medical field, business and many other application domains like bio surveillance. Within statistical quality control we can distinguish acceptance sampling, statistical process control (SPC), design of experiments and capability analysis. Control charts are one of the most important and commonly used tools of the SPC tool box, first proposed by Walter Shewhart in 1920. One of the main goals of control charts is to distinguish between the common variation due to chance causes and the variation from assignable causes in order to

†Local Score Control Chart

know when to act on the process or not. A control chart consists in a graphical representation of a succession of statistical tests. Three lines are most of the time represented: two of them are called control limits and are usually placed at plus or minus three times the standard deviation of the plotted statistic above and below a central line which represents the mean of the statistic or the target. Sample numbers taken over time are placed on the horizontal axis and the statistic of interest is placed on the vertical axis: It can be some quality characteristic, like the mean or the standard deviation of measures, number or percentage of defective units, number of units or samples between two occurrences of a given event like the apparition of a defect unit etc. As long as the plotted points fall inside the control limits, the process is considered as statistically in-control or otherwise out-of control. Extra rules have also been created to increase the sensitivity of the charts (see Section 2.4.1 for example). There are at the present time many control charts and Ali proposed in 2016 a large review [1].

For location process monitoring one can distinguish three usual control charts which are the Shewhart chart proposed by Walter A. Shewhart in 1920 ; the CUSUM chart introduced by Page in 1954 [2], and the EWMA chart proposed by Roberts in 1959 [3]. A Shewhart chart plots the mean $\bar{X}$ of samples and requires the hypothesis of a Gaussian model $\mathcal{N}(\mu, \sigma)$, the upper and lower control limits are $LCL = m_0 - 3\sigma/\sqrt{n}$ and $UCL = m_0 + 3\sigma/\sqrt{n}$ with n the sample size. If a sample mean verifies $\bar{X} \notin [LCL; UCL]$ the hypothesis $H_0 : \mu = m_0$ is rejected with a nominal value $\alpha = 0.27\%$. CUSUM and EWMA charts have been constructed to take into account information both from the present and the past. They are more detailed in Section 2.3.

The goal is to create a chart which detects a change in the observed signal. As long as the behavior of the observations is consistent with the target, there is no need to modify the state of the process, and if it changes, the interest is to detect it as quickly as possible. But modifying the settings of the chart to detect the change rapidly also increases the number of false alarm under no change conditions. On the other hand, trying to strongly avoid false alarms leads to long delays between the real time of a change occurrence and its detection. The objective is thus to propose a method that minimizes the average delay to detect the change and that also deals with an average time delay to a false alarm which is large enough.

In biological sequence analysis, an equivalent process as the CUSUM process is used and is called the Lindley process. It was defined in 1952 [4] for queuing theory. It is also used to highlight atypical regions in genomes, proteins or others kinds of biological sequences (see Mercier and Daudin (2001) [5] and also Fariello *et al.* (2017) [6] for an example on loci sequences). More precisely, the biologists used the maximum of the Lindley process, called the Local Score, to extract atypical segments of the sequences. A generalization of the Local Score to the comparison of two sequences is also defined and implemented in a famous software used by biologists all over the world: Basic Alignment Search Tool (BLAST, which can be used via the following link `https://blast.ncbi.nlm.nih.gov/Blast.cgi`.). At the present time, many results on the Local Score distribution exist for diverse contexts of biological sequence analysis (see Section 4 or Lagnoux *et al.* (2017) [7] for a review of the case of independent and identically distributed variables). However at the present time, those results are not exploited in Statistical Process Control and Monte Carlo simulation are achieved to estimate thresholds or percentiles of the CUSUM process.

We propose in this article to use the maximum of the CUSUM statistic, combining with an alarm based on a Local Score p-value lower than a given threshold $\alpha \in]0, 1[$. We also propose a statistic other than the classical Local Score which is also used to analyze biological sequences: the on-going excursion score which allows to focus on the information of the immediate past. See Fariello *et al.* (2017) [6] for an example of application on real sequences using a sequence model taking into account dependance between the sequence components. For the application of this second statistic. Its exact distribution is established in Section C.

The main difference between biological sequence analysis and SPC is that biological sequences are not sequentially (on-line) observed.

Plan: We recall the classical performance criterion and the classical control charts EWMA, CUSUM, and also the enhanced and mixed version of these previous charts in Section 2. The statistics used in biological sequence analysis are presented in Section 3 and we verify that their transfer to control chart context has sense. More precisely, as biological sequence analysis is based on a scoring function choice, Section 3.1 is dedicated to this purpose and we present two different kinds of scoring schemes. The definition of the stopping time of the proposed control charts is defined in Section 3.4. The main results on the Local Score distribution are recalled in Section 4. Section 5 proposes an evaluation of the performance of the proposed charts and a comparison with existing charts using classical performance criteria, the average run length to an alarm (ARL) for both in-control (ARL_0) and out-of control (ARL_1) processes. We consider here Gaussian model in our simulation. An illustration is given in Section 6. The main results of this work are gathered in Section 7. Appendix section present some practical approaches for the use of Local Score and a method to establish the exact distribution of the score for the on-going excursion, the second statistic proposed herein.

2 | USUAL CONTROL CHARTS

2.1 | Some definitions and notation

A control chart graphically represents a succession of statistical tests for which the hypothesis H_0 corresponds to an in-control process, that is with a given signal density f_0 of mean μ_0 and standard deviation σ_0; and the alternative hypothesis H_1 to an out-of control process, that is with a signal density f_1 different from f_0 of mean $\mu_1 = \mu_0 + \delta\sigma_0$, and standard deviation $\sigma_1 = \sigma_0/q$, where $q \in \mathbb{R}^{+*}$ and $\delta \in \mathbb{R}^*$.

A_i is the received signal or the measure at index i for individual values
n is the sample size if a data sample is observed at index i
$\bar{A}_i$ is the empirical mean of the sample at index i
I is the sequence length for finite sequences $(A_i)_{1\leqslant i\leqslant I}$
δ is called the shift in location
q corresponds to a change in variation

An alarm corresponds to the rejection of the hypothesis H_0 and is usually defined by a plotted point out of the control limits. Examples of extra rules for an alarm are given in Section 2.4.1. Sometimes, alarms are defined by the mean of a stopping time T. For example: $T = \inf\{i : \bar{A}_i \notin [LCL; UCL]\}$ with LCL (respectively UCL) the lower (respectively upper) control limit for a Shewhart chart.

2.2 | Performance criterion

The usual performance criterion are the average run length (ARL) to an alarm for an in-control (ARL_0) and an out-of control process (ARL_1); and the corresponding standard deviation $SdRL_0$ and $SdRL_1$. In Zaman *et al.* (2015)[8] the authors present an alternative measure to the ARL, the extra quadratic loss (EQL), defined as a weighted average of ARLs over a range of values of the shift δ.

$$EQL = \frac{1}{\delta_{\max} - \delta_{\min}} \int_{\delta_{\min}}^{\delta_{\max}} \delta^2 ARL(\delta) d\delta \, . \tag{1}$$

Note that there are also other interesting comparative performance criterion as the relative ARL ($RARL$)[9,10]; the performance comparison index (PCI)[11] which facilitates the ranking based on the EQL.

As we will only consider the best chart highlighted in the literature for each domain of shift (small, moderate and large ones) we use a discrete adaptation of the EQL for 4 different values of δ: 0.25, 0.5, 1 and 2.

2.3 | Classical Control Charts

We focus here on existing charts used to monitor the location of the process and which take into account the past information as the one we propose in Section 3.2. The Shewhart chart is not discussed here as it is outperformed by the other charts we present here.

The classical **EWMA** control charts are proposed by Roberts in 1959 and are defined as follows. Let λ in $[0, 1]$ be the weight of the present compared to the past and $Z_0 = \mu_0$. The test statistic of the EWMA chart is

$$Z_i = \lambda \cdot \bar{A}_i + (1 - \lambda) \cdot Z_{i-1} \tag{2}$$

where $\bar{A}_i$ is the average of n current observations, but it can be individual observations with $n = 1$. We also have $Z_i = \lambda \cdot \sum_{j=0}^{i-1}(1 - \lambda)^j \bar{Z}_{i-j} + (1 - \lambda)^i \cdot Z_0$. Mean and standard deviation of the EWMA statistic are

$$\mathbb{E}[Z_i] = \mathbb{E}[\bar{Z}_i] = \mu_0 \text{ and } \sigma_{\bar{Z}_i} = \sigma \cdot \sqrt{\frac{\lambda[1 - (1 - \lambda)^{2i}]}{n(2 - \lambda)}}$$

with σ the standard deviation of the process. The usual control limits are defined as

$$CL_Y = \mu_0 \pm L\sigma_Y \tag{3}$$

for $Y = A_i$ or $\bar{A}_i$. In practice, L is chosen according to the choice of λ and a prefixed ARL_0 value. For large i the control limits converge to $CL = m_0 \pm L\sigma\sqrt{\frac{\lambda}{n(2-\lambda)}}$. EWMA charts are known to quickly detect small to moderate shifts in the process mean and are recognized to have a good performance [12]. For $\lambda = 0.1$, small shifts can be detected, however there is a bad performance for the large ones; $\lambda = 0.2$ is an intermediate choice and $\lambda = 0.4$ bring results similar to a Shewhart chart. EWMA charts are easy to interpret and are also well adapted for individual value context. Like the Shewhart chart, the EWMA chart is also based on the normality of the quality characteristic under study, but it is not very sensitive to the Gaussian assumption.

CUSUM charts: Page in 1954 [2] introduced the CUSUM process which accumulates signal derives from a target as follows:

$$CUSUM_0 := 0 \text{ and } CUSUM_k := \max(0, CUSUM_{k-1} + x_k) \tag{4}$$

with x_k a score assigned to the sample k. The CUSUM process is also used in the quickest detection theory (see Egea *et al.* (2017) [13] for an overview).

The CUSUM charts are based on two statistics called upper and lower one-sided CUSUM respectively, defined as follows with $C_0^+ = C_0^- = 0$:

$$\begin{aligned} C_i^+ &= \max[0, (\bar{A}_i - \mu_0) - K + C_{i-1}^+] \\ C_i^- &= \min[0, -(\bar{A}_i - \mu_0) - K + C_{i-1}^-] \end{aligned}$$

and are more appropriate for Gaussian characteristics. Usually $K = k\sigma_0$ with $k = \delta/2$ and $\delta = \frac{|\mu_1 - \mu_0|}{\sigma_0/\sqrt{n}}$ the mean shift. CUSUM charts can be used for individual observations instead of the means. The control limits are given by $CL = H = h \cdot \sigma_0$ with a different value of h for the two statistics. The values of h are the fixed critical thresholds chosen by considering a desired performance of the chart. The selection of the (k, h) pair, greatly influences the ARL performance of the chart. In practice, h is set up for an expected ARL_0 and an ARL_1 is deduced (see Granjon (2013) [14], p.17 for more details). CUSUM charts are known to detect more rapidly small to moderate shift.

Another way to present the CUSUM statistics is presented in Section 3.3.

2.4 | Improved Control Charts

2.4.1 | Extra-rule EWMA control chart

As said previously, the main rule used to declare the out of control process is at least one plotted point out of the control limits. There are some additive rules for the Shewhart chart [15] to improve the sensitivity of the chart. These additive rules are usually restricted to the Shewhart chart but works have been done to extend them to CUSUM [16] and EWMA [17] charts improving their performance for small and moderate shifts, without inflating the pre-specified false alarm rate. For an extra rule example, a process can be said to be out-of-control if there exists more than 7 consecutive increasing points. In the work of Abbas *et al.* (2011) [17] the authors propose to apply such extra rules for EWMA charts and they show that it increases the performance of the chart. Scheme II proposed by the author is based on the two following conditions. If one of the conditions is satisfied, the process is declared out-of control: (1) At least two out of three consecutive points fall below a LSL and the point above the LSL (if any) falls between the CL and the LSL. (2) At least two out of three consecutive points fall above a USL and the point below the USL (if any) falls between the CL and the USL with

$$CL = \mu_0, \quad LSL = \mu_0 - L_S\sigma\sqrt{\frac{\lambda}{n(2-\lambda)}}, \quad USL = \mu_0 + L_S\sigma\sqrt{\frac{\lambda}{n(2-\lambda)}} \tag{5}$$

where L_S is the signaling limit coefficient which is set according to the pre-specified value of ARL_0. According to the work of Abbas *et al.* (2011) [17] Scheme II exhibits dominance in general as compared with all the other schemes and charts covered in their article (Shewhart, classical CUSUM and EWMA, enhanced EWMA with scheme I and II, FIR CUSUM, FIR EWMA, weighted CUSUM, double CUSUM, and distribution free CUSUM charts). So we restrict in Section 5 our comparison of our method to the EWMA scheme II performance for small to moderate shifts.

2.4.2 | Mixed EWMA-CUSUM (MEC) and CUSUM-EWMA Control Chart (MCE)

Zaman *et al.* (2015) [8] and Abbas *et al.* (2013) [18] the authors propose to combine the features of CUSUM and EWMA charts. The MEC feature consists in using the statistics of the EWMA of Equation (2) as input for the CUSUM chart and hence, the plotting

statistic for the MEC chart is as follows:

$$\begin{aligned} MEC_i^+ &= \max[0, (Z_i - \mu_0) - K_{Z_i} + MEC_{i-1}^+] \\ MEC_i^- &= \min[0, -(Z_i - \mu_0) - K_{Z_i} + MEC_{i-1}^-] \,, \end{aligned}$$

where K_{Z_i} is the time-varying reference value defined as:

$$K_{Z_i} = k_Z \cdot \sqrt{Var(Z)} = k_Z \cdot \sigma_{\bar{X}} \cdot \sqrt{\frac{\lambda[1-(1-\lambda)^{2i}]}{n(2-\lambda)}} \,,$$

and these statistics are plotted against the control limit H_{Z_i} defined as follows:

$$H_{Z_i} = h_Z \cdot \sqrt{Var(Z)} = h_Z \cdot \sigma_{\bar{X}} \cdot \sqrt{\frac{\lambda[1-(1-\lambda)^{2i}]}{n(2-\lambda)}} \,,$$

with h_Z the coefficient used to fix the predefined false alarm rate.

The MCE feature also consists in a mixture of the two statistics but in a reverse order. The two resulting MCE statistics are

$$\begin{aligned} MCE_i^+ &= (1-\lambda_C) \cdot MCE_{i-1}^+ + \lambda_C \cdot C_i^+ \\ MCE_i^- &= (1-\lambda_C) \cdot MCE_{i-1}^- + \lambda_C \cdot C_i^- \,, \end{aligned}$$

with C_i^+ and C_i^- are the classical CUSUM statistics (see Equation (5)) and $\lambda_C \in]0,1[$ the sensitivity parameter of the chart, and $MCE_0^+ = MCE_0^- = \mu_0$. The control limits are

$$CL_i = \mu_{C_i} \pm L_C \cdot \sigma_{C_i} \cdot \sqrt{\frac{\lambda_C[1-(1-\lambda_C)^{2i}]}{n(2-\lambda_C)}} \,,$$

with $\mu_{C_i} = \mathbb{E}[C_i^+] = \mathbb{E}[C_i^-]$ and $\sigma_{C_i}^2 = Var(C_i^+) = Var(C_i^-)$ and L_C the width coefficient, like L in Equation (3).

These two previous charts have been compared [18,8] with existing charts from the literature, for different parameter values of δ and for different prefixed ARL_0 that determines the parameters of the charts. The studies conclude that the MCE control chart is very sensitive for the detection of small and moderate shifts. More precisely, EWMA scheme II with $\lambda = 0.1$ becomes a bit superior to the MEC chart [18], and a bit superior when $\lambda \leqslant 0.25$ than the MCE chart [8].

3 | THE LOCAL SCORE CHART (LS CHART)

3.1 | Score functions

Biological sequence analysis and more precisely highlighting atypical regions in biological sequences, is based on the choice of a scoring function. See as an example the quite well known Kyte and Doolittle hydrophobic score scale [19] to highlight trans membrane regions, or the website `https://web.expasy.org/protscale/` which gathers numerous scoring functions to study proteins. Score scales attribute to each possible component of the sequence (the four nucleotides A, C, G, T for nucleic sequences for example, or the 20 amino acids for proteins) a real or an integer value which quantifies a level of a physico chemical property. These scoring functions are established mostly by biological experimental manipulation and deeply depend on the kind of segments to be highlighted. Mercier and Nuel (2020) [20] propose a mathematical method to learn scoring function from a data set.

In order to use the biological approach to highlight change-point in process monitoring, adapted scoring functions must be built. We present here two classical scoring functions used in signal process. The log likelihood ratio (LLR) scoring function is defined as follows [2]:

$$LLR_i = \ln\left(\frac{f_1(A_i)}{f_0(A_i)}\right) \tag{6}$$

where A_i is the observed signal at time i and f_0 (respectively f_1) is the H_0 (resp. H_1) distribution signal before (resp. after or during) the change in mean and/or in variability. It is usual to consider the signal distribution to be Gaussian but other models can also be used as exponential ones for example.

Tartakovsky *et al.* (2012)[21] propose a generalization of the LLR scoring function which allows to detect bias both in mean and variance at the same time

$$X_i = C_1 \cdot Y_i + C_2 \cdot Y_i^2 - C_3 \tag{7}$$

with $Y_i = (A_i - \mu_0)/\sigma_0$ the centered and standardized data under H_0, $C_1 = \delta \cdot q^2$, $C_2 = (1 - q^2)/2$, $C_3 = \delta^2 q^2/2 - \log(q)$, $\delta = (\mu_1 - \mu_0)/\sigma_0$ and $q = \sigma_0/\sigma_1$. Parameter δ corresponds to the minimum level of change in the mean that is required to be detected and no change-point detection on the mean brings $C_1 = 0$. Parameter q corresponds to the minimum level of change in the variance that is required to be detected and no change-point detection on the variance implies $C_2 = 0$. As mentioned in Tartakovsky *et al.* (2012)[21], this scoring function equals the LLR scoring scheme for Gaussian model.

3.2 | The Local Score statistic

Let $\mathbb{A} = (A_i)_{1 \leqslant i \leqslant I}$ be an observed signal of length I and s a scoring function defined on the possible components of the sequence $\mathbb{A}$. The Local Score was first defined in 1990 for biological sequence analysis as follows

$$M_I := \max_{0 \leqslant k \leqslant \ell \leqslant I} \sum_{t=k}^{\ell} s(A_t) \text{ with } s(A_0) := 0 \text{ by convention.} \tag{8}$$

This definition leads to the following properties.

Property 1 (Local Score)**.**

- $\forall I \geqslant 1,\ M_I \geqslant 0.$
- M_I is growing with I, thus $(\forall h > 0)$, $\mathbb{P}(M_I \geqslant h)$ is decreasing with I.

The Local Score corresponds to the maximal value of a given property, linked to the chosen scoring function, that can be found locally in the sequence. To each segment (i, j) of the sequence, with $1 \leqslant i \leqslant j \leqslant I$, a score $\sum_{t=i}^{j} s(A_t)$ can be associated to the segment; and the Local Score is defined as the maximal value over all possible segments of the sequence with every possible begin index and every possible end index. Sliding windows do the same but with a given and fixed window/segment size. Scan statistics[22] as well, for which we could define a scoring function which corresponds to the number of occurrences of a given event which occurs in the window.

In 2001, Mercier and Daudin gave an equivalent definition of the Local Score based on the Lindley process[5].

$$M_I := \max_{0 \leqslant k \leqslant I} W_k \tag{9}$$

$$\text{with} \quad W_0 := 0 \text{ and } W_k := \max(0, W_{k-1} + s(A_k)) \tag{10}$$

The process $(W_k)_{0 \leqslant k \leqslant I}$ is called the Lindley process[4] and was first defined in 1952 for Queuing Theory and transposed to biological sequence analysis in 2001[5].

Remark 1. The process $(W_k)_k$ defined in (10) can also be defined as follows:

$$W_0 := 0 \text{ and } W_k := S_k - \min_{0 \leqslant j \leqslant k} S_k \tag{11}$$

with $S_0 := 0$ and $S_i = \sum_{k=1}^{i} X_k$ the partial sums of the process $(X_k)_k$, with in our case $X_k := s(A_k)$. Such a definition is more used in Brownian motion theory.

3.3 | Relation between CUSUM process and Local Score

The CUSUM process has been defined in continuous inspection scheme or sequential tests[2,13] as in Equation (10) of the previous section.

The Local Score is the maximum of the CUSUM process.

Note that the two CUSUM chart statistics of Equation (5) can be linked to Equation (10) defining two different scoring functions: $s^+(A_i) = A_i - \mu_0 - K$ for $(C_i^+)_i$ and $s^-(A_i) = -A_i + \mu_0 - K$ for $(C_i^-)_i$.

The principal difference between biological sequence analysis and process monitoring is that a biological sequence $\mathbb{A} = (A_k)_{1 \leqslant k \leqslant I}$ is globally given and has a fixed length I and one Local Score value is calculated for a given sequence $\mathbb{A}$. In signal context or process monitoring, the signal sequence evolves with a "length" increasing with the time index. At time $i \leqslant I$, we observe the sequence $\mathbb{A}^i = (A_k)_{1 \leqslant k \leqslant i}$ and sometimes i can increase indefinitely.

3.4 | Stopping time definition

Let $(A_i)_{i\geqslant 1}$ be an observed signal over time i. Sequential analysis leads to define a Local Score process $(M_i)_{i\geqslant 1}$ with M_i the local score of the signal sequences $\mathbb{A}^i = (A_k)_{1\leqslant k\leqslant i}$.

Let $\alpha \in]0,1[$ and let us define the following stopping time for the Local Score statistic:

$$T_{LS} = \inf\{i \,:\, \mathbb{P}(M_i \geqslant m) < \alpha\} \tag{12}$$

with m the observed Local Score value of the sequence $(A_k)_{1\leqslant k\leqslant i}$. The transfer between the tools of a global analysis in biological context to an on-line/sequential analysis in process monitoring has sense. Indeed, let us consider $A_1, \dots, A_j$ with an observed Local Score value m realized by a segment with an end index $i < j$. The Local Score of the sequence $A_1, \dots, A_i$ is also m. Thus, using Property 1, if $\mathbb{P}(M_j \leqslant m) < \alpha$, then $\mathbb{P}(M_i \leqslant m) < \alpha$ and the alarm will be previously given at i without any extra delay.

In continuous inspection schemes[2,13], the properties and the definitions of Section 3.2 and 3.3 lead to the fact that the considered statistic is exactly the same as the one on which the Local Score chart is based.

Property 2. Let us consider on-line change-point detection, with no alarm before i. A potential alarm at time i for the CUSUM process induces that the CUSUM value at time i is higher than every ones before i. This induces that the CUSUM value at i corresponds to the Local Score.

$$CUSUM_i = M_i \,.$$

The difference stands on the manner to declare an alarm. Indeed, an action is taken at index i if the CUSUM value is larger than a given threshold H independent to the index i[2,13]. Such an approach is used for the CUSUM chart (see Section 2.3) with two thresholds, one for each statistic C_i^+ and C_i^-. It is the same for Wald's method[23,13] where the stopping time is defined as follows:

$$T_W = \inf\{i \,:\, M_i \geqslant H\}$$

with H deduced from the Wald's inequality[23]: $\alpha < e^{-H}$, so $H = -\ln(\alpha)$. But such a threshold does not depend on the parameters of the process at all. The approach proposed in the sequel consists in taking a threshold m which depends on i because the Local Score statistic depends on the length of the sequence i. It also depends on the chosen nominal value $\alpha \in]0,1[$ of the test. In our approach, α corresponds to a classical type I error for a statistical test at time i without considering the test issues before i. The α value must therefore be distinguished from the conditional alarm rate[24] which considered the previous test issues to be not significant (no alarm). The stopping time defined in (12) is equivalent to

$$T_{LS} = \inf\{i \,:\, M_i \geqslant m(\alpha, i)\}$$

with $m(\alpha, i)$ verifying $\mathbb{P}(M_i \geqslant m(\alpha, i)) = \alpha$. In 1995[24] a similar stopping time is proposed

$$T_M = \inf\{i \,:\, M_i \geqslant H'(\alpha, i)\}$$

with $H'(\alpha, i)$ verifying the following conditional equations:

$$\mathbb{P}(M_1 \geqslant h'_1) = \alpha \,, \quad \mathbb{P}(M_2 \geqslant h'_2 | M_1 < h'_1) = \alpha$$
$$\text{and} \quad \mathbb{P}(M_i \geqslant h'_i | M_1 < h'_1, \dots, M_{i-1} < h'_{i-1}) = \alpha \text{ for } i \geqslant 3$$

This choice is motivated to have a geometric stopping time, so the ARL_0, also called the mean time between two false alarms (MTBFA), corresponds to $1/\alpha$ as for the Shewhart control chart. In Sahki *et al.* (2020)[25] the authors also propose to use the maximum of the CUSUM process $\mathbb{W}$. They propose a stopping time based on a constant threshold defined as the empirical percentile of order $(1 - I\alpha)$ of the Local Score of sequence of a given length I, with $I \geqslant 100$. This threshold is then used for every $i \geqslant I$. Of course, as said by the authors[25], for large I, this is not possible to achieve. They also propose to use the CUSUM process itself and a threshold depending on i, called the empirical instantaneous threshold, which corresponds to an empirical percentile or order α of the CUSUM process at time i.

The different thresholds proposed in the literature[25,24,26,27] are established using simulation.

We propose here to use different theoretical results on the distribution of the maximum of the CUSUM process to apply the Local Score stopping time T_{LS} defined in (12) which differs from all the discussed methods.

Figure 1 illustrates the correspondence between CUSUM, or Lindley, process $(W_i)_i$ (top), the Local Score process $(M_i)_i$ (middle) and the $1 - p$-value process (bottom). We prefer to plot $1 - p_{value}$ rather than the p-value because it makes the figure clearer to read. In the figure, the simulated sequence is under H_0 and does not possess any atypical segment. The Local Score of the whole sequence is 12 and is achieved at index 10 . Before index 10, the Local Score is increasing with the index i and

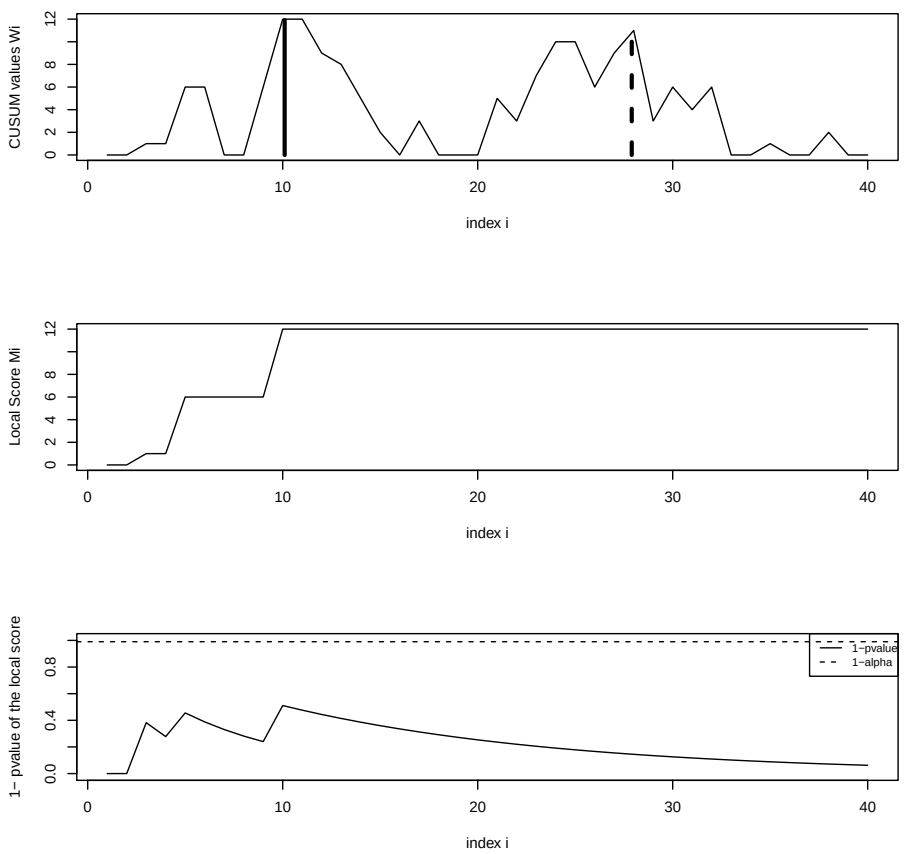

FIGURE 1 CUSUM, or Lindley, process $(W_i)_i$ (top), the Local Score process $(M_i)_i$ (middle) and the $1-p$-value process (bottom).

after it is still 12 even considering the rest of the sequence, as the successive excursions of the CUSUM process are not higher. We can see that this Local Score is not statistically significant, with $1-\mathbb{P}(M_{10} \geqslant 12) \simeq 0.55$, even considering only the portion $(A_1, \dots, A_{10})$ of the sequence; and then $1-\mathbb{P}(M_i \geqslant 12)$ decrease with i as the sequence length is increasing and so this gives more chance to be achieved. We can see in this example that longer the sequence is, the more difficult it is to have a significant Local Score. This figure leads us to also consider another statistic of the CUSUM process, called score of an excursion. Indeed, the segment in dash line at the bottom of the figure corresponds to a segment with a score of 11. It may be an atypical observation by itself, only considering this part of the sequence (from the last zero of the CUSUM process to the top of the on-going excursion) without taking into account the whole past of the sequence. It corresponds to consider that the sequence begins at index 20.

3.5 | Score of the on-going excursion Q

Another statistic is also used to highlight atypical regions in biological sequence. We will note Q the height, also called score, of the on-going excursion. Indeed, the CUSUM process defines non-negative excursions above 0. Figure 1 (top) represents the CUSUM process with 6 successive excursions above 0. Let us denote $Q^{(k)}$ as the maximal height of the k-th excursions. Considering the time index i, we define the on-going excursion the one to which the time i belongs. For example, the on-going excursion at index 10 is the second excursion of maximal height equals to 12. At index 9, the on-going excursion will be still the second excursion but with a maximal height with a value around 10, because for sequential analysis the considered sequence is cut at index i. Let us denote i_Q the begin index of the on-going excursion. Let us also define the on-going excursion stopping time as:

$$T_Q = \inf\{i \geqslant i_Q \ : \ \mathbb{P}(Q \geqslant \ell) < \alpha\} \tag{13}$$

with ℓ the height of the on-going excursion of the observed sequence at index i.

Property 3 (Q and M_i: two identical variables or not)**.** Let us suppose that the on-going excursion at the index i realizes for the first time a significant score value equal to ℓ, that it verifying $\mathbb{P}(Q \geqslant \ell) < \alpha$. Then, the Local Score of the sequence $(A_k)_{k=1,\dots,i}$ is also equal to ℓ and realized by the same segment.

Therefore, for on-line detection, the two statistics Q and M_i have an equal values. But the two variables on global analysis are different and so are their distribution and respective p-value.

Indeed, if a higher or equal excursion exists before index i, its p-value is thus smaller than or equal to the on-going one and so the alarm would have been given before and the detection process initialized.

Property 4. If a CUSUM value appears at i to be higher than at every index before i, then $M_i = Q = CUSUM$.

On-line detection implies that if your are at index i, there is no alarm before i. The property can easily be proved using an absurd reflexion.

Property 5.

$$ARL(M_i) > ARL(Q)$$

Indeed for $\ell > 0$, we have $\{M_i < \ell\} = \{\forall k,\ Q^{(k)} < \ell\} \subset \{Q < \ell\}$. So $\mathbb{P}(M_i < \ell) < \mathbb{P}(Q < \ell)$ and $\mathbb{P}(M_i \geqslant \ell) > \mathbb{P}(Q \geqslant \ell)$. So an excursion can have a p-value less than the threshold α but a Local Score of a same height can have a p-value larger than the threshold α. With the Property 4, stopping time T_Q will happen before that of the Local Score T_{LS}.

Sahki *et al.* (2020)[25] also propose a similar statistic as Q, with a threshold called Dynamic Empirical Instantaneous threshold. Their proposed stopping time is based on a threshold which corresponds to the percentile or order $1 - \alpha$ of the statistic Q established by Monte Carlo simulation. We establish in Section C the exact distribution of the statistic Q.

4 | DISTRIBUTION OF THE TEST STATISTICS

The stopping time definitions in (12) and (13) are based on the distribution of the Local Score and the on-going excursion score. To establish these distributions we chose among the different existing possible methods. Indeed, there exist many results on the Local Score distribution: authors have considered both I.I.D. model and Markovian one and they proposed approximations for long sequences and exact methods for medium and small ones. See Lagnoux *et al.* (2017)[7] for a detailed overview in I.I.D. model.

For the I.I.D. model, the exact method is very well adapted for sequences with lengths lower than 10^3. They require integer scores[5] but some generalization can be made for practical application. Asymptotic results[28,29] are performed for sequences with length $\geqslant 300$. Asymptotic results require a non-positive average score which is recovered with the scoring function defined in this sequel. For Markovian model, exact method is very well adapted for sequences with length $\leqslant 10^2$ and require integer scores[30]. Asymptotic results also exist in Markovian model with the same constraints as the ones in I.I.D. model[28,31].

We focus in this article on I.I.D. model where the observations $(A_k)_{1\leqslant k\leqslant i}$ are considered to be a realization of independent and identically distributed variables. We consider a Gaussian distribution in our simulation of Section 5 and 6.

At the present time work is in progress for creating a package R which gathers the different methods to establish the Local Score distribution.

Mercier and Daudin (2001)[5] have proven that

$$\mathbb{P}(M_I \geqslant m) = (1, 0, \dots, 0) \cdot \Pi^I \cdot (0, \dots, 0, 1)', \tag{14}$$

where Π is a $(m+1)$-square matrix linked to the distribution of $(s(A_i))$ and the sign $'$ stands for the transpose of a matrix. Relation (14) is called the exact method. In the case where the mean score is negative, the distribution of M_n minus a logarithmic term converges to a Gumbel distribution[32,28]:

$$\lim_{I\to\infty} \mathbb{P}(M_I \leqslant m + \frac{\log I}{\lambda}) = \exp\left\{ - K^* \cdot e^{-\lambda m}\right\}, \tag{15}$$

where λ is the unique positive root of $\mathbb{E}[e^{xs(A_i)}] = 1$ and K^* depends on the distribution of $(s(A_i))$. Cellier *et al.* (2003)[29] proposed an improved approximation for $\mathbb{P}(M_I \leqslant m + \frac{\log I}{\lambda})$ which is more accurate than (15) for smaller sequences. Note that

using such a scoring function as given by (6) assures that the average score is non positive under the hypothesis of an in-control process. So the two methods (15) and (14) can be used.

Both the two previous methods to compute the local score distribution need the score distribution to be given. This one depends on the choice of the model for the sequence and the scoring function. The score distribution is easily theoretically established in a Gaussian model and one can recover that LLR and Tar scores are the same.

4.1 | Distribution of the score of the on-going excursion

The approximation of the distribution of the score of the first excursion proposed by Karlin and Dembo (1992)[28] is not accurate for this context because the approximation is asymptotic with the value of the on-going excursion score growing to the infinity. We verify by simulation that it is not suitable for non-extreme score. Maybe the improved one of Cellier *et al.* (2003)[29] can be a solution for a fast estimation of the p-value. But we prefer to establish an exact method given in Section C.

The distribution of Q depends on the distance d between the begin index of the on-going excursion and the time index i which is considered, but it converges very fast. We will denote Q_d when it is necessary to consider the distance d.

5 | SIMULATIONS

Establishing theoretically the ARL for the Local Score statistic is quite a hard task. We then chose to use Monte Carlo simulations with 10^5 repetitions to carry out the calculation of this quantity as in the literature [17,16,18,8]. Indeed, if the distribution of the Local Score M_i for one sequence of length i has been already established in diverse contexts, there exits a dependence between M_i and M_{i+1} and to establish the distribution of the run length to an alarm we must consider the distribution of M_i conditioned to the events $(M_j)_{j<i}$ not significant.

The lines codes for the computation are developed in R language.

We use Gaussian model; δ values equal to 0.25, 0.5, 1 and 2; the LLR score defined in (6) for the Local Score chart and the Q-chart. Parameters of the Gaussian model are $\mu_0 = 0$, $\sigma_0 = 1$ and $q = 1$. We simulate individual values, so $\sigma_{Z_i} = \sigma_0 \cdot \sqrt{\frac{\lambda[1-(1-\lambda)^{2i}]}{(2-\lambda)}}$ for the EWMA chart.

5.1 | Score distribution

For Gaussian model Equation (6) leads to $LLR(a) = D_2 \cdot a^2 + D_3 \cdot a + D_1$ with

$$D_2 = \frac{1}{2} \cdot \left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2} \right) \quad D_3 = \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2} \right) \quad D_1 = \ln\left(\frac{\sigma_0}{\sigma_1} \right) + \frac{1}{2} \cdot \left(\frac{\mu_0^2}{\sigma_0^2} - \frac{\mu_1^2}{\sigma_1^2} \right).$$

For $q = 1$, $D_2 = 0$ and $D_3 = \delta/\sigma$. To get integer scores, we first multiply the LLR by $E = 10$ to get a larger range of final scores and then we take the integer part of the values. This method can be justified by the fact that we get a Local Score also multiplied by E and that it does not change the p-value of the event. We verified by simulation that taking the integer part does not change significantly the waiting time of the alarm (see Section 7). We have

$$p_k = \mathbb{P}(\lfloor E \cdot LLR(a) \rfloor = k) = \mathbb{P}\left(\frac{k}{E \cdot \delta} + \frac{\delta}{2} \leqslant U < \frac{k+1}{E \cdot \delta} + \frac{\delta}{2} \right)$$

with U a reduced and centered Gaussian variable. We can see here that for location change monitoring the score distribution is independent of μ_0, σ_0 and only depends on δ, so the p-value matrices in the next section will be fixed for a given δ.

5.2 | p-value matrices

We chose the following methodology. After establishing theoretically the score distribution for each δ, we compute in a first step the different p-values for (i, m) taking their values in $\mathcal{C} = [1, \dots, I_{max}] \times [0, \dots, m_{max}]$ using both methods, the exact and the approximated one. We divide the set $\mathcal{C}$, and the corresponding matrix $\left(\mathbb{P}(M_i \leqslant m)\right)_{i,m}$ in two blocks: for small i values ($i < 500$) we use the exact method and for large ones (above 500) we use the approximation of Karlin and Dembo (1992)[28]. We verify

that for $m = m_{max}$ all the p-values are lower enough to assume that for upper m it would be extreme observations, for which the p-value can be rapidly computed using the approximation of Karlin and Dembo (1992)[28].

For each observation (i, m) the Local Score is calculated: for $(i, m) \in C$ the p-value is given by the computation done previously and for the observed pair (i, m) out of the bound of (i_{max}, m_{max}), the approximation of Karlin and Dembo (1992)[28] is used to get the p-value.

Establishing the p-value matrix for the Q-statistic is faster as the p-values $\mathbb{P}(Q_d \geqslant m)$ converge very rapidly with d (several dozens are enough).

The p-value threshold α is chosen to be 0.05, 0.01 and 0.0027.

Filling the Local Score p-value matrix can take time depending on the computer you are using: for $\mu_0 = 0, \sigma_0 = 1, \delta = q = 1$, $i_{max} = 2000$ and $n_{max} = 150$ it can take one or two minutes. Note that these maximal values can be reduced. It is very fast for Q. Note that for location monitoring, the matrix only depends on δ and that the application of the method is immediate after this step as seen in the previous section.

5.3 | Comparison with the existing charts

Shewhart charts are known to have an $ARL_0 = 1/\alpha \simeq 370$ with $\alpha = 0.27\%$ the probability of a false alarm obtained with the classical control limit $CL_{\bar{X}} = m_0 \pm 3\sigma_{\bar{X}}$. That is the reason why we also used this threshold for our Local Score stopping time parameter of Equation (12). The corresponding ARL_0 of the Local Score chart is much larger than 370 as shown in Table 1. It is due to the fact that the sequential statistics $(M_i)_{i\geqslant 1}$ are not independent as EWMA and CUSUM charts in contrast to the Shewhart chart.

As we first observe that the LS chart has a very competitive ARL_0, we chose to compare it with charts with ARL_1 values of the same order and with the largest ARL_0 value given in studies (which is around 500). For this we use the work of Does *et al.*[17,18,8]. Abbas *et al.* (2011)[17] show the superiority of the EWMA scheme II chart over classical EWMA, EWMA with scheme I, classical CUSUM, FIR CUSUM, FIR EWMA, weighted CUSUM, double CUSUM, distribution-free CUSUM and CUSUM with scheme I and II. See Abbas *et al.* (2011)[17] for the detailed explanation of all these charts. The mixed EWMA-CUSUM charts (MEC) are studied in the work of Abbas *et al.* (2013)[18] and compared to the previous charts as well. The work of Zaman *et al.* (2015)[8] deals with mixed CUSUM-EWMA charts (MCE). Considering those previous studies, we chose to compare the charts we propose to the EWMA chart scheme II with parameters $\lambda = 0.1$ and $L_S = 2.3$ which has a good global performance and is very sensitive especially for small shift $\delta = 0.25$. We also recall the performance of the two following charts: the MEC chart with $\lambda = 0.25$, $k_Z = 0.5$, $h_Z = 11.2$ which performs than every charts covered in Abbas *et al.* (2013)[18] for small shifts $\delta = 0.25$; MCE chart[8] with $\lambda_C = 0.75$ and $L_C = 6.08$ which has a better performance than the MEC charts for $\delta \geqslant 0.5$ and $\lambda_C > 0.5$ and has a better performance than the Shewhart chart for every δ and λ values. The ARL values of the literature of these charts is recalled in Table 1 and we calculate their EQL for the δ values we cover (see Table 5).

5.4 | Numerical results and interpretation

Numerical results for ARL computation of the LS chart and Q-chart with different α values are gathered in Table 1. Table 2 gives the standard deviation and maximum of the run length over the 10^5 values. Percentiles are given in Table 4 for the best proposed charts. For the LS charts, we cut the sequences at index 10^4 if no alarm occurs before and we give, when survival data occur, the percentage of uncut sequences instead of the maximum RL value (which is therefore evidently equal to 10^4). For the proposed charts, even the ARL_0 and $SdRL_0$ criterion depend on δ because the δ value is taken into account in the scoring function definition. Tables 3 and 4 compare the best proposed chart with the competitive ones from the literature we have given in the previous section. The EQL criterion is given in Table 5.

Some expected observations can first be checked in Table 1. Indeed, first Property 5 is verified. Moreover, both ARL_0 and ARL_1 values are increasing with α decreasing because the stopping time is less severe. ARL_1 values are decreasing with δ increasing as the change is easier to be observed. Moreover, we can see that for a fixed α and with varying the δ value, the ARL_0 of the LS chart have quite equivalent values: 6400 for $\alpha = 5\%$, 9000 for $\alpha = 1\%$ and 9700 for $\alpha = 0.27\%$. It may be due to the fact that we have to manage with survival data here. Table 1 also shows that the statistic Q has a better performance to detect the change (see the smaller values of ARL_1) which has also been highlighted[25]. But both ARL_0 and ARL_1 must be taken into account. The proposed Q chart must not be considered as the one with the best competitive chart as the ARL_0 values are not large enough. Table 1 also shows that the ARL_0 of the LS chart is very large for every α value and every δ. Among all

TABLE 1 ARL of our proposed charts for Gaussian model with the following parameters: $\mu_0 = 0$, $\sigma_0 = 1$, $q = 1$ and different values of δ and LLR scoring function.

		LS(α)			Q(α)		
δ		5%	1%	0.27%	5%	1%	0.27%
0.25	ARL_0	6455	9100	9713	20.0	112.1	395
	ARL_1	47.7	318	618	10.9	36.0	74.5
0.5	ARL_0	6250	9066	9715	20.0	94.1	343
	ARL_1	16.7	33.3	47.8	6.7	15.0	24.7
1	ARL_0	6353	9072	9730	19.1	82.9	289
	ARL_1	4.8	8.0	10.8	3.4	5.8	8.2
2	ARL_0	6428	9151	9750	19.3	85.9	278
	ARL_1	1.6	2.4	3.0	1.5	2.2	2.8

TABLE 2 SdRL and maximum value (or the percentage of survival data stopped at index 10^4) for the different proposed statistics and different values of δ in Gaussian model with parameters $\mu_0 = 0$, $\sigma_0 = 1$, $q = 1$ and LLR scoring function.

		LS(α)			Q(α)		
δ		5%	1%	0.27%	5%	1%	0.27%
0.25	$SdRL_0$	4583.1	2772	1630.3	22.0	132.3	495.5
	max or %	(60%)	(90%)	(97%)	375	1892	7076
0.5	$SdRL_0$	4658	2827	1623	21.9	112	442
	max or %	(60%)	(90%)	(97%)	438	1787	6807
1	$SdRL_0$	4631.6	2816	1585.0	19.2	92.4	347
	max or %	(60%)	(90%)	(97%)	280	1119	4973
2	$SdRL_0$	4604	2700	1522	19.0	87.8	301.5
	max or %	(60%)	(90%)	(97%)	254	1160	4571.0
0.25	$SdRL_1$	41.8	469.2	838.6	10.6	35.4	74.5
	max or %	490	7137	(< 0.02%)	148	477	1029
0.5	$SdRL_1$	20.6	33.3	42.2	6.0	12.9	20.2
	max	255	394	633	72	162	239
1	$SdRL_1$	4.6	6.7	8.2	2.52	4.2	5.6
	max	60	82	97	29	44	62
2	$SdRL_1$	1.1	1.6	1.9	0.9	1.31	1.6
	max	15	20	20	10	16	18

the proposed charts, LS chart with $\alpha = 5\%$ outperformed the others. In Sakhi *et al.* (2020)[25] they estimate the ARL_0 values in a totally different manner. The authors make the hypothesis that the false alarm rate, which we denote in this sequel $\tilde{\alpha}$ to avoid confusion with the non conditional nominal value α of the test at time i without considering the test issues before i, is constant. Using censored data with simulated data of length $I = 100$, they propose using survival methods an estimation $\hat{\tilde{\alpha}}$ of $\tilde{\alpha}$ and deduce by the hypothesis of a constant false alarm rate the ARL_0 value by $ARL_0 = 1/\hat{\tilde{\alpha}}$. This hypothesis is in fact not verified without conditioning as in the work of Margavio *et al.* (1995)[24].

TABLE 3 Comparison of the ARL of our best proposed chart with competitive charts for Gaussian model with the following parameters $\mu_0 = 0$, $\sigma_0 = 1$, $q = 1$ and different values of δ. Values for the EWMA type II chart come from Table VII of Abbas *et al.* (2011)[17]. Values for the MEC chart come from Table I of Zaman *et al.* (2015)[8].

δ		LS $\alpha = 5\%$	EWMA II $\lambda = 0.1$ $L_S = 2.3$	MEC $\lambda = 0.25$ $h_z = 0.5$ $k_z = 11.2$	MCE $\lambda_C = 0.75$ $L_C = 6.08$
0.25	ARL_0	6455	503	501	$\simeq 500$
	ARL_1	47.7	66.68	80.26	142.7
0.5	ARL_0	6250	503	501	$\simeq 500$
	ARL_1	16.7	21.4	35.7	37.7
1	ARL_0	6353	503	501	$\simeq 500$
	ARL_1	4.8	7.6	18.45	9.6
2	ARL_0	6428	503	501	$\simeq 500$
	ARL_1	1.64	3.5	11.19	3.44

$SdRL$ values given in Table 2 for the proposed charts are quite large for the in-control process. It is particularly the case for $SdRL_0$ values, but this is due to the fact that very large RL are possible. Moreover, we can observe in Table 4 that the k-th percentiles denoted P_k with $k = 25, 50, 75$ are quite of the same order for $k = 25$ as the EWMA scheme II chart, which makes our proposed chart still competitive. We can observe in Table 2 that $SdRL_0$ values are decreasing with α decreasing whereas $SdRL_1$ values are increasing. Table 2 also shows that the variability for out-of control process ($SdRL_1$) is decreasing with δ increasing as it is observed for EWMA scheme II chart in Table 4. In this table, we can also see that the $SdRL_1$ of the LS chart has an equivalent value as the one of the EWMA scheme II chart; a much larger one and an equivalent P_{25} percentile for an in-control process. The fact that large values of run length for in-control process is a good point and possible small values seem to appear as the same as for the EWMA scheme II chart.

We can see in Table 3 that the LS chart for $\alpha = 5\%$ outperforms EWMA scheme II, MCE and MCE chart with parameters of these charts chosen to have the best performance, with both more competitive $ARL0$ (larger values) and ARL_1 (smaller values).

Considering in Table 5 the EQL criterion , we can see that the LS chart with $\alpha = 5\%$ or $\alpha = 1\%$ has globally a better performance compared to the EWMA scheme II, MCE and MEC charts. The EQL values for Q chart must not be considered because of a too low ARL_0 which is not taken into account in the EQL computation (cf. $\delta = 0$).

5.5 | Unknown parameters under H_1

Scoring functions defined in Section 3.1 relies entirely on the instantaneous log-likelihood ratio LLR defined in Equation (6) which depends on the distribution of the signal $(A_i)_i$, and thus on its different parameters. One may encounter the problem where different parameters are unknown. As explained in Granjon (2013)[14] the optimal method to overcome this consists in using the generalized likelihood ratio test principle which consists in replacing all the unknown parameters by their maximum likelihood estimates[33]. But this cannot be written in a recursive manner and its complexity grows with the number of available samples, which is not the case of the Local Score, CUSUM or Lindley processes. This explains that the GLR algorithm cannot be used for online applications. Possibilities exist to keep the process recursive but lead to suboptimal algorithms.

Considering the most common practical case with parameters under H_1, μ_1 and σ_1, unknown, the usual solution is to use values *a priori* set by the user as additional parameters. A logical and efficient way to set this parameter is to choose the most likely values that should have been taken after the change. Some examples are proposed in the next section.

TABLE 4 $SdRL$, percentiles and maximum values (or the percentage of survival data stopped at index 10^4) for the LS chart with $\alpha = 5\%$ and the EWMA scheme II chart for different values of δ, Gaussian model with parameters $\mu_0 = 0$, $\sigma_0 = 1$, $q = 1$ and LLR scoring function. Values for the EWMA type II chart have been taken from Table IX of Abbas *et al.* (2011)[17].

δ		LS $\alpha = 5\%$	EWMA II $\lambda = 0.1$ $L_S = 2.3$
0.25	$SdRL_0$ $(P_{25}, P_{50}, P_{75}, \max)$	4583.1 $(206, 10^4, 10^4, (60\%))$	501.9 (141,350,699,-);$P_{90} = 1165$
	$SdRL_1$ $(P_{25}, P_{50}, P_{75}, \max)$	41.8 (18,36,65,490)	61.0 (23,48,91,-);$P_{90} = 145$
0.5	$SdRL_0$ $(P_{25}, P_{50}, P_{75}, \max)$	4658 $(108, 10^4, 10^4, (60\%))$	501.9 $P_{90} = 1165$
	$SdRL_1$ $(P_{25}, P_{50}, P_{75}, \max)$	20.6 (3,9,22,273,255)	17.2 (10,17,28,-);$P_{90} = 43$
1	$SdRL_0$ $(P_{25}, P_{50}, P_{75}, \max)$	4631.6 $(142, 10^4, 10^4, (60\%))$	501.9 (141,350,699,-);$P_{90} = 1165$
	$SdRL_1$ $(P_{25}, P_{50}, P_{75}, \max)$	4.6 (2,3,6,60)	4.2 (4,10,10,-);$P_{90} = 13$
2	$SdRL_0$ $(P_{25}, P_{50}, P_{75}, \max)$	4604 $(167, 10^4, 10^4, (60\%))$	501.9 (141,350,699,-);$P_{90} = 1165$
	$SdRL_1$ $(P_{25}, P_{50}, P_{75}, \max)$	1.1 (1,1,2,15)	0.87 (3,3,4,-);$P_{90} = 5$

TABLE 5 EQL, see Equation (1), with the following parameters: $\mu_0 = 0$, $\sigma_0 = 1$, $q = 1$ and different values of δ. The computation has been achieved using the values of Table 1 and 3.

	LS(α)			Q(α)			EWMA II	MEC	MCE
	5%	1%	0.27%	5%	1%	0.27%	$\lambda = 0.1$ $h_Z = 11.2$	$\lambda = 0.25$ $k_Z = 0.5$ $L_S = 2.3$	$\lambda_C = 0.75$ $L_C = 6.08$
EQL	6.1	11.8	17.2	4.7	7.5	10.3	11.5	32.8	13.2
ARL_0	≥ 6000			$\simeq 20$			$\simeq 500$	$\simeq 500$	$\simeq 500$

6 | ILLUSTRATION

We want to illustrate in this section the fact that the LS charts avoid false alarms compared to the other charts proposed in the literature. Due to the previous consideration exposed in Section 5.3 we restrict our comparison to the EWMA scheme II chart. As the main improvement of the proposed LS chart stands on a much larger ARL_0 with an equivalent sensitivity to detect a change-point, we do not compute the extra rules of the EWMA scheme II chart. Indeed, it would provide more alarms for the EWMA chart. Our comparison for in-control process is then under estimated for our proposed chart. We add the signaling limits of the EWMA scheme II on the illustrative figures for out-of control process to visualize the alarms.

We first simulate Gaussian sequences under H_0 with parameters $\mu_0 = 0$, $\sigma_0 = 1$, $\delta = 1$, $q = 1$. We use LLR score with the same parameters and $\alpha = 5\%,\ 1\%$ for the Local Score chart. For the classical EWMA chart we select $\lambda = 0.1$ and $L = 2.814$ which corresponds to the best classical EWMA chart with a large ARL_0 and a good sensitivity[17]. We use the time variant limits given in Equation (3). We then count the number of false alarms for a classical EWMA chart and for the Local Score chart we propose. We compute the average number of alarms per sequence on 10^3 sequences of length I for each chart and different

TABLE 6 Average false alarm per sequence for the LS chart and the classical EWMA chart, computed on 10^3 Gaussian sequences of length I simulated under H_0 with $\mu_0 = 0$, $\sigma_0 = 1$, $\delta = 1$, $q = 1$. Local Score parameter choice: LLR scores with the same parameters than the ones used to simulate the sequences and $\alpha = 5\%$; EWMA parameter choice: $\lambda = 0.1$ and $L = 3$.

I	1000	500	300	100
LS	0.16	0.12	0.13	0.12
EWMA	2.74	1.38	0.86	0.3

values of I: for $I = 1000$ the classical EWMA chart gives an average number of alarms per sequence equal to 2.74 and the Local Score chart gives an average number of alarm per sequence equal to 0.16 which illustrates the superiority of the LS chart for in-control process. We can see in Table 6 that even for not so long sequences the false alarm rate per sequence for Local Score chart is far lower than for the other chart.

We also simulate data under H_1 with $\mu_0 = 0$, $\sigma_0 = 1$, $\delta = 0.5$, $q = 1$ and a change-point at index $i = 250$. For the EWMA chart we chose to represent both classical EWMA chart with $\lambda = 0.1$ and $L = 2.814$ and EWMA scheme II chart with $\lambda = 0.1$ and $L_S = 2.3$ on the same figure. Figure 2 shows at the top the LS chart and the two EWMA charts at the bottom. For the LS chart, we chose to ease the reading of the LS chart, to plot $1 - p_{value}$ ($\mathbb{P}(M_i \leqslant m_i)$ with m_i the observed value of the Local Score for the sequence $\mathbb{A}^i$) and $1-\alpha$, instead of $\mathbb{P}(M_i \geqslant m_i)$ and α: parts in the figure with a slowly decreasing value (from index $\simeq 10$ to $\simeq 40$ for example) correspond to the fact that the growing sequence does not achieve a larger Local Score with this additional part. Indeed, as the sequence length has also increased and thus the chance to achieve such a local score too, $1 - p_{value}$ decreases (see Property 1). The plotted points increase when a better Local Score is achieved. When the plotted value is above the $1 - \alpha$ line, an alarm is made. For the EWMA charts, Figure 2 (bottom) represents the Z_i values and the control limit of Equation (3) for the classical EWMA chart and the signaling limits of Equation (5) for the EWMA scheme II. In this figure we see that the change-point is rapidly detected in the two cases. But for the EWMA charts, several false alarms occur before the detection (see index $\simeq 80$, just before index 200 and after). It is easy to get such illustrative examples.

To illustrate the case of unknown parameters, we simulate two cases: one using $\mu_0 = 0$, $\sigma_0 = 1$, $\delta = 0.5$, $q = 1$ for the sequence after the change-point but we use $\delta = 1$ for the scoring function of the LS chart; and secondly we inverse the values of δ, meaning $\delta = 1$ for the simulated sequences after the change and $\delta = 0.5$ for the scoring function. The change-point index is still 250 in both cases. For 20 consecutive simulations we count the number of change detections without false alarm (D); the number of detections with false alarm (D+FA), the number of non detections of the change with false alarm (FA) and the number of non detections without false alarm for both charts (∅). The results are given in Table 7. The LS chart has much more correct detections without false alarm in both parameter cases than the EWMA scheme II chart, whereas this last chart often declares a false alarm before the change-point.

Figure 3 proposes an illustration to the case where $\delta = 1$ for the sequence simulation after the change point ; $\delta = 0.5$ for the sequence simulation before the change point and for the scoring function all over the sequence. In this figure, the LS chart detects correctly the change without prior false alarm whereas the EWMA scheme II chart declares false alarms before detecting the change (see index $\simeq 155$ and just before 250). As the Table 7 can show it is easy to obtain such an example.

7 | CONCLUSION

The CUSUM process defined in 1954 is equivalent to the Lindley process defined in 1952. The Local Score statistic used to detect atypical regions in biological sequences is the maximum of the CUSUM, or Lindley, process. A lot of works on the distribution of the Local Score exist and one is able to establish easily the p-values of the observed local score values. We define in this schedule a "Local Score" chart for which observations with 1-p-value greater than a given threshold $1 - \alpha$ lead to an alarm. Such a Local Score chart takes into account the past information. Using ARL and EQL performance criterion, we show that the Local Score chart has a much larger ARL_0 than all the performing charts covered in Abbas *et al.* (2011)[17] (2013)[18] and Zaman *et al.* (2015)[8]. The achieved comparison focuses on the best charts taking into account the past information, Gaussian model, location monitoring and diverse shift ranges. Such a large ARL_0 can be very appreciated especially in big data context. Moreover, the ARL_1 values are also performing very well whatever the shift values. Globally the EQL values exhibit dominance of the Local Score chart for a parameter $\alpha = 5\%$: The ARL values of the Local Score chart decrease with a growing α. As the

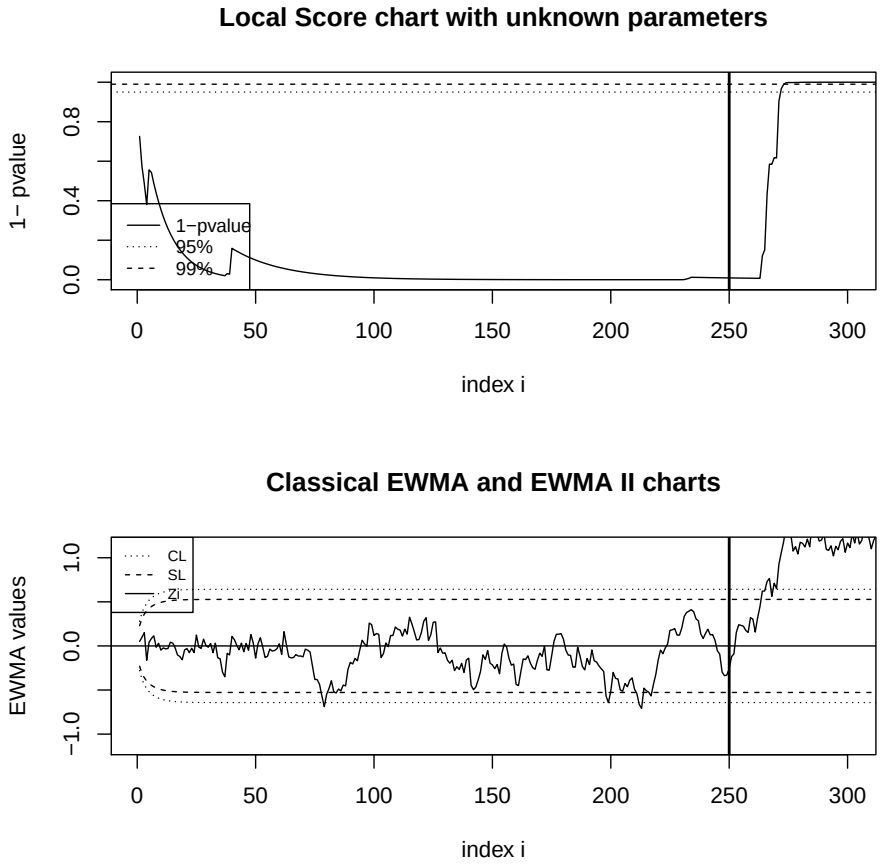

FIGURE 2 Local Score and EWMA charts for a Gaussian sequence with a change-point at index $i = 250$ (see vertical line). Score function is LLR with $\mu_0 = 0$, $\sigma_0 = 1$, $\delta = 0.5$, $q = 1$ and $\alpha = 1\%$ and 5% for the Local Score chart; and $\lambda = 0.1$ and $L = 2.814$ for the classical EWMA and $\lambda = 0.1$ and $L_S = 2.3$ for EWMA scheme II.

ARL_0 is above 6000 for $\alpha = 5\%$, 1% and 0.27% it would not be a real problem to decrease it in order to catch a better ARL_1. We have observed a large $SdRL_0$ value but this is due to possible large run length and we can observe similar 25 percentile values as in the EWMA scheme II chart, the LS chart is then still competitive.

The Q chart realizes very good ARL_1 but the false alarm rate for in-control process is not large enough. It could be of interest using a larger α threshold, but for $\alpha > 10\%$ we lose in our point of view the sense of the statistical test with a large nominal value which represents the instantaneous probability of a false alarm.

The Local Score chart has other qualities such as a very large ARL_0 for similar sensitivity to detect change-point. It can also be used for both location and dispersion monitoring at the same time with an adapted scoring function using $q \neq 1$. This needs to pre-compute the p-value matrices for each different q value. There are also other very interesting openings. Indeed, the distribution of the Local Score has also been established for Markov model, so it is possible to construct charts which could take into account an eventual dependance between the signal at time i and $i - 1$. Other models on the signal, rather than a Gaussian one can also be considered like exponential or geometric ones. Extra rules as given in Section 2.4.1 with defined signaling limits, with $\alpha = 10\%$ for example, could be used to increase the performance of detecting a change-point.

All these points, performance values and possible openings, allow us to conclude that the proposed Local Score chart can be qualified as a very interesting and promising method for high quality control monitoring and especially in big data contexts.

TABLE 7 Counts of correct detections of the change-point without false alarm (D); of correct detections with false alarm before (D+FA); of non detection of the change-point with false alarm (FA) and the number of non detection without false alarm (∅) for both Local Score and EWMA scheme II charts. Parameters are $\mu_0 = 0$, $\sigma_0 = 1$, $q = 1$. The values of δ for the simulation of the sequences, noted δ_{H_1}, is different than the one used in the scoring scheme, noted δ_{LLR}.

$\delta_{H_1} = 0.5$;$\delta_{LLR} = 1$	D	D+FA	∅	FA
LS	14/20	3/20	2/20	1/20
EWMA	5/20	14/20	0/20	1/20
$\delta_{H_1} = 1$;$\delta_{LLR} = 0.5$	D	D+FA	∅	FA
LS	13/20	7/20	0/20	0/20
EWMA	4/20	16/20	0/20	0/20

Acknowledgements

The author is thankful to the Agence pour les Mathématiques en Interaction avec l'Entreprise et la Société (AMIES) for providing a financial fund to achieve this work.

References

1. S. Ali, A. Pievatolob, and R. Gob. An overview of control charts for high-quality processes. Quality and Reliability Engineering International, 32:2171–2189, 2016.

2. E. S. Page. Continuous inspection schemes. Biometrika, 41(1):100–115, 1954.

3. S.W. Roberts. Control chart tests based on geometric moving averages. Technometrics, 1(3):239–250, 1959.

4. D.V. Lindley. The theory of queues with a single server. Mathematical Proceedings of the Cambridge Philosophical Society, 48(2):277–289, 1952.

5. S. Mercier and J.J. Daudin. Exact distribution for the local score of one i.i.d. random sequence. Jour. Comp. Biol, 8(4):373–380, 2001.

6. M.I. Fariello, S. Boitard, S. Mercier, D. Robelin, T. Faraut, C. Arnould, E. Le Bihan-Duval, J. Recoquillay, G. Salin, G. Dahais, F. Pitel, G. Leterrier, and M. Sancristobal. Accounting for linkage disequilibrium in genome scans for selection without individual genotypes : the local score approach. Molecular Ecology, 26(14):3700–3714, 2017.

7. A. Lagnoux, S. Mercier, and P. Vallois. Statistical significance based on length and position of the local score in a model of i.i.d. sequences. Bioinformatics, 33(5):654–660, 2017.

8. B. Zaman, M. Riaz, N. Abbas, and R.J.M.M. Does. Mixed cumulative sum-exponentially weighted moving average control charts: An efficient way of monitoring process location. Quality and Reliability Engineering International, 31(8):1407–1421, 2015.

9. Zhang Wu, Jianxin Jiao, Mei Yang, Ying Liu, and Zhaojun Wang. An enhanced adaptive cusum control chart. Iie Transactions, 41:642–653, 05 2009.

10. R. Jong-Hyun, W. Hong, and K. Sujin. Optimal design of a cusum chart for a mean shift of unknown size. Journal of Quality Technology, 42(3):311–326, 2010.

11. Y. Ou, Z. Wu, and F. Tsung. A comparison study of effectiveness and robustness of control charts for monitoring process mean. International Journal of Production Economics, 135(1):479–490, January 2012.

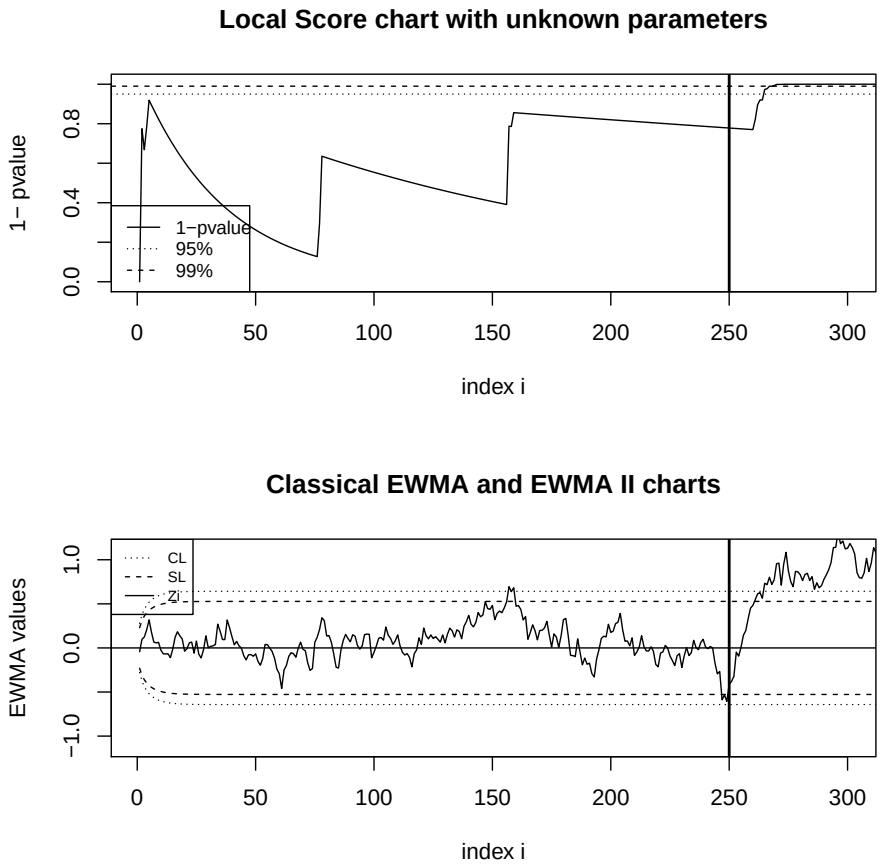

FIGURE 3 Example for Local Score and EWMA charts with unknown parameters: Gaussian sequence simulated with a change-point at index $i = 250$ (see vertical line) and $\mu_0 = 0$, $\sigma_0 = 1$, $\delta = 1$, $q = 1$ and $\alpha = 1\%$. Score function is LLR with $\mu_0 = 0$, $\sigma_0 = 1$, $\delta = 0.5$, $q = 1$ and $\alpha = 1\%$ and 5% for the Local Score chart; the parameters for the classical EWMA are $\lambda = 0.1$ and $L = 2.814$ and $\lambda = 0.1$ and $L_S = 2.3$ for the EWMA scheme II chart.

12. J. M. Lucas and M.S. Saccucci. Exponentially weighted moving average control schemes: Properties and enhancements. Technometrics, 32(1):1–12, 1990.

13. D. Egea-Roca, G. Seco-Granado, and J.A. Lopez-Salcedo. Comprehensive overview of quickest detection theory and its application to gnss threat detection. Gyroscopy and Navigation, 8(1):1–14, 2017.

14. P. Granjon. The cusum algorithm - a small review. hal-00914697, 2013.

15. M. Klein. Two alternatives to the Shewhart x control chart. Journal of Quality Technology, 32(4):427–431, 2000.

16. M. Riaz, N. Abbas, and R.J.M.M. Does. Improving the performance of cusum charts. Quality and Reliability Engineering International, 27:415–424, 2011.

17. N. Abbas, M. Riaz, and R. Does. Enhancing the performance of ewma charts. Quality and Reliability Engineering International, 27:821–833, 10 2011.

18. N. Abbas, M. Riaz, and R.J.M.M. Does. Mixed exponentially weighted moving average - cumulative sum charts for process monitoring. Quality and Reliability Engineering International, 29(3):345–356, 2013.

19. J. Kyte and R.F. Doolittle. A simple method for displaying the hydropathic character of a protein. Journal of molecular biology, 157(1):105–132, 1982.

20. S. Mercier and G. Nuel. Probabilizing the segmentation space in local score approaches. Methodology and Computing in Applied Probability, 2020. In revision.

21. A.G. Tartakovsky, A.S. Polunchenko, and G. Sokolov. Efficient computer network anomaly detection by changepoint detection methods. IEEE Journal of Selected Topics in Signal Processing, 7(1):4–11, 2012.

22. J. Glaz, V. Pozdnyakov, and S. Wallenstein. Scan statistics Methods ans Applications. Statistics for Industry and Technology, Birkauser, 2009.

23. A. Wald. Sequential Tests of Statistical Hypotheses. Ann. Math. Statist., 16(2):117–186, 1945.

24. W.H. Woodall, T.M. Margavio, M.D. Conerly and L.G. Drake. Alarm rates for quality control charts. Statistics & Probability Letters, 24(3):219–224, 1995.

25. N. Sahki, A. Gégout-Petit, and S. Wantz-Mézières. Performance study of detection thresholds for cusum statistic in a sequential context. In revision at Quality and Reliability Engineering International, 2020.

26. X. Shen, C. Zou, and F. Jiang, W. Tsung. Monitoring poisson count data with probability control limits when sample sizes are time varying. Naval Research Logistics, 60(8):625–636, 2013.

27. W. Huang, L. Shu, W.H. Woodall, and K.L. Tsui. Cusum procedures with probability control limits for monitoring processes with variable sample sizes. IIE Transactions, 48(8):759–771, 2016.

28. S. Karlin and A. Dembo. Limit distributions of maximal segmental score among Markov-dependent partial sums. AdAP, 24:113–140, 1992.

29. D. Cellier, F. Charlot, and S. Mercier. An improved approximation for assessing the statistical significance of molecular sequence features. Jour. Appl. Prob., 40:427–441, 2003.

30. C. Hassenforder and S. Mercier. Exact distribution of the local score for markovian sequences. AISM, 59(4):741–755, 2007.

31. S. Grusea and S. Mercier. Improvement on the distribution of maximal segmental score in a Markovian sequence. Journal of Applied Probability, 57.1, March 2020.

32. S. Karlin and S.-F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. PNAS, 87:2264–2268, 1990.

33. G. Lorden. Procedures for reacting to a change in distribution. Ann. Math. Statist., 42(6):1897–1908, 1971.

□

APPENDIX

A HOMOTHETIC TRANSFORMATION OF THE SCORING FUNCTION

Let us consider a scoring function s and $E \in \mathbb{R}^{+*}$. Let us denote $s' = E \cdot s$ and $M_i(s)$ (respectively $M_i(s')$) the Local Score variables associated to the scoring function s (resp. s'). Due to the definition of the local score given in Equations (8) and (9), we have $M_i(s') = E \cdot M_i(s)$. For a studied sequence of length i with respectively a Local Score a using s and a' using s', we have thus $a' = E \cdot a$ and

$$\mathbb{P}(M_i(s') \geqslant a') = \mathbb{P}(E \cdot M_i(s) \geqslant E \cdot a) = \mathbb{P}(M_i(s) \geqslant a) .$$

The statistical significance of the sequence Local Score does not change. This property allows to transform a rational scoring scheme into an integer one and thus to use the exact method to establish the p-value. Indeed, the method proposed by Mercier and Daudin (2001)[5] to compute the distribution of the Local Score, and recalled in (14) stands on a matrix Π that is fulfilled using the score distribution. It is a square matrix of size $(a+1) \times (a+1)$ with a the observed Local Score value and a must be an integer.

In our work we chose to rescale the scoring function to have a larger range which allows us to conserve a sufficient distinction between the scores after using the integer part and to have a scoring function which we consider "richer".

B INTEGER OR REAL SCORING FUNCTION

Using the exact method to establish the distribution of the Local Score or the score of an excursion requires integer scores to be able to consider the matrix Π of Equation (14)[5]. We study the effect of using the integer part of the LLR scoring function. Even if we have the possibility to define such a scoring function and to use it, we think it is pertinent to illustrate the induced changes. For this we simulate 10^4 sequences of length 3000 and we compute the Local Score values for each sequence using the scoring scheme $E \cdot LLR$ and the $\lfloor E \cdot LLR \rfloor$ one; with LLR defined in Equation (6), Gaussian signal with $\mu_0 = 0$, $\sigma_0 = 1$, $q = 1$, $E = 10$ and the different values of δ. We extracted the empirical p-values of the observed Local Score values in both cases. We declare significant a sequence for which the p-value is less than 0.05 and not significant in the other cases. We then compute the percentage of sequences which are declared the same in both case. We observe that the larger is δ, the larger the percentage is: for $\delta = 0.25$, 0.5, 1 and 2, we get 94.7% 96.8%, 99.4% and 99.7% sequences that are identically declared.

C EXACT DISTRIBUTION OF $Q^{(K)}$ IN THE I.I.D. CASE

The following method mimics the work of Mercier and Daudin[5] on the Local Score. The main idea is to create a Lindley process and to stop it with an adequate stopping time. Let $a > 0$, $\tau_a = \inf\{k \geqslant 1 : W_k \geqslant a\}$ and $\tau_- = \inf\{k \geqslant 1 : W_k < 0\}$. Let us define $\mathbb{W}^* = (W_k^*)_{1 \leqslant k \leqslant n}$ the stopped process by

$$W_k^* = W_k \text{ for } k < \inf\{T_-, \tau_a\}, \quad W_k^* = 0 \quad \forall k \geqslant T_- \text{ when } T_- < \tau_a,$$
$$\text{and} \quad W_k^* = a \quad \forall k \geqslant \tau_a \text{ when } \tau_a < T_-.$$

Let us consider the X_i i.i.d. Let p be the distribution of the X_i and f the corresponding c.d.f. The process $\mathbb{W}^*$ is a Markov chain taking its values in $\{0, 1, \dots, a\}$. Let $\Lambda = (p_{i,j})_{i,j}$ be the transition probability matrix with $p_{ij} = \mathbb{P}(W_{k+1}^* = j | W_k^* = i)$ given by

$$\begin{aligned} p_{i0} &= \mathbb{P}(X_{k+1} \geqslant -i) = f(-i) \quad \text{for } i = 1, \dots, a-1 \\ p_{ia} &= \mathbb{P}(X_{k+1} \geqslant a - i) = 1 - f(a - i - 1) \quad \text{for } i = 1, \dots, a-1 \\ p_{ij} &= \mathbb{P}(X_{k+1} \geqslant j - i) = p(j - i) \quad \text{for } i, j \in \{1, \dots, a-1\}\,. \end{aligned}$$

We then have

$$\Lambda = \left(\begin{array}{c|ccc|c} 1 & 0 & \dots & 0 & 0 \\ \hline f(-1) & p(0) & \dots & p(a-2) & 1 - f(a-2) \\ \vdots & & & & \vdots \\ f(-h) & \vdots & p(\ell - h) & \vdots & 1 - f(a-h-1) \\ \vdots & & & & \vdots \\ f(1-a) & p(2-a) & \dots & p(0) & 1 - f(0) \\ \hline 0 & 0 & \dots & 0 & 1 \end{array}\right).$$

Let us denote $p^* = (p^*(0), \dots, p^*(a))$ the distribution of W_1^*. We have

$$\begin{aligned} p^*(0) &= \mathbb{P}(X_1 \leqslant 0) = f(0) \\ p^*(a) &= \mathbb{P}(X_1 \geqslant a) = 1 - f(a-1) \\ p^*(k) &= \mathbb{P}(X_1 = k) = p(k) \quad \text{for } 1 \leqslant k \leqslant a-1\,. \end{aligned}$$

Result 1 (Exact law for $Q^{(k)}$ in i.i.d. case)**.** $\forall I \geqslant 1$

$$P(Q^{(1)} \geqslant a) = \mathbb{P}(W_I^* = a) \quad \text{with}$$

$$\begin{aligned} \mathbb{P}(W_1^* = a) &= 1 - f(a-1) \text{ and} \\ \mathbb{P}(W_I^* = a) &= p^* \cdot \Lambda^{I-1} \cdot (0, \dots, 0, 1)' \quad \forall I \geqslant 2 \end{aligned}$$

and under the i.i.d. model all the $(Q^{(k)})_k$ are i.i.d.

Author's biography

Sabine MERCIER is graduated in applied mathematics from the University of Rouen (PhD 1999). She presented her Habilitation defense in December 2018 titled "Local score distribution to highlight atypical segments in sequences" at the University of Toulouse Paul Sabatier. Currently, she is a teacher at the University of Toulouse Jean Jaurès, involved in the Master ISMAG specialized in computer science, applied mathematics and statistics for production management, for which she has supervised Statistical Process Control (SPC) courses and internships on the subject for nearly 20 years. She is member of the Mathematics Institute of Toulouse. Her main research activities are bio statistic, atypical region detection, stochastic processes and extremes values. She is also regularly interacting with companies. At the present time she is the leader of the project "Highlight" of the CIMI Labex (International Center of Mathematics and Computer Science of Toulouse) and of the project "Local Score and SPC" in collaboration with Ippon Innovation company.